David DERUE
Valentin CRIGNON

25 juin 2010

M8 – Dossier de Projet

Recherche de liens entre les épreuves d'un Décathlon

Usain Bolt – Recordman du monde du 100m (9,58s)

A l'attention de Stéphane CANU

SOMMAIRE

INTRODUCTION : Genèse du projet

Au début de notre EC de M8 sur le principe du traitement de l'information, nous avons dû faire des recherches afin de choisir un sujet de projet nous permettant d'effectuer une étude statistique sur une base de données. Nous avions tout d'abord pensé à faire une étude sur l'évolution de la démographie en France, en comparant les natalités et mortalités depuis les années 1960 jusqu'à nos jours, puis de l'étendre à d'autres pays de continents différents pour enfin essayer de prédire les évolutions futures. Cependant, après avoir rassemblé les données et commencé notre étude sous *Matlab*, nous nous sommes rendu compte que les évolutions de la démographie étaient les mêmes dans la plupart des pays étudiés et augmentaient de manière similaire. Ceci n'aurait donc pas été très intéressant à étudier.

Après l'aval de notre bienveillant professeur de statistiques, nous avons donc cherché à nouveau des données sur un autre thème qui nous intéressait. Nous avons finalement trouvé une base de données sur des résultats obtenus par des athlètes sportifs lors d'un décathlon. L'idée est désormais d'étudier les aptitudes sportives des athlètes dans chaque discipline. Ceci nous permettra de répondre à de multiples questions, comme de définir si certaines épreuves sont complémentaires, dans le sens où un athlète ayant obtenu un bon score dans une discipline aura de la même manière un bon score dans l'autre, quelles épreuves sont les plus importantes pour terminer sur le podium, ou encore quelle stratégie faut-il statistiquement adopter pour finir la compétition le mieux classé possible.

1. DESCRIPTION DES DONNEES

1.1 Présentation des données

Nous disposons, grâce à notre base de données, de **374 mesures**. Celles-ci sont réparties de manière suivante : nous possédons les résultats de **34 athlètes** dans **10 disciplines** différentes ainsi que le **score final** du décathlon pour chaque individu. Ces résultats sont présentés dans le tableau ci-dessous.

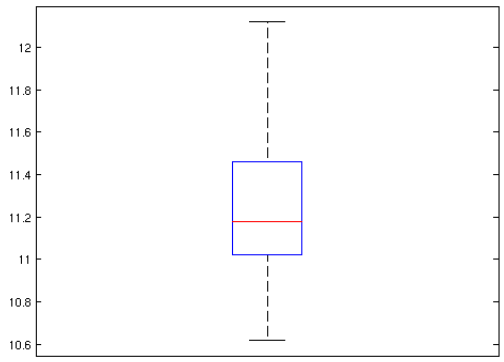
Candidat	100 Mètres	Longueur	Poids	Hauteur	400 Mètres	110 Mètres H	Disque	Perche	Javelot	1500 Mètres	Score
1	11,25	7,43	15,48	2,27	48,9	15,13	49,28	4,7	61,32	268,95	8488
2	10,87	7,45	14,97	1,97	47,71	14,46	44,36	5,1	61,76	273,02	8399
3	11,18	7,44	14,2	1,97	48,29	14,81	43,66	5,2	64,16	263,2	8328
4	10,62	7,38	15,02	2,03	49,06	14,72	44,8	4,9	64,04	285,11	8306
5	11,02	7,43	12,92	1,97	47,44	14,4	41,2	5,2	57,46	256,64	8286
6	10,83	7,72	13,58	2,12	48,34	14,18	43,06	4,9	52,18	274,07	8272
7	11,18	7,05	14,12	2,06	49,34	14,39	41,68	5,7	61,6	291,2	8216
8	11,05	6,95	15,34	2	48,21	14,36	41,32	4,8	63	265,86	8189
9	11,15	7,12	14,52	2,03	49,15	14,66	42,36	4,9	66,46	269,62	8180
10	11,23	7,28	15,25	1,97	48,6	14,76	48,02	5,2	59,48	292,24	8167
11	10,94	7,45	15,34	1,97	49,94	14,25	41,86	4,8	66,64	295,89	8143
12	11,18	7,34	14,48	1,94	49,02	15,11	42,76	4,7	65,84	256,74	8114
13	11,02	7,29	12,92	2,06	48,23	14,94	39,54	5	56,8	257,85	8093
14	10,99	7,37	13,61	1,97	47,83	14,7	43,88	4,3	66,54	268,97	8083
15	11,03	7,45	14,2	1,97	48,94	15,44	41,66	4,7	64	267,48	8036
16	11,09	7,08	14,51	2,03	49,89	14,78	43,2	4,9	57,18	268,54	8021
17	11,46	6,75	16,07	2	51,28	16,06	50,66	4,8	72,6	302,42	7869
18	11,57	7	16,6	1,94	49,84	15	46,66	4,9	60,2	286,04	7860
19	11,07	7,04	13,41	1,94	47,97	14,96	40,38	4,5	51,5	262,41	7859
20	10,89	7,07	15,84	1,79	49,68	15,38	45,32	4,9	60,48	277,84	7781
21	11,52	7,36	13,93	1,94	49,99	15,64	38,82	4,6	67,04	266,42	7753
22	11,49	7,02	13,8	2,03	50,6	15,22	39,08	4,7	60,92	262,93	7745
23	11,38	7,08	14,31	2	50,24	14,97	46,34	4,4	55,68	272,68	7743
24	11,3	6,97	13,23	2,15	49,98	15,38	38,72	4,6	54,34	277,84	7623
25	11	7,23	13,15	2,03	49,73	14,96	38,06	4,5	52,82	285,57	7579
26	11,33	6,83	11,63	2,06	48,37	15,39	37,52	4,6	55,42	270,07	7517
27	11,1	6,98	12,69	1,82	48,63	15,13	38,04	4,7	49,52	261,9	7505
28	11,51	7,01	14,17	1,94	51,16	15,18	45,84	4,6	56,28	303,17	7422
29	11,26	6,9	12,41	1,88	48,24	15,61	38,02	4,4	52,68	272,06	7310
30	11,5	7,09	12,94	1,82	49,27	15,56	42,32	4,5	53,5	293,85	7237
31	11,43	6,22	13,98	1,91	51,25	15,88	46,18	4,6	57,84	294,99	7231
32	11,47	6,43	12,33	1,94	50,3	15	38,72	4	57,26	293,72	7016
33	11,57	7,19	10,27	1,91	50,71	16,2	34,36	4,1	54,94	269,98	6907
34	12,12	5,83	9,71	1,7	52,32	17,05	27,1	2,6	39,1	281,24	5339

Tableau des données *decathlon.txt*

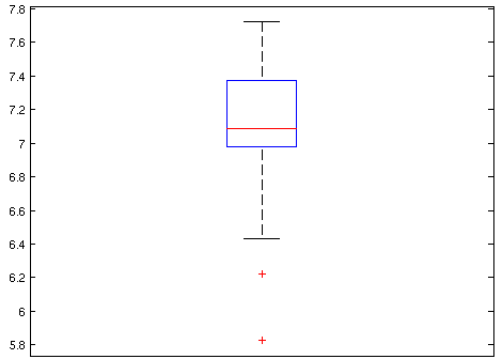
Bien évidemment, jongler avec un si grand nombre de variables est très compliqué en les analysant une par une et sans aucun logiciel d'aide aux statistiques. C'est pourquoi nos outils principaux lors de cette étude seront ***Matlab R2010a*** et ***R 2.11.1***.

1.2 Boîtes à moustache

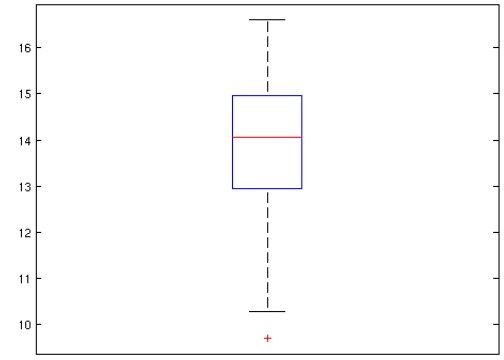
Nous allons nous intéresser aux disciplines, une par une. Afin de les analyser, nous avons tracé à l'aide du logiciel *Matlab* les boîtes à moustache de chaque discipline. Ceci va nous permettre de définir le genre de répartition des résultats pour chaque épreuve, et s'il y a des anomalies observées.



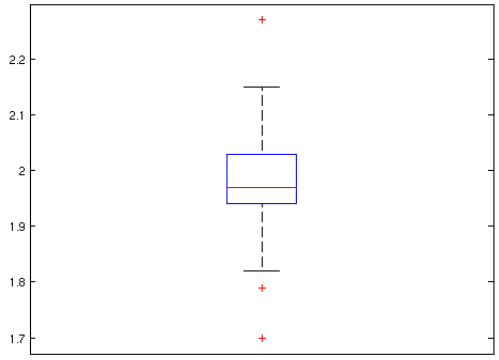
1 - ***100 mètres***



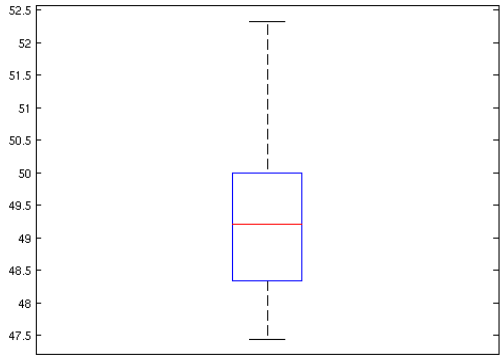
2 - ***Longueur***



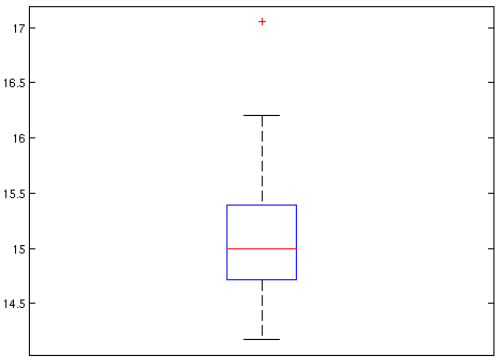
3 - ***Poids***



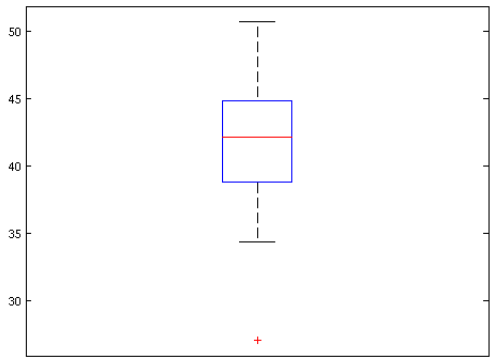
4 - ***Hauteur***



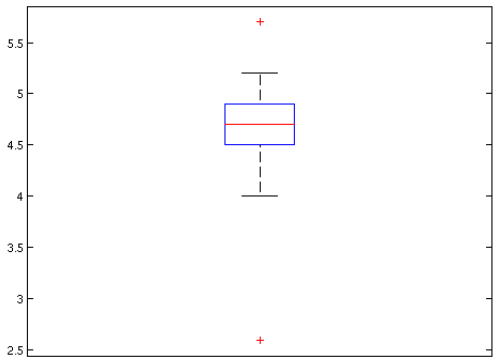
5 - ***400 mètres***



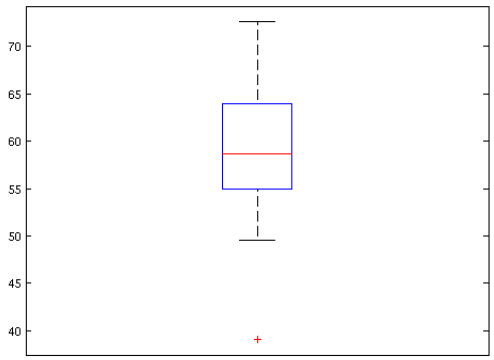
6 - ***110 mètres haies***



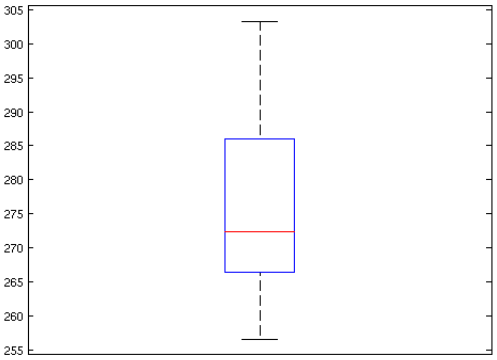
7 - ***Disque***



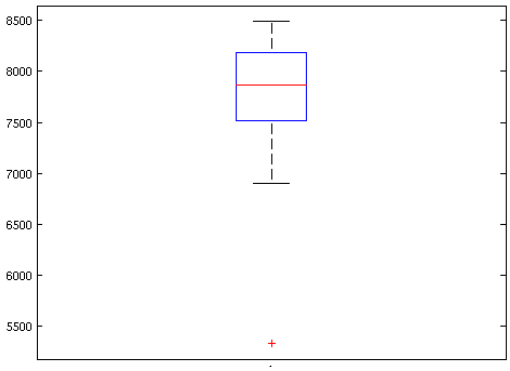
8 - ***Perche***



9 - ***Javelot***



10 – ***1500 mètres***



11 - ***Score final***

Nous pouvons alors observer le fait que la répartition des mesures (résultats de disciplines) est assez hétérogène en fonction de l'épreuve choisie. En effet, certaines mesures se trouvent parfois hors épure, ce qui signifie que certains athlètes ont eu de réelles difficultés par rapport aux autres candidats sur des épreuves, ou qu'ils ont réussi avec brio (ce qui est plus rare). Par exemple, notre individu 34 fut dernier dans la plupart des épreuves (hormis au lancer de javelot). On peut observer que ces résultats sont pour la plupart hors épure dans la majorité des boîtes à moustache que nous avons créées. Inversement, la boîte à moustache 4 concernant le saut en hauteur nous prouve que des athlètes peuvent aussi avoir des résultats très bons se trouvant hors épure. Cela signifie qu'il a survolé l'épreuve et

surclassé ses rivaux. C'est le cas de l'individu 1 en saut en hauteur ou encore l'individu 7 à la perche.

De plus, en observant la boîte à moustache correspondant au score final du décathlon, nous constatons qu'une mesure se trouve de loin hors épure. Cela signifie qu'un athlète fut largement devancé durant la compétition par les autres participants, puisque la mesure de son score final ne se trouve pas dans la boîte à moustache : c'est bien évidemment l'individu 34...

Par ailleurs, nous observons que les résultats au-dessus de la médiane sont de plus en plus proches comparés à ceux en-dessous. En effet, plus de la moitié des scores des athlètes sont supérieurs à la médiane. Or la répartition des scores sous la médiane est beaucoup plus importante. Cela signifie donc que la compétition fut serrée.

1.3 Analyse en Composantes Principales

On se propose de réaliser l'ACP de nos données avec le logiciel *R*, les graphiques étant réalisés sous *Office Excel 2010*.

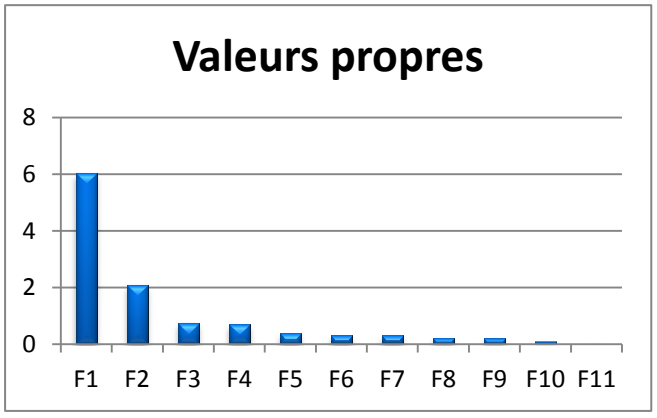
Valeurs propres :

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
V. propre	6,01	2,08	0,74	0,69	0,38	0,3	0,29	0,22	0,2	0,08	0,00
% Variance	54,63	18,91	6,74	6,26	3,43	2,75	2,6	2,03	1,86	0,75	0,03
% Cumulé	54,63	73,54	80,28	86,54	89,97	92,72	95,32	97,35	99,22	99,97	100

Nous allons nous intéresser **à F1 & F2** qui représentent **73,54 %** de la variance.

Contribution des variables (%) :

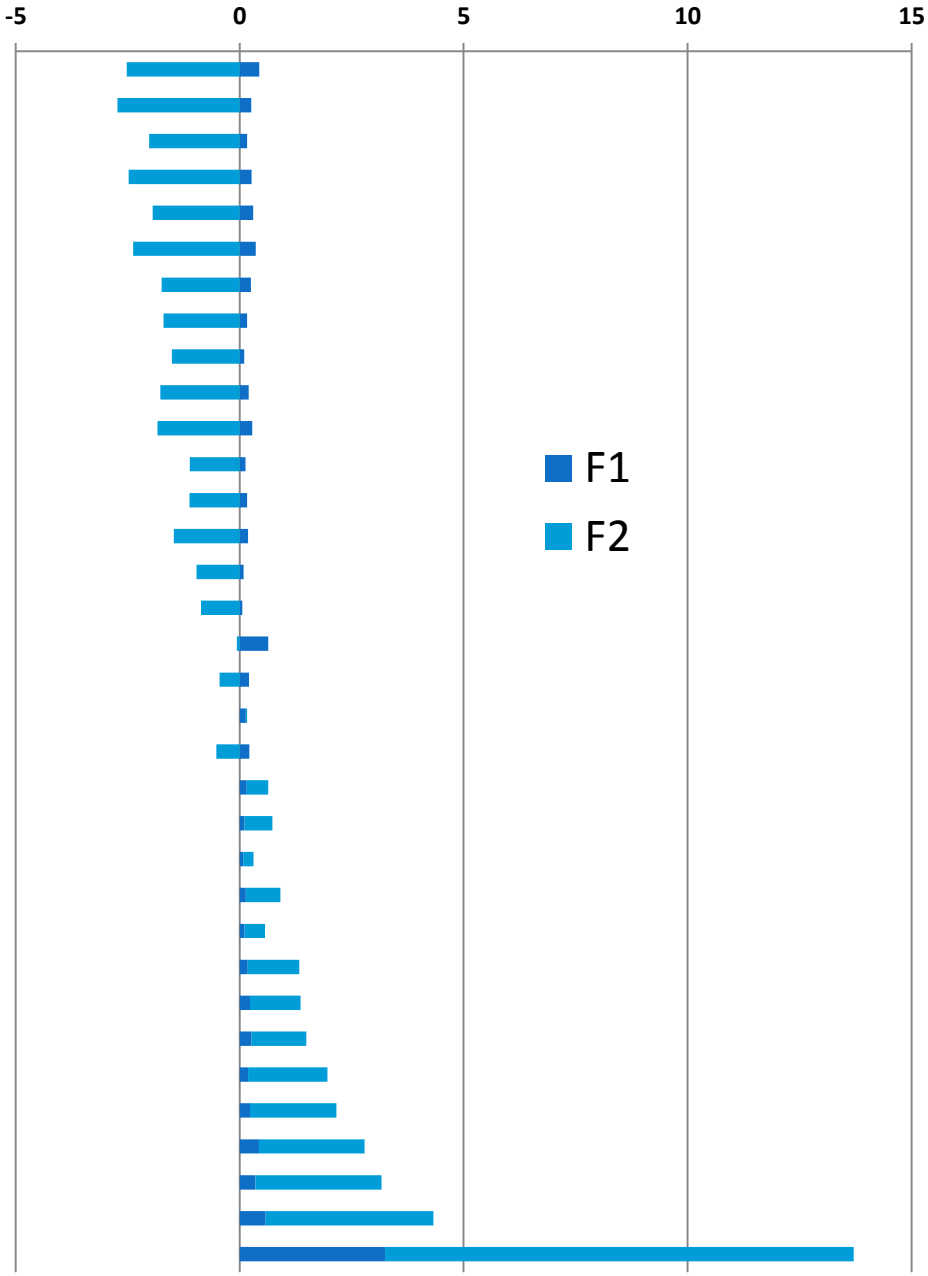
	F1	F2
100 mètres	0,8	-0,29
Longueur	-0,81	0,28
Poids	-0,72	-0,57
Hauteur	-0,61	0,01
400 mètres	0,66	-0,61
110 mètres H	0,83	-0,19
Disque	-0,69	-0,6
Perche	-0,87	-0,09
Javelot	-0,66	-0,43
1500 mètres	0,2	-0,79
Score final	-0,99	0,01



Les signes des contributions s'expliquent par le fait que deux sortes d'épreuves existent : les épreuves de distance, où il faut réaliser la plus grande valeur possible (lancers et sauts) et les épreuves de vitesse, où il faut réaliser le chrono le plus faible (voir schéma plus loin). Selon F1, on remarque que les contributions des courses sont négatifs, et ceux des sauts et lancers, positifs.

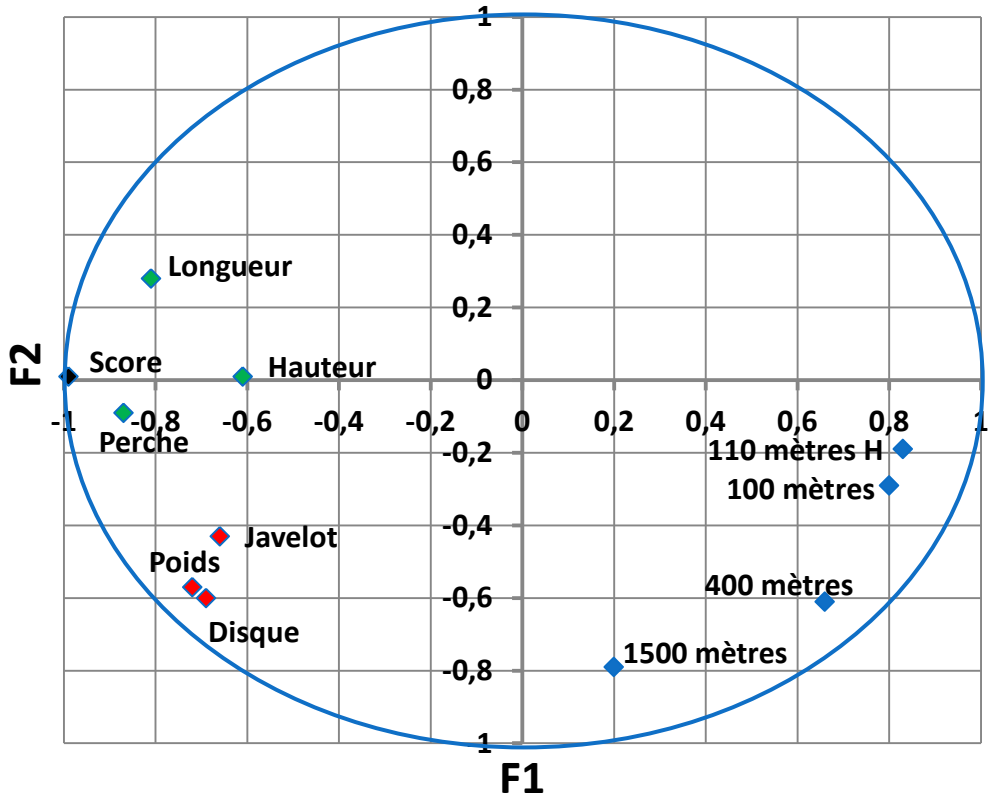
Contribution des individus (%) :

	F1	F2
1	0,44	-2,52
2	0,26	-2,73
3	0,17	-2,02
4	0,27	-2,48
5	0,3	-1,94
6	0,36	-2,38
7	0,25	-1,74
8	0,17	-1,7
9	0,1	-1,51
10	0,2	-1,77
11	0,28	-1,83
12	0,13	-1,11
13	0,17	-1,12
14	0,19	-1,47
15	0,09	-0,96
16	0,06	-0,86
17	0,64	-0,06
18	0,21	-0,45
19	0,14	0,03
20	0,22	-0,52
21	0,16	0,48
22	0,11	0,62
23	0,08	0,23
24	0,14	0,77
25	0,11	0,46
26	0,18	1,15
27	0,23	1,13
28	0,27	1,22
29	0,19	1,77
30	0,23	1,93
31	0,43	2,36
32	0,35	2,82
33	0,58	3,75
34	3,25	10,46



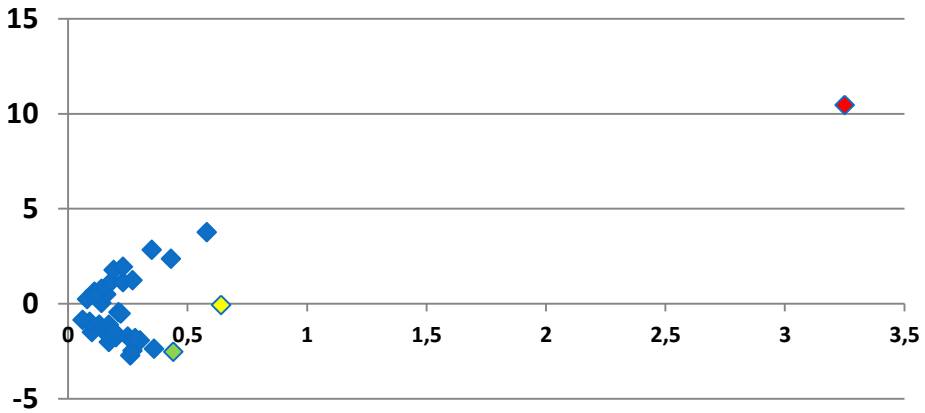
Les meilleurs individus contribuent à l'amélioration du groupe, tandis que les moins bons, contribuent à faire baisser le groupe. Mais on remarque quand même qu'un individu possède une contribution de près de **15%** à lui-seul…

Variables (axes F1 et F2 : 73,54% variance)



Dans ce graphique, on distingue bien les trois régions : les **lancers**, les **sauts**, et les **courses**. Les résultats aux épreuves d'un même groupe sont donc liés entre eux. *Quelles « lois » peut-on établir entre ces épreuves ?* De plus, le score final se trouve dans la zone des sauts. *Quel lien y a-t-il entre les résultats aux sauts et le score final ?*

Position des individus selon les axes F1 et F2



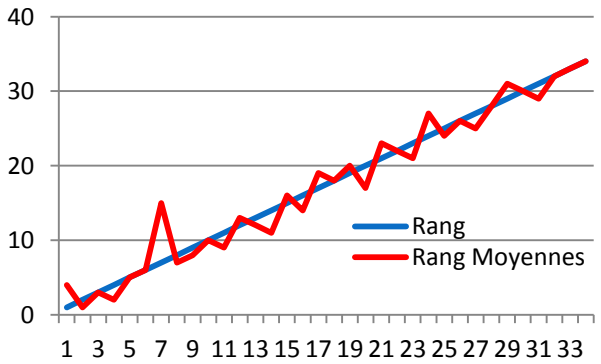
On remarque ici qu'un concurrent est très largement hors du groupe. Ce concurrent a eu des résultats très en deçà de ceux des autres participants (n°**34**, en rouge). *Mais, bien que ses résultats soient éloignés, sont-ils liés en eux, et obéissent-ils aux mêmes « lois » que pour les autres membres du groupe ?* Le **n°1** est en vert, et le **n°17** (médian) est en jaune.

Pour tenter d'expliquer certaines lois ou liens, nous allons utiliser la régression.

1.4 Rang et Moyenne des Rangs

Nous avons cette fois dans le tableau ci-dessous classé les athlètes non plus à partir des points accordés selon chaque épreuve, mais par le classement qu'ils ont obtenu dans chaque discipline. Nous avons alors reporté les résultats sur un graphe ci-dessous.

Individus	100m	Longueur	Poids	Hauteur	400m	110 m	Disque	Perche	Javelot	1500m	Moyenne	Rang moyenne
1	20	7	4	1	13	21	2	16	13	12	10,9	**4**
2	3	2	9	17	2	6	10	5	11	19	8,4	**1**
3	16	5	15	19	8	12	12	3	7	7	10,4	**3**
4	1	8	8	7	16	9	9	9	8	24	9,9	**2**
5	8	6	27	21	1	5	22	4	19	1	11,4	**5**
6	2	1	22	3	9	1	14	11	31	20	11,4	**6**
7	17	21	17	4	19	4	19	1	12	27	14,1	**15**
8	11	28	6	13	5	3	21	15	10	8	12	**7**
9	15	16	10	8	17	7	16	12	5	14	12	**8**
10	19	13	7	16	11	10	3	2	17	28	12,6	**10**
11	5	3	5	15	24	2	18	14	3	32	12,1	**9**
12	18	11	12	23	15	19	15	17	6	2	13,8	**13**
13	9	12	28	5	6	13	24	6	22	3	12,8	**12**
14	6	9	21	20	3	8	11	31	4	13	12,6	**11**
15	10	4	14	18	14	27	20	18	9	10	14,4	**16**
16	13	18	11	9	23	11	13	10	21	11	14	**14**
17	26	31	2	12	33	32	1	13	1	33	18,4	**19**
18	32	25	1	22	22	17	4	7	16	26	17,2	**18**
19	12	22	23	26	4	14	23	27	32	5	18,8	**20**
20	4	20	3	33	20	24	8	8	15	21	15,6	**17**
21	31	10	19	25	26	30	26	23	2	9	20,1	**23**
22	28	23	20	10	29	23	25	19	14	6	19,7	**22**
23	24	19	13	14	27	16	5	29	24	18	18,9	**21**
24	22	27	24	2	25	25	28	24	27	22	22,6	**27**
25	7	14	25	11	21	15	29	28	29	25	20,4	**24**
26	23	30	32	6	10	26	32	25	25	16	22,5	**26**
27	14	26	29	32	12	20	30	20	33	4	22	**25**
28	30	24	16	24	31	22	7	22	23	34	23,3	**28**
29	21	29	30	30	7	29	31	30	30	17	25,4	**31**
30	29	17	26	31	18	28	17	26	28	30	25	**30**
31	25	33	18	28	32	31	6	21	18	31	24,3	**29**
32	27	32	31	27	28	18	27	33	20	29	27,2	**32**
33	33	15	33	29	30	33	33	32	26	15	27,9	**33**
34	34	34	34	34	34	34	34	34	34	23	32,9	**34**



Nous pouvons donc observer que les classements ne sont donc pas les mêmes. Ceci signifie donc bien que certaines disciplines rapportent plus de points que d'autres. Nous nous sommes aussi attardés sur une assez importante anomalie qui est l'individu 7. En effet, alors qu'il était 7ème dans le premier classement, il se retrouve 15ème. Or lorsqu'on prête plus attention à ses résultats, il a terminé 4ème en saut en hauteur et 1er en saut à la perche ! Ceci appuie l'hypothèse selon laquelle les sauts sont les épreuves qui rapportent le plus de points.

2 REGRESSION

2.4 Matrice de corrélation

Nous allons dans cette partie de notre dossier essayer de définir s'il existe un lien entre les aptitudes des athlètes dans deux disciplines différentes. Pour se faire, nous allons utiliser la régression linéaire sur certaines paires de variables, en fonction du coefficient de corrélation trouvé pour chaque paire.

Après de nombreux essais, nous avons finalement trouvé certaines informations très utiles, se traduisant par des liens entre plusieurs disciplines. Nous avons tout d'abord calculé la matrice de corrélation des mesures grâce au logiciel *Matlab* avec le code ci-dessous :

```
X=data;
[n,p]=size(X);
cov(X);
Xn = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X)+eps);
matricecorrelation = cov(Xn)
```

Nous avons finalement trouvé la matrice suivante :

$$\begin{bmatrix} 1 & -0{,}69 & -0{,}42 & -0{,}36 & 0{,}70 & 0{,}75 & -0{,}35 & -0{,}63 & -0{,}34 & 0{,}25 & -0{,}76 \\ -0{,}69 & 1 & 0{,}39 & 0{,}47 & -0{,}64 & -0{,}65 & 0{,}37 & 0{,}63 & 0{,}45 & -0{,}36 & 0{,}80 \\ -0{,}42 & 0{,}39 & 1 & 0{,}32 & -0{,}14 & -0{,}49 & 0{,}86 & 0{,}64 & 0{,}70 & 0{,}20 & 0{,}72 \\ -0{,}36 & 0{,}47 & 0{,}32 & 1 & -0{,}28 & -0{,}49 & 0{,}38 & 0{,}47 & 0{,}34 & -0{,}13 & 0{,}63 \\ 0{,}70 & -0{,}64 & -0{,}14 & -0{,}28 & 1 & 0{,}65 & -0{,}15 & -0{,}52 & -0{,}15 & 0{,}55 & -0{,}65 \\ \mathbf{0{,}75} & -0{,}65 & -0{,}49 & -0{,}49 & 0{,}65 & 1 & -0{,}40 & -0{,}71 & -0{,}35 & 0{,}15 & -0{,}80 \\ -0{,}35 & 0{,}37 & \mathbf{0{,}86} & 0{,}38 & -0{,}15 & -0{,}40 & 1 & 0{,}62 & 0{,}62 & 0{,}29 & 0{,}67 \\ -0{,}63 & 0{,}63 & 0{,}64 & 0{,}47 & -0{,}52 & -0{,}71 & 0{,}62 & 1 & 0{,}56 & -0{,}07 & 0{,}87 \\ -0{,}34 & 0{,}45 & 0{,}70 & 0{,}34 & -0{,}15 & -0{,}35 & 0{,}62 & 0{,}56 & 1 & 0{,}05 & 0{,}67 \\ 0{,}25 & -0{,}36 & 0{,}20 & -0{,}13 & 0{,}55 & 0{,}15 & 0{,}29 & -0{,}07 & 0{,}05 & 1 & -0{,}26 \\ -0{,}76 & \mathbf{0{,}80} & 0{,}72 & 0{,}63 & -0{,}65 & -0{,}80 & 0{,}67 & \mathbf{0{,}87} & 0{,}67 & -0{,}26 & 1 \end{bmatrix}$$

Matrice de corrélation

Nous pouvons donc bien observer que la diagonale de cette matrice est composée des mêmes valeurs 1 ce qui est normal puisqu'il s'agit de la corrélation entre une même variable.

Les conclusions que nous pouvons alors en tirer sont les suivantes. D'importants coefficients de corrélation sont obtenus entre les variables :

- **100 mètres** (1) avec **110 mètres haies** (6) *r = 0,75*
- **Saut en longueur** (2) avec le **Score** (11) *r = 0,80*
- **Lancer de poids** (3) avec **Lancer de Disque** (7) *r = 0,86*
- **Perche** (8) avec le **Score** (11) *r = 0,87*

Les autres disciplines ayant eu des coefficients inférieurs à 0,70, nous n'en tiendront plus compte dans le reste de ce rapport. En effet, un coefficient de corrélation aussi petit signifie qu'il n'existe pas réellement une linéarité entre les variables.

Inversement, nous observons que certaines disciplines possèdent un coefficient de corrélation très faible entre elles, ce qui empêche d'utiliser une régression linéaire puisque le modèle serait erroné. On en déduit qu'il n'existe aucun lien concernant les résultats sportifs d'un même individu dans les deux épreuves observée. C'est le cas des variables **javelot** et **1500m** qui ne sont nullement corrélées (r = 0,05).

Par la suite, pour les variables possédant un coefficient de corrélation assez élevé, nous devons nous assurer que le modèle calculé est correct. C'est pourquoi nous allons calculer le coefficient de détermination R^2. Ce dernier se détermine comme suit :

$$R^2 = \frac{SC_{Expliqué}}{SC_{Total}} = \frac{\sum_{i=1}^{n} (zi - \bar{y})^2}{\sum_{i=1}^{n} (yi - \bar{y})^2}$$

Plus le coefficient de détermination va se rapprocher de 1, plus le modèle pourra être considéré comme étant bon. Plus il se rapproche de 0, plus le modèle sera considéré comme étant mauvais.

Le code que nous avons utilisé sous *Matlab* pour calculer nos coefficients de détermination est le suivant :

```
Z=X*A;
% Calcul du coefficient de corr R2
% Erreurs (Résidus):
e  = Y-Z;
% Variance totale
SCT=(Y-mean(Y))'*(Y-mean(Y));
% Variance Expliquée
SCM=(Z-mean(Y))'*(Z-mean(Y));
% coefficient de détermination R2 :
R2=SCM/SCT
```

2.5 Recherche de liens linéaires entre les variables

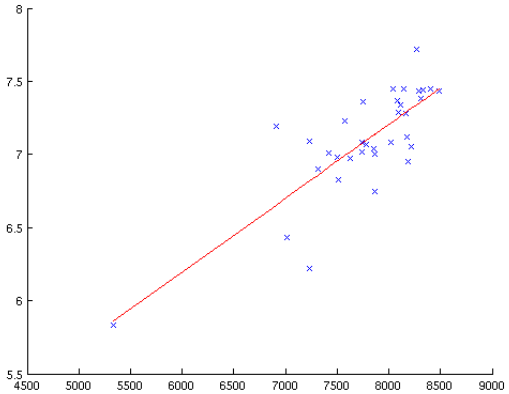
2.5.1 Saut en longueur – Score

Nous allons tout d'abord commencer par calculer le coefficient de détermination des variables **Saut en longueur** (2) et **Score** (11).

R2 =

0.6432

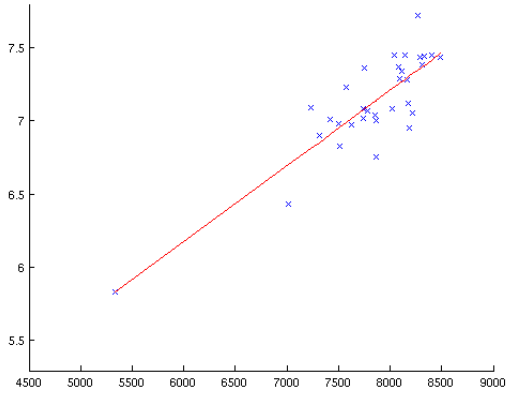
X	Y	e	c
8488.00	7.43	-0.02	0.00
8399.00	7.45	0.04	0.00
8328.00	7.44	0.07	0.00
8306.00	7.38	0.02	0.00
8286.00	7.43	0.08	0.00
8272.00	7.72	0.38	0.07
8216.00	7.05	-0.26	0.03
8189.00	6.95	-0.35	0.05
8180.00	7.12	-0.18	0.01
8167.00	7.28	-0.01	0.00
8143.00	7.45	0.17	0.01
8114.00	7.34	0.08	0.00
8093.00	7.29	0.04	0.00
8083.00	7.37	0.12	0.01
8036.00	7.45	0.23	0.02
8021.00	7.08	-0.14	0.01
7869.00	6.75	-0.39	0.05
7860.00	7.00	-0.13	0.01
7859.00	7.04	-0.09	0.00
7781.00	7.07	-0.02	0.00
7753.00	7.36	0.28	0.02
7745.00	7.02	-0.06	0.00
7743.00	7.08	0.01	0.00
7623.00	6.97	-0.04	0.00
7579.00	7.23	0.24	0.02
7517.00	6.83	-0.13	0.01
7505.00	6.98	0.03	0.00
7422.00	7.01	0.10	0.00
7310.00	6.90	0.04	0.00
7237.00	7.09	0.27	0.04
7231.00	**6.22**	**-0.60**	**0.21**
7016.00	6.43	-0.28	0.07
6907.00	**7.19**	**0.54**	**0.32**
5339.00	5.83	-0.03	0.03



Nous trouvons après calcul $\mathbf{R^2 = 0.64}$. Nous allons alors pour tenter d'améliorer ce coefficient calculer les résidus et contributions de chaque point pour déterminer des points que nous pourrions considérer comme aberrant pour notre régression comme suit ci-dessous.

Ainsi grâce au tableau ci-dessus, nous constatons que les individus **31** et **33** ont tous deux une forte contribution et un fort résidu comparé aux autres individus (**|e|> 0,50**). Nous décidons alors de les enlever et de calculer le coefficient de détermination sans ces points :

```
p_ab =

    31    33


A =

    0.0005
    3.0695


R2 =

    0.7382
```

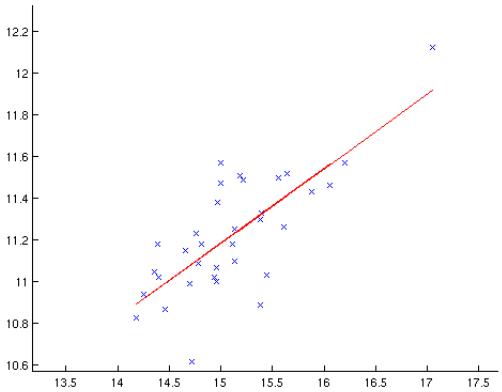


Nous trouvons alors **R^2 = 0.74**, ce qui est le meilleur résultat que nous puissions obtenir avec ces variables. Bien qu'il ne soit pas très proche de 1, nous allons considérer que le modèle est bon, car nous avons des données basées sur des résultats sportifs. En effet, de nombreux facteurs sont en jeu lors de l'étude de telles variables, ce qui fait qu'il est beaucoup plus difficile de trouver des coefficients de détermination très proche de 1. L'équation du modèle que nous avons obtenue est finalement :

$$y = 0,0005x + 3,0695$$

2.5.2 100 mètres – 110 mètres haies

Ensuite, nous réalisons les mêmes calculs entre les variables **100 mètres** et **110 mètres**. Nous trouvons alors le résultat **R^2 = 0.56**. Ce résultat est vraiment très faible et trop éloigné de 1 pour que l'on puisse admettre que le modèle soit bon. Il faut enlever beaucoup trop de valeurs pour pouvoir enfin trouver un coefficient de détermination convenable pour la suite de notre étude. Nous concluons donc que ce modèle est mauvais. En conséquence, nous n'admettons plus de lien entre les variables **100 mètres** et **110 mètres**.

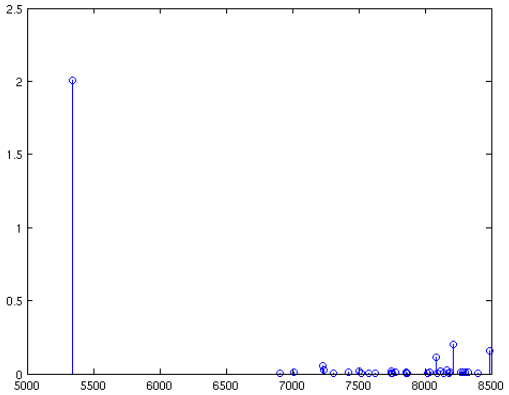


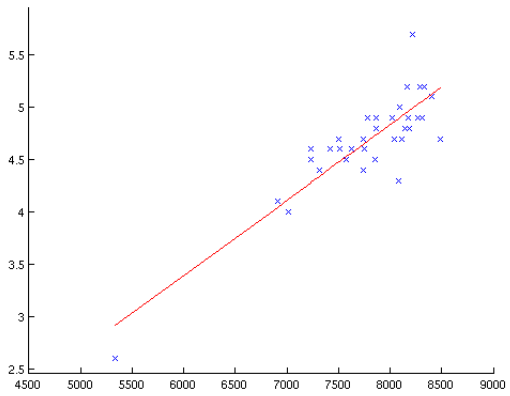
2.5.3 Perche – Score final

Le calcul du coefficient de détermination avec les variables **Perche** et **Score** nous donne **R^2 = 0.75**. Cependant, en étudiant le graphe avec le nuage de points, nous avons observé que les individus **7** et **14** possèdent un résidu plus important que les autres (**|e|> 0,50**). Cette hypothèse nous est confirmée par le tableau ci-dessous que nous avons calculé :

```
R2 =

    0.7573
```

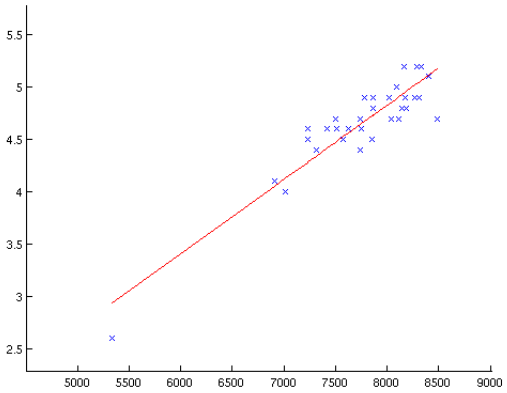
X	Y	e	c
8488.00	4.70	-0.49	0.16
8399.00	5.10	-0.02	0.00
8328.00	5.20	0.13	0.01
8306.00	4.90	-0.15	0.01
8286.00	5.20	0.16	0.01
8272.00	4.90	-0.13	0.01
8216.00	**5.70**	**0.71**	**0.20**
8189.00	4.80	-0.17	0.01
8180.00	4.90	-0.06	0.00
8167.00	5.20	0.25	0.02
8143.00	4.80	-0.14	0.01
8114.00	4.70	-0.22	0.02
8093.00	5.00	0.10	0.00
8083.00	**4.30**	**-0.59**	**0.11**
8036.00	4.70	-0.16	0.01
8021.00	4.90	0.05	0.00
7869.00	4.80	0.06	0.00
7860.00	4.90	0.17	0.01
7859.00	4.50	-0.23	0.01
7781.00	4.90	0.22	0.01
7753.00	4.60	-0.05	0.00
7745.00	4.70	0.05	0.00
7743.00	4.40	-0.25	0.02
7623.00	4.60	0.04	0.00
7579.00	4.50	-0.03	0.00
7517.00	4.60	0.12	0.00
7505.00	4.70	0.22	0.02
7422.00	4.60	0.18	0.01
7310.00	4.40	0.06	0.00
7237.00	4.50	0.22	0.02
7231.00	4.60	0.32	0.05
7016.00	4.00	-0.12	0.01
6907.00	4.10	0.06	0.00
5339.00	2.60	-0.31	2.01





Nous avons alors tenté de retirer ces deux individus pour observer les changements que cela aurait sur le coefficient de détermination. Le nouveau coefficient de détermination calculé est désormais de **0.84** !

```
p_ab =

     7    14


A =

    0.0007
   -0.8527


R2 =

    0.8405
```



X	Y	e	c
8488.00	4.70	-0.47	0.27
8399.00	5.10	-0.01	0.00
//	//	//	//
6907.00	4.10	0.05	0.00
5339.00	**2.60**	**-0.34**	**4.04**

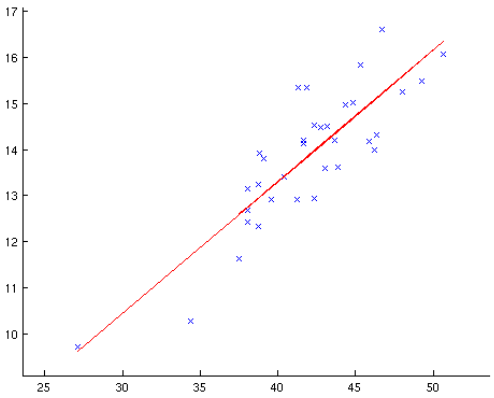
Avec le graphe des contributions, nous remarquons par ailleurs que l'individu 34 possède la contribution la plus élevée. Cela est dû au fait que son score est très éloigné de ceux des autres candidats, ce qui lui confère un **levier très important**. En revanche son résidu étant relativement faible, le point reste selon nous acceptable. On peut donc en conclure que le modèle sans les individus **7** et **14** est bon. L'équation de la droite est alors :

$$\boldsymbol{y = 0,0007\,x - 0,8527}$$

2.5.4 Lancers de poids et de disque

Les recherches suivantes portées sur les variables **Lancer de poids** et **Lancer de disque**. Son coefficient de détermination était à l'origine **R^2 = 0.73**. Avec la même méthode que précedemment, nous avons tenté d'améliorer ce coefficient. Cependant, aucune amélioration ne pu être observée, car aucun point ne se démarquait plus comme étant un point aberrant. Ainsi l'équation de la droite pour ces deux variables sera :

$$\boldsymbol{y = 0,28x + 1,8779}$$



Sur nos quatre modèles de départ possédant un coefficient de corrélation élevé, nous en avons donc finalement retenu trois, dont les modèles sont bons et approuvés par le coefficient de détermination. La régression nous aura donc permis de trouver un lien linéaire entre les aptitudes d'un sportif dans deux disciplines différentes.

Nous constatons finalement que les disciplines ayant un réel lien sont le **lancer de poids** et le **lancer de disque**. Ceci peut paraître logique, puisque même si chacune de ces disciplines nécessite une technique qui lui est propre, le lancer est basé sur la force d'impulsion que l'on donne à l'objet. Au contraire, il n'existe pas de lien évident entre l'épreuve du **100m** et celle du **1500m** car même si cela reste de la course, l'une fait appel à de la rapidité, alors que l'autre fait appel à de l'endurance.

Enfin, nous avons trouvé un lien entre le fait de réussir certaines disciplines, et d'être bien classé à la fin du décathlon. En effet, nous remarquons qu'une linéarité existe entre les résultats obtenus en **saut à la perche** ou alors en **saut en longueur** et le **score final** du décathlon. Nous nous sommes alors demandés si ces résultats étaient dus au pur hasard ou s'il y avait une explication à cela. Finalement, après avoir effectué plusieurs recherches, nous nous sommes rendu compte que le nombre ***de points obtenus dans ces épreuves était plus élevé*** ! En effet, les scores sont calculés de telle manière qu'un écart « proportionnel » entre deux athlètes sur une épreuve de course et de lancer sera à l'avantage du lancer. C'est-à-dire qu'il vaut mieux être très bon dans des disciplines de saut plutôt que dans celles de course, car cela rapporte plus de points au final. La relation entre le score final et les disciplines de saut en longueur et de saut à la perche est donc justifiée, puisque ces épreuves rapportent en général un plus grand nombre de points que les autres. Cet hypothèse est d'autant plus confirmé par le fait que si l'on observe les résultats des athlètes qui se retrouvent sur le podium, on constate qu'ils ont obtenu les meilleurs résultats notamment dans les disciplines de sauts, alors que leurs résultats en course se situent être dans la moyenne des résultats du décathlon.

Résumé des lois :

$$\boldsymbol{ResultatSautEnLongueur = 0,0005 \times ScoreFinal + 3,07}$$

$$\boldsymbol{ResultatSautPerche = 0,0007 \times ScoreFinal - 0,85}$$

$$\boldsymbol{LancerPoids = 0,29 \times LancerDisque + 1,88}$$

(programme intégral MatLab en annexe)

3 TESTS

Test de Student sur la pente de la régression

Nous allons maintenant nous attaquer à la partie test de notre étude. Cette partie aura pour but de valider nos hypothèses selon lesquelles il existe un lien entre les aptitudes d'un athlète sur deux disciplines.

Ayant trouvé précédemment des modèles, équations de droite après régression concernant certaines disciplines, nous allons procéder à des tests de **Student**, afin de s'assurer que les coefficients a et b de nos modèles sont acceptables.

Nous commençons par nos variables **Perche** et **Score**. L'équation que nous a donné la régression au sens des moindres carrés, est $\boldsymbol{y = 0,0007\,x - 0,8527}$.

Le but du test de Student va être de définir si notre pente est dûe au hasard.

Les hypothèses sont : $\begin{cases} H_0 : Indépendance & \boldsymbol{a = 0} \\ H_1 : Dépendance & \boldsymbol{a \neq 0} \end{cases}$

Avec *MatLab*, nous avons déjà calculé : $\begin{cases} \boldsymbol{\hat{a} = 0,0007} \\ \boldsymbol{\hat{b} = -0,8527} \end{cases}$

Calcul de la variance estimée :

$$\boldsymbol{\hat{\sigma}^2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2 = \boldsymbol{0,0426}$$

Calcul de la variance : $\boldsymbol{S_x^2 = 366836,64}$

Calcul de :

$$\boldsymbol{t} = \frac{\hat{a}}{\sqrt{\frac{\hat{\sigma}^2}{S_x^2}}} = \boldsymbol{2,05}$$

Nous avons ici **32** degrés de liberté. Avec la fonction `tpdf(2.05,32)` de MatLab, on trouve que la probabilité de Student, la **p-valeur**, est est égale à **0,05**.

On peut donc conclure que notre coefficient est correct, car il y a environ 5 chances sur 100 qu'il soit dû au hasard, ce qui nous semble relativement bas.

De la même manière, nous trouvons pour les couples de variables suivants :

- **Longueur** et **Score**
  $\boldsymbol{t = 1,26}$ donc $\boldsymbol{pval = 0,1779} > 0{,}05$
- **Poids** et **Disque**
- $\boldsymbol{t = 1,63}$ donc $\boldsymbol{pval = 0,1062} > 0{,}05$

En prenant un risque de première espèce de **5%**, nous ne pouvons pas conclure à la dépendance. Néanmoins, avec des chances d'erreurs de 10 à 20 %, nous pouvons toujours conserver nos lois à titre indicatif et avec prudence.

CONCLUSION

Pour conclure, nous pouvons affirmer à partir de ce dossier et à l'aide d'une ACP, d'une régression linéaire et enfin d'un test de Student que des relations existent réellement entre les aptitudes des sportifs pour certaines épreuves. Grâce au compte des points, nous savons désormais que la discipline du **saut à la perche** est une épreuve majeure lors du décathlon qui rapporte un nombre de points plus important que les autres. Ainsi, on peut supposer qu'après avoir obtenu les résultats de l'épreuve de perche, nous pourrons en déduire approximativement le classement final du décathlon.

De plus, le fait que le thème de notre projet soit libre, cela nous a permis de nous pencher sur un domaine qui nous intéressait, plutôt que d'avoir un sujet abstrait imposé. Ce rapport fut aussi bénéfique pour nous permettre d'apprendre par nous-même à faire des recherches et une étude statistique, d'utiliser des logiciels mathématiques et de restituer ce que l'on a appris durant tout notre semestre grâce à l'EC de M8.

Special thanks à M^{r} CANU.

ANNEXE

Notre programme *MatLab*
decathlon.m

```
Y=[data(:,3)]
X=[data(:,7) ones(34,1)]

indn=(1:size(X,1))';

p_ab=[]
X(p_ab,:)=[];
Y(p_ab)=[];
n=size(X);
p=2;

% Calcul des coefficients de la régression
A=inv((X')*X)*(X')*Y

figure(1)
hold on
plot(X(:,1),Y,'x')
plot(X,A(1)*X+A(2),'r')
hold off

Z=X*A;
% Erreurs (Résidus):
e  = Y-Z;
% Variance totale
SCT=(Y-mean(Y))'*(Y-mean(Y));
% Variance Expliquée
SCM=(Z-mean(Y))'*(Z-mean(Y));

% coefficient de dÃ©termination R2 :
R2=SCM/SCT

Yp=A(1)*X(:,1) + A(2);

e=Y-Yp;
figure(2)
stem(X(:,1),e,'or');

s=((e')*e)/(n(1,1)-1-p)

H = X*inv(X'*X)*X';
h = diag(H);
c=(h./(p*(1-h).*(1-h))).*((e.*e)/s);

stem(X(:,1),c,'o');

fprintf(1,'   X       Y       e       c \n');
fprintf(1,'-----------------------------------\n');
fprintf(1,'%6.2f  %6.2f  %6.2f  %6.2f  \n',[X(:,1) Y e c]');
```

Dossier édité le 25 juin 2010