Charles ODEND'HAL Groupe 4

Je déclare sur l'honneur que ce mémoire a été écrit 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

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

2nde partie

Table des matières :

I. Sélection du modèle optimal

La 6ème hypothèse des moindres carrés ordinaires implique une absence de colinéarité entre les variables explicatives du modèle, signifiant que $(X'X)$ est régulière et que la matrice inverse $(X'X)^{-1}$ existe. Lorsque les colonnes de X sont très corrélées, cela pose un problème de multicolinéarité qui voit cette hypothèse violée. Celui-ci a pour conséquences une augmentation de la variance estimée de certains coefficients. Si on décide alors d'ajouter une variable ou d'augmenter la taille de l'échantillon, on s'expose à l'instabilité des coefficients des MCO.

Pour remédier à ce problème, une des méthodes consiste à spécifier un nouveau modèle dans lequel les variables fortement corrélées entre elles ont été éliminées.

a) Méthode toutes les régressions possibles

Dans un premier temps, nous créons 7 modèles des MCO en testant toutes les combinaisons des variables explicatives possibles.

Tableau récapitulatif des t de Student et critères d'Akaike

Modèle	X1	X2	X3	AIC
1	9,262			8908,944
2		11,88		8857,932
3			-2,956	8899,551
4	11,13	13,44		8741,263
5	8,556		-0,9263	8830,146
6		11,81	-1,764	8768,571
7	10,79	13,6	0,9185	8658,585

Parmi les modèles ci-dessus, on retient ceux dont toutes les variables sont significatives. On compare chaque t de Student à une statistique de test

On compare chaque statistique de test à une loi de Student à n-k-1 degrés de liberté au seuil de 5%. On a n=1322 et k compris entre 1 et 3 donc on considère la valeur dans la table de Student correspondant au nombre maximal de degrés de liberté soit : **1,95996**

D'après cette règle de décision, on retient les modèles 1, 2, 3 et 4. On remarque que la variable X3 n'est pas significative dans les modèles de régression multiple.

Parmi cet ensemble restreint, on retient le modèle qui minimise le critère d'Akaike. Il s'agit du modèle 4.

b) Méthode de sélection par étage

Etape 1

On calcule les coefficients de corrélation simple entre toutes les variables explicatives potentielles et Y.

```
Coefficients de corrélation, utilisant les observations 1 - 1325
(sans prendre en compte les valeurs manquantes)

Two-tailed critical values for n = 1322: 5% 0,0539, 1% 0,0708

          Y           X1           X2           X3
     1,0000       0,2468       0,3104      -0,0811  Y
                  1,0000      -0,0971      -0,2365  X1
                               1,0000      -0,1121  X2
                                            1,0000  X3
```

On retient la variable Xi dont le coefficient de corrélation simple avec Y est le plus élevé, soit X2 ($0{,}3104 > 0{,}2468 > -0{,}0811$). Avant d'estimer le modèle, on vérifie que le coefficient de corrélation $r_{Y,X2}$ est significativement différent de 0. On effectue un test de significativité du paramètre a2. D'après les résultats précédemment obtenus, on a $\left|t_{\hat{a}_2}{}^*\right| = 11{,}88 > t_{n-k-1} = 1{,}96$, donc a2 est significative. Ainsi $r_{Y,X2}$ est significativement différent de 0.

Etape 2

On estime Y en fonction de X2. On calcule le résidu $e_2 = y - \hat{a}_0 - \hat{a}_2 x_2$.

Modèle 2: MCO, utilisant les observations 1-1325
Variable dépendante: Y

	coefficient	éc. type	t de Student	p. critique	
const	25,3078	0,710807	35,60	2,73e-195	***
X2	0,555379	0,0467576	11,88	5,45e-031	***

Moyenne var. dép.	33,45019	Éc. type var. dép.	7,193725
Somme carrés résidus	61914,14	Éc. type régression	6,840928
R2	0,096363	R2 ajusté	0,095680
F(1, 1323)	141,0828	P. critique (F)	5,45e-31
Log de vraisemblance	-4426,966	Critère d'Akaike	8857,932
Critère de Schwarz	8868,311	Hannan-Quinn	8861,823

Coefficients de corrélation, utilisant les observations 1 - 1325
(sans prendre en compte les valeurs manquantes)

Two-tailed critical values for n = 1322: 5% 0,0539, 1% 0,0708

residu2	X1	X3	
1,0000	0,2913	-0,0479	residu2
	1,0000	-0,2365	X1
		1,0000	X3

On a donc :

$|\rho_{e,X1}| = 0{,}2913$

$\left|\rho_{e,X3}\right| = 0{,}0479$

En valeur absolue, $\rho_{e,X1}$ est donc le coefficient le plus élevé. On vérifie si ce coefficient est significativement différent de 0.

<u>Test de Student</u>

i) Hypothèses

$H_0 : r_{e,X1} = 0$

$H_1 : r_{e,X1} \neq 0$

ii) Statistique de test

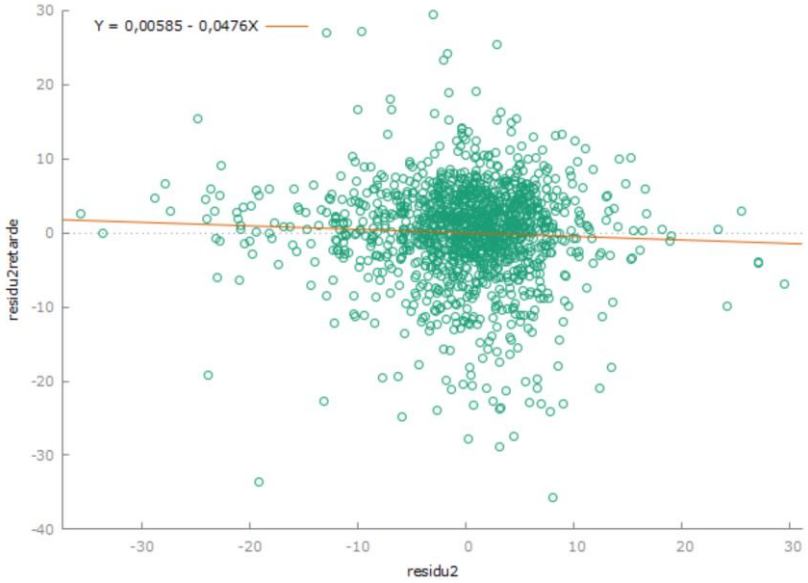
$$t^* = \frac{\rho_{e,X1}}{\sqrt{\frac{1-{\rho_{e,X1}}^2}{n-2}}} = \frac{0{,}2913}{0{,}0263} = 11{,}076$$

iii) Règle de décision

On compare la statistique de test (en valeur absolue) à la valeur critique $t_{n-k-1} = 1{,}96$. H_0 est rejetée. $\rho_{e,X1}$ est significativement différent de 0. Le modèle optimal contient donc 2 variables explicatives X1 et X2.

II. Tests de détection d'autocorrélation des erreurs

On commence par modifier la structure de notre jeu de données en le transformant en séries temporelles. On construit ensuite le nuage de points e(t) en fonction de e(t-1).



On observe une très faible autocorrélation négative des erreurs. Le signe des résidus a plutôt tendance à alterner d'une période à l'autre.

Test de Durbin-Watson

Le test de Durbin-Watson permet de détecter une autocorrélation des erreurs d'ordre 1 telle que $\varepsilon_t = \rho\varepsilon_t + v_t$ avec $v_t \to N(0, {\delta_v}^2)$.
On vérifie si les condition d'utilisation sont satisfaites.

- $n = 1325 > 15$
- Y_{t-1} n'est pas parmi les variables explicatives
- On trie les données avec Y de façon croissante

i) Hypothèses

$H_0: \rho = 0$

$H_1: \rho \neq 0$

ii) Statistique de test

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} {e_t}^2} = \frac{129649}{61914{,}1} = 2{,}094$$

iii) Règle de décision

On cherche les valeurs d1 et d2 dans la table de Durbin-Watson. On a n = 1325 et k = 3.

On obtient au seuil de 5%:

$$d_1 = 1{,}61$$

$$d_2 = 1{,}74$$

Ainsi $d_2 < DW < 4 - d_2$. H_0 est acceptée. On peut donc dire que $\rho = 0$, il n'y a pas d'autocorrélation des erreurs.

Test de Breush-Godfrey

On commence par calculer e_t en estimant le modèle par les MCO. On estime ensuite par les MCO l'équation auxiliaire :

$e_t = a_0 + a_1 x_{1t} + a_2 x_{2t} + a_3 x_{3t} + \rho_1 e_{t-1} + \cdots + \rho_p e_{t-p} + v_t$

L'équation auxiliaire s'estime pour n=1322-1=1321 observations car e_0 n'existe pas.

```
? ols residu2 const X1 X2 X3 residu2retarde

Modèle 1: MCO, utilisant les observations 1-1325 (n = 1321)
Suppression d'observations manquantes ou incomplètes: 4
Variable dépendante: residu2

                 coefficient   éc. type    t de Student  p. critique
  ------------------------------------------------------------------
  const          -4,80850     1,41139       -3,407       0,0007     ***
  X1              0,00253680  0,000235615    10,77        5,75e-026  ***
  X2              0,0486508   0,0443863       1,096       0,2732
  X3              0,0113459   0,0118785       0,9552      0,3397
  residu2retarde -0,0366350   0,0257267      -1,424       0,1547

Moyenne var. dép.      0,072528   Éc. type var. dép.    6,666454
Somme carrés résidus   53692,03   Éc. type régression   6,387442
R2                     0,084736   R2 ajusté             0,081954
F(4, 1316)             30,45929   P. critique (F)       2,83e-24
Log de vraisemblance -4321,488    Critère d'Akaike      8652,976
Critère de Schwarz     8678,907   Hannan-Quinn          8662,698

Constante mise à part, la probabilité critique est la plus élevée pour la variable 3 (X3)
```

On procède ensuite à un test du multiplicateur de Lagrange sur l'équation auxiliaire.

i) Hypothèses

$H_0: \rho = 0$

$H_1: \rho \neq 0$

ii) Statistique de test

$LM = n \times R^2 = 1321 \times 0{,}0847 = 111{,}889$

iii) Règle de decision

On compare la valeur de la statistique de test à celle obtenue dans la table du Chi-Deux à p=1321 degrés de liberté au seuil de 5%, soit 2665.13. H_0 n'est pas rejetée donc les erreurs ne subissent pas un processus autorégressif d'ordre 1. On observe une absence d'autocorrélation des erreurs.

III. Tests de détection de l'hétéroscédasticité des erreurs

Test de Goldfeld-Quandt

Le nombre d'observations est important (n=1322>30). On soupçonne la variable X1 d'être la cause de l'hétéroscédasticité.

On commence donc par ordonner les observations en fonction de X1 (ordre croissant).

On dispose de 1322 observations. On omet donc $C = \frac{1322}{4} \approx 330$ observations centrales C. On estime ensuite le modèle sur 2 sous-échantillons.

L'échantillon 1 comporte les 496 premières observations. L'échantillon 2 comporte les 496 dernières observations.

```
? ols Y const X1 X2 X3

Modèle 6: MCO, utilisant les observations 1-496 (n = 494)
Suppression d'observations manquantes ou incomplètes: 2
Variable dépendante: Y

             coefficient    éc. type    t de Student   p. critique
  ------------------------------------------------------------------
  const      23,0123       3,02994        7,595        1,57e-013  ***
  X1          0,00423969   0,00153525     2,762        0,0060     ***
  X2          0,470697     0,0898027      5,241        2,37e-07   ***
  X3         -0,00701239   0,0270166     -0,2596       0,7953

Moyenne var. dép.      31,34150   Éc. type var. dép.     8,618829
Somme carrés résidus   34414,01   Éc. type régression    8,380494
R2                     0,060294   R2 ajusté              0,054541
F(3, 490)              10,47995   P. critique (F)        1,08e-06
Log de vraisemblance -1749,145    Critère d'Akaike       3506,291
Critère de Schwarz     3523,101   Hannan-Quinn           3512,891
```

$SCR1 = 34414{,}01$

```
? ols Y const X1 X2 X3

Modèle 7: MCO, utilisant les observations 827-1322 (n = 496)
Variable dépendante: Y

             coefficient    éc. type    t de Student  p. critique
  -----------------------------------------------------------------
  const      20,2037       2,06038        9,806       7,47e-021  ***
  X1          0,000881581  0,000627639    1,405       0,1608
  X2          0,848703     0,0535075     15,86        4,63e-046  ***
  X3          0,0117046    0,0120793      0,9690      0,3330

Moyenne var. dép.      35,21512   Éc. type var. dép.     5,461790
Somme carrés résidus   9733,956   Éc. type régression    4,447973
R2                     0,340804   R2 ajusté              0,336785
F(3, 492)              84,78810   P. critique (F)        3,12e-44
Log de vraisemblance -1442,040    Critère d'Akaike       2892,080
Critère de Schwarz     2908,906   Hannan-Quinn           2898,685
```

$SCR2 = 9733{,}956$

i) Hypothèses

$H_0 : E({e_t}^2) = {\sigma_{e_t}}^2$

$H_1 : E({e_t}^2) \neq {\sigma_{e_t}}^2$

ii) Statistique de test

$$F^* = \frac{SCR1/ddl_2}{SCR2/ddl_1} = \frac{34414{,}01/492}{9733{,}956/492} = 3{,}535$$

On place toujours la SCR la plus grande au numérateur.

iii) Règle de decision

On compare F^* à la valeur obtenue dans la table de Fisher à (492,492) degrés de libertés, soit 1.

$F^* = 3{,}535 > 1$ donc H_0 est rejetée. L'hypothèse d'homoscédasticité est rejetée au seuil de 5%.

Test de White

Dans un premier temps, on teste Y en fonction de X et on extrait les résidus e.

On teste l'hypothèse nulle d'homoscédasticité des erreurs :

$H_0 : a_1 = a_2 = b_1 = \cdots = a_k = b_k$

Modèle 11: MCO, utilisant les observations 1-1325 (n = 1322)
Suppression d'observations manquantes ou incomplètes: 3
Variable dépendante: esq

	coefficient	éc. type	t de Student	p. critique	
const	378,170	38,9809	9,701	1,54e-021	***
X1	-0,119474	0,0115223	-10,37	2,87e-024	***
X1sq	3,40481e-05	4,23541e-06	8,039	2,00e-015	***
X2	-28,9941	3,15282	-9,196	1,41e-019	***
X2sq	0,818876	0,109716	7,464	1,53e-013	***
X3	-0,920921	0,936815	-0,9830	0,3258	
X3sq	0,00697460	0,00628508	1,110	0,2673	

Moyenne var. dép.	40,67776	Éc. type var. dép.	92,56932
Somme carrés résidus	9010858	Éc. type régression	82,77901
R2	0,203970	R2 ajusté	0,200338
F(6, 1315)	56,15811	P. critique (F)	6,64e-62
Log de vraisemblance	-7710,510	Critère d'Akaike	15435,02
Critère de Schwarz	15471,33	Hannan-Quinn	15448,63

La statistique de test est la suivante :

$LM = n \times R^2 = 1322 \times 0{,}204 = 269{,}688$

On la compare à la valeur lue dans la table du Chi-Deux à p=6 degrés de liberté, soit 12,5916.

$269{,}688 > 12{,}5916$ donc l'hypothèse d'homoscédasticité est rejetée.

IV. Corrections

Correction de l'autocorrélation des erreurs

Considérons le modèle initial à 3 variables explicatives :

$$Y_t = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + \varepsilon_t$$

```
Modèle 3: MCO, utilisant les observations 1-1325 (n = 1318)
Suppression d'observations manquantes ou incomplètes: 7
Variable dépendante: res0

            coefficient   éc. type    t de Student   p. critique
  ------------------------------------------------------------
  res0_1      0,0321169   0,0273968      1,172         0,2413

Moyenne var. dép.      0,004658   Éc. type var. dép.    6,348971
Somme carrés résidus   53032,21   Éc. type régression   6,345662
R2 non-centré          0,001042   R2 centré             0,001042
F(1, 1317)             1,374257   P. critique (F)       0,241294
Log de vraisemblance  -4305,024   Critère d'Akaike      8612,047
Critère de Schwarz     8617,231   Hannan-Quinn          8613,991
```

A l'observation t-1, le modèle s'écrit :

$$Y_{t-1} = a_0 + a_1 X_{1_{t-1}} + a_2 X_{2_{t-1}} + a_3 X_{3_{t-1}} + \varepsilon_{t-1}$$

$$Y_t - \rho Y_{t-1} = a_0(1-p) + a_1\left(X_1 - \rho X_{1_{t-1}}\right) + a_2\left(X_2 - \rho X_{2_{t-1}}\right) + a_3\left(X_3 - \rho X_{3_{t-1}}\right) + \varepsilon_t - \varepsilon_{t-1}$$

Le modèle transformé s'écrit :

$$Y_t^* = b_0 + a_1 X_{1_t}^* + a_2 X_{2_t}^* + a_3 X_{3_t}^* + v_t$$

On estime le modèle en quasi-différences, en prenant

$\hat{\rho} = \hat{\rho}^0$

```
Modèle 4: MCO, utilisant les observations 1-1325 (n = 1318)
Suppression d'observations manquantes ou incomplètes: 7
Variable dépendante: d0Y

             coefficient    éc. type     t de Student   p. critique
  ---------------------------------------------------------------
  const       19,7563       1,35601        14,57        1,10e-044
  d0X1         0,00253646   0,000241922    10,48        9,34e-025
  d0X2         0,606312     0,0442438      13,70        4,74e-040
  d0X3         0,0120208    0,0117368       1,024       0,3059

Moyenne var. dép.      32,45295   Éc. type var. dép.    6,980866
Somme carrés résidus   53031,56   Éc. type régression   6,352863
R2                     0,173715   R2 ajusté             0,171828
F(3, 1314)             92,08325   P. critique (F)       4,34e-54
Log de vraisemblance  -4305,016   Critère d'Akaike      8618,031
Critère de Schwarz     8638,767   Hannan-Quinn          8625,806
```

On obtient ensuite les résidus e^1.

```
Modèle 5: MCO, utilisant les observations 1-1325 (n = 1318)
Suppression d'observations manquantes ou incomplètes: 7
Variable dépendante: res1

            coefficient   éc. type    t de Student   p. critique
  ------------------------------------------------------------
  res1_1      0,0321003   0,0273965      1,172         0,2415

Moyenne var. dép.     -0,002630   Éc. type var. dép.    6,348931
Somme carrés résidus   53031,58   Éc. type régression   6,345625
R2 non-centré          0,001041   R2 centré             0,001041
F(1, 1317)             1,372863   P. critique (F)       0,241533
Log de vraisemblance  -4305,016   Critère d'Akaike      8612,032
Critère de Schwarz     8617,215   Hannan-Quinn          8613,975
```

On estime ensuite le modèle en quasi-différences après avoir actualisé la valeur :

$\hat{\rho} = \hat{\rho}^1$

```
Modèle 6: MCO, utilisant les observations 1-1325 (n = 1318)
Suppression d'observations manquantes ou incomplètes: 7
Variable dépendante: dlY

            coefficient    éc. type     t de Student   p. critique
  ------------------------------------------------------------------
  const      19,7563       1,35601        14,57        1,10e-044
  dlX1        0,00253646   0,000241922    10,48        9,34e-025
  dlX2        0,606312     0,0442438      13,70        4,74e-040
  dlX3        0,0120208    0,0117368       1,024       0,3059

Moyenne var. dép.      32,45294   Éc. type var. dép.     6,980866
Somme carrés résidus   53031,56   Éc. type régression    6,352863
R2                     0,173715   R2 ajusté              0,171828
F(3, 1314)             92,08324   P. critique (F)        4,34e-54
Log de vraisemblance -4305,016    Critère d'Akaike       8618,031
Critère de Schwarz     8638,767   Hannan-Quinn           8625,806
```

On observe une stabilité des estimateurs â issus de 2 itérations successives. Le modèle transformé est le suivant :

$$Y_t^* = 19{,}7563 + 0{,}00254\, X_{1_t}^* + 0{,}606312\, X_{2_t}^* + 0{,}01202\, X_{3_t}^* + v_t$$

Toutes les relations observées sont positives. La constante a le t de Student le plus élevé et est donc la variable la plus significative. La probabilité critique de X3 est élevée. Cette variable n'est pas significative au seuil de 5%.

<u>Correction de l'hétéroscédasticité des erreurs</u>

Pour corriger de l'hétéroscédasticité, on utilise la régression pondérée. On transforme les variables initiales pour estimer le modèle des MCO en utilisant les variables transformées.

D'après le test de White, X1 est la variable qui a le t de Student le plus élevé (en valeur absolue). C'est donc la variable la plus significative. On divise les variables de notre modèle initial par $\sqrt{X_1}$.

Notre modèle devient :

$$\frac{Y}{\sqrt{X_1}} = \frac{a_0}{\sqrt{X_1}} + \frac{a_1 X_1}{\sqrt{X_1}} + \frac{a_2 X_2}{\sqrt{X_1}} + \frac{a_3 X_3}{\sqrt{X_1}} + \frac{\varepsilon_t}{\sqrt{X_1}}$$

On l'estime par les MCO :

```
Modèle 2: MCO, utilisant les observations 1-1325 (n = 1322)
Suppression d'observations manquantes ou incomplètes: 3
Variable dépendante: YT

            coefficient    éc. type    t de Student   p. critique
  ------------------------------------------------------------------
  const       0,790348    0,0859133       9,199       1,37e-019  ***
  X1T        -0,0101611   0,00171111     -5,938       3,68e-09   ***
  X2T         0,995919    0,0478352      20,82        1,85e-083  ***
  X3T         0,0621700   0,0123313       5,042       5,26e-07   ***

Moyenne var. dép.      1,220358   Éc. type var. dép.     0,789533
Somme carrés résidus   221,4571   Éc. type régression    0,409909
R2                     0,731066   R2 ajusté              0,730454
F(3, 1318)             1194,277   P. critique (F)        0,000000
Log de vraisemblance -694,8464    Critère d'Akaike       1397,693
Critère de Schwarz     1418,440   Hannan-Quinn           1405,471
```

$$\frac{Y}{\sqrt{X_1}} = 0{,}79 - \frac{0{,}01X_1}{\sqrt{X_1}} + \frac{0{,}996X_2}{\sqrt{X_1}} + \frac{0{,}062X_3}{\sqrt{X_1}} + \frac{\varepsilon_t}{\sqrt{X_1}}$$

On peut écrire le modèle estimé en utilisant les données initiales, soit :

$$Y = 0{,}79 - 0{,}01X_1 + 0{,}996X_2 + 0{,}062X_3 + \varepsilon_t$$

Dans ce modèle, X2 est la variable la plus significative. La distance des tirs (X2) et la fréquence de ceux-ci réalisés dans le corner (X3) influencent positivement le pourcentage de réussite à 3 points. Le nombre de minutes jouées (X1) l'influence négativement de manière très légère.

V. Conclusion

Dans un premier temps, cette étude a permis de sélectionner le modèle optimal d'explication du pourcentage de réussite à 3 points des joueurs de NBA. Celui-ci a pour variables explicatives le nombres de minutes jouées (X1) et la distance moyenne des tirs tentés (X2). L'observation du nuage de points entre e(t) et e(t-1), ainsi que les tests de Durbin-Watson et Breusch Godfrey permettent de conclure une absence d'autocorrélation des erreurs. Néanmoins, les tests de Goldfeld-Quandt et White rejettent l'hypothèse d'homoscédasticité. Dans un dernier temps, nous avons élaboré un modèle optimal permettant de répondre à cette problématique en stabilisant les variances des perturbations. Pour améliorer la capacité explicative de notre modèle, nous pourrions intégrer de nouvelles variables à celui-ci. On pense notamment au poste du joueur sur le terrain ou à son nombre d'années d'expérience en NBA.