Projet de R - M1/ISF

5 janvier 2025

Les salaires des joueurs de Premier League sont-ils corrélés avec leurs perfomances sportives et leurs valeurs marchande ?

Julien Lassus

PANTHÉON-ASSAS
UNIVERSITÉ
PARIS

Table des matières

1 Partie I : Introduction

1.1 Présentation du sujet

Nous allons nous intéressé à l'économie du sport, et en particulier aux données financières de la Premier League (Ligue de football nationale d'Angleterre). L'objectif de l'analyse est de s'intéresser, à l'aide d'étude statistiques / économétriques, aux liens entre salaires, valeurs marchandes et performances sportives. Il faut prendre en compte que la performances sportives est observées d'un point de vue statistiques, ce qui ne reflète parfois pas réellement la qualité d'un joueur de football.

L'intérêt de l'analyse est donc de tirer une conclusion sur les liens entre ses variables, tout en tenant compte des cas particulier tels que des joueurs disposant d'un salaire élevée mais d'une performances statistiques faibles ou bien des joueurs disposant d'un faible salaire / valeurs marchandes mais ayant des performances statistiques élevée. Nous tenteront d'en comprendre les raisons en conclusion.

1.2 Description de la base de données

Nous avons initialement trois bases de données (toutes des **données en panel**). La première, depuis Kaglle, reprenant l'ensemble des joueurs de Premier - League 2020/2021 avec leurs statistiques sportives (base1). La deuxième recense les valeurs marchandes de l'ensemble des joueurs européen en 2020 / 2021 (base2). Il fallait donc garder uniquement les joueurs de premier - league. Enfin, la dernière base (base3), concernant les salaires, est issue du site web `https://www.capology.com/uk/premier-league/salaries/2020-2021/` qui empêche l'acquisition de ses données en dehors d'un abonnement payant. Ainsi, les techniques basique d'extraction de données (via API ou Scrapling) ne fonctionnait pas. Pour acquérir les données nous avons donc du faire 25 screenshot contenant les différents parties du tableau de données, puis utiliser chatgpt qui dispose d'un OCR (Reconnaissance optique de caractères) afin de transformer ces documents png en tableau Excel (qui ont été par la suite fusionner). La base finale contient **485 observations** (joueurs de foot) et **64 variables** dont les plus importantes sont le salaire annuel, la valeur marchande ainsi que la variable performances (composante des statistiques sportives). **L'ensemble des valeurs financières sont exprimé en euros en valeurs 2021.**

1.3 Traitement spécifique des données

1.3.1 Fusion de plusieurs bases :

Nous avions donc trois bases. de données sur Excel (format csv). Nous avons fait le choix de les importer séparément afin de pouvoir apporter des modifications à la base complète sans dégrader les bases initiale.

```r
# 1- Import de la base sur les joueurs avec info et stats utile pour performances

base1 <- read.csv("~/Desktop/dataset - 2020-09-24.csv")

# 2- Import de la base avec noms et valeur marchande des joueurs

base2 <- read.csv("~/Desktop/players.csv")

#3- Import de la base avec nom et salaire via un code HTML
base3 <- read_xlsx("/Users/julienlassus/Desktop/Projet R M1ISF/premier_league_salaries_2020_2021_updated.xlsx")
```

FIGURE 1 – Import des bases

Ensuite, nous avons donc du fusionner ces 3 bases en utilisant le nom des joueurs (unique) comme clé. Préalablement, il a fallut s'assurer que le nom des joueurs était bien écrit de la même façon sur les 3 bases (R prenant en compte les majuscules). Nous avons donc utilisé la commande suivante sur les trois bases :

```r
# Assemblage des 3 bases pour une seule
#Suppression de tous les accents pour pouvoir fusionner les bases à l"aide des noms
base1$Name <- stri_trans_general(base1$Name,"Latin-ASCII")
base2$Name <- stri_trans_general(base2$Name,"Latin-ASCII")
base3$Name <- stri_trans_general(base3$Name,"Latin-ASCII")
```

FIGURE 2 – Mise au même format

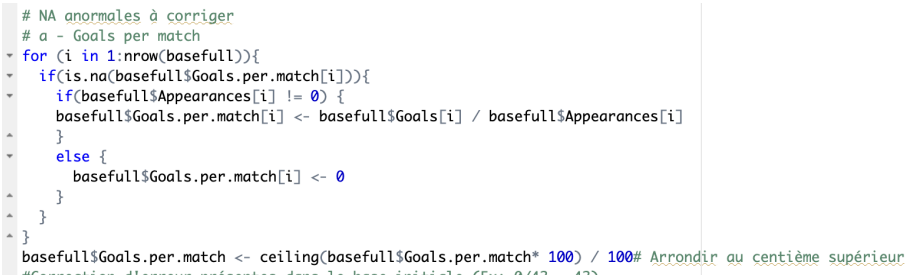
Enfin, nous avons fusionner les trois bases à l'aide de la commande left join, ce qui à donné basefull contenant les statistiques sportives, les salaires annuelles et hebdomadaire ainsi que la valeur marchande pour chaque joueur.

1.3.2 Nettoyage de la base

i) Régularisation des valeurs manquantes (NA) :
Nous avons remplacé les valeurs manquantes (NA) présentes dans les variables numériques par 0 lorsque cela avait du sens c'est-à-dire que la valeur NA signifiait 0. Lorsque la variable contenait des NA dites anormales nous avons utiliser une boucle pour la remplacer par la bonne valeur (cas par cas selon le calcul de la variable). Par exemple pour le ratio

$$Goals.per.match = \frac{buts}{matchs}$$

certains joueurs avait une valeur NA alors que le nombre de buts et de matchs été bien renseignés.



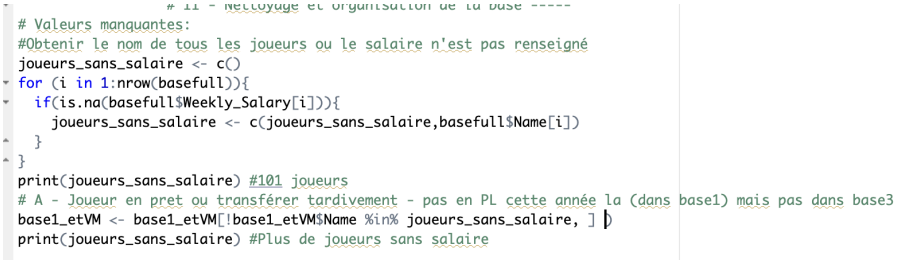
```r
# NA anormales à corriger
# a - Goals per match
for (i in 1:nrow(basefull)){
  if(is.na(basefull$Goals.per.match[i])){
    if(basefull$Appearances[i] != 0) {
    basefull$Goals.per.match[i] <- basefull$Goals[i] / basefull$Appearances[i]
    }
    else {
      basefull$Goals.per.match[i] <- 0
    }
  }
}
basefull$Goals.per.match <- ceiling(basefull$Goals.per.match* 100) / 100# Arrondir au centième supérieur
```

FIGURE 3 – Boucle pour Goals.per.match

Cette méthode a entraîné quelques petites correction car lorsque le dénominateur vaut 0, la fonction renvoyait la valeur du numérateur.

ii) Régularisation des salaires :

Pour les variables de Salaire hebdomadaire et annuelle, certains joueurs avait la valeur NA. Après des recherches sur les dates d'actualisation de base3 et base1, nous nous sommes rendu compte que tous les joueurs concernés était des joueurs qui avait été transférés ou prêtés dans un autre championnat assez tardivement dans la saison. Comme base1 à été créer plus tôt en 2020 ces joueurs sont donc présent dans base1 mais pas dans la liste officiel des salaires des joueurs de premier league de la saison 2020 - 2021, car il n'y était plus. Ces joueurs ont donc été placés dans une liste, puis supprimer de basefull.



```r
                # II - Nettoyage et organisation de la base -----
# Valeurs manquantes:
#Obtenir le nom de tous les joueurs ou le salaire n'est pas renseigné
joueurs_sans_salaire <- c()
for (i in 1:nrow(basefull)){
  if(is.na(basefull$Weekly_Salary[i])){
    joueurs_sans_salaire <- c(joueurs_sans_salaire,basefull$Name[i])
  }
}
print(joueurs_sans_salaire) #101 joueurs
# A - Joueur en pret ou transférer tardivement - pas en PL cette année la (dans base1) mais pas dans base3
base1_etVM <- base1_etVM[!base1_etVM$Name %in% joueurs_sans_salaire, ] )
print(joueurs_sans_salaire) #Plus de joueurs sans salaire
```

FIGURE 4 – Suppression des joueurs

iii) Régularisation de la valeur marchande :

Pour les joueurs disposant d'une valeur marchande inférieur à 15 Millions d'euros, l'information n'est pas disponible. Nous avons donc remplacé les NA par la valeur moyenne des joueurs de PL de la saison 2020 - 2021 dont la valeur marchande est inférieur à 14 millions (statistiques trouvé via un site spécialisé) : 8.4 millions.

```r
base1_etVM <- base1_etVM %>%
  mutate(Markey.Value.In.Millions...= ifelse(is.na(Markey.Value.In.Millions...),8.4,Markey.Value.In.Millions...))
base1_etVM$Markey.Value.In.Millions...[1:10] #Fonctionne
```

FIGURE 5 – Transfo VM

1.3.3 Modification et création de variable

i) Détection de valeur aberrantes et correction :

A l'aide de la commande boxplot et donc des boites à moustaches suivantes nous avons pu détecter des valeurs aberrantes pour les variables salaires et valeurs marchandes. Toutefois, étant donnée la faible quantité de données et l'importance des salaires, valeurs marchande et performances élevées (joueur sortant du lot) nous avons jugés bon de **ne pas les supprimer** (perte de sens économique).

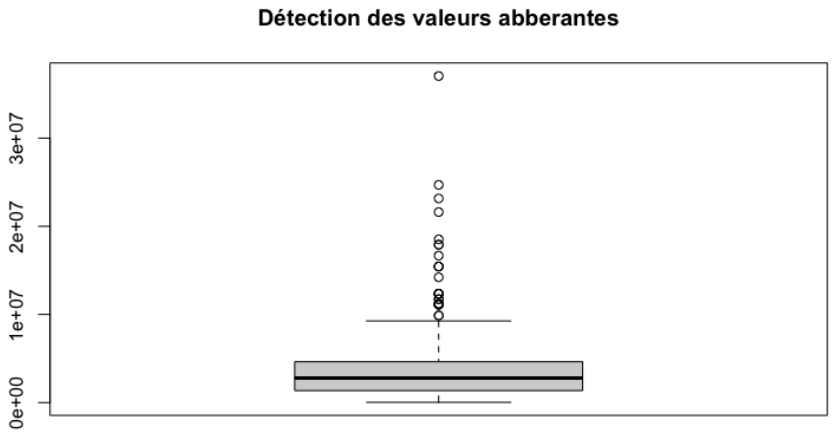


FIGURE 6 – Salaires

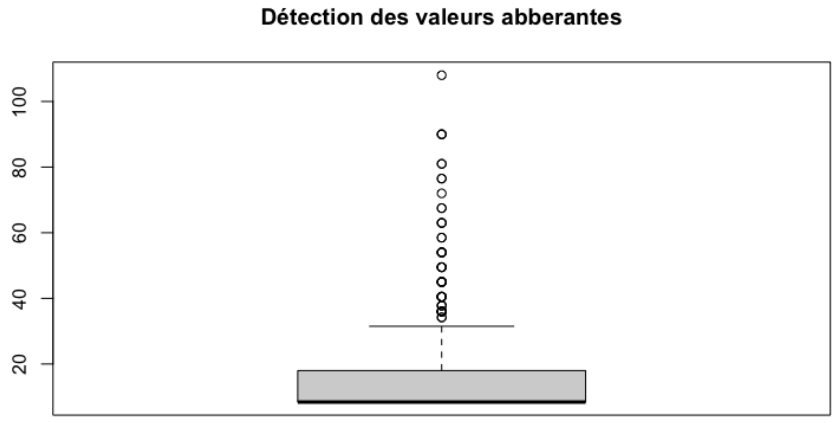


FIGURE 7 – Valeurs marchandes

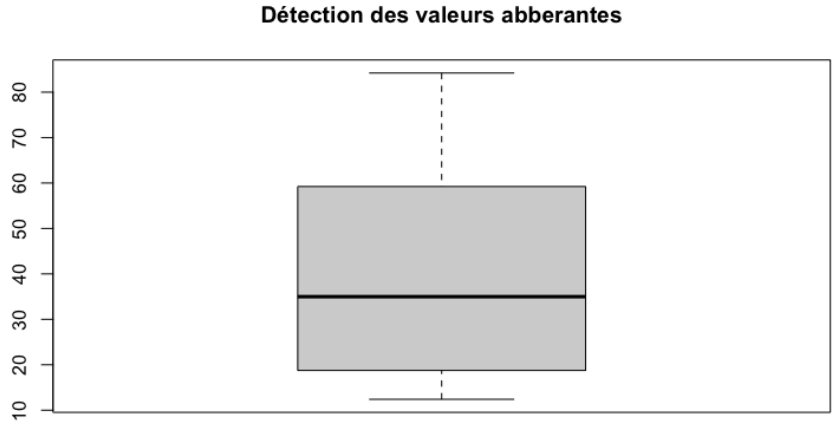


FIGURE 8 – Performances

ii) Mise au format numérique des variables salaires :
Pour les variables des salaires hebdomadaire et annuel, le type de donnée avait été défini comme characters. Or cela entraînait de nombreuses complications dans l'utilisation de ses variables. Nous avons donc du modifier le type de ces

variables ce qui impliquait de réécrire ces valeurs sous une forme adéquate.

```r
#Pour salaire annuel
# 1. Retirer le symbole de l'euro (€) et remplacer la virgule par un point
basefull$Yearly_Salary <- gsub("€", "", basefull$Yearly_Salary)  # Retirer €
basefull$Yearly_Salary <- gsub(",", ".", basefull$Yearly_Salary)  # Remplacer , par .
basefull$Yearly_Salary_num <- gsub("\\.","",basefull$Yearly_Salary)
# 2. Convertir en numérique
basefull$Yearly_Salary_num <- as.numeric(basefull$Yearly_Salary_num)
summary(basefull$Yearly_Salary_num) #Fonctionne
```

FIGURE 9 – exemple pour Salaire annuel

Nous avons par la suite supprimer les anciennes variables pour les remplacer par leurs équivalent de type numeric.

iii) Rajout de la variable Performances :

Afin de pouvoir réaliser un modèle économétrique basé sur la performances des joueurs, nous avons du créer un score de performance (noté sur 100) dont les variables statistiques utilisées changent selon la position du joueur : Goals, Defender, Midfielder and Forward.

Nous allons expliqué la création de la note pour une position, le mécanisme étant le même pour les autres position, en changeant simplement les variables utilisées en fonction des critères footballistiques que nécessite le poste.

Tout d'abord, on introduit une fonction normalise qui convertit les valeurs sur une échelle de 0 à 1. On normalise toutes les valeurs que l'on souhaite intégrer par rapport à la position. Puis, on introduit un coefficient pour chaque variable en fonction du poids que l'on veut donner à la variable. Par exemple, la variable goals aura un poids plus important que celle du nombre de tirs car son influence sur le match est plus importante. Enfin, on applique la formule suivante pour n variable :

$$(\sum_{i=1}^{n} variable_i normaliser * poids_i) * 100$$

Nous avons donc après ces opérations, une colonne pour chaque position que nous allons fusionner en une seule colonne Performances.

2 Partie II : Statistiques descriptives

2.1 Résume statistiques des variables les plus importantes du modèles :

Var2	Freq
Min.	29687
1st Qu.	1358467
Median	2778682
Mean	3708873
3rd Qu.	4631136
Max.	37049089

FIGURE 10 – Summary Salaire

Var2	Freq
Min.	12.39878
1st Qu.	18.74755
Median	34.98610
Mean	37.93255
3rd Qu.	59.23099
Max.	84.21255

FIGURE 11 – Summary score de Performances

Var2	Freq
Min.	8.40000
1st Qu.	8.40000
Median	8.40000
Mean	16.03175
3rd Qu.	18.00000
Max.	108.00000

FIGURE 12 – Summary de la valeur marchande

Quelques interprétations utile :
- Le salaire maximum est de 37.049.089 millions d'euros pour la saison 2020 - 2021
- Le salaire median (2778682) est inférieure d'environ un millions d'euros du salaire moyen (3708872) ce qui est représentatif de l'écart important entre les plus haut salaires et les plus bas. De même, 75% des joueurs ont un salaire inférieur ou égales à 4631136 (d'après le trosième quantile). - Le joueur le plus performant dispose d'une note de 84.21 sur 100 (Sergio Aguero), cela est cohérent avec les avis des journalistes sur cette saison de premier league.
- La valeur marchande maximale est de 108 millions d'euro, correspondant au joueur Harry Kane (note de 70 sur 100 en performances).

TABLE 1 – Écarts-types des variables d'intérêts

	Variables	Ecart.type
Yearly_Salary_num	Yearly_Salary_num	3,749,111
Markey.Value.In.Millions...	Markey.Value.In.Millions...	15
Performances	Performances	19

Ces écarts-types montrent une grande variation dans les trois variables analysées : les salaires, la valeur marchande et les performances des joueurs. Cela est cohérent avec la réalité du football professionnel, où certains joueurs ont des niveaux de rémunération et de performance bien supérieurs à la moyenne. Un écart-type élevé traduit une plus grande dispersion des données, signalant une large hétérogénéité au sein de chaque variable.

2.2 Analyse graphique élémentaire

Avant d'effectuer des tests statistiques, afin de bien comprendre notre base, nous allons nous intéresser aux liens graphiques entre différentes variables afin d'avoir une première intuition des corrélations que l'on pourrait avoir.

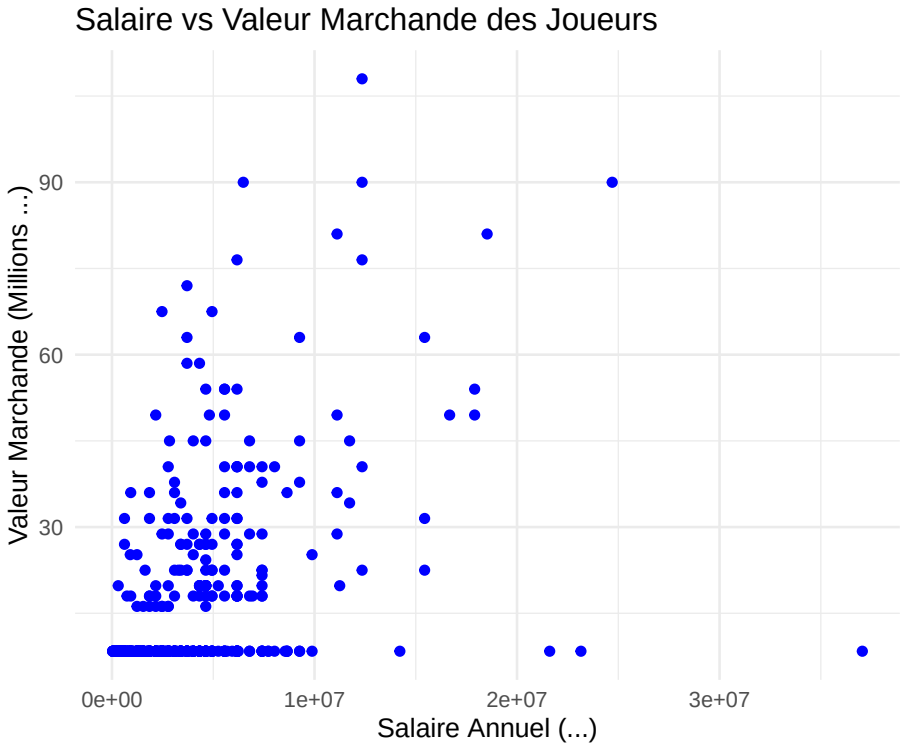


FIGURE 13 – Graphique salaire et valeur marchande

Ici l'on remarque que les variables valeur marchande et salaire semblent être relativement corrélé puisque en général le salaire augmente lorsque la valeur marchande augmente.

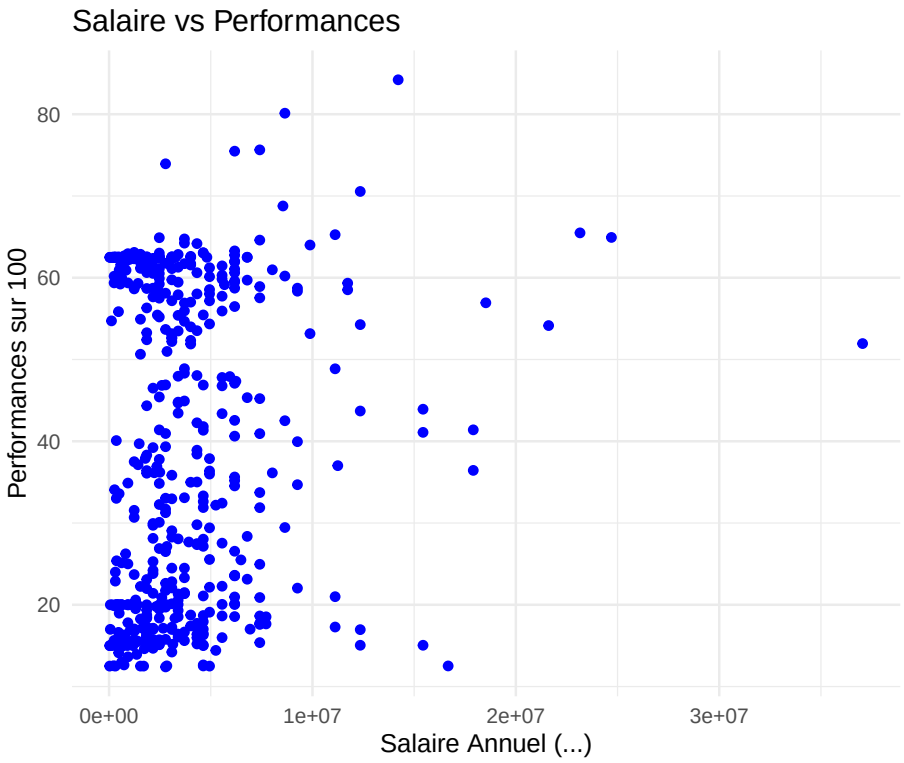


FIGURE 14 – Graphique salaire et performances

A l'inverse, nous avons ici une forte concentration des salaires dans un intervalle assez faible (entre O et un peu plus d'un million annuel) alors que les performances augmente énormément (entre 0 et 80). Cela nous pousse à penser que ces variables ne seraient pas corrélé ou alors très faiblement.

On peut donc s'intéresser aux graphiques des autres variables afin de repérer les variables qui sont susceptible d'être liée avec la variable des salaires.

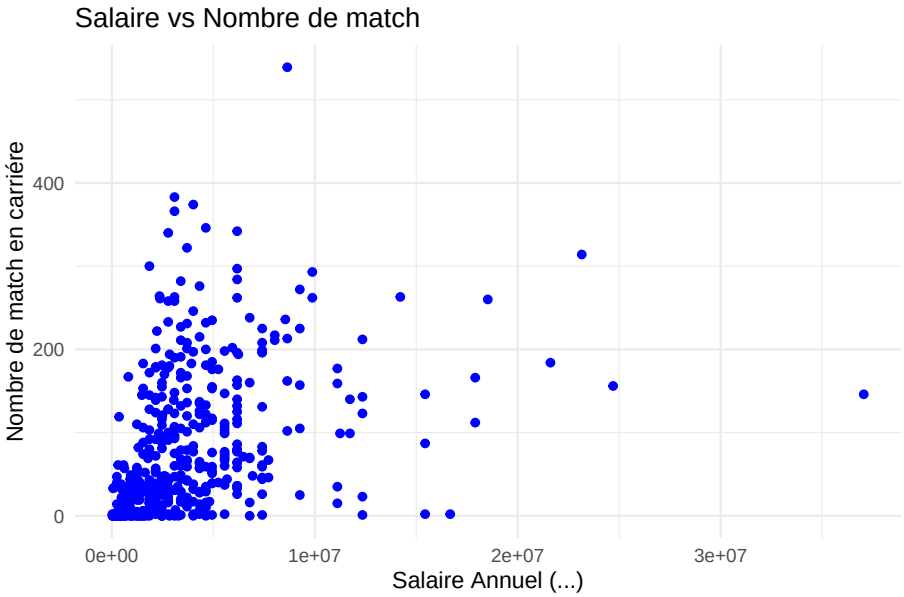


FIGURE 15 – Graphique salaire et nombre de match

On peut supposer une corrélation entre ces deux variables.

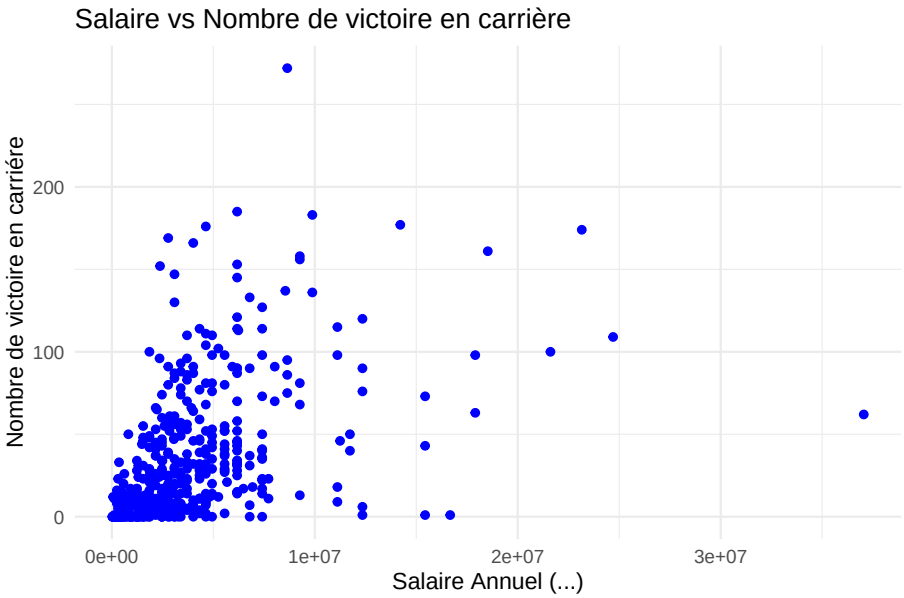


FIGURE 16 – Graphique salaire et nombre de victoire en carrière

On peut également supposer une légères corrélation entre les deux variables . On constate également la forte ressemblance entre ce graphique et le précédent, ce qui s'explique simplement : plus un joueur à joué de match plus son nombre de victoire pourra être élevée.
En effet, cela est flagrant lorsqu'on observe le graphique entre nombre de victoire et nombre de match joué.

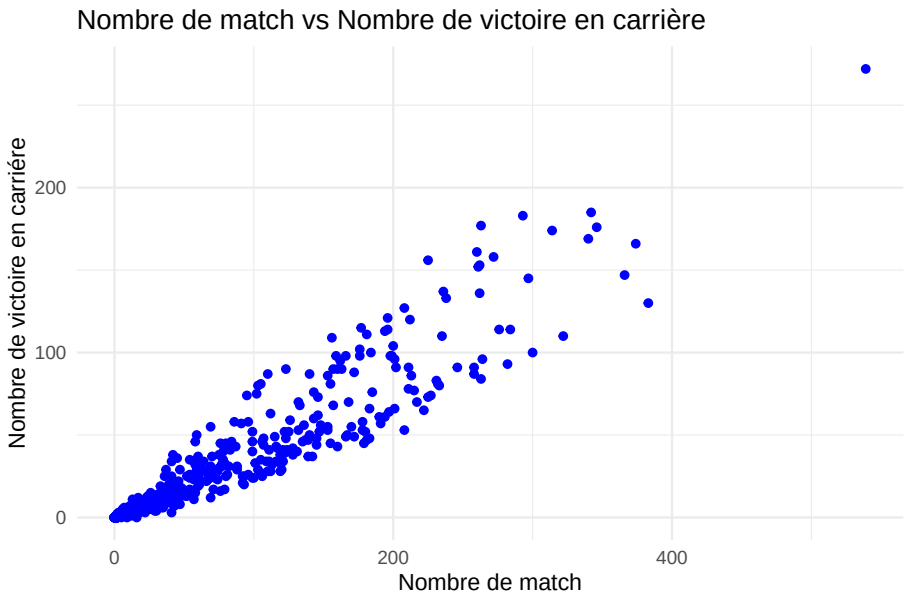


FIGURE 17 – Graphique nombre de match et nombre de wins en carrière

2.3 Matrice de corrélation - coefficient de Pearson

Pour nos 3 variables d'intérêts on remarque une cohérence entre le coefficient de corrélation et les graphiques analysés. En effet, Performances dispose effectivement d'un coefficient de corrélation très bas avec la variable salaire et la valeur marchande semble être corrélé mais pas fortement.

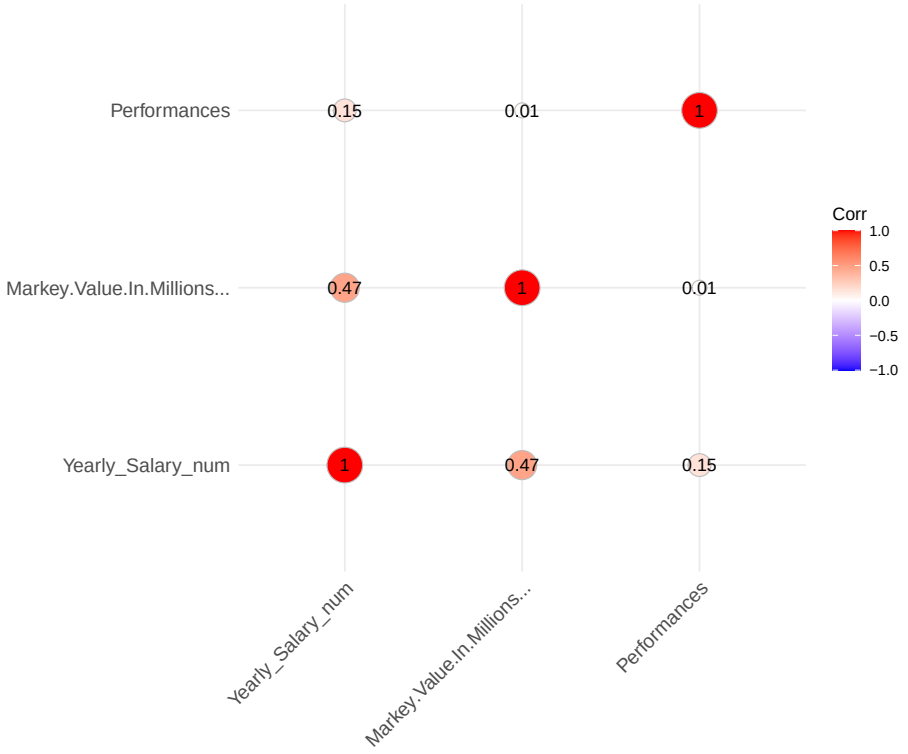


FIGURE 18 – Matrice de corrélation de nos 3 variables d'intérêts

On peut d'or et deja s'intéresser à des variables qui serait plus corrélés à la variable que l'on souhaite expliqué (salaire). Pour cela nous utlisons un programme nous donnant la matrice de corrélation des 10 variables disposant du plus haut coefficient de corrélation avec salaire.

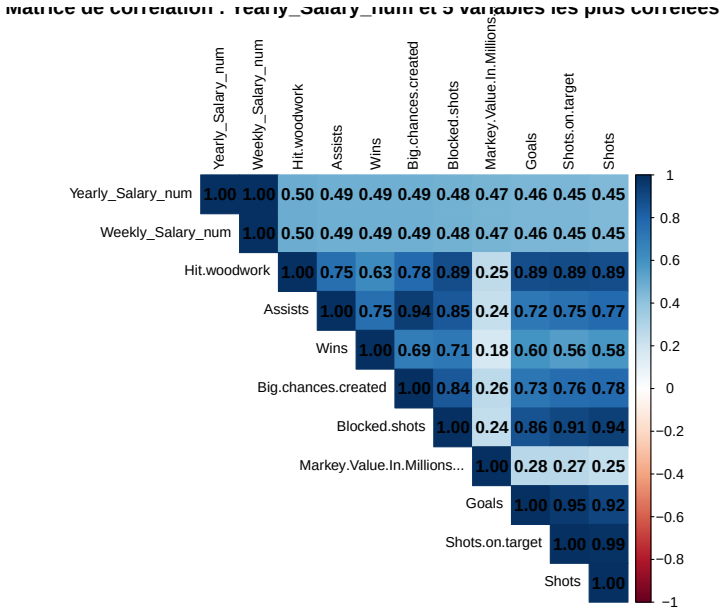


FIGURE 19 – Matrice de corrélation supérieur des 10 variables les plus corrélées avec salaire.

On retrouve bien la valeur marchande. Les variables Wins, goals et Big chanced created nous paraissent cohérentes avec la réalité sportive. Nous allons donc examiner ces variables en regardant en même temps la significativité de ces variables.

2.4 Les variables suivent-elles une loi normales ?

Shapiro-Wilk Test for Selected Variables

	Shapiro-W Statistic	p-value	Normal Distribution
Yearly_Salary_num	0.7355	0	No
Markey.Value.In.Millions...	0.5728	0	No
Performances	0.8699	0	No
Wins	0.7767	0	No
Goals	0.5084	0	No
Big.chances.created	0.6480	0	No
Appearances	0.8353	0	No

FIGURE 20 – Résultats et interprétation du shapiro test

On remarque que nos variables d'intérêts ne suivent pas de loi normales car on rejette l'hypothèse nul de normalité par le test de Shapiro-Wilk.

2.5 Test de comparaison de moyenne et de variance entre différent sous groupes : Anova

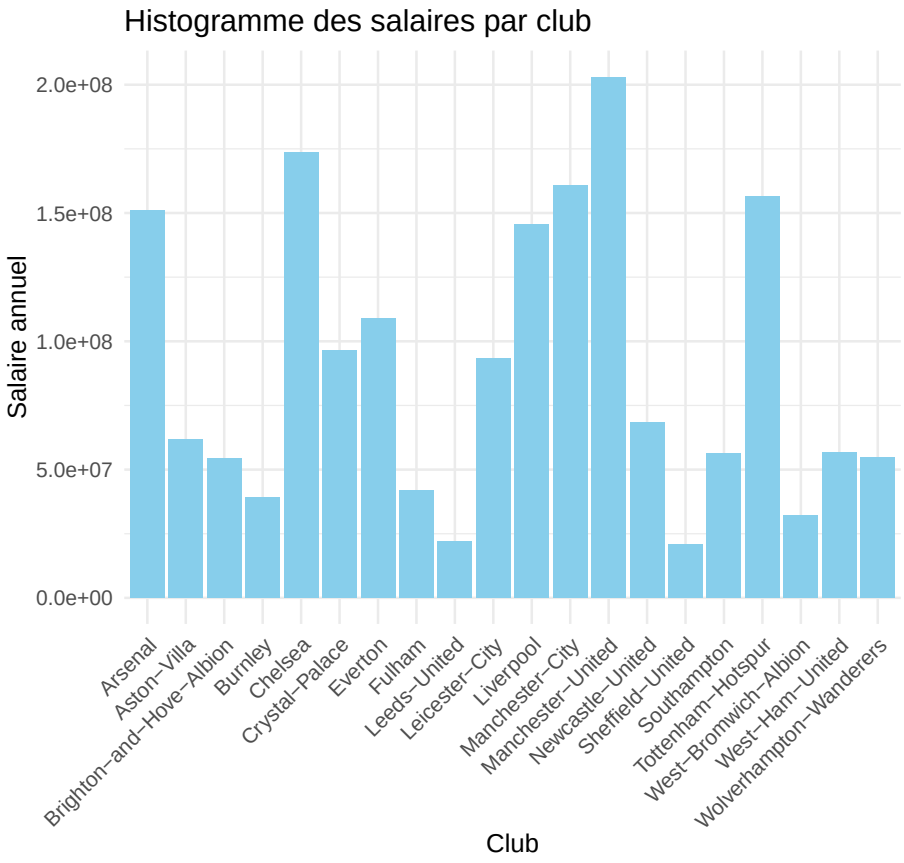
Après avoir transformé les variables en sous groupe afin d'avoir des variables catégorielles, on remarque que du point de vue de ce test la variable performance a un impact significatif sur la variable dépendante (le salaire annuel).

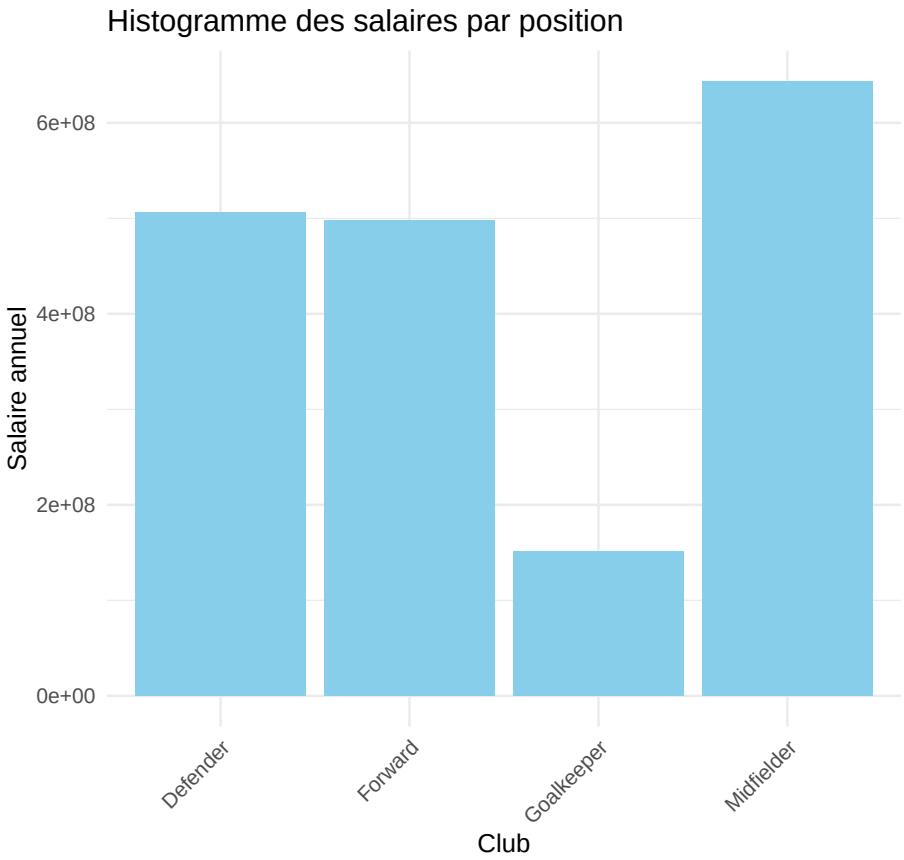
TABLE 2 – Tableau des résultats ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Performances_Cat	3	3.773584e+14	1.257861e+14	10.6086	0.0000
Wins_Cat	3	7.541754e+14	2.513918e+14	21.2019	0.0000
Performances_Cat :Wins_Cat	4	5.126184e+13	1.281546e+13	1.0808	0.3653
Residuals	474	5.620228e+15	1.185702e+13	NA	NA

2.6 Graphique complémentaire :

Nous allons réaliser des histogrammes pour visualiser la répartition des salaires en fonction des club et du poste. Cela pourra nous être utile par la suite.


Histogramme des salaires par club
Salaire annuel
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
Arsenal
Aston−Villa
Brighton−and−Hove−Albion
Burnley
Chelsea
Crystal−Palace
Everton
Fulham
Leeds−United
Leicester−City
Liverpool
Manchester−City
Manchester−United
Newcastle−United
Sheffield−United
Southampton
Tottenham−Hotspur
West−Bromwich−Albion
West−Ham−United
Wolverhampton−Wanderers
Club



On remarque une répartition cohérente des salaires en fonction du prestige et des revenus financiers des clubs. La position semble jouer un rôle également, en particulier pour les gardiens qui semblent disposer d'un salaire moindre que les autres postes.

3 Partie III : Économétrie

3.1 Estimation économétrique

Nous allons donc étudier notre modèles économétriques afin de proposer une réponse à notre problématique.
Modèle 1 : Modèle simple de régression multiple (MCO)

$$\text{Salary}_i = \beta_0 + \beta_1 \times \text{Performance}_i + \beta_2 \times \text{MarketValue}_i + \epsilon_i \quad (1)$$

Voici les données du modèles 1 :

[]

TABLE 3 – Modéle 1

	Dependent variable :
	Yearly_Salary_num
Markey.Value.In.Millions...	113,439.800***
	(9,672.866)
Performances	28,285.650***
	(7,700.017)
Constant	817,287.200**
	(361,383.600)
Observations	485
R^2	0.240
Adjusted R^2	0.236
Residual Std. Error	3,275,996.000 (df = 482)
F Statistic	75.946*** (df = 2 ; 482)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

• Lorsque la valeur marchande et les performances sont nulles (ce qui est théorique et peu probable), le salaire annuel est estimé à 817,287.

• Pour chaque augmentation de 1 million dans la valeur marchande, le salaire annuel augmente en moyenne de 113,440. Cela montre une relation positive et significative entre la valeur marchande et le salaire.

• Pour chaque augmentation d'un point dans l'indicateur de performances, le salaire annuel augmente en moyenne de 28286. Cela indique également une relation positive significative (mais peu importante) entre les performances et le salaire.

• Les coefficients pour Markey.Value.In.Millions... et Performances sont tous deux très significatifs, ce qui renforce la crédibilité de leurs impacts sur le salaire.

• Residual standard error : 3,276,000. Cela mesure l'erreur moyenne entre les valeurs observées et prédites. Une valeur élevée peut indiquer que le modèle n'explique pas bien la variation des salaires.

• Multiple R-squared : 0.2396. Environ 24% de la variance dans le salaire annuel est expliquée par les variables incluses dans le modèle. Ce n'est pas très élevé,

indiquant qu'il pourrait y avoir d'autres facteurs non inclus qui influencent le salaire.

• Adjusted R-squared : 0.2365. Cette valeur ajustée tient compte du nombre de variables dans le modèle et est légèrement inférieure à R^2 , ce qui est normal lorsque l'on ajoute des variables.

• F-statistic : 75.95 avec un $p - value < 2.2e - 16$. Cela indique que le modèle global est significatif, c'est-à-dire que les variables explicatives combinées expliquent significativement la variance du salaire.

Les résultats montrent que la valeur marchande et les performances sont des déterminants significatifs des salaires des joueurs de football dans le modèle. Cependant, l'ajustement pourrait être amélioré, et il serait utile d'explorer d'autres variables qui pourraient avoir un impact sur le salaire.

Intéressons nous maintenant aux test des propriétés indispensables à la réalisation correcte d'un modèle économétriques : on remarquera que, comparer à R, SAS dispose de moins de commande déjà concu pour réaliser des tests et afficher des graphiques cohérent. Cela nécessite plus de programmation.

1. Hétéroscédasticité : Test de Breusch-Pagan et Goldfeld et Quandt

Le test de Breusch-Pagan teste l'hypothèse nulle d'homoscédasticité (variance constante des résidus). Un p-value supérieur à 0.05 indique qu'il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle (donc les résidus serait homoscédastique), mais ici, le p-value est proche de 0.05, ce qui suggère une légère tendance vers l'hétéroscédasticité. Il serait prudent d'explorer cela davantage, car une hétéroscédasticité pourrait affecter la fiabilité des estimations des coefficients. De plus, la fonction checkheteroscedasticity nous indique qu'il y a de l'hétéroscédasticité.

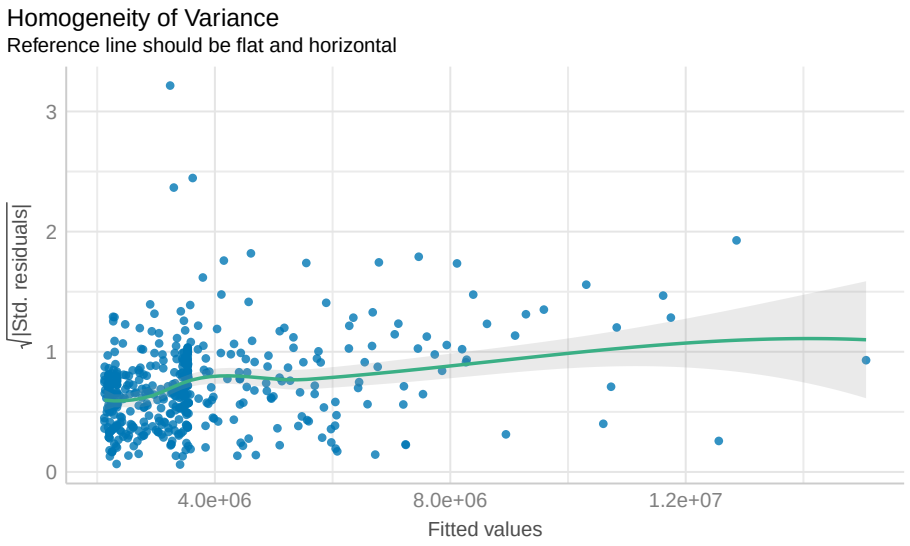


FIGURE 21 – Graphique variance des résidus et salaires

Graphiquement on remarque que la variance des résidus n'est pas de variance constante, ce qui est révélateur d'hétéroscédasticité.
On utilise le test de GQ sur la variable explicatives Performances afin de savoir si on a une relation du type

$$\sigma_\epsilon^2 = a * X_j^2$$

avec Xj les valeurs croissantes de Performances. Le résultat du test nous indique une pvalue inférieur à 0.01 indiquant que la variance des résidus augmente de manière significtaive entre les deux segments de données (l'un correspondant aux faibles valeurs et l'autre aux valeurs élevées de X). Ainsi, l'hétéroscédasticité semble liées à la variable Performances.

2. Autocorrélation des résidus : Test de Durbin-Watson et ACF

Le test de Durbin-Watson teste l'hypothèse nulle d'absence d'autocorrélation des résidus (c'est à dire que le terme d'erreurs en t n'est pas liée avec le terme d'erreur en t'). Une valeur de DW proche de 2 indique aucune autocorrélation, tandis qu'une valeur inférieure à 2 indique une autocorrélation positive. Avec une p-value très faible (0.001152) et une statistique de test DW inférieur à 2, il y a des preuves solides d'autocorrélation positive des résidus, ce qui peut indiquer que le modèle est mal spécifié ou qu'il manque des variables explicatives pertinentes.
On le remarque d'ailleurs sur le graphique des fonctions d'autocovariance.

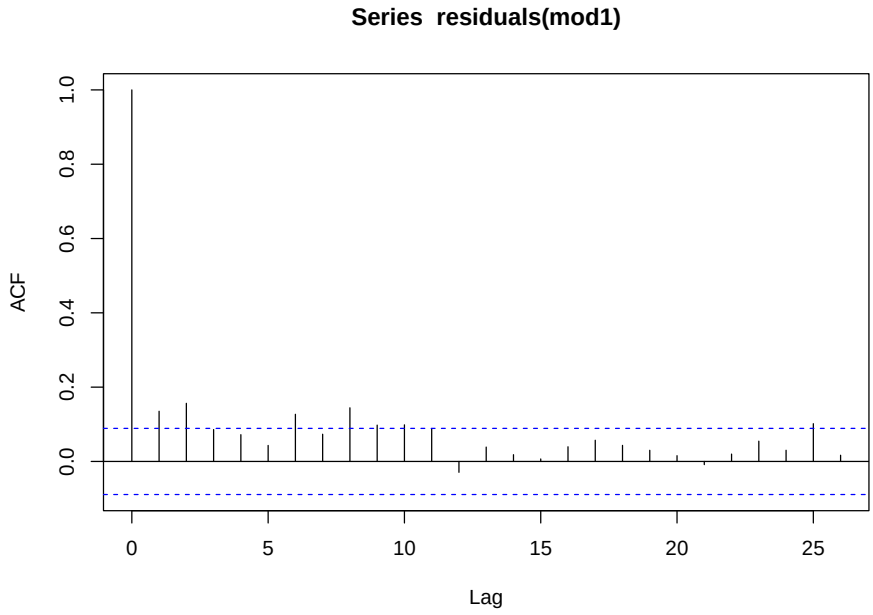


FIGURE 22 – Autocorrélation en fonction des retards (Lag)

En effet, pour 6 retard on retrouve des barres dépassant le seuil de significativité de manière positive. Ainsi, on en déduit également de l'autocorrrélation positive.

3. Variance Inflation Factor :
On veut tester la colinéarité des variables explicatives, c'est-à-dire les corrélations qui existent entre les différentes variables explicatives. Ici nous avons des résultats proche de 1 donc il n'y a pas de colinéarité préoccupante pour nos 2 variables explicatives.

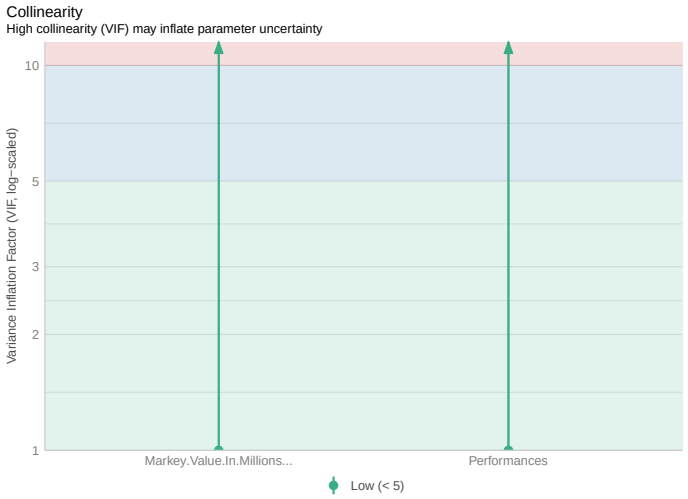


FIGURE 23 – VIF graphique

4. Les résidus suivent ils une loi normales ? Test de shapiro et graphique

Par le biais du test de Shapiro, nous avons une pvalue inférieur à 0.01 ce qui signifie que les résidus ne suivent pas une loi normales. Cela se confirme par le graphique suivant.

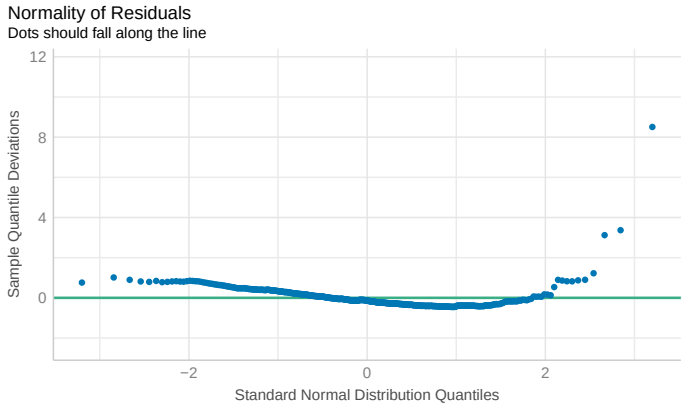


FIGURE 24 – test de normalité des résidus

En effet, le tracé bleu devrait suivre le vert si les résidus suivaient une loi normales.

Ces résultats nous permettent d'affiner nos interprétations. En effet, les ré-

sultats de notre modèle de régression montrent que la valeur marchande et les performances sont des déterminants significatifs des salaires des joueurs. Toutefois, nous rencontrons des problème sur les propriétes nécessaire à une bonne estimation du modèle : les résultats du test de Breusch-Pagan suggèrent qu'il pourrait y avoir un problème **d'hétéroscédasticité**, ce qui peut affecter la précision des estimations. De plus, le test de Durbin-Watson indique une **autocorrélation positive**, ce qui pourrait signaler un problème dans le modèle, peut-être en raison de variables explicatives manquantes. Enfin, la corrélation entre la valeur marchande et les performances est faible mais significative. Cela suggère que bien qu'il existe une relation, elle n'explique qu'une petite partie de la variance des salaires. En conclusion, nous devons corriger les problèmes d'hétéroscédasticité et d'autocorrélation positive. De plus, **d'autres variables doivent être examinés** pour améliorer le modèle !

Pour faire face aux problèmes évoquées nous allons apporter deux modifications :
- Une transformation logarithmique sur les variables salaires, valeurs marchandes et Performances qui vas permettre de réduire l'impact des valeurs extrême et d'obtenir un modèle en pourcentage.
- L'ajout des variables age et Position (sous la forme de variable dummy) car la valeur age étant intuitivement un facteur explicatif pertinent de la variation des salaires (et non compris dans le calcul de la note de Performance), et la variable Position sous forme de dummies (variable intialement characters) vas permettre de capturer la variation de salaire liées aux rôles spécifiques des joueurs.

Modèle 2 :

$$\begin{aligned}\log(\text{Salary})_i = \beta_0 + \beta_1 \times \log(\text{Market Value})_i \\ + \beta_2 \times \log(\text{Performance})_i + \beta_3 \times \text{Age}_i \\ + \beta_4 \times \text{Position}_i + \epsilon_i \end{aligned} \tag{2}$$

Lors de la transformation de la variable Position en dummies, la Position Defender est utilisée en tant que variable référence afin d'éviter une situation où toutes les variables sont fortement corrélées entre elles. Cela à une incidence sur la façon d'interpréter les valeurs des coefficients que nous avons ci-dessous :

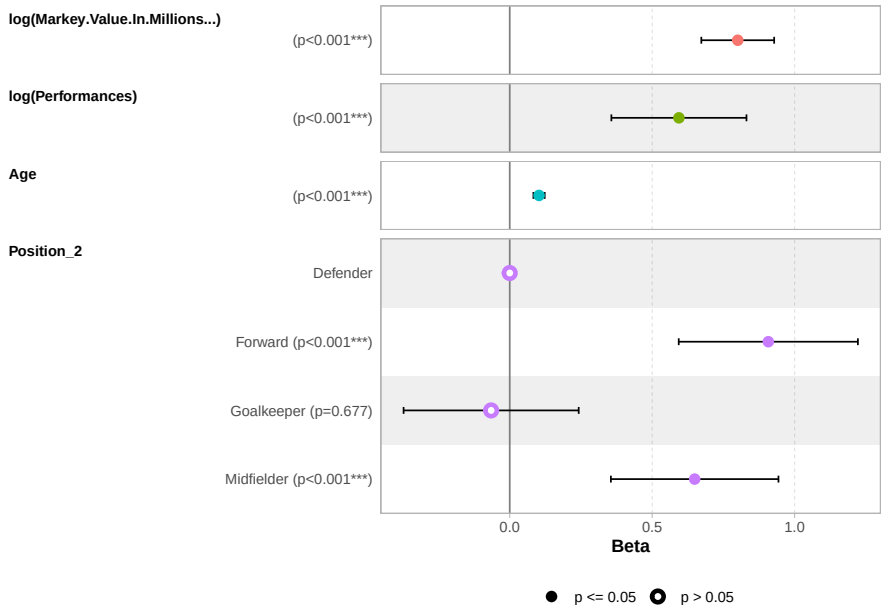


FIGURE 25 – Graphique des valeurs des coeffients, leurs pvalue et leur intervalle de confiance

• **log(Markey.Value.In.Millions. . .)** : 0.80028
Ce coefficient positif signifie qu'une augmentation de 1 % de la valeur marchande prévue entraîne une augmentation de 0,8 % du salaire annuel moyen. Ce prédicteur est statistiquement **très significatif**, avec une grande influence sur la variable dépendante.

• **log(Performances)** : 0.59383 Ce coefficient positif indique qu'une augmentation de 1 % des performances entraîne une augmentation de 0,59 % du salaire, avec un niveau de **signification élevé**.
Cela nous confirme l'importance de ces deux variables pour estimer le salaire.

• **Age** : 0.10295 Une année supplémentaire d'âge est associée à une augmentation de 10,3 % du salaire, toutes choses égales par ailleurs (ce coefficient est **très significatif**). Cela pourrait suggérer que l'expérience ou la longévité dans la ligue contribue à des salaires plus élevés.

• **Position** (variable catégorielle avec "Defender" comme référence) :

• **Forward** : 0.90758
Être attaquant (plutôt que défenseur) est associé à une augmentation de 90,8 % du salaire, un effet statistiquement **très significatif.**

• **Goalkeeper** : -0.06523

Être gardien de but (plutôt que défenseur) est associé à une légère baisse du salaire (-6,5 %), mais cet effet n'est **pas statistiquement significatif**.

• **Midfielder** : 0.64912

Être milieu de terrain (plutôt que défenseur) est associé à une augmentation de 64,9 % du salaire, avec un effet **significatif**.
L'on retrouve des résultats en corrélation avec l'histogramme 2.6 concernant l'influence de la Position sur le salaire d'un joueur.

• **Residual standard error** : 0.8428

L'écart-type des résidus est de 0,8428, indiquant à quel point les valeurs observées diffèrent des valeurs prévues.

• **Multiple R-squared** : 0.4475

Le modèle explique environ 44,75 % de la variabilité des salaires annuels en fonction des prédicteurs, ce qui est modérément élevé. On remarque que ce nouveau modèle à permis d'expliquer **une part plus importante de la variabilité** des salaires annuel (seulement 24% pour le modèle 1)

• **Adjusted R-squared** : 0.4406

Avec une valeur ajustée de 0,4406, ce modèle a permis de réduire la dispersion des résidus, signalant des estimations plus stables et mieux adaptées aux données.

• **F-statistic** : 64.54 ($p < 2.2e\text{-}16$)

Ce F-statistique très significatif montre que le modèle, dans son ensemble, a une valeur explicative et que les variables indépendantes ont un effet combiné sur la variable dépendante.

Ce modèle montre que la valeur marchande, les performances, l'âge et la position du joueur expliquent de manière significative les variations du salaire annuel des joueurs de Premier-League. La valeur marchande et la position (en particulier pour les attaquants et les milieux de terrain) sont des prédicteurs fortement liés aux salaires. Cependant, il reste une portion de la variance non expliquée par ces variables, suggérant qu'il pourrait y avoir d'autres facteurs influençant les salaires, ou qu'une transformation du modèle pourrait encore l'améliorer. On pense notamment à des variables non présentes dans nos bases de données comme le palmarès des joueurs, la durée du contrat qui les unis avec leur club ou bien le coûts d'achat du joueur pour le club qui peut avoir une influence sur sa capacité à payer un salaire plus ou moins élevées.

Quels sont les conséquences sur les propriétés des termes d'erreurs de ce nouveau modèle ?

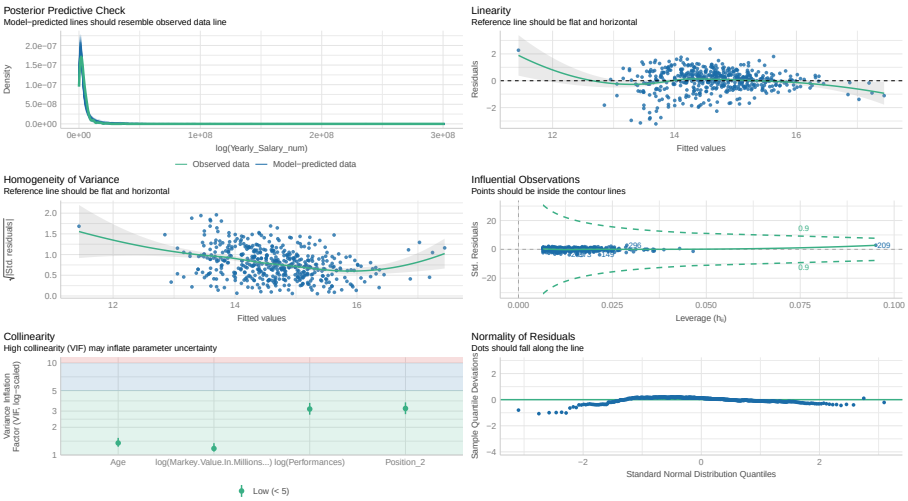


FIGURE 26 – Visualisation graphique des propriétés statistiques

Nous allons utiliser ces graphiques pour soumettre nos impressions puis les confirmer via les tests statistiques :
- Sur le graphique concernant l'homogénéité de la variance on observe une tendance dans la dispersion de la variance des résidus ce qui est révélateur **d'hétéroscedasticité des résidus**. En effet, les tets statistiques confirment cette hypothése (test de BP et check heteroscedasticity du package performance).
- Sur le graphique de la Colinéarité, l'ensemble des indicateurs VIF sont en dessous de 5 ce qui signifie que la **multicolinéarité est faible et acceptable**.
- Sur le graphique de la normalité des résidus on remarque les résidus ne suivent pas une loi normales mais en sont assez proche puisque la partie entre -2 et 2 suit la ligne de référence. Cela indique que la **non-normalité des résidus** aura une influence que très légère sur les tests de significativité des coefficient. Les tests statistiques confirment cette intuition.
- Ces graphiques ne nous permetre pas de formuler des hypothése sur l'autocorrélation des résidus mais le test de Durbin Watson ainsi que la commande check autocorrelation confirme qu'il y a de l'**autocorrélation positive** au sein des résidus.
Ainsi ce nouveau modèle n'apporte pas de correction sur les propriétés statistiques mais nous permet de disposer d'un modèle **plus représentatif du salaire** et qui en explique une part plus importante. **Apportons désormais quelque corrections.**

3.2 Tentatives de corrections

3.2.1 Erreur robuste de Newey et West

Nous avons choisis cette correction car elle est efficace en présence d'hétéroscédasticité et d'autocorrélation. Elle permet de corriger les estimateurs des variances et des covariances des estimateurs des MCO. Ansi, les valeurs estimées des coefficients sont didentiques à celles figurant dans le modéle 2 initiale mais les écarts-types des coefficients (et donc les t de Student) sont différents.

Comparaison des Statistiques de Test et p−values (Classique vs. Newey−West)

Variable	Estimate	Std..Error..Classique.	t.value..Classique.	p.value..Classique.	Std..Error..Newey.West.	t.value..Newey.West.	p.value..Newey.West.
(Intercept)	7.47655315	0.4598	16.2601	0.0000	0.4875	15.3378	0.0000
log(Markey.Value.In.Millions...)	0.80027647	0.0651	12.3005	0.0000	0.0649	12.3224	0.0000
log(Performances)	0.59382507	0.1207	4.9209	0.0000	0.1186	5.0056	0.0000
Age	0.10294955	0.0101	10.2290	0.0000	0.0142	7.2666	0.0000
Position_2Forward	0.90757998	0.1600	5.6719	0.0000	0.1350	6.7237	0.0000
Position_2Goalkeeper	−0.06522537	0.1564	−0.4171	0.6768	0.1582	−0.4123	0.6803
Position_2Midfielder	0.64911999	0.1497	4.3374	0.0000	0.1334	4.8670	0.0000

FIGURE 27 – Comparaison modèle 2 sans et avec correction de NW

- Coefficients sensibles à l'hétéroscédasticité/auto-corrélation : L'intercept et la variable Age montrent des erreurs standard significativement plus élevées après la correction, ce qui suggère une variabilité non prise en compte par les erreurs classiques.
- Coefficients peu affectés : log(Markey.Value.In.Millions...), log(Performances), et Position_2Goalkeeper ne changent que très peu après la correction, indiquant qu'ils sont robustes aux effets d'hétéroscédasticité et d'autocorrélation

Effet global : La correction Newey-West apporte une meilleure estimation de l'incertitude sur certains coefficients, ce qui rend les tests de significativité plus fiables, surtout pour les coefficients où des problèmes potentiels d'hétéroscédasticité et d'autocorrélation sont présents.

3.2.2 Les moindres carrés généralisés

Table 4 – Modéle 2 MCG

	Dependent variable :
	log(Yearly_Salary_num)
log(Markey.Value.In.Millions...)	0.720***
	(0.064)
log(Performances)	0.507***
	(0.118)
Age	0.105***
	(0.010)
Position_2Forward	0.840***
	(0.164)
Position_2Goalkeeper	−0.145
	(0.178)
Position_2Midfielder	0.590***
	(0.159)
Constant	7.961***
	(0.455)
Observations	485
Log Likelihood	−606.270
Akaike Inf. Crit.	1,236.541
Bayesian Inf. Crit.	1,286.576
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Les résultats montrent que le modèle MCG capture l'hétéroscédasticité et une faible autocorrélation dans les données, améliorant la fiabilité des coefficients. Les prédicteurs log(Markey.Value.In.Millions...), log(Performances), et Age restent significatifs et positivement associés au salaire, tandis que la position influence le salaire de manière significative pour les attaquants et les milieux de terrain.

4 Conclusion

4.1 Introspection

Nous pensons que la création de la variable Performances qui inclue une multitude de statistique peut-être à l'origine des différents problèmes que nous rencontrons. En particulier, la forte corrélation entre nombre de match et âge et la présence de l'indice nombre de match dans le calcul de performances (indirectement) pourrait être à l'orgine de multicolinéarité. De plus, la faible quantité de données dont nous disposions (peu de joueurs) a probablement eu un impact sur la fiabilité de notre modèle. Cela aurait pu être plus pertinent de réaliser cette étude sur un ensemble plus important de joueur. Enfin, certaines données ayant certainement un impact sur le salaire sont absentes de nos données : durée du contrat liant les joueurs à leur club, palmarès du joueur, lien de parenté avec des anciens joueurs... Un modèle contenant ces données serait plus pertinent. En conclusion, pour améliorer ce modèle il faudrait :
- Approfondir nos connaissances en matières de correction d'hétéroscédasticité et d'autocorrélation, nous avons eu du mal à interpréter les corrections faites.
- Repenser la manière de quantifier la performances des joueurs
- Trouver des bases de données contenant les données manquantes

4.2 Commentaires

Nous avons souhaité faire varier les manière de présenter les résultats afin de développer notre manipulation de R. C'est donc pour cela que les résultats de contenu similaire sont présentés avec des graphiques différent. Nous avons utilisé parfois des programmes R nous donnant directement le code LaTeX créant le tableau souhaité (exemple : 3).