Faculté des sciences **Année : 2023/2024**

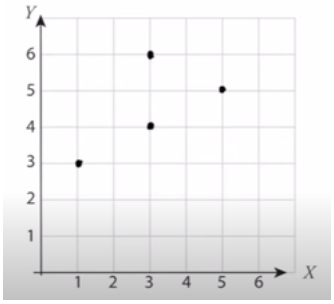
1ère année Master Informatique (ASD)

Matière : Analyse de données **TD2**

Corrigé type de la série 2 d'exercices

Exercice 1

Représentation graphique des données (nuage de points)



Détecter une relation linéaire entre les heures passées à étudier et les notes du test implique la recherche d'une droite qui passe au mieux parfaitement par toutes les données.

Donc, trouver les valeurs des paramètres du modèle (a et b) de la régression linéaire simple de telle façon que la droite passe au mieux par tous les points.

$$F(x) = ax + b$$

$$a = \frac{Cov(X,,Y)}{Var(X)} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{Y} - a.\bar{X}$$

x	y	xy	x^2	y^2
1	3	3	1	9
3	4	12	9	16
3	6	18	9	36
5	5	25	25	25
Total=12	Total=18	Total=58	Total=44	Total=86
Moy=3	Moy=4.5	Moy=14.5	Moy=11	Moy=21.5

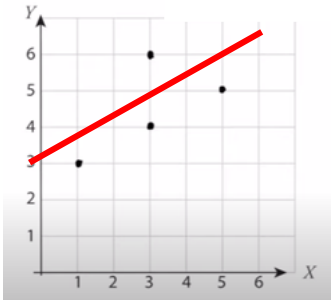
$$a = \frac{14.5 - (3 * 4.5)}{11 - 9} = \frac{1}{2}$$

$$b = 4.5 - (0.5 * 3) = 3$$

Y= 0.5 x+3

Pour dessiner la droite : on suppose x=0, y=3 et x=6, y=6

Y= 0.5 x+3



Oui, il existe une relation entre le nombre des heures et les notes test, plus on passe de heures à étudier, plus a de bonne note.

Exercice 2

1) Calculer les paramètres de la régression linéaire multiple qui permet de modéliser ces données.

$$x=\begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix} \quad \mathbf{y}=\begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} *$$

$$\begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix} = \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

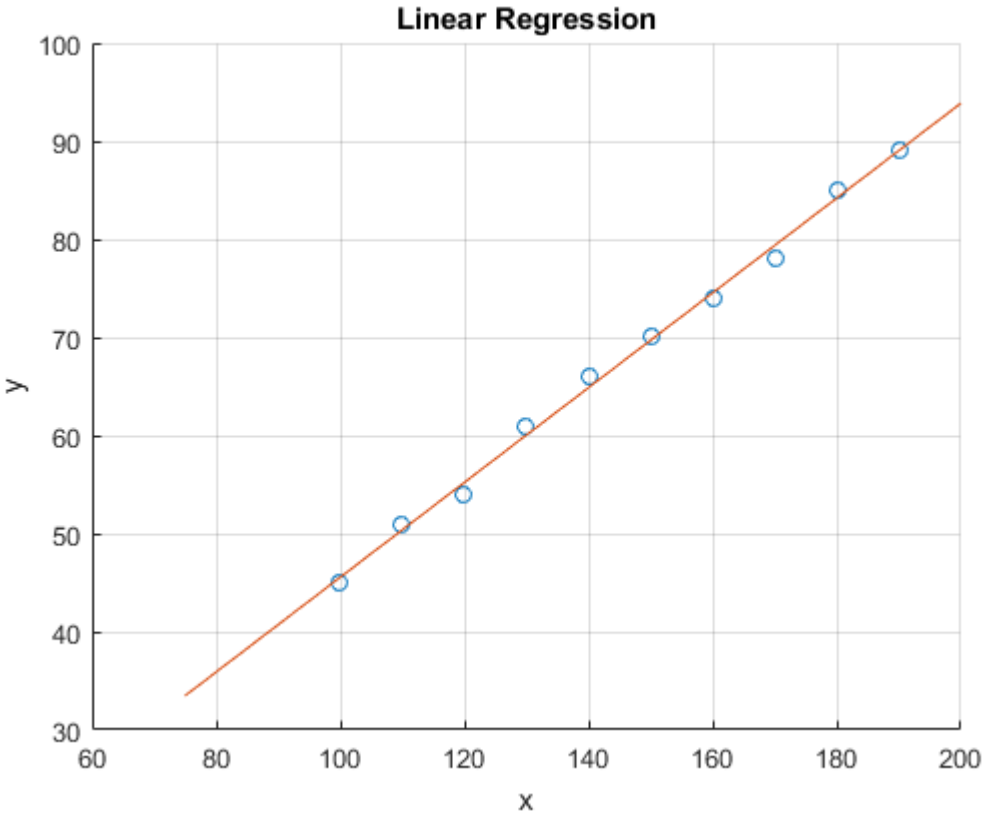
$$* \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix} = \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T Y = \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}^{-1} * \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 0.68 \\ 1.74 \\ 0.73 \end{bmatrix}$$

$$f(x) = 0.73 + 1.74\, x_1 + 0.68\, x_2$$

Exercice 3

1) Donner la représentation graphique de ces données.



2) Modèle de régression linéaire :

On a x = [100; 110; 120; 130; 140; 150; 160; 170; 180; 190];

et l'étiquette y = [45; 51; 54; 61; 66; 70; 74; 78; 85; 89];

Dans le modèle de la régression linéaire multiple, l'ensemble des paramètres est calculé par la formule suivante : $\theta = (X^T X)^{-1} X^T Y$

$$X = \begin{bmatrix} 100 & 1 \\ 110 & 1 \\ 120 & 1 \\ 130 & 1 \\ 140 & 1 \\ 150 & 1 \\ 160 & 1 \\ 170 & 1 \\ 180 & 1 \\ 190 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 45 \\ 51 \\ 54 \\ 61 \\ 66 \\ 70 \\ 74 \\ 78 \\ 85 \\ 89 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 100 & 110 & 120 & 130 & 140 & 150 & 160 & 170 & 180 & 190 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 100 & 1 \\ 110 & 1 \\ 120 & 1 \\ 130 & 1 \\ 140 & 1 \\ 150 & 1 \\ 160 & 1 \\ 170 & 1 \\ 180 & 1 \\ 190 & 1 \end{bmatrix} = \begin{bmatrix} 218500 & 1450 \\ 1450 & 10 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 100 & 110 & 120 & 130 & 140 & 150 & 160 & 170 & 180 & 190 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 45 \\ 51 \\ 54 \\ 61 \\ 66 \\ 70 \\ 74 \\ 78 \\ 85 \\ 89 \end{bmatrix} = \begin{bmatrix} 101570 \\ 673 \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T Y = \begin{bmatrix} 0.0001 & -0.0176 \\ -0.0176 & 2.6485 \end{bmatrix} * \begin{bmatrix} 101570 \\ 673 \end{bmatrix} = \begin{bmatrix} 0.4830 \\ -2.7394 \end{bmatrix}$$

$$y = -2.7394 + 0.4830x$$

3) Prédire la valeur de rendement pour la température 80°C :

$\hat{y} = -2.7394 + 0.4830*80 = 35.9006$

4) La valeur de rendement sera supérieure à 100 si on a :

$y = -2.7394 + 0.4830x > 100$

$$x > \frac{100 + 2.7394}{0.4830}$$

$$x > 212\,°\,C$$

Exercice 4

On suppose qu'on a régression quadratique (m=2) avec une seule variable x, chercher le meilleur ajustement de la fonction f par rapport aux données d'apprentissage suivantes.

x_i	y_i	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$
2	17	4	8	16	34	68
3	34	9	27	81	102	306
6	121	36	216	1296	726	4356
10	321	100	1000	10000	3210	4356
14	617	196	2744	38416	8638	120932
16	801	256	4096	65536	12816	205056
Total		601	5091	115345	25526	362818

$$\begin{pmatrix} 115345 & 8091 & 601 \\ 8091 & 601 & 51 \\ 601 & 51 & 6 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 362818 \\ 25526 \\ 1911 \end{pmatrix}$$

$$M = \begin{pmatrix} 115345 & 8091 & 601 \\ 8091 & 601 & 51 \\ 601 & 51 & 6 \end{pmatrix}$$

$$Y = \begin{pmatrix} 362818 \\ 25526 \\ 1911 \end{pmatrix}$$

X=$\begin{pmatrix} a \\ b \\ c \end{pmatrix}$??

$$X = M^{-1}Y$$

a=3

b=2

c=1

$$f(x) = 1 + 2\,x^1 + 3x^2$$

Exercice 5

1. Chercher les paramètres du modèle linéaire en utilisant les données d'apprentissage suivantes :

   X= [1, 2, 3, 4, 5]

   Y= [65, 70, 75, 80, 85]

Le modèle de régression linéaire simple pour ce jeu de données serait donc : Y = 5X + 60

x_i	y_i	X_i^2	R_i^2
1	65	1	0
2	70	4	0
3	75	9	0
4	80	16	0
5	85	25	0

2. Calculer le résidu (R) pour chaque individu

$\hat{Y}_1 = 5 * 1 + 60 = 65$, $R_1 = 65 - 65 = 0$

$\hat{Y}_2 = 5 * 2 + 60 = 70$, $R_2 = 70 - 70 = 0$

$\hat{Y}_3 = 5 * 3 + 60 = 75$, $R_3 = 75 - 75 = 0$

$\hat{Y}_4 = 5 * 4 + 60 = 80$, $R_4 = 80 - 80 = 0$

$\hat{Y}_5 = 5 * 5 + 60 = 85$, $R_5 = 85 - 85 = 0$

3. Calculer la somme des carrés des résidus (SCR)

   $SCR = R_1^2 + R_2^2 + R_3^2 + R_4^2 + R_5^2$

   $SCR = 0$

4. Calculer l'estimation σ^2

   $\sigma^2 = SCR / (n - 2)$, $n=5$

   $\sigma^2 \approx 0$

5. Calculer l'erreur type. Quelle est votre conclusion.

   Erreur type $(\sigma) = \sqrt{\sigma^2}$

   Erreur type $(\sigma) = \sqrt{0}$

   Erreur type $(\sigma) \approx 0$

   Elle représente l'écart-type de l'erreur dans les prédictions du modèle.s

   Plus l'erreur type est faible, plus les prédictions du modèle sont précises.

Exercice 6

x_i	y_i	$X_i = \ln x_i$	$Y_i = \ln y_i$	X_i^2	X_iY_i
8	55	2.079	4.007	4.32	8.33
8.9	120	2.186	4.787	4.78	10.5
10.7	275	2.37	5.617	5.62	13.3
12.8	610	2.549	6.413	6.5	16.4
15	1356	2.708	7.219	7.33	19.5
18	3050	2.89	8.023	8.35	23.2
Total=14.78	36.1			36.9	912

1) Le nuage de points a une forme de puissance (car il possède le plus grand coefficient de corrélation)

Ajustement puissance : $r(ln\ x_i, \ln y_i) = 0.998$

Ajustement linéaire : $r(x_i,\ y_i) = 0.941$

Ajustement exponentiel : $r(x_i, \ln y_i) = 0.988$

Ajustement logarithmique : $r(ln\ x_i,\ \ y_i) = 0.897$

2) L'équation du modèle :

$$a = \frac{Cov(X,Y)}{Var(X)} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}$$
$$b = \bar{Y} - a.\bar{X}$$

$\bar{X} = \frac{14.78}{6} = 2.46$

$\bar{Y} = \frac{36.1}{6} = 6.01$

$Cov(X,Y) = 0.39$

$Var(X) = 0.08$

$a = \frac{0.39}{0.08} = 4.84$

b=6.01-(4.84*2.46)=-5.90

Y=4.84 x-5.90

On pose X=$ln\ x_i\ et\ Y = \ln y_i$

$\ln y_i = 4.84\ ln\ x_i$-5.9

y=$e^{4.84\ ln\ x_i - 5.9}$ // $(e^x)^y = e^{xy}$

y=$(e^{\ ln\ x_i})^{4.84} * e^{-5.9}$ // $e^{x+y} = e^x * e^y,\ e^{x-y} = e^x * e^{-y} = \frac{e^x}{e^y}$

y=$x^{4.84} * 0.0027$