UNIVERSITE MOHAMMED PREMIER
ECOLE NATIONALE DES SCIENCES APPLIQUEES
OUJDA-MAROC

Filière :GC3 & GI3 — Année universitaire : 2018 - 2019
Elément de module : Analyse des données — Enseignant : M. Derouich

Fiche TD N° : 2

Exercice 1. :

1. Montrer que :$|\rho(X,Y)| \leq 1$.
2. On considère deux séries statistiques (x_i) et (y_i) de taille n
   Soient α_i et β_i deux séries statistiques liées aux séries statistiques (x_i) et (y_i) par les relations suivantes :
   $\forall i\ \alpha_i = \dfrac{x_i - c}{d}$ avec $d \neq 0\ c, d \in \mathbb{R}$
   $\forall i\ \beta_i = \dfrac{y_i - c'}{d'}$ avec $d' \neq 0\ c', d' \in \mathbb{R}$
   Montrer les propriétés suivantes :
   $i)\ cov(\alpha,\beta) = \dfrac{1}{dd'} cov(x,y)$ et $ii)\ \rho(\alpha,\beta) = \dfrac{|dd'|}{dd'}\rho(x,y)$

Exercice 2. :
Le tableau suivant représente des âges de patients X et les pressions systoliques Y de 9 malades.

L'âge X	56	42	72	36	63	47	55	49	38
Tension artérielle Y	147	125	160	118	149	128	150	145	115

1. 1. Représenter le nuage de points $M(x_i; y_i)$ dans le repère orthogonal ci-dessous.
2. Calculer la moyenne et l'écart-type de chacun des deux caractère X et Y.
3. Placer le point $G(\overline{X}, \overline{Y})$ dans le repère précédent.
4. Calculer la covariance et le coefficient de corrélation du couple (X,Y). Que peut-on conclure ?
5. Trouver la droite de régression de X en Y.
6. Lorsque l'âge est 75 ans , quelle Tension artérielle Y peut-on prévoir ?

Exercice 3. :
sur un échantillon de 100 étudiants, on relevé la taille X en centimètre, ainsi que le poids Y en kilogrammes comme l'indique le tableau suivant

$X \diagdown Y$	$[50,60[$	$[60,70[$	$[70,80[$	$[80,90[$	$[90,100[$
$[150,160[$	10	3	1	0	0
$[160,170[$	2	12	6	7	2
$[170,180[$	1	7	11	17	4
$[180,190[$	0	2	2	4	9

1. Calculer la moyenne et l'écart-type de chacun des deux caractères X et Y
2. Calculer la covariance et le coefficient de corrélation du couple (X,Y). Que peut-on conclure ?
3. Trouver la droite de régression de Y en X.

Exercice 4. :
On dispose pour un secteur industriel donné et sur une période de 8 années du nombre de salariés Y (en milliers) et du chiffre d'affaires X (en dizaines de milliards) :

Année	1	2	3	4	5	6	7	8
X	3	4	5	6	8	9	11	13
Y	3.5	4.2	5	5.5	6	6.5	6.7	7.2
Ln(X)	1,1	1,4	1,6	1,8	2,1	2,2	2,4	2,6

1. Représenter le nuage de points (x_i, y_i).
2. Calculer la moyenne et l'écart-type de chacun des deux caractères X et Y.
3. Calculer la covariance du couple (X, Y).
4. a) Donner le coefficient de corrélation linéaire $\rho(X, Y)$ de la série statistique (x_i, y_i) . Un ajustement affine est-il justifié ?
   b) Ecrire une équation de la droite de régression D de Y en X . Représenter D dans le repère précédent
5. Calculer la moyenne et l'écart-type de variable Z.
6. Calculer la covariance du couple (Z, Y).
7. a) Donner le coefficient de corrélation linéaire $\rho(Z, Y)$ de la série statistique (z_i, y_i) . Un ajustement affine est-il justifié ?
   b) Ecrire une équation de la droite de régression Δ de Y en Z .
8. En l'an 2010, on prévoit pour le secteur étudié un chiffre d'affaires de 400 milliards.
   i)Utiliser les droites $(D) : Y = aX + b$ et $(\Delta) : Y = a'Z + b'$ pour proposer deux prévisions du nombre d'employés de ce secteur à l'horizon 2010.
   ii) Quelle prévision vous semble la plus appropriée

UNIVERSITE MOHAMMED PREMIER
ECOLE NATIONALE DES SCIENCES APPLIQUEES
OUJDA-MAROC

Filière : GC3 & GI3 — Année universitaire : 2019-2020
Elément de module : Analyse des données — Enseignant : M. Derouich

Correction fiche TD N° : 2

Exercice 1. :

1. On a

$$\begin{aligned} |\rho(X,Y)| &= \frac{|cov(X,Y)|}{\sigma_X \sigma_Y} \\ &= \frac{1}{\sigma_X \sigma_Y} \frac{1}{n} \left| \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) \right| \end{aligned}$$

Par application de l'inégalité de Cauchy-Schwarz,

$$\left| \sum_{i=1}^{n} a_i b_i \right| \leq \sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}$$

avec $a_i = x_i - \overline{X}$ et $b_i = y_i - \overline{Y}$ on obtient

$$\begin{aligned} \left| \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) \right| &\leq \sqrt{\sum_{i=1}^{n} (x_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (y_i - \overline{Y})^2} \\ \frac{1}{n} \left| \sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) \right| &\leq \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X})^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{Y})^2} \\ |cov(X,Y)| &\leq \sigma_x \sigma_y \\ \frac{|cov(X,Y)|}{\sigma_x \sigma_y} &\leq 1 \\ \Rightarrow |\rho(X,Y)| &\leq 1 \end{aligned}$$

2.

$$\begin{aligned} i) cov(\alpha,\beta) &= \frac{1}{n} \sum_{i=1}^{n} (\alpha_i - \overline{\alpha})(\beta_i - \overline{\beta}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - c}{d} - \frac{\overline{x} - c}{d} \right) \left(\frac{y_i - c'}{d'} - \frac{\overline{y} - c'}{d'} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{d} \right) \left(\frac{y_i - \overline{y}}{d'} \right) \\ &= \frac{1}{dd'} \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{cov(X,Y)}{dd'} \end{aligned}$$

ii) On a $\rho(\alpha,\beta) = \dfrac{cov(\alpha,\beta)}{\sigma_\alpha \sigma_\beta}$

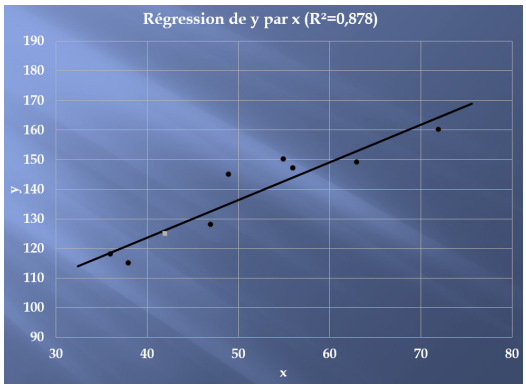
Or $\sigma_\alpha = \dfrac{\sigma_x}{|d|}$, $\sigma_\beta = \dfrac{\sigma_y}{|d'|}$ et $cov(\alpha,\beta) = \dfrac{cov(X,Y)}{dd'}$

$$\begin{aligned} \rho(\alpha,\beta) &= \frac{cov(\alpha,\beta)}{\sigma_\alpha \sigma_\beta} \\ &= \frac{|dd'|}{dd'} cov(X,Y) \end{aligned}$$

Exercice 2.

										somme
âge(x)	56	42	72	36	63	47	55	49	38	458
T.A(y)	147	125	160	118	149	128	150	145	115	1237
X^2	3136	1764	5184	1296	3969	2209	3025	2401	1444	24428
Y^2	21609	15625	25600	13924	22201	16384	22500	21025	13225	172093
XY	8232	5250	11520	4248	9387	6016	8250	7105	4370	64378

1.



2.

$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \frac{458}{9} \\ &= 50.89 \\ \overline{Y} &= \frac{1}{n}\sum_{i=1}^{n} y_i. \\ &= \frac{1237}{9} \\ &= 137.44 \end{aligned}$$

$$
\begin{aligned}
V(X) = \sigma_X^2 &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{X}^2 \\
&= \frac{24428}{9} - 50.89^2 \\
&= 124.54 \\
\Rightarrow \sigma_X &= 11.16 \\
V(Y) = \sigma_Y^2 &= \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{Y}^2. \\
&= \frac{172093}{9} - 137.44^2 \\
&= 230.47 \\
\Rightarrow \sigma_Y &= 15.18
\end{aligned}
$$

3. Voir figure
4.

$$
\begin{aligned}
\sigma_{XY} = cov(X,Y) &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\,\overline{Y} \\
&= \frac{64378}{9} - 50.89 \times 137.44 \\
&= 158.72 \\
\rho(X,Y) &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \\
&= \frac{158.72}{11.16 \times 15.18} \\
&= 0,94
\end{aligned}
$$

5. l'équation de la droite de régression de Y en X :

$$Y - \overline{Y} = a\left(X - \overline{X}\right) \text{ avec } a = \frac{cov(X,Y)}{V(X)}.$$

Donc :$Y - 137,44 = a(X - 50,89)$ avec : $a = 1,274385408$
l'équation de la droite de régression de X en Y :

$$X - \overline{X} = a'\left(Y - \overline{Y}\right) \text{ avec } a' = \frac{cov(X,Y)}{V(Y)}.$$

Donc :$X - 50,89 = a'(Y - 137,44)$ avec : $a' = 0,688665095$

6. Lorsque l'âge est 75 ans c.à.d $X = 75$ donc :
$Y = 137,44 + 1,274385408(75 - 50,89) = 168,1654322.$

Exercice 3.

X/Y	55	65	75	85	95	$n.i$	$n_{i.}\ x_i$	$n_{i.}\ x_i^2$	$\sum_j n_{ij}\ y_j$	$x_i \sum_j n_{ij}\ y_j$
155	10	3	1	0	0	14	2170	336350	820	127100
165	2	12	6	7	2	29	4785	789525	2125	350625
175	1	7	11	17	4	40	7000	1225000	3160	553000
185	0	2	2	4	9	17	3145	581825	1475	272875
$n_{.j}$	13	24	20	28	15	100	17100	2932700		1303600
$n_{.j}\ y_j$	715	1560	1500	2380	1425	7580				
$n_{.j}\ y_j^2$	39325	101400	112500	202300	135375	590900				
$\sum_i n_{ij}\ x_i$	2055	4040	3440	4870	2695					
$y_j \sum_i n_{ij}\ xi$	113025	262600	258000	413950	256025	1303600				

1. On a $n = 100$, $p = 5$ et $q = 4$

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i \\
&= \frac{17100}{100} \\
&= 171,00 \\
\overline{Y} &= \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j \\
&= \frac{7580}{100} \\
&= 75,80 \\
V(X) &= \left(\frac{1}{n}\sum_{i=1}^{p} n_{i.}X_i^2\right) - \overline{X}^2 \\
&= \frac{2932700}{100} - 171^2 \\
&= 86,00 \\
\Rightarrow \sigma_X &= 9,27 \\
&= \left(\frac{1}{n}\sum_{j=1}^{q} n_{.j}Y_j^2\right) - \overline{Y}^2 \\
&= \frac{590900}{100} - 75.80^2 \\
&= 163,36 \\
\Rightarrow \sigma_y &= 12,78
\end{aligned}$$

2.

$$\begin{aligned}
\sigma_{XY} = cov(X,Y) &= \left(\frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j\right) - \overline{X}\,\overline{Y} \\
&= \frac{1303600}{100} - 171 \times 75,80 \\
&= 74,20 \\
\rho(X,Y) &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \\
&= \frac{74,20}{9,27 \times 12,78} \\
&= 0,63
\end{aligned}$$

3. l'équation de la droite de régression de Y en X :

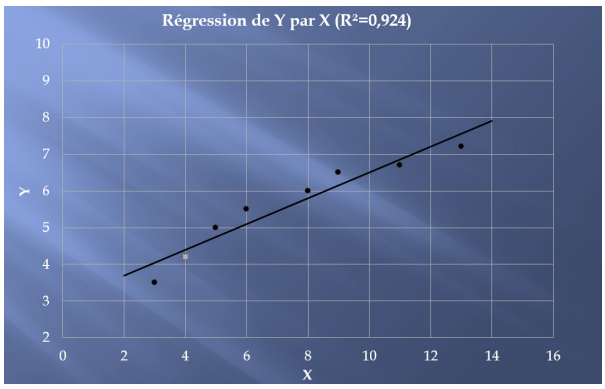
$$Y - \overline{Y} = a\left(X - \overline{X}\right) \text{ avec } a = \frac{cov(X,Y)}{V(X)}.$$

Donc : $Y - 75,80 = a(X - 171)$ avec : $a = 0,862790698$

Exercice 4.

Année	1	2	3	4	5	6	7	8	somme
X	3	4	5	6	8	9	11	13	59
Y	3,5	4,2	5	5,5	6	6,5	6,7	7,2	44,6
$Z = Ln(X)$	1,0986122	1,3862943	1,6094379	1,7917594	2,0794415	2,1972245	2,3978952	2,5649493	15,12561478
X^2	9	16	25	36	64	81	121	169	521
Y^2	12,25	17,64	25	30,25	36	42,25	44,89	51,84	260,12
Z^2	1,2069489	1,9218120	2,5902903	3,2104019	4,3240771	4,8277958	5,7499017	6,5789652	30,41019332
XY	10,5	16,8	25	33	48	58,5	73,7	93,6	359,1
YZ	3,845143	5,8224363	8,0471895	9,854677	12,476649	14,281959	16,065898	18,467635	88,86158867

1.



2.

$$\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{59}{8} \\
&= 7,375 \\
\overline{Y} &= \frac{1}{n}\sum_{i=1}^{n} y_i. \\
&= \frac{44,6}{8} \\
&= 5,575
\end{aligned}$$

$$\begin{aligned}
V(X) = \sigma_X^2 &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{X}^2 \\
&= \frac{521}{8} - 7,375^2 \\
&= 10,734375 \\
\Rightarrow \sigma_X &= 3,276335606 \\
V(Y) = \sigma_Y^2 &= \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{Y}^2. \\
&= \frac{260,12}{8} - 5,575^2 \\
&= 1,434375 \\
\Rightarrow \sigma_Y &= 1,197653957
\end{aligned}$$

3.

$$\begin{aligned}
\sigma_{XY} = cov(X,Y) &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\,\overline{Y} \\
&= \frac{359,1}{8} - 7,375 \times 5,575 \\
&= 3,771875
\end{aligned}$$

4. **a)On a**

$$\begin{aligned}
\rho(X,Y) &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \\
&= \frac{3,771875}{3,276335606 \times 1,197653957} \\
&= 0,961252664 \\
\Rightarrow \rho^2(X,Y) &= 0.92 > 0.75
\end{aligned}$$

donc l'ajustement affine est justifié

b) l'équation de la droite de régression : de Y en X :

$$Y - \overline{Y} = a\left(X - \overline{X}\right) \text{ avec } a = \frac{cov(X,Y)}{V(X)}.$$

Donc :$Y - 5.575 = a(X - 7.375)$ avec : $a = 0,35138$

5.

$$\begin{aligned}
\overline{Z} &= \frac{1}{n}\sum_{i=1}^{n} Z_i \\
&= \frac{15,12561478}{8} \\
&= 1,891 \\
V(Z) = \sigma_Z^2 &= \left(\frac{1}{n}\sum_{i=1}^{n} Z_i^2\right) - \overline{Z}^2 \\
&= \frac{30,41019332}{8} - 1,891^2 \\
&= 0,226520689 \\
\Rightarrow \sigma_Z &= 0,475941896
\end{aligned}$$

$$\begin{aligned}
\sigma_{ZY} = cov(Z,Y) &= \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i\right) - \overline{Z}\,\overline{Y} \\
&= \frac{88,86158867}{8} - 1,891 \times 5,575 \\
&= 0,567035784
\end{aligned}$$

6. **a)On a**

$$\begin{aligned}
\rho(Z,Y) &= \frac{cov(Z,Y)}{\sigma_Z \sigma_Y} \\
&= \frac{0,567035784}{0,475941896 \times 1,197653957} \\
&= 0,99477572 \\
\Rightarrow \rho^2(Z,Y) &= 0,989578733 > 0.75
\end{aligned}$$

donc l'ajustement affine est justifié

b) l'équation de la droite de régression : de Y en Z :

$$Y - \overline{Y} = a\left(Z - \overline{Z}\right) \text{ avec } a = \frac{cov(X,Y)}{V(X)}.$$

Donc :$Y - 5.575 = A(Z - 1,9)$ avec : $A = 2,503241$

7. *i*) ✓ Prévision fournie par$Y = aX + b$:
$X = 40$en dizaines de milliards $\Longrightarrow Y(40) = 17,03886$ donc $Y \simeq 17039$ salariés
✓ Prévision fournie par $Y = AZ + B$:
$X = 40$ en dizaines de milliards
$Z = Ln(X) \simeq 3,688879454 \Longrightarrow Y(\ln(40)) = 10,07627$ donc $Y \simeq 10076$ salariés

ii) On a $\rho^2(Z,Y) > \rho^2(X,Y)$ donc $Y = AZ + B$ fourni une qualité d'ajustement meilleure que celle du $Y = aX + b$
par conséquent la prévision la plus appropriée est : $Y \simeq 10076$ salariés