MASTER 1 SIDE **Pôle Lamartine - ULCO**

INFORMATIQUE ET MODÈLES LINÉAIRES Novembre 2010 - Contrôle Continu, Semestre 1

Durée de l'épreuve : 2h00 Documents autorisés. Calculatrice autorisée

(Les deux exercices sont indépendants. Un soin tout particulier sera apporté à la rédaction des réponses)

Exercice 1

On dit que dans une famille, les aînés ont tendance à être plus indépendants que leurs cadets. Un chercheur élabore une échelle d'indépendance en 25 points et procède à l'évaluation de 20 aînés et du frère ou de la sœur qui suit directement chacun des aînés. Imaginons qu'il obtienne les résultats suivants :

i	Aîné	Cadet	i	Aîné	Cadet
1	8	9	11	17	13
2	13	15	12	12	8
3	8	10	13	2	7
4	5	7	14	13	8
5	12	10	15	19	14
6	15	13	16	18	12
7	5	8	17	14	8
8	15	12	18	17	11
9	16	13	19	18	12
10	5	9	20	20	10

Un collaborateur du premier chercheur suggère que la différence observée sur une paire dépend essentiellement du score de l'aîné

1. Pour chaque paire, on appelle x_i le score de l'aîné et y_i la différence (algébrique) des scores entre l'aîné et le cadet. Calculer le coefficient de corrélation R entre les deux séries étudiées. On pourra utiliser les résultats intermédiaires suivants :
$$\sum_i x_i = 252 \quad ; \quad \sum_i y_i = 43 \quad ; \quad \sum_i x_i^2 = 3722 \quad ; \quad \sum_i y_i^2 = 415 \quad ; \quad \sum_i x_i y_i = 920.$$
2. À l'aide d'un test de Student-Fisher au seuil de $\alpha = 5\%$, montrer que l'hypothèse H_0 : « $R = 0$ » n'est pas acceptable et que par conséquent, la corrélation est significative. On utilisera pour cela la statistique
$$T = \sqrt{n-2}\frac{R}{\sqrt{1-R^2}}$$
qui suit une loi de Student-Fisher à $\nu = n - 2$ degrés de liberté, ainsi que la table de l'annexe.
3. Déterminer une équation de la droite de régression des (y_i) par rapport aux (x_i).
4. Représenter sur un graphique le nuage de points $(x_i; y_i)$ et la droite déterminée au 3.
5. Donner une estimation du niveau d'indépendance d'un cadet dont le frère (aîné) a obtenu une note de 22
6. Suggérer d'autres facteurs de variation que l'on aurait pu prendre en compte dans une telle étude.

Correction :

1. On a $\overline{x} = \dfrac{252}{20} = 12,6$, $\overline{y} = \dfrac{43}{20} = 2,15$, $\sigma_X^2 = \dfrac{37222}{20} - (12,6)^2 = 27,34$ (donc $\sigma_X = 5,23$), $V(Y) = \sigma_Y^2 = \dfrac{415}{20} - (2,15)^2 = 16,13$ (donc $\sigma_Y = 4,015$) et $Cov(X,Y) = \dfrac{920}{20} - (12,6 \times 2,15) = 18,91$. Finalement, $R = \dfrac{18,91}{5,23 \times 4,015} = 0,9002$.

2. On détermine ensuite la valeur observée de la statistique T soit $T_{obs} = \sqrt{n-2}\dfrac{R}{\sqrt{1-R^2}} = 8,76$. Pour un nombre de degrés de liberté (d.d.l.) égal à 18, un seuil $\alpha = 5\%$, et un test bilatéral, on obtient grâce à la table de la loi de Student Fisher, $T_{lu} = 2,101$. On remarque que $T_{lu} < T_{obs}$ donc on rejette l'hypothèse H_0 selon laquelle R est négligeable. Le coefficient R est donc significatif (la corrélation est significative).

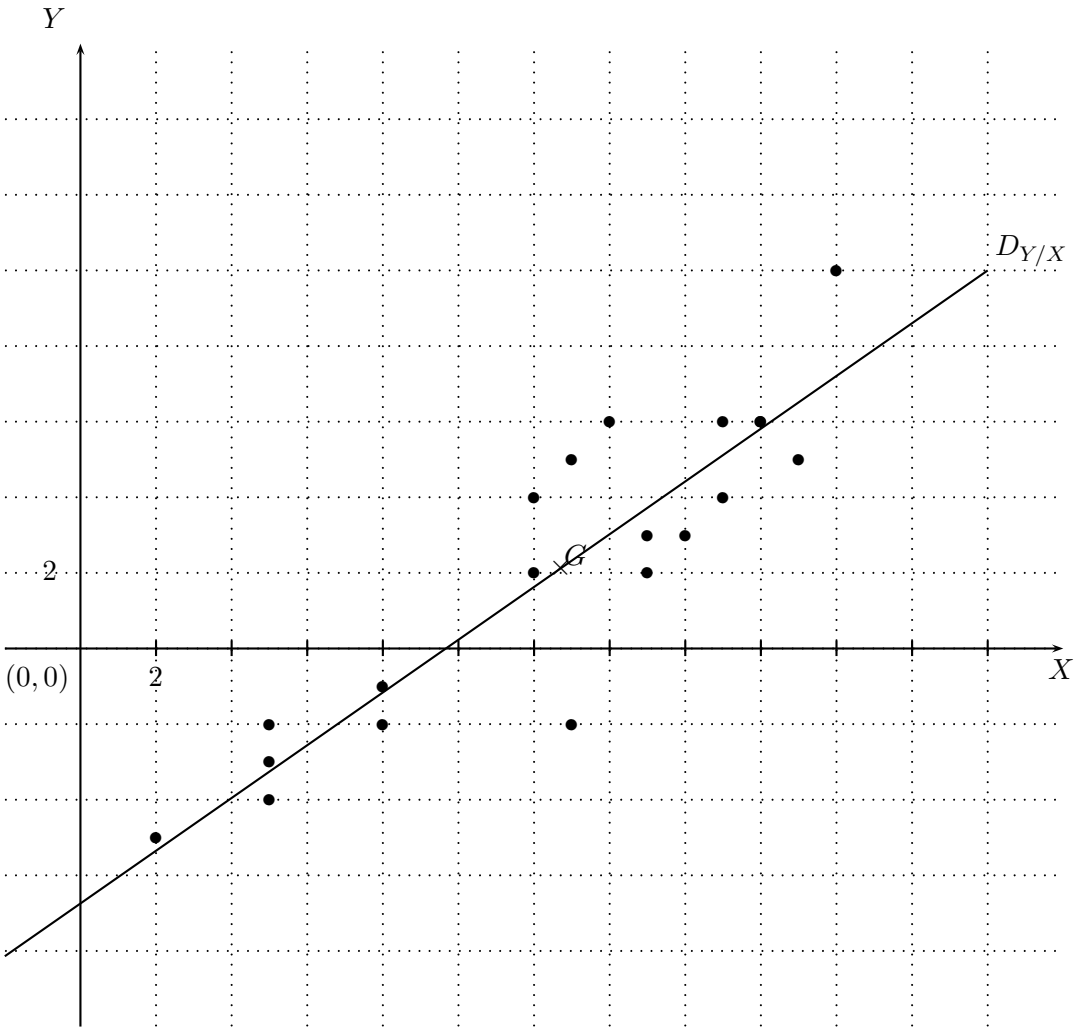
3. Montrons que la droite de régression admet pour équation

$$D_{Y/X} \ : \ Y = 0,69X - 6,56.$$

Les coefficients de l'équation $Y = bX + a$ de la droite des moindres carrés sont donnés par

$$\left\{ \begin{array}{rcl} b & = & \dfrac{Cov(X,Y)}{V(X)} = \dfrac{18,91}{27,34} = 0,69 \\ a & = & \overline{y} - b\overline{x} = 2,15 - 0,69 \times 12,6 = -6,56 \end{array} \right. .$$

4. On a le nuage de point suivants.



On a utilisé le tableau suivant qui présente x_i et y_i pour tout $i \in \{1, \ldots, 20\}$:

i	Aîné x_i	y_i	i	Aîné x_i	y_i
1	8	-1	11	17	4
2	13	-2	12	12	4
3	8	-2	13	2	-5
4	5	-2	14	13	5
5	12	2	15	19	5
6	15	2	16	18	5
7	5	-3	17	14	6
8	15	3	18	17	6
9	16	3	19	18	6
10	5	-4	20	20	10

5. Si on pose $X = 22$ dans l'équation trouvée précédemment, on trouve $Y = 0,69 \times 22 - 6,56 = 8,62$. Une estimation du niveau d'indépendance d'un cadet dont le frère (aîné) a obtenu une note de 22 est de $22 - 8,62 = 13,38$.
6. L'étude ne tient pas compte de facteurs de variation tels que le sexe, la différence d'âge, la composition de la famille,...

Exercice 2

Le tableau suivant donne le classement de 100 individus suivant les deux caractères « masse corporelle » (Y) et « taille » (X) :

X \ Y	$[40;45[$	$[45;50[$	$[50;55[$	$[55;60[$
$[150;155[$	20	9	1	0
$[155;160[$	2	18	4	1
$[160;165[$	0	5	12	6
$[165;170[$	0	1	7	14

1. Déterminer et reporter sur un graphique, l'ensemble des points qui constituent les lignes de régression de Y en X et X en Y.
2. Montrer que les rapports de corrélation sont égaux à $h^2_{X/Y} = 0,6732$ et $h^2_{Y/X} = 0,6692$.
   Montrer à l'aide du test de Fisher-Snédécor (au seuil de $\alpha = 5\%$) que ces rapports sont significatifs.
3. Calculer le coefficient de corrélation linéaire entre X et Y.
   À l'aide d'un test de Student-Fisher au seuil de $\alpha = 5\%$, montrer que ce coefficient est significatif.
4. Déterminer les équations des droites de régression et les tracer sur le graphique.

Correction :

1. • La ligne de régression de Y en X est une ligne polygonale joignant les points ayant :
     – pour abscisses les centres de classes de valeurs de X,
     – pour ordonnées les moyennes conditionnées de Y pour X fixé.
   • La ligne de régression de X en Y est une ligne polygonale joignant les points ayant :
     – pour abscisses les moyennes conditionnées de X pour Y fixé,
     – pour ordonnées les centres de classes de valeurs de Y.

   Pour pouvoir tracer ces lignes de régression, il faut d'abord calculer les moyennes conditionnées de X en Y et de Y en X. L'hypothèse faite pour la répartition des valeurs à l'intérieur des classes est

l'hypothèse habituelle de concentration de toutes les valeurs d'une classe au centre de la classe.

Le tableau des calculs se trouve ci-dessous :

x_i \ y_j	42,5	47,5	52,5	57,5	$n_{i.}$	$\overline{Y_i}$	$n_{i.}\overline{Y_i}$	$n_{i.}\overline{Y_i}^2$	$n_{i.}x_i$	$n_{i.}x_i^2$
152,5	20	9	1	0	30	44,33	1330,00	58963,33	4575,00	697687,50
157,5	2	18	4	1	25	48,30	1207,50	58322,25	3937,50	620156,25
162,5	0	5	12	6	23	52,72	1212,50	63919,84	3737,50	607343,50
167,5	0	1	7	14	22	55,45	1220,00	67654,55	3685,00	617237,50
$n_{.j}$	22	33	24	21	100		4970,00	248859,97	15935,00	2542425,00
$\overline{X_j}$	152,95	157,20	162,71	165,60						
$n_{.j}\overline{X_j}$	3365	5187,5	3905	3477,5	15935				$\overline{x} = 159,35$	$s_T^2(X) = 31,83$
$n_{.j}\overline{X_j}^2$	514692,05	815459,28	635376,04	575857,44	2541384,81				$s_m^2(X) = 21,43$	$h_{X/Y}^2 = 0,6732$
$n_{.j}y_j$	935,00	1567,50	1260,00	1207,50	4970,00		$\overline{y} = 49,70$	$s_m^2(Y) = 18,51$		
$n_{.j}y_j^2$	39737,50	74456,25	66150,00	69431,25	249775,00		$s_T^2(Y) = 27,66$	$h_{Y/X}^2 = 0,6692$		

Les formules utlisées dans ce tableau sont :

$$Cov(X,Y) = \frac{1}{n}\sum_{i,j} n_{ij}x_i y_j - \overline{xy} \text{ avec } n = \sum_{i,j} n_{ij} = 100,\ n_{i.} = \sum_j n_{ij},\ n_{.j} = \sum_i n_{ij},$$

$$\overline{x} = \frac{1}{n}\sum_i n_{i.}x_i = \frac{1}{n}\sum_{i,j} n_{ij}x_i = \frac{1}{n}\sum_j n_{.j}\overline{X_j} \text{ avec } \overline{X_j} = \frac{1}{n_{.j}}\sum_i n_{ij}x_i,$$

$$\overline{y} = \frac{1}{n}\sum_j n_{.j}y_j = \frac{1}{n}\sum_{i,j} n_{ij}y_j = \frac{1}{n}\sum_i n_{i.}\overline{Y_i} \text{ avec } \overline{Y_i} = \frac{1}{n_{i.}}\sum_j n_{ij}y_j,$$

$$s_T^2(X) = \frac{1}{n}\sum_i n_{i.}x_i^2 - \overline{x}^2 \text{ et } s_T^2(Y) = \frac{1}{n}\sum_j n_{.j}y_j^2 - \overline{y}^2,$$

$$s_m^2(X) = \frac{1}{n}\sum_j n_{.j}\overline{X_j}^2 - \left(\frac{1}{n}\sum_j n_{.j}\overline{X_j}\right)^2 \text{ et } s_m^2(Y) = \frac{1}{n}\sum_i n_{i.}\overline{Y_i}^2 - \left(\frac{1}{n}\sum_i n_{i.}\overline{Y_i}\right)^2,$$

$$h_{Y/X}^2 = \frac{s_m^2(Y)}{s_T^2(Y)} \text{ et } h_{X/Y}^2 = \frac{s_m^2(X)}{s_T^2(X)}.$$

De ce tableau on extrait les données suivantes qui permettent de tracer les lignes de régression.

Régression de Y en X	
y_j	$\overline{X_j}$
42,5	152,95
47,5	157,20
52,5	162,71
57,5	165,50

Régression de X en Y	
x_i	$\overline{Y_i}$
152,5	44,33
157,5	48,30
162,5	52,72
167,5	55,45

d'où les courbes de régression :

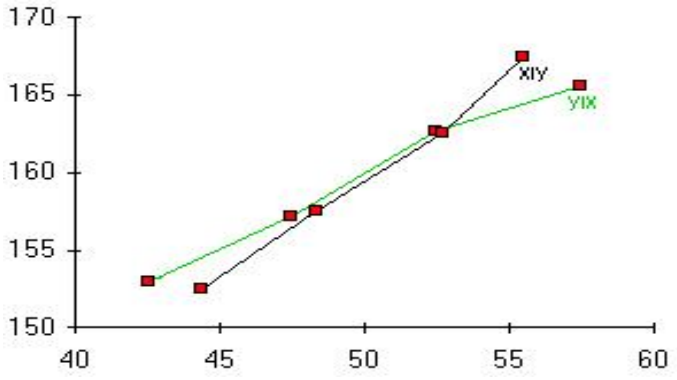


FIG. 1 – Courbes de régression.

2. Les rapports de corrélation ont été calculés dans le tableau précédent :
   - $h_{Y/X}^2 = \frac{s_m^2(Y)}{s_T^2(Y)} = 0,6692$ et
   - $h_{X/Y}^2 = \frac{s_m^2(X)}{s_T^2(X)} = 0,6732.$

   Testons leur significativité : on utilise un test de Fisher-Snédécor (au seuil de $\alpha = 5\%$).
   - On a pour $h_{Y/X}^2$:

$$F_{obs} = \frac{(n-k)}{k-1}\frac{h_{Y/X}^2}{1-h_{Y/X}^2} = \frac{(100-4)}{4-1}\frac{0,6692}{1-0,6692} = 64,74$$

   - On a pour $h_{X/Y}^2$:

$$F_{obs} = \frac{(n-k)}{k-1}\frac{h_{X/Y}^2}{1-h_{X/Y}^2} = \frac{(100-4)}{4-1}\frac{0,6732}{1-0,6732} = 65,92$$

Dans les deux cas, on compare F_{obs} à la valeur lue $F_{lu} = F_{3;96;0,95} \simeq 3,25$ et dans les deux cas, on remarque que $F_{obs} > F_{lu}$ donc on rejette l'hypothèse H_0 selon laquelle les rapports de corrélation ne sont pas significatifs.

3. Le coefficient de corrélation linéaire R de X et Y est donné par la formule : $R = \dfrac{Cov(X,Y)}{s_T(X)s_T(Y)}$.
Afin de déterminer la covariance $Cov(X,Y)$, donnée par la formule $Cov(X,Y) = \dfrac{1}{n}\sum_{ij} n_{ij}x_iy_j - \overline{xy}$, il faut calculer $\sum_{ij} n_{ij}x_iy_j$ ce qui peut se faire en complétant le tableau précédent :

x_i	$n_{i.}$	$\overline{Y_i}$	$n_{i.}\overline{Y_i}x_i$
152,5	30	44,33	202825,00
157,5	25	48,30	190181,25
162,5	23	52,72	197031,25
167,5	22	55,45	204350,00
Total	$n = 100$		794387,50

Donc

$$Cov(X,Y) = \frac{794387,50}{100} - 159,35 \times 49,70 = 24,18$$
$$\Rightarrow R = \frac{24,18}{\sqrt{27,66} \times \sqrt{31,83}} = \frac{24,18}{5,26 \times 5,64} = 0,8149.$$

Le coefficient de détermination est le carré du coefficient de corrélation linéaire : $R^2 = 0,6641$. Le fait que les rapports de corrélation et le coefficient de détermination aient des valeurs proches montre que les régressions optimales sont pratiquement linéaires.
Testons maintenant la significativité de ce coefficient à l'aide d'un test de Student-Fisher au seuil de $\alpha = 5\%$. On a $T_{obs} = \sqrt{100-2}\dfrac{0,8149}{\sqrt{1-0,8149}} = 18,75$ et $T_{lu} = T_{98;0,05} = 1,99$ donc $F_{obs} > F_{lu}$, on rejette ainsi l'hypothèse H_0 selon laquelle le coefficient de corrélation linéaire R n'est pas significatif.

4. Les droites de régression sont déterminées par la méthode des moindres carrés ordinaire.
   - Régression linéaire de Y par rapport à X : on a $b = \dfrac{Cov(X,Y)}{s_T^2(X)} = \dfrac{24,18}{27,66} = 0,8742$ et $a = \overline{y} - b\overline{x} = 159,35 - 0,8742 \times 49,70 = 115,90$ ce qui donne l'équation de $D_{Y/X}$ soit
   $$D_{Y/X} \ : \ Y = 0,8742X + 115,90.$$
   - Régression linéaire de X par rapport à Y : on a $b = \dfrac{Cov(Y,X)}{s_T^2(Y)} = \dfrac{Cov(X,Y)}{s_T^2(Y)} = \dfrac{24,18}{31,83} = 0,7597$ et $a = \overline{x} - b\overline{y} = 49,70 - 0,7597 \times 159,35 = -71,36$ ce qui donne l'équation de $D_{X/Y}$ soit
   $$D_{X/Y} \ : \ X = 0,7597Y - 71,36 \Leftrightarrow Y = 1,3163X + 93,93.$$

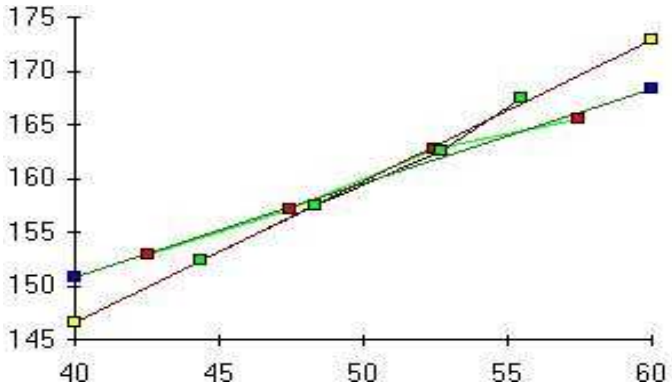


FIG. 2 – Courbes de régression et droite de corrélation.

Une droite de régression est très proche le la courbe de régression correspondante, comme on s'y attendait.