Charles ODEND'HAL Groupe 4

Sujet : Le pourcentage de réussite au tir à 3 points des joueurs de NBA

1ère partie

« Je déclare sur l'honneur que ce mémoire a été rédigé de ma main, sans aide extérieure non autorisée, qu'il n'a pas été présenté auparavant pour évaluation et qu'il n'a jamais été publié, dans sa totalité ou en partie. Toutes parties, groupes de mots ou idées, aussi limités soient-ils, y compris des tableaux, graphiques, cartes etc. qui sont empruntés ou qui font référence à d'autres sources bibliographiques sont présentés comme tels, sans exception aucune. »

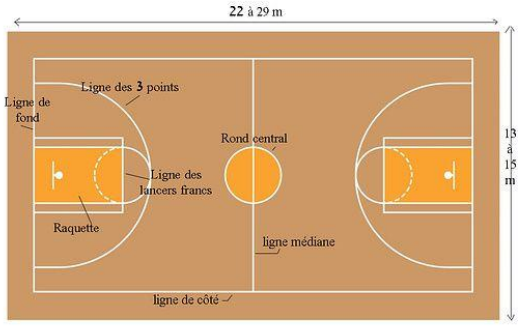
1. Introduction

La National Basketball League (NBA) est la principale ligue de basketball au monde. Chaque année, elle regroupe 30 équipes canadiennes et américaines pour 82 matchs de saison régulière. A l'issue de celle-ci, le classement détermine les équipes qualifiées pour le tournoi des 'playoffs' à l'issue duquel le vainqueur deviendra champion NBA. Au-delà de la forte compétitivité des équipes et du spectacle proposé à chacune des rencontres, la NBA est un championnat très plaisant à suivre pour les observateurs férus de statistiques. En effet, à l'image des autres ligues américaines majeures, comme la NFL ou la MLB, les équipes intègrent l'analyse des données dans leurs stratégies depuis la fin des années 2000, très en avance sur de nombreux autres sports internationaux comme le football, le rugby ou le tennis. Cette approche statistique est particulièrement relayée dans les médias et constitue le critère principal de performance d'un joueur de basketball. Contrairement à d'autres sports, les récompenses individuelles comme le MVP (Most Valuable Player) sont accordées presque exclusivement au vu des moyennes de points, rebonds et passes décisives par match, avec presque aucune considération pour le nombre de victoires de l'équipe. Cette importance accordée à la statistique m'a permis de constituer ma base de données de manière très sélective. L'échantillon retenu comporte 1325 observations correspondantes au statistiques sur une saison de tous les joueurs de NBA en 2018-2019, 2021-2022 et 2022-2023. Je n'ai pas pu étudier les données des saisons 2019-2020 et 2020-2021 car celles-ci ont été écourtées en raison de la pandémie de COVID-19, faussant les observations sur ma variable des minutes jouées par saison. Toutes mes observations sont issues du site spécialisé américain basketball-reference.com.

2. Définitions et statistiques descriptives

A. Variable endogène

La variable que je cherche à expliquer dans ce projet est le pourcentage de réussite aux tirs à 3 points des joueurs de NBA sur une saison. Au basketball, un panier marqué comptabilise 3 points pour son équipe lorsqu'il est lancé depuis la zone située strictement derrière la ligne des 3 points, qui prend une forme d'aimant dont la distance varie entre 6m75 et 7m23 du panier.



Un terrain de basketball

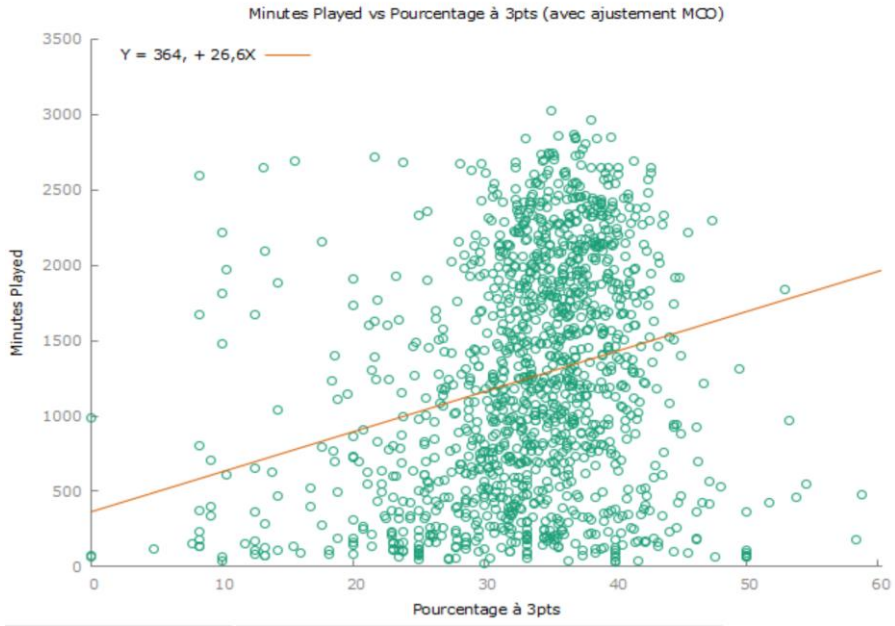
Le tir à 3 points est un des gestes les plus attendus d'un match de basket. En effet, c'est le tir qui rapporte le plus de points possible. De plus, il est particulièrement spectaculaire et sa réalisation provoque systématiquement une réaction de la part du public.

Sur l'échantillon observé, on observe une moyenne de 33,45% de réussite, une médiane de 34,3% et un écart-type de 7,19 points de pourcentage. La valeur minimum est à 0% (sur 14 tentatives) et la valeur maximum de 58% (49,4% pour les joueurs à plus de 30 tentatives dans la saison).

B. Variables explicatives

➢ Variable X1 : Le nombre de minutes jouées dans la saison

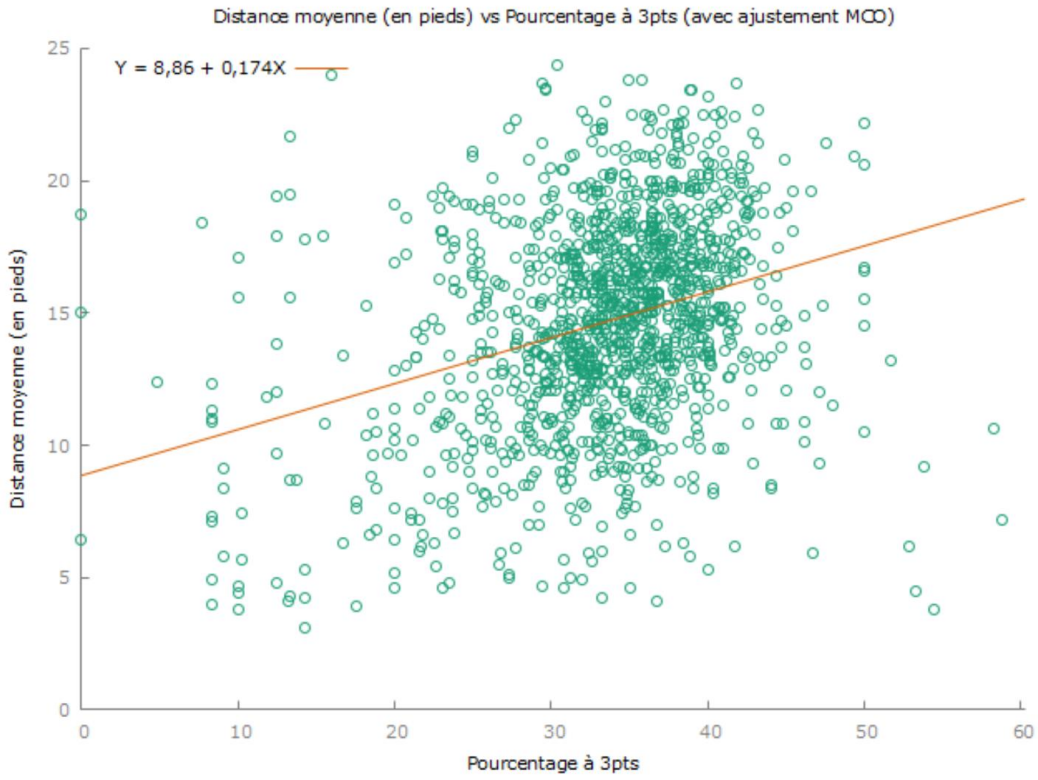
La première variable explicative est le nombre de minutes jouées par le joueur dans la saison. X1 est mesurée en minutes. La valeur moyenne est de 1253,4 minutes. Une saison durant 82 matchs, cela correspond à une moyenne de 15,29 minutes par match. La valeur médiane est de 1208 minutes. L'écart-type 774,97 minutes, illustrant une grande disparité entre les joueurs. La valeur minimum est de 25 minutes, ce qui est assez impressionnant sachant que le joueur est parvenu à effectuer 10 tentatives à 3pts lors de ce très court intervalle de temps. La valeur maximum est de 3028 minutes réalisées en 82 matchs soit une moyenne de 36,9 minutes par match au maximum, sur 48 possibles. Ces statistiques descriptives illustrent la rotation importante effectuée par les équipes de NBA. Avec seulement 5 joueurs par équipe sur le terrain simultanément, les entraîneurs préservent généralement la forme de leurs joueurs qui doivent effectuer de longs déplacements et d'importants efforts tout au long de l'année.



Le coefficient de corrélation entre y et x1 est égal à 0,2468. Il existe une corrélation positive entre nombre de minutes jouées par saison et pourcentage de réussite au tir à 3pts.

- **Variable X2 : La distance moyenne (en pieds) des tirs tentés**

La deuxième variable explicative est la distance moyenne au panier de tous les tirs tentés par le joueur lors de la saison. Cette distance est mesurée en pieds. La distance au panier d'un tir varie de 0 pieds, lors d'un 'dunk', qui consiste à jeter la balle directement dans le panier, à plus de 35 pieds (10,67m) pour les tirs les plus lointains. Sur notre échantillon, la moyenne est de 14,661 pieds (4,47m) et la valeur minimum de 3,1 pieds (0,95m). La valeur maximum est de 24,4 pieds (7,44m), soit un joueur placé en moyenne derrière la ligne des 3pts lors de ses tirs. L'écart-type est de 4,02m.
Le coefficient de corrélation est de 0,31. Il existe une corrélation positive entre distance moyenne des tirs et pourcentage de réussite des tirs à 3pts.



3. Estimation du modèle par les MCO

A. Résultats de l'estimation

```
Modèle 1: MCO, utilisant les observations 1-1325
Variable dépendante: Y

             coefficient    éc. type     t de Student   p. critique
  ------------------------------------------------------------------
  const      21,3432       0,767600        27,81        2,35e-134
  X1          0,00259497   0,000233162     11,13        1,45e-027
  X2          0,603952     0,0449393       13,44        1,09e-038

Moyenne var. dép.      33,45019   Éc. type var. dép.    7,193725
Somme carrés résidus   56610,07   Éc. type régression   6,543816
R2                     0,173776   R2 ajusté             0,172526
F(2, 1322)             139,0247   P. critique (F)       1,59e-55
Log de vraisemblance  -4367,632   Critère d'Akaike      8741,263
Critère de Schwarz     8756,831   Hannan-Quinn          8747,099
```

L'estimation du modèle s'écrit :

$\hat{y} = 21{,}3432 + 0{,}0026\ X_1 + 0{,}604\ X_2$

Le modèle semble montrer une corrélation positive entre la variable Y et X_1 de même que entre Y et X_1.

B. Test de significativité globale du modèle

i) Hypothèses

$H_0: a_1 = a_2 = 0$

H_1 : au moins un des paramètres est non nul

ii) Calcul de la statistique de test

Sous H_0 : $F^* = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} = \frac{\frac{0{,}174}{2}}{\frac{1-0{,}174}{1325-2-1}} = 139{,}24$

iii) Règle de décision (avec $\alpha = 5\%$)

On compare F^* à $F^{\alpha}(k, n-k-1) = F^{0{,}05}(2, 1325-2-1) = F^{0{,}05}(2, 1322) = 2{,}99$

On voit que $F^* = 139{,}24 > 2{,}99 = F^{\alpha}$; H_0 est rejeté au seuil $\alpha = 5\%$.

Le modèle est significatif.

C. Tests sur les coefficients

➢ **Coefficient $\hat{a}_1$**

$\hat{a}_1 = 0{,}00259497 = 2{,}59 * 10^{-3} > 0$

i) Hypothèses

On définit l'hypothèse nulle H_0 et l'hypothèse alternative H_1

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

ii) Calcul de la statistique de test

Nous calculons pour ce test le ratio de Student.

Sous H_0 : $t^*_{\hat{a}1} = \frac{\hat{a}_1 - a_1}{\hat{\sigma}_{\hat{a}_1}} = \frac{2{,}59 * 10^{-3} - 0}{2{,}33 * 10^{-4}} = 11{,}116$

iii) Règle de décision (avec α = 5%)

Nous savons que $\frac{\hat{a}1 - a_1}{\hat{\sigma}_{\hat{a}1}}$ suit une loi de Student à n-k-1 degrés de liberté.

Sous H0, on compare donc $|t^*_{\hat{a}1}|$ à $t^{\alpha/2}$ (n – k – 1) = $t^{0,05}$ (1325 – 2 – 1) = $t^{0,05}$ (1322) = 1,96

On observe que $|t^*_{\hat{a}1}|$ = 11,116 > 1,96 = $t^{0,05}$ (1322), H_0 est rejetée au seuil *α = 5%.*

Ainsi, nous pouvons dire que le coefficient $\hat{a}_1$ est significativement différent de 0 d'après le test de Student. La variable explicative X1 est donc contributive à l'explication de la variable Y.

- **Coefficient $\hat{a}_2$**

$\hat{a}_2$ = 0,604 > 0

i) Hypothèses

On définit l'hypothèse nulle H_0 et l'hypothèse alternative H_1

$$\begin{cases} H_0 : a_2 = 0 \\ H_1 : a_2 \neq 0 \end{cases}$$

ii) Calcul de la statistique de test

Nous calculons pour ce test le ratio de Student.

Sous H_0 : $t^*_{\hat{a}2} = \frac{\hat{a}_2 - a_2}{\hat{\sigma}_{\hat{a}_2}} = \frac{0{,}604 - 0}{0{,}045} = 13{,}42$

iii) Règle de décision (avec α = 5%)

Nous savons que $\frac{\hat{a}_2 - a_2}{\hat{\sigma}_{\hat{a}_2}}$ suit une loi de Student à n-k-1 degrés de liberté.

Sous H0, on compare donc $|t^*_{\hat{a}2}|$ à $t^{\alpha/2}$ (n – k – 1) = $t^{0,05}$ (1325 – 2 – 1) = $t^{0,05}$ (1322) = 1,96

On observe que $|t^*_{\hat{a}2}|$ = 13,42 > 1,96 = $t^{0,05}$ (1322) ; H_0 est rejetée au seuil $\alpha = 5\%$.

Le coefficient $\hat{a}_2$ est donc significativement différent de 0 d'après le test de Student. La variable explicative X2 est donc contributive à l'explication de la variable Y.

- **Test d'égalité des variables**

i) Hypothèses

$$\begin{cases} H_0 : a_1 = a_2 \\ H_1 : a_1 \neq a_2 \end{cases} \qquad \begin{cases} H_0 : a_1 - a_2 = 0 \\ H_1 : a_1 - a_2 \neq 0 \end{cases}$$

i) Calcul de la statistique de test

On note $\hat{d} = \hat{a}_1 - \hat{a}_2$ et $d = a_1 - a_2$

Sous H_0 : $t^* = \frac{\hat{d}-d}{\hat{\sigma}\,\hat{d}} = \frac{\hat{d}}{\hat{\sigma}\,\hat{d}} = \frac{\hat{a}1 - \hat{a}2}{\hat{\sigma}_{\hat{a}1 - \hat{a}2}}$

Avec $\hat{\sigma}_{\hat{a}1} = 2{,}33*10^{-4}$ et $\hat{\sigma}_{\hat{a}2} = 0{,}045$

Et d'après la matrice de covariance des coefficients de régression, on a :

$cov(\hat{a}_1, \hat{a}_2) = 1{,}018*10^{-6}$

Ainsi :

$\hat{\sigma}_{\hat{a}1 - \hat{a}2} = \hat{\sigma}^2\hat{a}_1 + \hat{\sigma}^2\hat{a}_2 + 2 * (1) * (-1) * cov(\hat{a}_1, \hat{a}_2)$

$\hat{\sigma}^2_{\hat{a}1 - \hat{a}2} = \hat{\sigma}^2\hat{a}_1 + \hat{\sigma}^2\hat{a}_2 - 2 * cov(\hat{a}_1, \hat{a}_2) = 2{,}023*10^{-3}$

Donc $t^* = \frac{\hat{a}1 - \hat{a}2}{\hat{\sigma}_{\hat{a}1 - \hat{a}2}} = \frac{2{,}59*10^{-3} - 0{,}604}{\sqrt{(2{,}023}*10^{-3})} = -13{,}37$

ii) Règle de décision (avec α = 5%)

$|t^*| = 13{,}37$

On compare cette valeur à t^{α} (n – k – 1) = $t^{0,05}$ (1325 – 2 – 1) = $t^{0,05}$ (1322) = 1,96

On observe que $|t^*| = 13{,}37 > 1{,}96 = t^{0,05}$ (1322), H_0 est rejetée au seuil $\alpha = 5\%$.

Les coefficients sont donc significativement différents.

4. Tests

A. Test de stabilité des paramètres (de Chow)

On dispose de données relatives aux postes des joueurs. Cette variable est à l'origine de la constitution des deux échantillons. Le premier sous-échantillon comporte les valeurs associées aux meneurs et aux arrières (respectivement postes 1 et 2). Il contient 619 observations (soit $n_1 = 619$).

Le deuxième sous-échantillon comporte les valeurs associées aux ailiers, ailiers forts et pivots (postes 3,4 et 5). Il regroupe 706 observations ($n_2 = 706$).

Cette séparation est intéressante car le poste du joueur influence le comportement du joueur sur le terrain. Il est probable que les données des sous-échantillons identifient des variables expliquant différemment le pourcentage de réussite au tir à 3pts.

L'équation de régression s'écrit :

$y = a_0^1 + a_1^1 x_1 + a_2^1 x_2 + \varepsilon$ pour la base 1

$y = a_0^2 + a_1^2 x_1 + a_2^2 x_2 + \varepsilon$ *pour la base 2*

i) Hypothèses

$H_0 : a_0^1 = a_0^2 ; a_1^1 = a_1^1 ; a_2^1 = a_2^2$

H_1 : au moins un des coefficient d'un modèle est différent de celui de l'autre modèle

ii) Calcul de la statistique de test

D'après l'estimation antérieure du modèle non-contraint, on a donc :

$$SCR^{NC} = 56610{,}07$$

Régression augmentée pour le test de Chow
MCO, utilisant les observations 1-1325
Variable dépendante: Y

	coefficient	éc. type	t de Student	p. critique
const	19,7077	1,42204	13,86	7,10e-041
X1	0,00303389	0,000332505	9,124	2,63e-019
X2	0,658186	0,0820502	8,022	2,28e-015
splitdum	2,44463	1,71438	1,426	0,1541
sd_X1	-0,000848735	0,000466339	-1,820	0,0690
sd_X2	-0,0618272	0,100157	-0,6173	0,5371

Moyenne var. dép.	33,45019	Éc. type var. dép.	7,193725
Somme carrés résidus	56397,64	Éc. type régression	6,538951
R2	0,176876	R2 ajusté	0,173756
F(5, 1319)	56,68633	P. critique (F)	1,71e-53
Log de vraisemblance	-4365,141	Critère d'Akaike	8742,282
Critère de Schwarz	8773,417	Hannan-Quinn	8753,953

Test de Chow pour rupture structurelle à l'observation 619
F(3, 1319) = 1,65607 avec p. critique 0,1747

D'après la sortie GRETL du test de Chow, on a $SCR^C = 56397{,}64$

On a :

$$ddl^{NC} = ddl^1 + ddl^2 = (n1 - k - 1) + (n2 - k - 1) = n - 2(k+1) = 1325 - 6 = 1319$$

On a : $ddl^n = (n - k - 1) - [(n_1 - k - 1) + (n_2 - k - 1)] = k + 1 = 3$

Sous H_0, $\frac{(SCR^C - SCR^{NC})/ddl^n}{\frac{SCR^{NC}}{ddl^{NC}}}$ suit une distribution de Fisher à (k+1) et (n-2(k+1)) degrés de liberté.

Sous H_0, $F^* = \frac{(SCR^C - SCR^{NC})/ddl^n}{SCR^{NC}/ddl^{NC}} = \frac{(56397{,}64 - 56610{,}07)/3}{56610{,}07/1319} = 1{,}65607$

On compare F^* à $F^{\alpha}(k+1, n-2(k+1)) = F^{0.05}(2+1, 1325-2(2+1))$

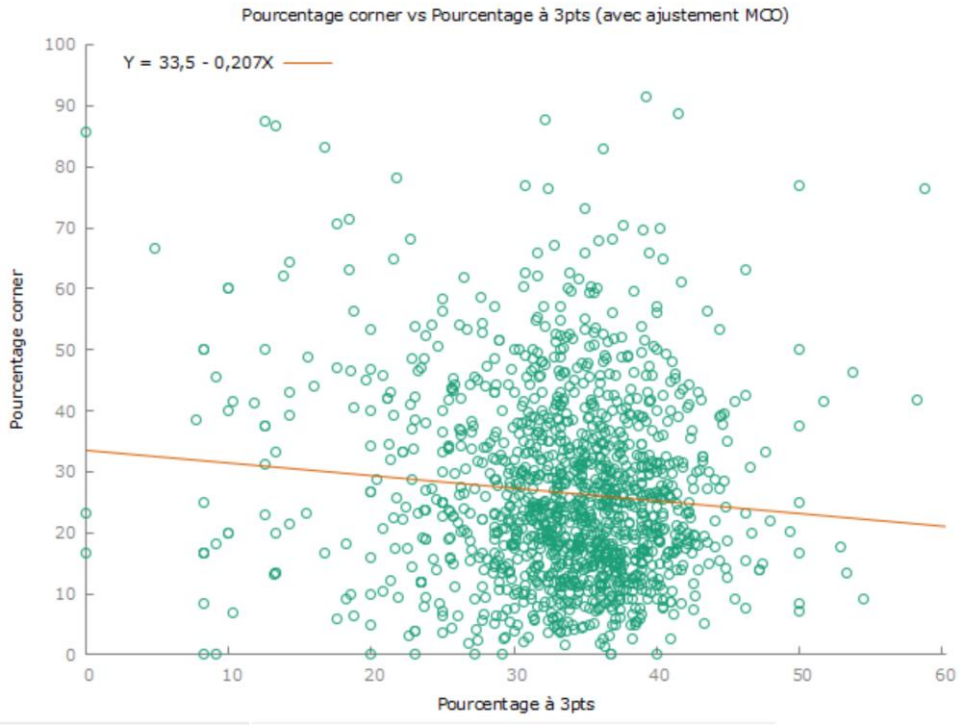
$= F^{0.05}(3{,}1319) = 2{,}6$

Or $F^* = 1{,}65607 < 2{,}6 = F^{0{,}05}(3{,}1319)$

L'hypothèse H0 n'est pas rejetée au seuil $\propto = 5\%$. Sous l'hypothèse d'homoscédasticité ($E(\varepsilon_t{}^2) = \sigma_\varepsilon{}^2 \ \forall t$), les coefficients sont significativement stables sur l'ensemble de la période.

B. Test d'ajout de variables

A l'occasion de ce test, on ajoute la variable X3 correspondante à la proportion des tirs à 3pts tirée depuis le 'corner'. Un 'corner' est une des 4 zones du terrain proches des angles et orientée à environ 0° du panier. C'est la zone la plus favorable à la réussite au tir à 3pts pour deux raisons : l'angle est celui avec lequel les joueurs sont le plus familiers ; la distance est la plus faible possible pour un tir rapportant autant de points. C'est aussi la seule zone du terrain ou le tir à 3pts semble avoir l'espérance la plus importante.



On observe une corrélation négative entre Y et X3. Cela pourrait s'expliquer par le fait que les joueurs moins adroits privilégient davantage le 'corner', une zone ou le tir semble moins difficile à réaliser. En revanche, les joueurs ayant davantage confiance en leur efficacité aux tir auraient davantage tendance à tirer depuis des zones plus compliquées, en raison des angles et des distances.

Dans ce test, nous cherchons à déterminer si l'ajout de la variable explicative X3 améliore de manière significative – au seuil de 5 % – le pouvoir explicatif du modèle.

i) Hypothèses

$$\begin{cases} H_0 : SCR^C = SCR^{NC} \\ H_1 : SCR^C \neq SCR^{NC} \end{cases}$$

ii) Calcul du Fisher empirique (F^*)

$$F^* = \frac{(SCR^1 - SCR)\ /(k - k')}{SCR/(n - k - 1)}$$

Avec :

k = nombre de variables explicatives du modèle complet = 3

k' = nombre de variables explicatives du modèle sans l'ajout de X3 = 2

$SCR = 56610{,}07$

Et :

Modèle 3: MCO, utilisant les observations 1-1325
Variable dépendante: Y

	coefficient	éc. type	t de Student	p. critique	
const	21,3352	0,912829	23,37	2,14e-101	***
X1	0,00259564	0,000236926	10,96	8,55e-027	***
X2	0,604084	0,0456881	13,22	1,42e-037	***
X3	0,000195125	0,0121181	0,01610	0,9872	

Moyenne var. dép.	33,45019	Éc. type var. dép.	7,193725
Somme carrés résidus	56610,06	Éc. type régression	6,546292
R2	0,173776	R2 ajusté	0,171899
F(3, 1321)	92,61315	P. critique (F)	2,13e-54
Log de vraisemblance	-4367,631	Critère d'Akaike	8743,263
Critère de Schwarz	8764,020	Hannan-Quinn	8751,044

$SCR^1 = 56610{,}06$

D'où, sous H_0, on a :

$$F^* = \frac{(56610{,}06 - \ 56610{,}07)\ /(3-2)}{56610{,}07/(1325-3-1)} = \ 2{,}33 * 10^{-4}$$

iii) Règle de décision (avec α = 5%)

On compare F* à $F^{\alpha}\,(\mathrm{q}, ddl^{nc}) \ = \ F^{0{,}05}\,(3, 1325 - 2 - 1) \ = \ F^{0{,}05}\,(3, 1322) \ = \ 2{,}6$

On voit que F* = $2{,}33 * 10^{-4}$ < 2,6 = $F^{0{,}05}$ (3, 1322); H_0 n'est pas rejetée au seuil α = 5%.

L'ajout de X3 n'améliore pas le pouvoir explicatif du modèle de manière significative au seuil de 5%.

5. Conclusion

D'après nos observations, le nombre de minutes jouées et la distance moyenne des tirs d'un joueur sont des variables significatives pour expliquer son pourcentage de réussite aux tirs à 3 points. Le modèle des moindres carrés ordinaires composé de ces deux variables est significatif, la relation qu'il propose permet d'expliquer en partie la variable endogène. On aurait pu penser dans un premier temps que le poste du joueur pouvait être amené à faire évoluer cette relation entre les variables. Néanmoins, d'après le test de Chow au seuil de 5%, le pourcentage de réussite des arrières et ailiers/pivots peut être expliqué de manière similaire par le modèle. La prise en compte du pourcentage des tirs à 3pts réalisés dans le corner n'améliore pas le pouvoir explicatif du modèle de manière significative. Dans le cadre d'une expérience plus extensive, de nombreuses autres variables auraient pu être considérées. Celle-ci a permis d'illustrer l'importance de la confiance dans la précision à 3pts des joueurs de NBA. La confiance des entraîneurs influe sur le nombre de minutes et la confiance des joueurs influe sur la distance de leurs tirs.

6. Bibliographie

Basketball-reference.com Totals (https://www.basketball-reference.com/leagues/NBA_2023_totals.html)
Basketball-reference.com Shooting (https://www.basketball-reference.com/leagues/NBA_2023_shooting.html)