POLYTECHNIQUE
MONTRÉAL

École Polytechnique de Montréal

Département Mathématiques et génie Industriel

MTH2302D

Probabilités et statistique

Laboratoire #2 : Travail de session 2/2

Soumis par

Joe Abdo (1939689)

Kenza Rami Yahyaoui (1903555)

Horby Brisseau (1881786)

Groupe #3

2 décembre 2019

Table des matières

1. Contexte d'acquisition des données

La NBA ou National Basketball Association est la ligue de basketball nord-américaine réunissant 29 équipes Américaines et une unique Canadienne. Considérée aujourd'hui comme la ligue de basketball la plus prestigieuse au monde.
Cette ligue porte avec elle plus de 50 ans d'expertise en enregistrement et études de statistiques. L'enregistrement a commencé en 1950 pour les rebonds et les minutes de jeux mais ce n'est qu'en 1979 que tous les aspects du jeu furent enregistrés avec le nombre de 3 points et nombre de paniers par match.
Avec la fin de la saison 2018-2019 en NBA, nous revenons avec les statistiques de cette saison splendide qui a vu la fin du règne des Golden States Warrios. Mais aussi l'émergence de la force canadienne grâce au Toronto Raptors qui a ramené la coupe à la maison.
Nos données réunissent les statistiques de la saison précédente des 50 meilleurs joueurs classées par le nombre de points par match ou « Point per Game » représenté dans notre tableau avec le symbole PPG.
Sur ceux-ci, se basent les bookers de paris surtout à Las Vegas [2] pour la saison à venir 2019-2020.
Il serait donc intéressant de pouvoir comprendre les bases de ces études.

2. Source

Les données du fichier Excel, ont été prise du site de la National Basketball Association [1], quelques colonnes ont été retirées (jugées inutiles pour la suite du travail et desquelles on peut disposer).

3. Glossaire [4]

PPG - Points Per Game – Point Par Partie
NAT – Nationality - Nationalité
MIN – Minute Per Game – Minutes Par Partie
FGM – Field Goals Made – Panier Marqués
FGA – Field Goals Attempted – Panier Tentés
3PM – 3 Point Field Goals Made – 3 Point Marqués
REB – Rebounds – Rebonds
AST – Assists – Passes décisives
STL – Steals – Interception
BLK – Blocked Shots – Paniers Bloqués
TOV – Turnovers - Perte De Balle
EFF – Efficiency – Efficacité

4. Explication des données

Les cinquante joueurs les plus prolifiques, classés selon leur nombre de points par match sur la saison entière, sont sujets de notre étude. Leurs chiffres dans différents secteur du jeu sont nos variables aléatoires (les acronymes des noms de variables sont en anglais, les mots que nous utilisons sont les appellations françaises) :

Les variables de nombre de 3 points sur la saison et la nationalité sont des variables discrètes (variable qualitative pour celle de la nationalité), toutes autres variables sont des variables continues car c'est une variable qui peut prendre n'importe quelle valeur entre sa valeur minimale et sa valeur maximale.

5. Questions Ouvertes

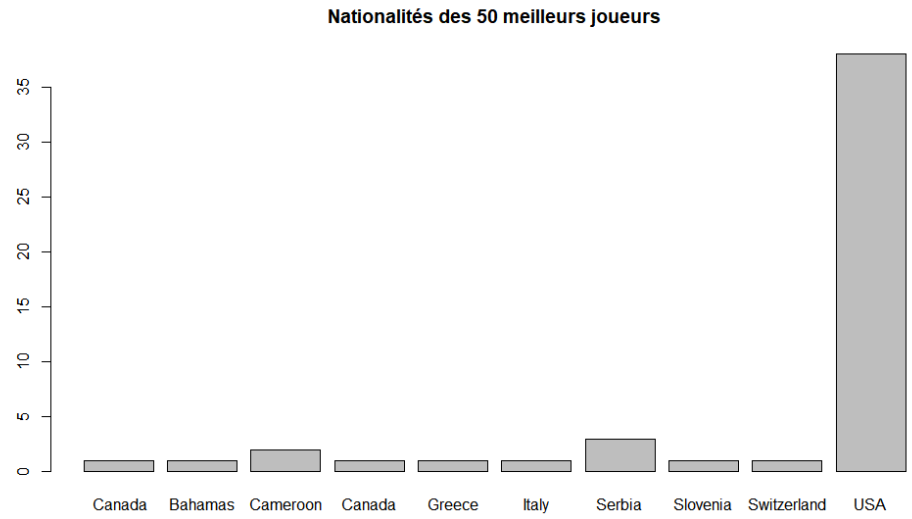
I. La nationalité d'un joueur donne-t-elle avantage pour être parmi les élites de la NBA?

II. Est-ce qu'un joueur ayant plus de perte de balle a-t-il toujours moins d'efficacité, existe-t-il une relation entre ses deux variables?

III. Ce peut-il que l'augmentation des minutes jouées par match diminue l'efficacité d'un joueur ?

6. Réponse aux questions et analyse des données

I. La nationalité d'un joueur donne-t-elle avantage pour être parmi les élites de la NBA?



Répartition des joueurs selon leur nationalité

Il est clairement visible que les joueurs du top 50 sont en grande majorité des joueurs américains.

II. Est-ce qu'un joueur ayant plus de perte de balle a-t-il toujours moins d'efficacité, existe-t-il une relation entre ses deux variables? [5]

Afin de répondre à la deuxième question émise dans la partie 1 du projet, nous avons conclu qu'il était préférable d'utiliser uniquement la technique de régression et de tests paramétriques usuels. Ainsi, en rassemblant les données des pertes de balles et de l'efficacité des cinquante joueurs, nous avons obtenu ce nuage de points :

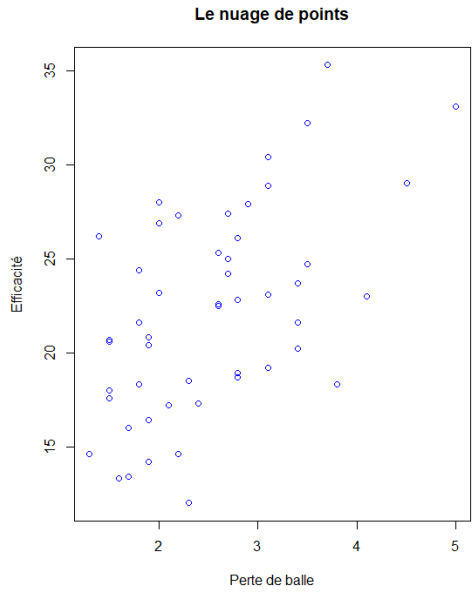


Figure 1 : Nuage de points entre la perte de balle et l'efficacité de chaque joueur

Ainsi, à l'aide de la fonction " lm(Y ~ X) " où X sont les valeurs liées aux pertes de balles des joueurs et Y les valeurs liées à l'efficacité de chaque joueur, nous pouvons observé la droite représentant la corrélation par régression linéaire entre ces données :

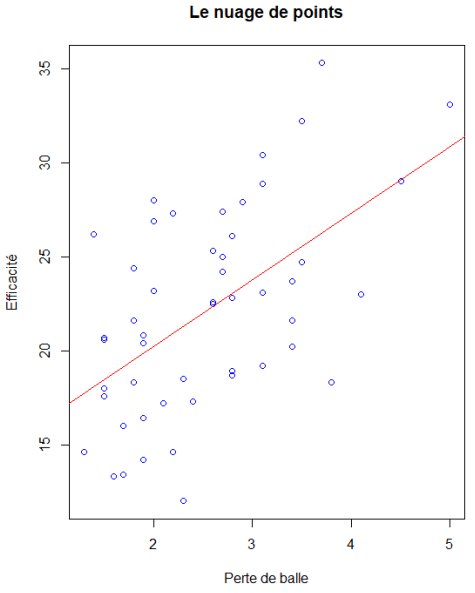


Figure 2 : Corrélation entre la perte de balle et l'efficacité de chaque joueur par régression linéaire

À première vue, il semble bien que la perte de balle ait un impact sur l'efficacité des joueurs. Toutefois, nous pouvons observer plusieurs valeurs s'éloignant de la droite de régression. Ainsi, il serait intéressant de vérifier la corrélation qui semblerait exister entre ces valeurs, afin de valider l'observation précédente. De plus, nous voudrions savoir, à quel point la perte de balle affecte l'efficacité des joueurs.

Afin de répondre aux questions précédentes et d'être rigoureux sur nos conclusions, nous avons décidé d'utiliser le test d'hypothèse de Student testant $B_i = 0$:

Pour B_0 :

$$\boldsymbol{H_0: \beta_0 = 0}$$
$$\boldsymbol{H_1: \beta_0 \neq 0}$$

Pour B_1 :

$$\boldsymbol{H_0: \beta_1 = 0}$$
$$\boldsymbol{H_1: \beta_1 \neq 0}$$

**Nous nous intéresserons plus particulièrement aux tests faits sur $\boldsymbol{\beta_0}$, car il représente le coefficient de régression*

À partir de ce test, on pourra arriver à la conclusion qu'il existe bel et bien ou non une corrélation linéaire entre les valeurs lorsqu'on rejettera l'hypothèse **H_0**. Donc, pour tester ces hypothèse, nous avons utilisé la fonction " summary(regLin)", où " regLin " représente la régression linéaire entre ces données :

```
> summary(regLin)

Call:
lm(formula = donnees$EFF ~ donnees$TOV)

Residuals:
    Min      1Q  Median      3Q     Max
-9.2690 -3.5370 -0.1509  2.9874  9.0720

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.1221     2.0634   6.360 7.06e-08 ***
donnees$TOV   3.5421     0.7719   4.589 3.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.582 on 48 degrees of freedom
Multiple R-squared:  0.3049,    Adjusted R-squared:  0.2905
F-statistic: 21.06 on 1 and 48 DF,  p-value: 3.216e-05
```

Figure 3 : Résultats du test d'hypothèse

À partir de ces résultats, nous trouvons les estimations suivantes :

$$\boldsymbol{\beta_0 \approx 13,1221}$$
$$\boldsymbol{t_0 \approx 13,1221 \div 2,0634 \approx 6,360}$$
$$\boldsymbol{valeur - p \approx 7,06e^{-8} \approx 0,002368}$$

Puisque la valeur-p est très petite (presque nulle), nous pouvons rejeter l'hypothèse $\mathbf{H_0}$. Ainsi, on peut conclure que le coefficient de régression n'est pas nul, ce qui implique qu'il existe bel et bien une corrélation entre la perte de balle et l'efficacité d'un joueur. Par contre, on peut aussi observer le coefficient de corrélation $\mathbf{R^2}$:

Lorsqu'on prend en compte les valeurs résiduelles

$$\boldsymbol{R^2 \approx 0,3049}$$

Lorsqu'on ne prend pas en compte les valeurs résiduelles

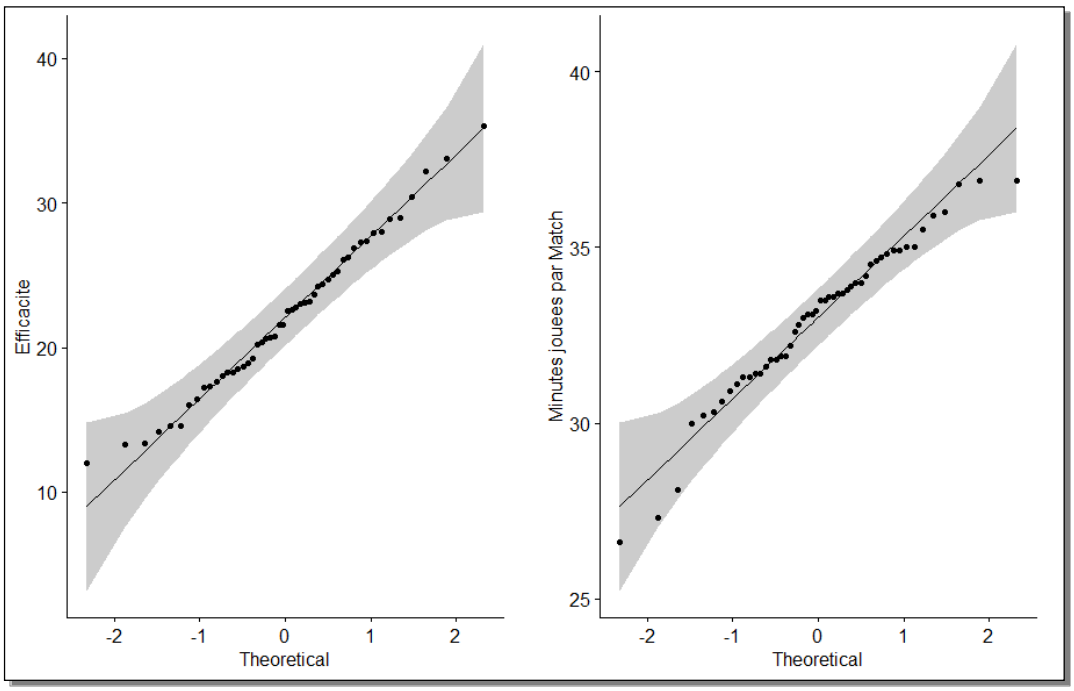
$$\boldsymbol{R^2 \approx 0,2905}$$

Selon ces résultats, il n'existerait qu'une faible corrélation entre la perte de balle et l'efficacité d'un joueur dû la faible valeur de $\mathbf{R^2}$.

Donc, à partir du nuage de points des valeurs de perte de balle et de leur efficacité durant la saison de chaque joueur de la NBA que l'on a obtenu, nous avons pu supposer qu'il existerait peut-être une corrélation entre ces variables et qu'elle pourrait être linéaire. Puis, nous avons poussé notre recherche par un test d'hypothèse sur le coefficient de régression et nous sommes arrivés à la conclusion qu'il existe bel et bien une corrélation linéaire entre les deux variables. Par contre, nous avons pu aussi déterminer que la perte de balle n'a pas un très grand impact sur l'efficacité d'un joueur de basket professionnel à cause de la faible valeur du coefficient de corrélation entre les variables étudiés.

III. Ce peut-il que l'augmentation des minutes jouées par match diminue l'efficacité d'un joueur? [7]

Nous cherchons la relation entre les deux variables minutes jouées par match et efficacité pour ce nous avons choisi d'effectuer une analyse complète, commençons par s'assurer de la normalité des données pour pouvoir effectuer ensuite des tests de corrélation paramétriques sur ces variables.



Tests de normalité sur les variables « efficiency » et « mpg »

```
> shapiro.test(data$efficiency);

        Shapiro-Wilk normality test

data:  data$efficiency
W = 0.98579, p-value = 0.8049

> shapiro.test(data$mpg);

        Shapiro-Wilk normality test

data:  data$mpg
W = 0.96378, p-value = 0.128
```

Test de Shapiro-Wilk pour la normalité des variables « efficiency » et « mpg »

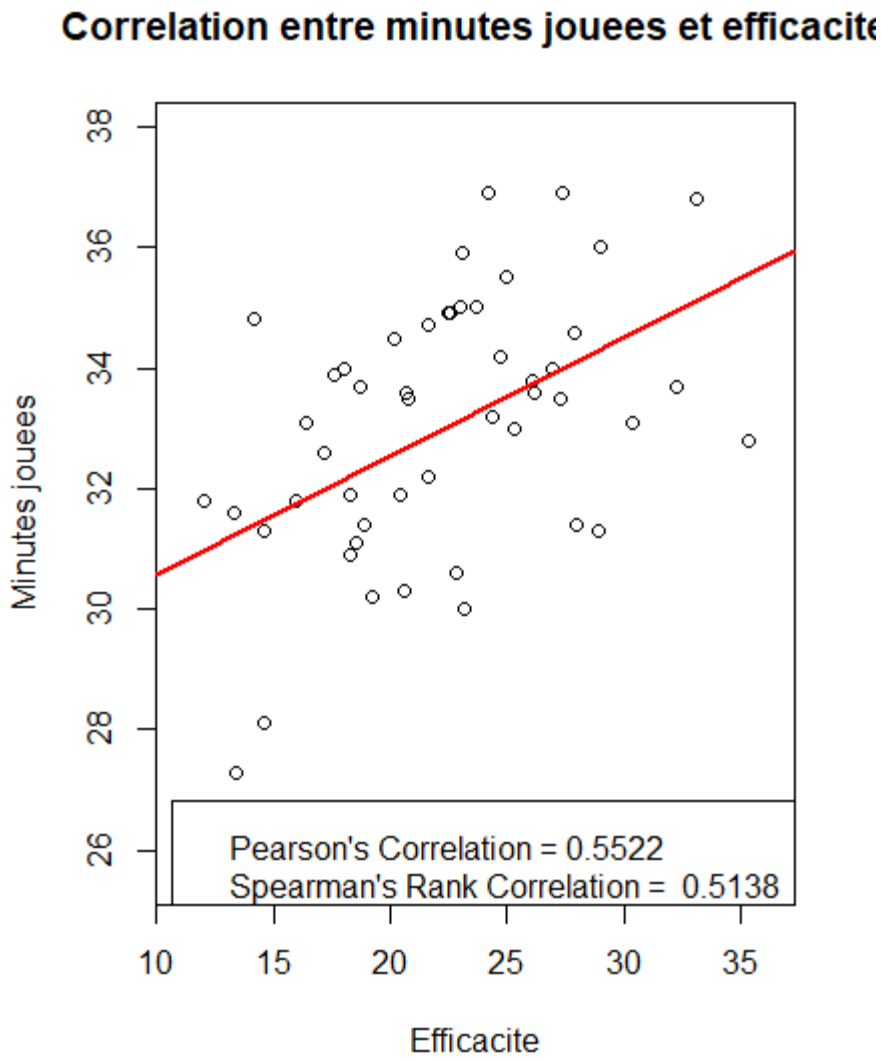
Voir le fichier « 1939689_1903555_1881786 – Question3 .r »
(Contient l'installation de librairies externes pour la représentation des graphes)

On voit donc bien que ces deux variables ont une répartition normale, de plus le test de Shapiro vient le montrer en effet avec des valeurs de « p » supérieurs à 0.05 (ou 5%). Ce test qui prend comme hypothèse nulle que les données suivent une distribution normale.
Dans notre cas on choisit un alfa de 5%.
On dit donc que la distribution des variables ne s'éloigne pas de la distribution normale, car p-value > alfa
On peut donc effectuer des tests de corrélation paramétriques, dont l'hypothèse nulle est l'absence de relation entre les deux variables. Dans notre cas les tests donnent les résultats suivants :

```
> cor.test(data$efficiency, data$mpg, method = "pearson");

        Pearson's product-moment correlation

data:  data$efficiency and data$mpg
t = 3.667, df = 48, p-value = 0.0006136
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2178158 0.6601854
sample estimates:
      cor
0.4678014

> cor.test(data$efficiency, data$mpg, method = "pearson", conf.level =
 0.9);

        Pearson's product-moment correlation

data:  data$efficiency and data$mpg
t = 3.667, df = 48, p-value = 0.0006136
alternative hypothesis: true correlation is not equal to 0
90 percent confidence interval:
 0.2611345 0.6334627
sample estimates:
      cor
0.4678014
```

Test de corrélations pour les variables « efficiency » et « mpg »

Les résultats des coefficients varient entre -1 (indique une corrélation négative) et 1 (corrélation positive). Ces résultats sont plus proches de 1 donc on a une corrélation positive entre ces deux variables. De plus les p-value sont inferieures a alfa (5%) on rejette donc l'hypothèse nulle et on accepte donc que ces variables sont liées.
Pour en être bien sur nous allons trouver l'intervalle de confiance pour la corrélation de Pearson.
Cet intervalle est donné dans le test de Pearson a 95% de confiance on trouve un intervalle [0.2611345; 0.6334627]



Trace des deux variables « efficiency » et « mpg »

7. Conclusion

L'analyse des données fournies par la NBA a permis de répondre aux questions que nous nous sommes posée au début de notre travail. On peut donc affirmer avec certitude :

I. Être de nationalité américaine donne une plus grande chance d'être parmi les élites de la NBA.

II. Les tests sur les variables de pertes de balles par match et efficacité donnent comme résultat que les deux variables pourraient être liées et il pourrait exister une relation de régression linéaire entre elles.

III. Les variables « mpg » et « efficiency pourraient avoir une relation de régression linéaire mais les coefficient de corrélation sont quand même faibles.

8. Références

[1]"Season Leaders", *NBA Stats*, 2019. [Online]. Available: https://stats.nba.com/leaders/. [Accessed: 16- Sep- 2019].

[2]"NBA Odds, 2019 USA Basketball Lines, NBA Betting | Online Vegas Odds NBA", *MyBookie Online Sportsbook*, 2019. [Online]. Available: https://mybookie.ag/sportsbook/nba/. [Accessed: 18- Sep- 2019].

[3]"FAQ", *NBA Stats*, 2019. [Online]. Available: https://stats.nba.com/help/faq/. [Accessed: 18- Sep- 2019].

[4]"Lexique du basket-ball", *Fr.wikipedia.org*, 2019. [Online]. Available: https://fr.wikipedia.org/wiki/Lexique_du_basket-ball/. [Accessed: 20- Sep- 2019].

[5]"Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA", *Sthda.com*, 2019. [Online]. Available: http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r#kendall-correlation-formula. [Accessed: 01- Nov- 2019].

[6]"Correlation Test Between Two Variables in R - Easy Guides - Wiki - STHDA", Sthda.com, 2019. [Online]. Available: http://www.sthda.com/english/wiki/wiki.php?id_contents=7312.. [Accessed: 06- Nov- 2019].

[7]"Linear Regression R", *DataCamp Community*, 2019. [Online]. Available: https://www.datacamp.com/community/tutorials/linear-regression-R. [Accessed: 12- Nov- 2019].