Université Echahid Hama Lakhdar, El-oued
Institut des sciences exactes
Département d'informatique
$1^{er} Master$: Intelligence Artificielle et Science de données
Semestre : 1.2024

— Contrôle : Classification et Analyse de données —

Le 13/01/2024 Durée : 1h :30.

1 Question de cours : (6.5 pts)

1.1 QCM

1. La droite de régression simple passe-t-elle par la point $(\hat{x}, \hat{y})$.{Oui/Non,Justifier} (1.5 pts)
2. Le coefficient de corrélation $r : 0 \leq r \leq 1$.{Oui/Non} (1 pt)
3. Le coefficient de déterminant $R^2 : 0 \leq R^2 \leq 1$.{Oui/Non} (1 pt)

1.2 ACP

— Décrire brièvement les étapes vue en cours et TD d'analyse en composant principal (1 pt).
— Comparer entre régression linéaire et ACP (2 pts).

2 Exercice (13.5 pts)

On considère le modèle de régression simple :$y = ax + b$.

n	1	2	3	4	5
y	10	20	20	25	25
x	1	2	2	2	3

Table 1 –

2.1 A)

— Dessiner le nuage des points (1 pt).
— Calculer la corrélation de Bravais-Pearson r_1 entre les variables y et x ; est elle significative (1.5 pt).
— Estimer les valeurs de paramètre a et b (1 pt).
— Calculer l'erreur standard résiduelle (1.5 pt).
— Compléter la table d'analyse ANOVA et calculer R^2 (3pts) :

Source de variation	Somme des carrés	ddl	carré moyen
régression (expliquée)	SCE=$\sum_{i=1}^{i=n}(\hat{y_i} - \bar{y})^2 = ...$	p=...	$\frac{1}{p}\sum_{i=1}^{i=n}(\hat{y_i} - \bar{y})^2 = ...$
Résiduelle	SCR=$\sum_{i=1}^{i=n}(y_i - \hat{y})^2 =$	n-p-1=...	$\frac{1}{n-p-1}\sum_{i=1}^{i=n}(y_i - \hat{y})^2 = ...$
Totale	STC=$\sum_{i=1}^{i=n}(y_i - \bar{y})^2 =$	n-1=...	$\frac{1}{n-1}\sum_{i=1}^{i=n}(y_i - \bar{y})^2 = ...$

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = ... \quad (1)$$

— Comment voyez vous le modèle, Expliquer (2 pt)

2.2 B)

— Calculer la corrélation de Rang de Sperman r_2 entre y et x (1 pt).
— Que vous pouvez conclure entre les deux coefficients de correlation r_1 et r_2. (2.5 pt)

********** ***Bon Courage*** **********

2 – Table de la corrélation linéaire de Bravais Pearson

Cette table indique le seuil inférieur de significativité d'un coefficient de corrélation linéaire : si r calculé est supérieur à r lu dans la table, on conclut qu'il existe une corrélation linéaire significative, avec un risque d'erreur α; r se lit en fonction de ν.

$\nu = n - p - 1$ où n : nombre de couples;
p : nombre de variables explicatives (une seule dans le cas d'une corrélation simple).

$\alpha \rightarrow$ / $\nu \downarrow$	0,10	0,05	0,02	$\alpha \rightarrow$ / $\nu \downarrow$	0,10	0,05	0,02
1	0,987 7	0,996 9	0,999 5	16	0,400 0	0,468 3	0,542 5
2	0,900 0	0,950 0	0,980 0	17	0,388 7	0,455 5	0,528 5
3	0,805 4	0,878 3	0,934 3	18	0,378 3	0,443 8	0,515 5
4	0,729 3	0,811 4	0,882 2	19	0,368 7	0,432 9	0,503 4
5	0,669 4	0,754 5	0,832 9	20	0,359 8	0,422 7	0,492 1
6	0,621 5	0,706 7	0,788 7	25	0,323 3	0,380 9	0,445 1
7	0,582 2	0,666 4	0,749 8	30	0,296 0	0,349 4	0,409 3
8	0,549 4	0,631 9	0,715 5	35	0,274 6	0,324 6	0,381 0
9	0,521 4	0,602 1	0,685 1	40	0,257 3	0,304 4	0,357 8
10	0,497 3	0,576 0	0,658 1	45	0,242 8	0,287 5	0,338 4
11	0,476 2	0,552 9	0,633 9	50	0,230 6	0,273 2	0,321 8
12	0,457 5	0,532 4	0,612 0	60	0,210 8	0,250 0	0,294 8
13	0,440 9	0,513 9	0,592 3	70	0,195 4	0,231 9	0,273 7
14	0,425 9	0,497 3	0,574 2	80	0,182 9	0,217 2	0,256 5
15	0,412 4	0,482 1	0,557 7	90	0,172 6	0,205 0	0,242 2
				100	0,163 8	0,194 6	0,230 1

3 – Table du rho de Spearman

$\alpha \rightarrow$ / $n \downarrow$	0,05	0,01	$\alpha \rightarrow$ / $n \downarrow$	0,05	0,01
4	1,00	—	24	0,34	0,49
5	0,90	1,00	26	0,33	0,47
6	0,83	0,94	28	0,32	0,45
7	0,71	0,89	30	0,31	0,43
8	0,64	0,83	35	0,28	0,40
9	0,60	0,78	40	0,26	0,37
10	0,56	0,75	45	0,25	0,35
12	0,51	0,71	50	0,24	0,33
14	0,46	0,64	55	0,22	0,32
16	0,42	0,60	60	0,21	0,30
18	0,40	0,56	70	0,20	0,28
20	0,38	0,53	80	0,19	0,26
22	0,36	0,51	100	0,17	0,23

n : Nombre de couples.

3 Solution

4 Question de cours

4.1 QCM

1. Oui,$\bar{y} = a\bar{x} + b$
2. Non.
3. Oui.

4.2 ACP

Les étapes vue en TD :

1. Matrice centre.
2. Matrice réduite
3. Matrice centrée réduite.
4. Recherche des axes principaux.
5. Recherche des valeurs propres et vecteurs propres.

ACP et régression linéaire :

— La recherche des paramètres de régression (Problème de régression).

— La recherche des axes principaux qui réduire les données (Réduction des données.).

5 Exercie 2

5.1 A

— $r = 0.86$,$v = 5 - 2 = 3$,$r_t h = 0.0.80$ Alors r est significatif avec un taux d'erreur 0.10.

— $a = 7.5$ et $b = 5$

— L'erreur standard résiduelle :3.536,

Source de variation	Somme des carrés	ddl	carré moyen
régression (expliquée)	SCE=$\sum_{i=1}^{i=n}(\hat{y_i} - \bar{y})^2 = 112.5$	p=1	$\frac{1}{p}\sum_{i=1}^{i=n}(\hat{y_i} - \bar{y})^2 = 112.5$
Résiduelle	SCR=$\sum_{i=1}^{i=n}(y_i - \hat{y})^2 = 37.5$	n-p-1=3	$\frac{1}{n-p-1}\sum_{i=1}^{i=n}(y_i - \hat{y})^2 = 12.5$
Totale	STC=$\sum_{i=1}^{i=n}(y_i - \bar{y})^2 = 150$	n-1=4	$\frac{1}{n-1}\sum_{i=1}^{i=n}(y_i - \bar{y})^2 = 37.5$

— $R^2 = 0.75$

— Le modèle explique les données par 70%.

5.2 B

— $r_2 = .la corrlation de Pearson est plus adapte aux relations linaires entre variables continues, tandis que la corrlation de Spearm$