Université Abdelmalek Essaâdi,
Faculté .S.J.E.S Tétouan,
Licence fondamentale en Sciences Economiques et Gestion.
Année universitaire 2020 - 2021

Travaux Dirigés 3 : Statistique Descriptive
Ajustement et corrélation

Exercice 1
On dispose du tableau statistique suivant:

x_i	1	3	5	7	8	12
y_i	6	9	12	14	19	27

1- Donner la représentation graphique de y en fonction de x, et s'assurer qu'un ajustement linéaire de y par x semble légitime.
2- Donner par la méthode géométrique directe l'équation d'ajustement de y par x.
3- Calculer les moyennes arithmétique $\bar{x}$ et $\bar{y}$ de x et y.
4- Effectuer ensuite l'ajustement par la droit des moindres carrés.
5- Calculer le coefficient de corrélation linéaire entre les variables x et y et conclure.

Exercice 2
Le prix et la quantité demandée d'un produit ont varié dans le temps comme suit:

Prix(x)	9	15	21	26	38	53
Demande(y)	35	47	67	88	121	152

1- Calculer le coeficient de corrélation linéaire entre les variables x et y.Interpreter le résultat.
2- Déterminer, par la méthode des moindres carrées, l'équation de la droite de regression de y par rapport à x.
3- Quelle sera la demande (y) lorsque le prix (x) aura atteint la valeur 60 ? .

Exercice 3
On a monté une série d'expérience dans une unité pilote en vue une d'étudier l'influence de la température sur le rendement d'une réaction chimique sous une pression donnée. Les données recueillies sont les suivantes (x est la température t-60°C ; y est le rendement en %):

x	1	2	3	4	5	6	7	8	9	10
y	4	6	8	11	12	15	16	18	21	22

Etudier la liaison entre y et x. On fera le graphique en "nuage de points" des valeurs de y en fonction des valeurs de x. On construira la droite de régression de y en x et on donnera la valeur du coefficient de corrélation linéaire r(X,Y). Pour calculer la valeur des deux coeficients de la droite de régression, on se servira des formules dans le cas d'une variable explicative et des relations matricielles (cas de plus d'une variable explicative).

Exercice 4
Un tableau statisique se présent comme suit:

x_i	-5	-1	3	10	13
y_i	33	25	17	3	-3

1- Calculer le coeficient de corrélation linéaire entre les variable x et y.
2- Déterminer l'équation de la fonction de régression permettant d'éstimer y à partir de la connaissance de x.
3- Reprrésenter graphiquement le nuage des points et de la fonction de régression.

Exercice 5

Une distribition statistique se présente de la façon suivante:

	Prix du Brut (en dollar par baril)	Prix de l'essence HT (en centime d'euros par litre)	Prix du diesel HT (en centime d'euros par litre)
Janvier	85	130	115
Février	95	140	125
Mars	105	130	130
Avril	115	140	135
Mai	125	150	145
Juin	135	155	155

1- En considérant d'une part la relation entre le prix du pétrole brut et le prix de l'essence et d'autre part la relation entre le prix du pétrole brut et le prix du diesel, et en supposant que le prix du pétrole brut a un impact sur les prix à la pompe des stations services: Tracer sur deux graphique différents les nuages de points correspondants.

2- Aprés avoir calculé pour chacune des variable (prix du pétrole brut, prix de l'essance et prix du diesel) la moyenne et la variance, déterminer les deux droites de régression. .

3- Comparer les deux coefficients directeurs des droites de régression. Que pouvez-vous en conclure quant à l'évolution des prix respectifs de l'essence et du diesel selon l'évolution future du prix du pétrole ?

4- Calculer et interpréter pour chacune des deux droites le coefficient de détermination. .

5- La croissance des coures du pétrole amène l'acheteur à se demander quels seront les prix des deux carburants à la pompe des stations services si le baril du brut atteignait un jour 200 dollars.Calculer ces prix.

6- En pleine réflexion, l'achateur se dit qu'il sera plus avantageux pour lui de prendre les transports en commun lorsque le prix au litre dépassera les deux euros. Pour quels cours du pétrole brut, devra-t-il s'appreter à vendre sa voiture selon qu'il ait opté pour une voiture essence ou diesel ?

Exercice 6: Espérance de vie et PIB dans le monde

Le tableau ci-dessous reproduit une partie d'un jeu de données donnant, pour 209 pays, le PIB par habitant (en milliers de dollars, estimation 2009) et l'espérance de vie à la naissance (en années, estimation 2010). La figure qui suit est le nuage de point de l'espérance de vie en fonction du PIB par habitant pour les 209 pays .

L'exercice porte sur l'intégralité des pays contenus dans le fichier initial que l'on n'a pas reproduits pour des raisons évidentes.

Pays	Espérance de vie	PIB/habitant
Albanie	77,22	7,7
Algérie	74,26	7,1
Samoa américaines	73,97	8,0
Andorre	82,36	4,49
Angola	38,48	8,3
Anguilla	80,77	12,2
.	.	.
.	.	.
.	.	.

Le but de l'exercice est de savoir si l'espérance de vie à la naissance d'un pays donné peut être déduite du PIB par habitant de ce pays.

1. Quelle est la population étudiée ? Quelle est sa taille ? Quelles sont les variables étudiées ? Quels sont leurs types ?
2. D'après la figure ci-dessous et le commentaire qui la suit, quelle est la régression d'intérêt ? (Cocher la bonne réponse)

☐ La régression de l'espérance de vie en le PIB par habitant ;

☐ La régression du PIB par habitant en l'espérance de vie.

3. On note X le "PIB par habitant" et Y "l'Espérance de vie à la naissance" et on donne les calculs intermédiaires suivants :

$$\sum_{i=1}^{N} x_i = 3419,9 \quad et \quad \sum_{i=1}^{N} y_i = 14728,68$$

$$\sum_{i=1}^{N} x_i^2 = 126676,8 \quad et \quad \sum_{i=1}^{N} y_i^2 = 1058461$$

et

$$\sum_{i=1}^{N} x_i y_i = 257436,3$$

(a) Déterminer les moyennes et les écarts types de X et Y ainsi que la covariance entre X et Y .
(b) En déduire le coefficient de corrélation linéaire entre X et Y , r(X,Y). Commenter la valeur de ce coefficient.
(c) Calculer l'équation de la droite de régression correspondant à la question 2 et la représenter en bleu sur la figure du nuage de points. Que pensez-vous de la qualité de ce modèle pour faire des prévisions ?
4. On donne les calculs intermédiaires suivants :

$$\sum_{i=1}^{N} log(x_i) = 459,237 \quad et \quad \sum_{i=1}^{N} log(x_i)^2 = 1290,884$$

et

$$\sum_{i=1}^{N} log(x_i) y_i = 33750,48$$

(a) Déterminer la moyenne et l'écart type de log(X) ainsi que la covariance entre log(X) et Y .
(b) En déduire le coefficient de corrélation linéaire entre log(X) et Y , r(log(X), Y). Commenter la valeur de ce coefficient en le comparant à celui trouvé dans la question 3b.
(c) Calculer l'équation de la droite de régression de Y en log(X) et représenter en rouge cette courbe de régression sur la figure du nuage de points.
(d) Même si la valeur de r(log(X),Y) n'est pas très élevée, des résultats de statistique inférentielle nous informent de la bonne qualité de la régression linéaire de Y en log(X). Quelle est l'estimation de l'espérance de vie pour un pays dont le PIB par habitant est égal à 20 000 doolar ?

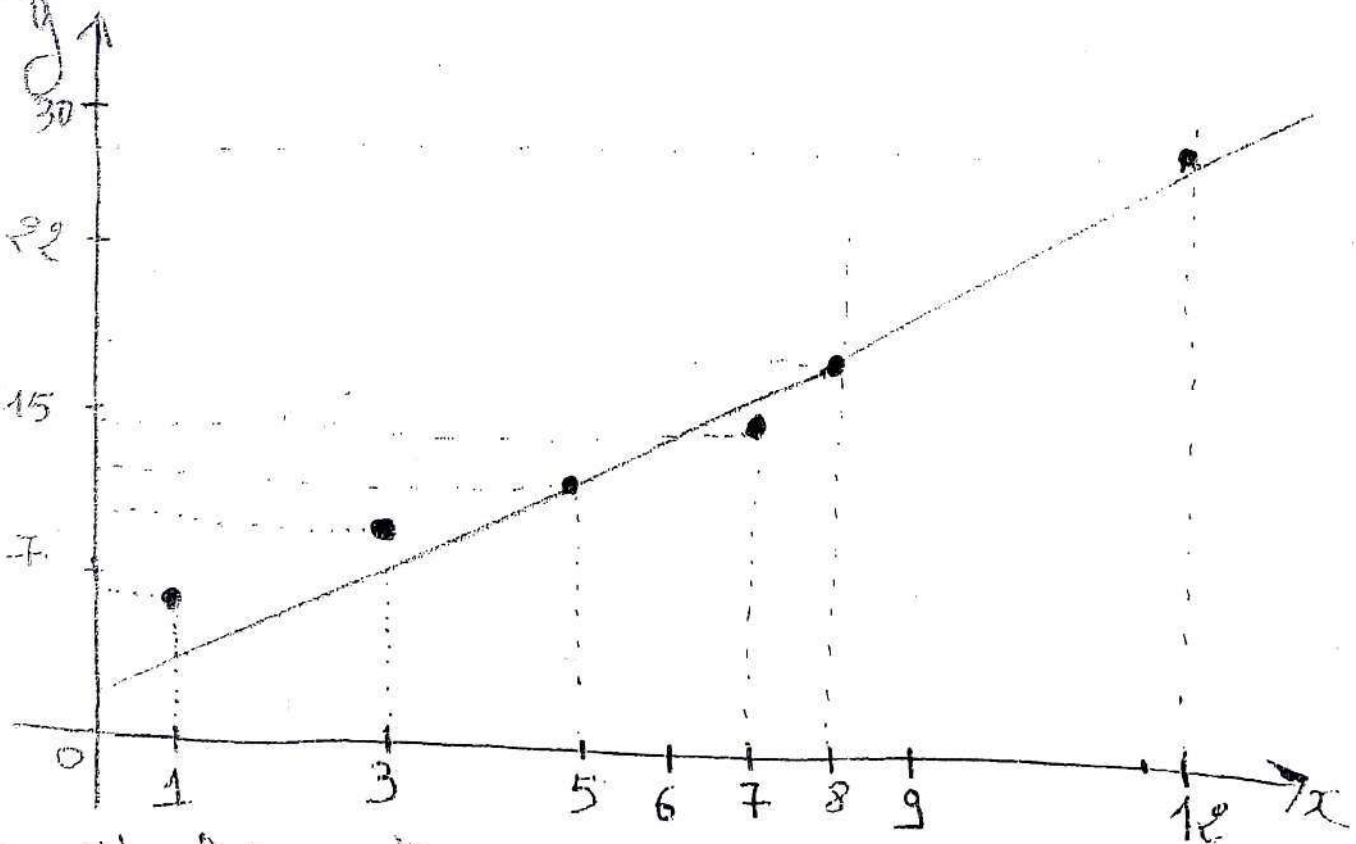
S1 : Correction des exercices de TD N°3 (Ajustement et corrélation)

2019-2020

Exercice n°1

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})^2$
1	6	−5	−5,5	25	42,5	72,25
3	9	−3	−5,5	9	16,5	30,25
5	12	−1	−2,5	1	2,5	6,25
7	14	1	−0,5	1	−0,5	0,25
8	18	2	4,5	4	9	20,25
12	27	6	12,5	36	75	156,25
36	87			76	146	285,5

1)



on voit clairement que le nuage de point s'étale linéaire
met et un ajustement linéaire de y par x semble légitime.

2°) Recherche de l'éq de la droite d'ajustement par la méthode géométrique : $\boxed{y = ax + b}$

cette droite passe par les points $(5,12)$ et $(8,19)$.

par exemple :

$$\Rightarrow \begin{cases} 5a + b = 12 & (1) \\ 8a + b = 19 & (2) \end{cases}$$

$(2) - (1) \Rightarrow 3a = 7 = \boxed{a = \frac{7}{3}}$

$(1) \Rightarrow b = 12 - 5a = 12 - 5 \times \frac{7}{3} = \frac{36 - 35}{3} = \boxed{\frac{1}{3}}$

Donc la droite cherché est l'équation :

$$\boxed{y = \frac{7}{3}x + \frac{1}{3}} \rightarrow y = 2,33x + 0,33$$

3°) les moyennes arithmétiques :

$$\bar{x} = \frac{1}{6}\sum_{i=1}^{6} x_i \quad \text{et} \quad \bar{y} = \frac{1}{6}\sum_{i=1}^{6} y_i$$

$$\bar{x} = \frac{36}{6} = (6) \quad \text{et} \quad \bar{y} = \frac{87}{6} = (14,5)$$

4°) Ajustement par la Méthode des moindres carrés :

(M. M. C)

$$a = \frac{\sum_{i=1}^{6} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{6} (x_i - \bar{x})^2} \quad \text{et} \quad b = \bar{y} - a\bar{x}$$

$$\Rightarrow a = \frac{147}{76} = \boxed{1,91} \quad \text{et} \quad b = 14,5 - 1,91 \times 6$$

$$= 14,5 - 11,46 = (3,04)$$

5/ le coefficient de corrélation :

$$r = \frac{\sum_{i=1}^{6} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{6} (x_i - \bar{x}) \sum_{i=1}^{6} (y_i - \bar{y})^2}}$$

$$r = \frac{145}{\sqrt{76 \times 285,5}} = \frac{145}{\sqrt{21698}} = \frac{145}{147,3}$$

$$\boxed{r = 0,98 \simeq 1}$$

⇒ Forte corrélation positive entre x et y

Ex : 2

1)

x Prix	y Demande (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
9	35	-18	-50	324	2500
13	47	-12	-38	144	1444
21	67	-6	-18	36	324
26	88	-1	3	1	9
38	121	11	36	121	1296
53	152	26	67	676	4489
$\sum_{i=1}^{6}$ 162	510			1302	10062

$\bar{x} = 27 \quad \bar{y} = 85$

$(x_i - \bar{x})(y_i - \bar{y})$
900
456
108
-3
396
1742
3599

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r = \frac{3599}{\sqrt{1302 \times 10062}}$$

$$r = \frac{3599}{\sqrt{13100724}} = \frac{3599}{3619,49}$$

$$r = 0,994$$

Interprétation : il y a entre le prix et la demande, une forte corrélation positive.

Ex : 21

1)

x Prix (€)	y Demande (g)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
9	35	-18	-50	324	2500
16	47	-12	-38	144	1444
21	67	-6	-18	36	324
26	88	-1	3	1	9
38	121	11	36	121	1296
53	152	26	67	676	4489
$\sum_{i=1}^{6}$ 162	510			1302	10062

$\bar{x} = 27$ $\bar{y} = 85$

$(x_i - \bar{x})(y_i - \bar{y})$
900
456
108
-3
396
1742
3599

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r = \frac{3599}{\sqrt{1302 \times 10062}}$$

$$r = \frac{3599}{\sqrt{13100724}} = \frac{3599}{3619,49} \approx$$

$r = 0,994$

Interprétation : il y a entre le prix et la demande, une forte corrélation positive.

2°) Soit $y = ax + b$

Méthode des moindres carrés :

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{et} \quad b = \bar{y} - a\bar{x}.$$

D'où $a = \frac{3599}{1302} = 2,76$

$b = 85 - 2,76 \times 27 = 85 - 74,52 = 10,48$.

La droite de regression par la MMC est :

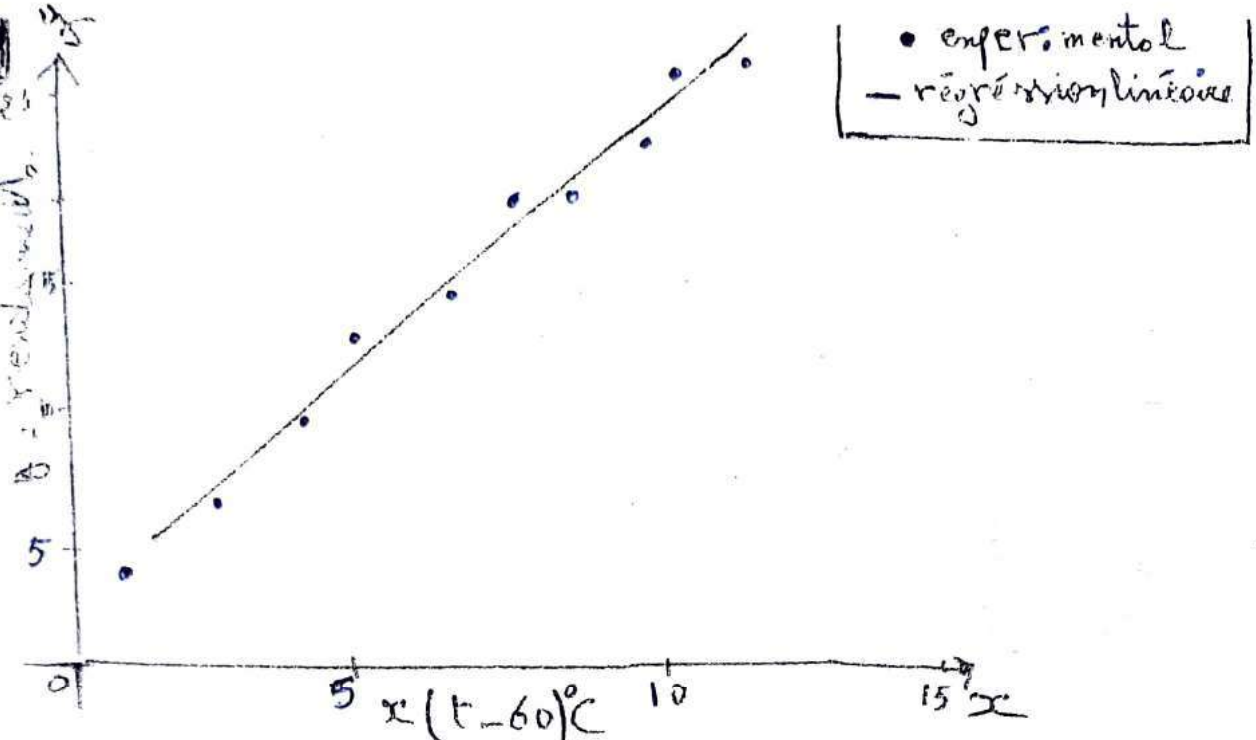
$$\boxed{y = 2,76x + 10,48}$$

3°) $x = 60 \Rightarrow y = 2,76 \times 60 + 10,48$

$\Rightarrow y = 165,6 + 10,48$

$\Rightarrow \boxed{y = 176,08}$

EX:31



1) Droite de régression.

* $n = 10$; $\sum_{i=1} x_i = 55$; $\sum_{i=1} x_i^2 = 385$; $\sum_{i=1} y_i = 133$; $\sum_{i=1} y_i^2 = 2111$

$\sum_{i=1} x_i y_i = 899$.

D'où

$$\sum_{i=1} (x_i - \bar{x})^2 = \sum_{i=1} x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$\sum_{i=1} (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$\sum_{i=1} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1} x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$\sum_{i=1} (x_i - \bar{x})^2 = 82,5$; $\sum_{i=1} (y_i - \bar{y})^2 = 342,1$; $\sum_{i=1} (x_i - \bar{x})(y_i - \bar{y}) = 167,5$

Donc la droite de régression des moindres carrés de y en x est :

$y = ax + b$; Avec $a = 2,0303$ et $b = 2,133$

et $r(x, y) = 0,9970$.

* Si on applique directement les relations matricielles, on obtient :

$$y = \begin{pmatrix} 9 \\ 6 \\ \vdots \\ 22 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \\ 1 & 10 \end{pmatrix} \Rightarrow X^T X = \begin{pmatrix} 10 & 55 \\ 55 & 385 \end{pmatrix}, \quad X^T y = \begin{pmatrix} 133 \\ 899 \end{pmatrix}$$

$$\Rightarrow \begin{cases} 10b + 55a = 133 \\ 55b + 385a = 899 \end{cases} \Rightarrow \text{la solution}$$

EX.N°: 4

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
-5	33	-9	18	81	324	-162
-1	25	-5	10	25	100	-50
3	17	-1	2	1	4	-2
10	3	6	-12	36	144	-72
13	-3	9	-18	81	324	-162
20	75	0	0	224	896	-448

1) Coefficient de corrélation linéaire entre X et Y :

$$r = \frac{\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{5} (x_i - \bar{x})^2 \sum_{i=1}^{5} (y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{20}{5} = 4 \qquad \bar{y} = \frac{75}{5} = 15$$

$$r = \frac{-448}{\sqrt{224 \times 896}} = \frac{-448}{\sqrt{200704}} = \frac{-448}{448} = -1$$

$\Rightarrow$ une forte corrélation négative.

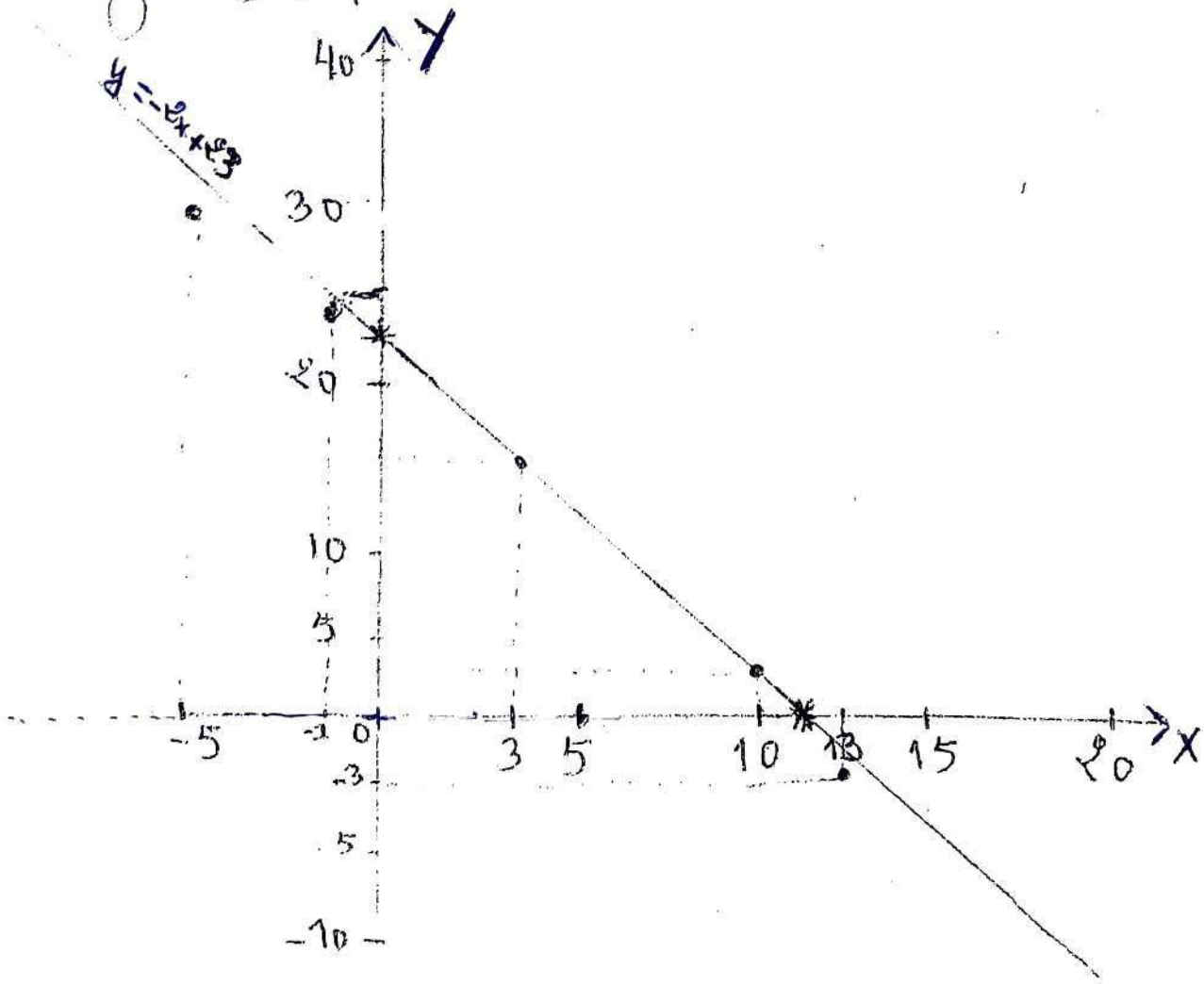
2) L'équation de la fonction (linéaire) de régression de y par rapport à x, de la forme : $y = ax + b$; par la méthode de moindres carrés :

$$a = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2} = \frac{-448}{224} = -2$$

et $b = \bar{y} - a\bar{x} = 15 + 2 \times 4 = 15 + 8 = 23$

D'où $\boxed{y = -2x + 23}$

3) Représentation du nuage de points et de la droite de régression :



Exercice

Un potentiel acheteur d'un véhicule automobile se questionne quant au choix du carburant qu'il va privilégier pour sa voiture. Il sait que le cours du pétrole brut influe directement sur les prix à la pompe dans les stations services. Il décide alors de comparer l'impact du cours du prix du pétrole sur le prix de l'essence ainsi que sur celui du diesel. Pour cela il effectue un relevé des prix de chaque produit sur les six premiers mois de l'année 2008.

	Prix du Brut (en dollar par baril)	Prix de l'essence HT (en centimes d'euros par litre)	Prix du diesel HT (en centimes d'euros par litre)
Janvier	85	130	115
Février	95	140	125
Mars	105	130	130
Avril	115	140	135
Mai	125	150	145
Juin	135	155	155

1) En considérant d'une part la relation entre le prix du pétrole brut et le prix de l'essence et d'autre part la relation entre le prix du pétrole brut et le prix du diesel,

et en supposant que le prix du pétrole brut a un impact sur les prix à la pompe des stations services : Tracer sur deux graphiques différents les nuages de points correspondants.

Pour représenter correctement les nuages de points : le prix du pétrole brut devait être placé en abcisses pour chacun des graphiques et le prix de l'essence (ou du gazole) en ordonnées.

2) Après avoir calculé pour chacune des variables (prix du pétrole brut, prix de l'essence et prix du diesel) la moyenne et la variance, déterminer les deux droites de régression.

Notons x le prix du pétrole brut, y le prix de l'essence et z le prix du diesel.

$$\begin{aligned}
\bar{x} &= \frac{1}{N}\sum_i n_i x_i = \frac{1}{6}(85 + \ldots + 135) = 110 \\
\bar{y} &= \frac{1}{N}\sum_i n_i y_i = \frac{1}{6}(130 + \ldots + 155) = 140,83 \\
\bar{z} &= \frac{1}{N}\sum_i n_i z_i = \frac{1}{6}(115 + \ldots + 155) = 134,17 \\
V(x) &= \frac{1}{N}\sum_i n_i x_i^2 - \bar{x}^2 = \frac{85^2 + \ldots + 135^2}{6} - 110^2 = 291,67 \\
V(y) &= \frac{1}{N}\sum_i n_i y_i^2 - \bar{y}^2 = \frac{130^2 + \ldots + 155^2}{6} - 140,83^2 = 87,74 \\
V(z) &= \frac{1}{N}\sum_i n_i z_i^2 - \bar{z}^2 = \frac{115^2 + \ldots + 155^2}{6} - 134,17^2 = 169,24
\end{aligned}$$

Les droites de régression ont pour équations : $y = ax + b$ avec $a = \frac{cov(x,y)}{V(x)}$ et $b = \bar{y} - a\bar{x}$ et $z = a'x + b'$ avec $a' = \frac{cov(x,z)}{V(x)}$ et $b' = \bar{z} - a'\bar{x}$. Il faut donc au préalable calculer les covariances :

$$cov(x,y) = \frac{1}{N}\sum_i\sum_j x_i y_j - \bar{x}\bar{y} = \frac{85 * 130 + \ldots + 135 * 155}{6} - 110 * 140,83 = 137,8$$

$$cov(x,z) = \frac{1}{N}\sum_i\sum_j x_i z_j - \bar{x}\bar{z} = \frac{85 * 115 + \ldots + 135 * 155}{6} - 110 * 134,17 = 220,46$$

Après calcul, les droites de régression ont pour équations : $y = 0,47x + 89,13$ et $z = 0,76x + 50,57$

a. Comparer les deux coefficients directeurs des droites de régression. Que pouvez-vous en conclure quant à l'évolution des prix respectifs de l'essence et du diesel selon l'évolution future du prix du pétrole ?

$a = 0,47 < a' = 0,76$: le coefficient directeur associé au gazole est plus élevé que celui associé à l'essence. Ainsi, si le prix du pétrole brut (x) augmente de 1, le prix de l'essence augmente de 0,47 et celui du gazole de 0,76. Le prix du gazole est donc davantage sensible aux fluctuations du prix du pétrole.

b. Calculer et interpréter pour chacune des deux droites le coefficient de détermination.

Le coefficient de détermination fournit une indication de la qualité de l'ajustement. Il est, par définition, toujours compris entre 0 et 1.

$$R^2(x,y) = \frac{cov(x,y)^2}{V(x)V(y)} = \frac{137,8^2}{291,67 * 87,74} = 0,74$$

Le modèle explique 74% de la réalité : le fait que le prix de l'essence diffère selon les mois peut être expliqué à 74% par le fait que le prix du pétrole brut diffère.

$$R^2(x,z) = \frac{cov(x,z)^2}{V(x)V(z)} = \frac{290,46^2}{291,67 * 169,26} = 0,98$$

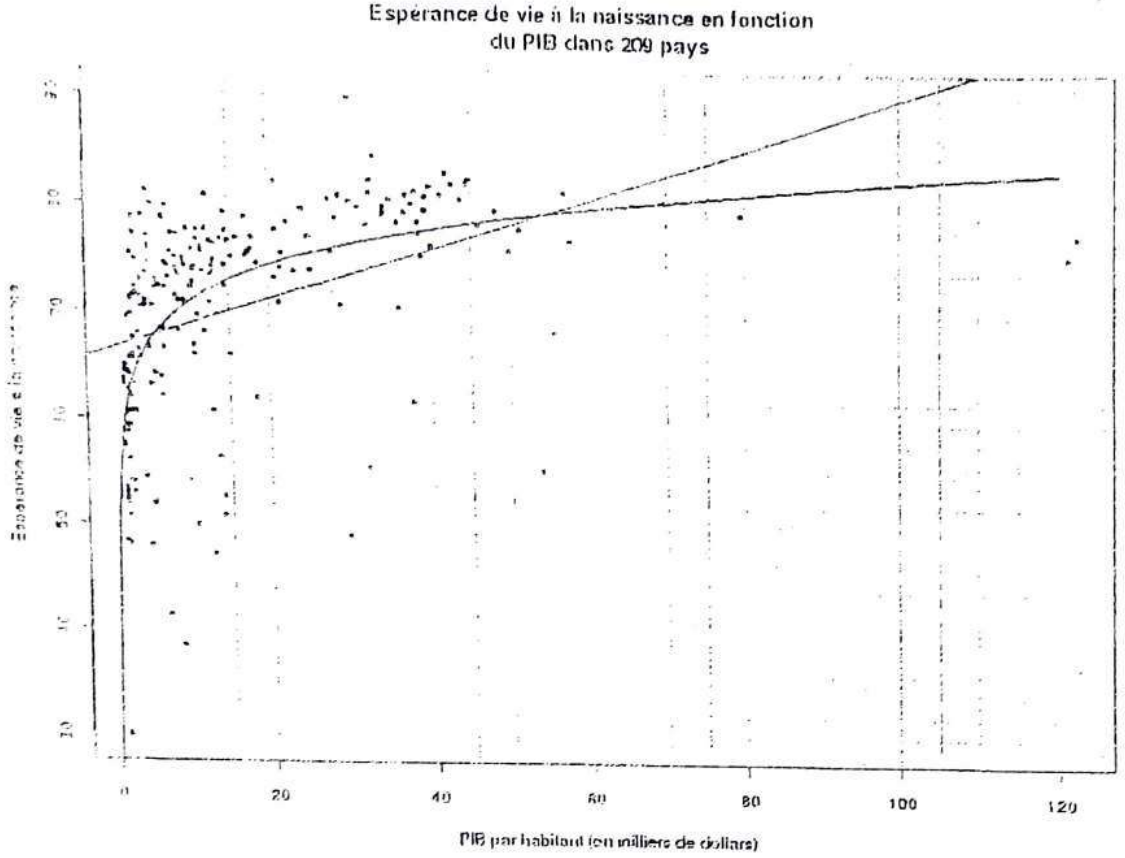
Le modèle explique 98% de la réalité : le fait que le prix du gazole diffère selon les mois peut être expliqué à 98% par le fait que le prix du pétrole brut diffère.

3) La croissance des cours du pétrole amène l'acheteur à se demander quels seront les prix des deux carburants à la pompe des stations services si le baril du brut atteignait un jour 200 dollars. Calculer ces prix.

Si le prix du pétrole brut (x) est égal à 200, calculons y et z : $y = 0,47 * 200 + 89,13 = 183,13$ et $z = 0,76 * 200 + 50,57 = 202,57$.

4) En pleine réflexion, l'acheteur se dit qu'il sera plus avantageux pour lui de prendre les transports en commun lorsque le prix au litre dépassera les deux euros. Pour quels cours du pétrole brut, devra-t-il s'apprêter à vendre sa voiture selon qu'il ait opté pour une voiture essence ou diesel?

Pour quels prix du pétrole brut, le prix de l'essence (y) et le prix du gazole (z) dépassent-ils les 2 euros, c'est à dire 200 centimes : $200 = 0,47x + 89,13 \Rightarrow x = 235,9$ et $200 = 0,76x + 50,57 \Rightarrow x = 196,6$.



Le but de l'exercice est de savoir si l'espérance de vie à la naissance d'un pays donné peut être déduite du PIB par habitant de ce pays.

1. Quelle est la population étudiée? Quelle est sa taille? Quelles sont les variables étudiées? Quels sont leurs types?

   Réponse :

   La population étudiée est l'ensemble des 209 pays pour lequel les données sont disponibles ; sa taille est donc $N = 209$.
   Les variables étudiées sont le PIB par habitant et l'espérance de vie à la naissance qui sont toutes les deux des variables quantitatives continues.

2. D'après la figure ci-dessous et le commentaire qui la suit, quelle est la régression d'intérêt? (Cocher la bonne réponse)
   - ☒ La régression de l'espérance de vie en le PIB par habitant ;
   - ☐ La régression du PIB par habitant en l'espérance de vie.

3. On note X le "PIB par habitant" et Y l'"Espérance de vie à la naissance" et on donne les calculs intermédiaires suivants :

$$\sum_{i=1}^{N} x_i = 3\,419,9 \qquad \text{et} \sum_{i=1}^{N} y_i = 14\,728,68,$$

$$\sum_{i=1}^{N} x_i^2 = 126\,676,8 \qquad \text{et} \sum_{i=1}^{N} y_i^2 = 1\,058\,461$$

et

$$\sum_{i=1}^{N} x_i y_i = 257\,436,3.$$

12

Exercice – Espérance de vie et PIB dans le monde

Le tableau ci-dessous reproduit une partie d'un jeu de données donnant, pour 209 pays, le PIB par habitant (en milliers de dollars, estimation 2009) et l'espérance de vie à la naissance (en années, estimation 2010). La figure qui suit est le nuage de point de l'espérance de vie en fonction du PIB par habitant pour les 209 pays [3]. L'exercice porte sur l'intégralité des pays contenus dans le fichier initial que l'on n'a pas reproduits pour des raisons évidentes.

Pays	Espérance de vie	PIB / habitant
Albanie	77,22	7,7
Algérie	74,26	7,1
Samoa américaines	73,97	8,0
Andorre	82,36	44,9
Angola	38,48	8,3
Anguilla	80,77	12,2
Antigua et Barduba	75,26	17,2
Argentine	76,76	13,9
Arménie	72,96	5,5
Aruba	75,51	21,8
⋮	⋮	⋮

3. Les données ont été obtenues sur le site de la CIA (Central Intelligence Agency, USA), "The World Factbook" : https://www.cia.gov/library/publications/the-world-factbook/index.html. Quelques valeurs atypiques ont été supprimées du jeu de données initial.

(a) Déterminer les moyennes et les écarts types de X et Y ainsi que la covariance entre X et Y.
Réponse :

Moyennes

$$\overline{X} = \frac{3\,419,9}{209} \simeq 16,363 \qquad \text{et } \overline{Y} = \frac{14\,728,68}{209} \simeq 70,472$$

Variances

$$\text{Var}(X) = \frac{126\,676,8}{209} - 16,363^2 \simeq 338,356 \qquad \text{et Var}(Y) = \frac{1\,058\,461}{209} - 70,472^2 \simeq 98,081$$

d'où $\sigma(X) = \sqrt{338,356} \simeq 18,39$ et $\sigma(Y) = \sqrt{98,081} \simeq 9,90$. **Covariance**

$$\text{Cov}(X,Y) = \frac{257\,436,3}{209} - 16,363 \times 70,472 \simeq 78,605.$$

(b) En déduire le coefficient de corrélation linéaire entre X et Y, $r(X,Y)$. Commenter la valeur de ce coefficient.
Réponse :

$$r(X,Y) = \frac{78,605}{18,39 \times 9,90} \simeq 0,431.$$

La corrélation linéaire entre X et Y a une intensité moyenne : l'espérance de vie à la naissance a une dépendance linéaire moyenne avec le PIB par habitant.

(c) Calculer l'équation de la droite de régression correspondant à la question 2 et la représenter en bleu sur la figure du nuage de points. Que pensez-vous de la qualité de ce modèle pour faire des prévisions ?
Réponse :

La droite de régression de Y en X a pour équation $Y = aX + b$ où

$$a = \frac{78,605}{338,356} \simeq 0,232 \qquad \text{et } b = 70,472 - 0,232 \times 16,363 \simeq 66,7.$$

D'après la valeur de $r(X,Y)$ et aussi d'après la figure, le modèle est de faible qualité pour faire des prévisions

4. On donne les calculs intermédiaires suivants :

$$\sum_{i=1}^{N} \log(x_i) = 459,237 \qquad \text{et } \sum_{i=1}^{N} \log(x_i)^2 = 1\,290,884,$$

et

$$\sum_{i=1}^{N} \log(x_i) y_i = 33\,750,48.$$

(a) Déterminer la moyenne et l'écart type de $\log(X)$ ainsi que la covariance entre $\log(X)$ et Y.
Réponse :

Moyenne et variance de $\log(X)$

$$\overline{\log(X)} = \frac{459,237}{209} \simeq 2,197 \qquad \text{et Var}(\log(X)) = \frac{1\,290,884}{209} - 2,197^2 \simeq 1,348$$

d'où $\sigma(\log(X)) = \sqrt{1,348} \simeq 1,16$. **Covariance**

$$\text{Cov}(\log(X), Y) = \frac{33\,750,48}{209} - 2,197 \times 70,472 \simeq 6,637.$$

(b) En déduire le coefficient de corrélation linéaire entre $\log(X)$ et Y, $r(\log(X), Y)$. Commenter la valeur de ce coefficient en le comparant à celui trouvé dans la question 3b.
Réponse :

$$r(\log(X), Y) = \frac{6,637}{1,16 \times 9,90} \simeq 0,577.$$

La corrélation linéaire entre $\log(X)$ et Y a une intensité moyenne mais plus forte que la corrélation entre X et Y. L'espérance de vie à la naissance a une dépendance linéaire meilleure avec le logarithme du PIB par habitant qu'avec le PIB par habitant lui-même.

(c) Calculer l'équation de la droite de régression de Y en $\log(X)$ et représenter en rouge cette courbe de régression sur la figure du nuage de points.

Réponse :

La droite de régression de Y en $\log(X)$ a pour équation $Y = a\log(X) + b$ où

$$a = \frac{6,637}{1,348} \simeq 4,92 \qquad \text{et } b = 70,472 - 4,92 \times 2,197 \simeq 59,66.$$

Pour représenter la courbe de régression sur le nuage de points, on calcule les valeurs prédites pour l'espérance de vie pour un certain nombre de valeurs du PIB par habitant selon la courbe de régression :

PIB/hab	2	5	10	15	20	30	40	60	80	120
Prévision	63,1	67,6	71,0	73,0	74,4	76,4	77,8	79,8	81,2	83,2

(d) Même si la valeur de $r(\log(X), Y)$ n'est pas très élevée, des résultats de statistique inférentielle nous informent de la bonne qualité de la régression linéaire de Y en $\log(X)$. Quelle est l'estimation de l'espérance de vie pour un pays dont le PIB par habitant est égal à \$20 000 ?

Réponse :

L'estimation de l'espérance de vie pour un pays dont le PIB par habitant est

$$y = 4,92 \times \log(20) + 59,66 \simeq 74,4 \text{ ans.}$$

[1] Il s'agissait du logarithme népérien, noté sur votre calculatrice "ln" ; je n'ai pas compté d'erreur pour la confusion qui aurait pu être détectée toutefois par la valeur trouvée.

15