Statistique descriptive

Année académique 2020–2021

Carole.Baum@uliege.be

Chapitre 4 : Série statistique bivariée

Exercice 1. Le Tableau de contingence 1 décrit la répartition d'une population constituée de 535 ménages selon les deux variables suivantes : X représente le nombre de pièces de l'habitation et Y correspond au nombre d'enfants du ménage.

TABLE 1 – Tableau de contingence pour la série double "Nombre de pièces - Nombre d'enfants".

	Valeurs de Y				
Valeurs de X	0	1	2	3	4
1	7	3	2	1	0
2	24	32	21	2	1
3	16	35	54	26	4
4	9	28	74	55	12
5	4	12	46	13	12
6	2	6	16	11	7

(a) Que représentent les effectifs n_{23} et n_{54} ? Calculer les fréquences correspondantes.

(b) Déterminer les distributions marginales des variables X et Y. Calculer les moyennes, médianes et variances marginales ainsi que les modes marginaux.

(c) Déterminer la distribution conditionnelle de la variable Y sachant que le nombre de pièces du logement est égal à 4. Calculer la moyenne et la variance de cette distribution conditionnelle.

(d) Calculer la covariance de la série double.

Solution :

(a) $n_{23} = 21$ il s'agit du nombre de ménages ayant 2 pièces dans l'habitation et 2 enfants.
$n_{54} = 13$ il s'agit du nombre de ménages ayant 5 pièces dans l'habitation et 3 enfants.
On a $f_{23} = 21/535 = 0.039$ et $f_{54} = 13/535 = 0.024$.

(b) La distribution marginale de la variable X et donnée par

x_i	1	2	3	4	5	6
n_i	13	80	135	178	87	42

On peut alors déterminer les paramètres demandés.
— $\bar{x} = \frac{1}{535}(13 \cdot 1 + ... + 42 \cdot 6) = 3.70$;
— $\frac{n}{2} = 267.5 \Rightarrow \tilde{x} = x_{(268)} = 4$ car $\begin{cases} N(3) = 228 \\ N(4) = 406 \end{cases}$;
— $s_x^2 = \frac{1}{535}(13 \cdot 1^2 + ... + 42 \cdot 6^2) - 3.70^2 = 1.42$;

— $x_M = 4$.

De la même manière, la distribution marginale de Y et les paramètres demandés sont donnés par

y_j	0	1	2	3	4
n_j	62	116	213	108	36

— $\bar{y} = \frac{1}{535}(62 \cdot 0 + ... + 36 \cdot 4) = 1.89$;
— $\tilde{y} = y_{(268)} = 2$;
— $s_y^2 = \frac{1}{535}(62 \cdot 0^2 + ... + 36 \cdot 4^2) - 1.89^2 = 1.13$;
— $y_M = 2$.

(c) La distribution conditionnelle de Y sachant que $X = 4$ est donnée par

$y_i \mid x = 4$	0	1	2	3	4	
n_i	9	28	74	55	12	178

On a alors
— $\bar{y}_{|x=4} = \dfrac{1}{178}(9 \cdot 0 + ... + 12 \cdot 4) = 2.185$;
— $s^2_{y_{|x=4}} = \dfrac{1}{178}(9 \cdot 0^2 + ... + 12 \cdot 4^2) - 2.185^2 = 0.906$.

(d) $s_{xy} = \dfrac{1}{535}(7 \cdot 1 \cdot 0 + 3 \cdot 1 \cdot 1 + ... + 0 \cdot 1 \cdot 4 + 24 \cdot 0 \cdot 2 + ... + 7 \cdot 6 \cdot 4) - 3.70 \cdot 1.89 = 0.47$.

Exercice 2. Le directeur d'une entreprise vinicole a l'habitude d'offrir à ses employés une prime de fin d'année de 10, 15, 20 ou 25 unités monétaires. Le tableau 2 représente le tableau de contingence mettant en rapport la variable Y = "montant de la prime" avec la variable X = "taille de la cave à vin exprimée en nombre de bouteilles".

TABLE 2 – Tableau de contingence pour la série double "Taille de la cave - Montant de la prime".

	Y			
X	10	15	20	25
$[0, 100]$	11	13	5	3
$]100, 200]$	14	21	15	6
$]200, 300]$	5	8	9	7
$]300, 400]$	3	6	8	9
$]400, 500]$	2	3	5	4

(a) Quelle est la prime le plus souvent distribuée ? Que représente cette valeur dans la distribution marginale de Y ?

(b) Quelle est la somme totale dépensée par le directeur pour offrir les primes de fin d'année ?

(c) Si le directeur voulait être équitable par rapport à tous ses employés, quel montant donnerait-il à chacun ? Cette valeur est-elle proche des moyenne et médiane marginales de Y ?

(d) Comparer les nombres moyens de bouteilles détenues par les employés en tenant compte des différentes primes perçues.

Solution :

(a) La prime la plus souvent distribuée est le mode de la distribution marginale de y :

y_i	10	15	20	25	
n_i	35	51	42	29	157

Ainsi, $y_M = 15$.

(b) La somme totale dépensée est donnée par $10 \cdot 35 + 15 \cdot 51 + 20 \cdot 42 + 25 \cdot 29 + 25 \cdot 29 = 2680$.

(c) Une somme "honnête" consisterait à diviser la somme totale par le nombre d'employés : $\frac{2680}{157} = 17.07$;
$\bar{y} = 17.07$ (la moyenne est, par définition, la valeur donnée ci-dessus) ;
$\tilde{y} = 15$.

(d) $\bar{x}_{|y=10} = \frac{11 \cdot 50 + 14 \cdot 150 + 5 \cdot 250 + 3 \cdot 350 + 2 \cdot 450}{35} = 167.14$;
$\bar{x}_{|y=15} = \frac{13 \cdot 50 + 21 \cdot 150 + 8 \cdot 250 + 6 \cdot 350 + 3 \cdot 450}{51} = 181.37$;
$\bar{x}_{|y=20} = \frac{5 \cdot 50 + 15 \cdot 150 + 9 \cdot 250 + 8 \cdot 350 + 5 \cdot 450}{42} = 233.33$;
$\bar{x}_{|y=25} = \frac{3 \cdot 50 + 6 \cdot 150 + 7 \cdot 250 + 9 \cdot 350 + 4 \cdot 450}{29} = 267.24$.

Exercice 3. Le tableau 3 reprend les âges de l'époux (variable X) et de l'épouse (variable Y) pour les 100 derniers mariages enregistrés dans une ville.

TABLE 3 – Ages des époux pour les 100 derniers mariages enregistrés dans une ville

	Classes de Y			
Classes de X	$[18,22]$	$]22,26]$	$]26,30]$	$]30,34]$
$[20,24]$	14	5	2	0
$]24,28]$	15	19	7	3
$]28,32]$	5	12	5	2
$]32,36]$	0	1	2	2
$]36,40]$	0	1	3	2

(a) Calculer l'âge moyen des époux et des épouses au moment du mariage.

(b) Comment peut-on obtenir à partir du tableau de contingence les informations suivantes : (1) 38% des femmes se marient entre 22 et 26 ans ; (2) 19% des mariages concernent des hommes âgés de 24 à 28 ans et des femmes âgées de 22 à 26 ans ; (3) 31,6% des femmes âgées de 22 à 26 ans se sont mariées avec des hommes de 28 à 32 ans.

Solution :

(a) Les distributions marginales sont données par

x_i	$[20;24]$	$]24;28]$	$]28;32]$	$]32;36]$	$]36;40]$
n_i	21	44	24	5	6

y_i	$[18;22]$	$]22;26]$	$]26;30]$	$]30;34]$
n_i	34	38	19	9

dont on déduit

$$\bar{x} = \frac{21 \cdot 22 + 44 \cdot 26 + 24 \cdot 30 + 5 \cdot 34 + 6 \cdot 38}{100} = 27.24 \quad \text{et} \quad \bar{y} = \frac{34 \cdot 20 + 38 \cdot 24 + 19 \cdot 28 + 9 \cdot 32}{100} = 24.12$$

(b) (1) Il s'agit de la fréquence marginale de la classe $]22;26]$ chez les femmes, à savoir 38/100.
(2) Il s'agit de la fréquence du couple $(]24;28];]22;26])$, à savoir 19/100.
(3) On cherche la proportion de femmes entre 22 et 26 ans qui se sont mariées avec des hommes de 28 à 32 ans. Parmi les 38 femmes vérifiant la première condition, 12 vérifient la seconde. On a donc $12/38 = 31.6\%$.

Exercice 4. Le gérant d'un magasin d'appareils électroménagers a enregistré chaque semaine le nombre x_i de centaines d'appels téléphoniques reçus de l'extérieur et le chiffre d'affaires y_i réalisé en unités monétaires. Les résultats sont repris dans le tableau 4 où n_{ij} indique le nombre de semaines où le magasin a reçu x_i centaines d'appels téléphoniques et a fait y_j UM comme chiffre d'affaires.

TABLE 4 – Tableau de contingence pour la série double "Nombre d'appels téléphoniques - Chiffre d'affaires".

Nombre d'appels téléphoniques	Chiffre d'affaires 1	2	3	4	5
2	9	5	3	1	0
3	4	5	7	3	1
4	0	6	9	6	3
5	0	5	14	14	5

(a) Combien de semaines cette enquête a-t-elle duré ?

(b) Combien d'appels téléphoniques le magasin reçoit-il en moyenne par semaine ?

(c) Déterminer les distributions marginales ainsi que les moyennes et variances marginales.

(d) Comparer les chiffres d'affaires moyens réalisés lorsque le magasin reçoit 2 ou 5 appels téléphoniques sur la semaine.

(e) Calculer la covariance et le coefficient de corrélation linéaire entre les deux variables.

Solution :

(a) Le nombre de semaines qu'a duré l'enquête est donné par la somme des n_{ij}.
On a donc $n = 9 + 5 + \ldots + 0 + 4 \ldots + 5 = 100$.

(b) La distribution marginale du nombre d'appels téléphoniques est donnée par

x_i	2	3	4	5
n_i	18	20	24	38

dont on déduit que $\bar{x} = \dfrac{18 \cdot 2 + 20 \cdot 3 + 24 \cdot 4 + 38 \cdot 5}{100} = 3.82$.

(c) La distribution et la moyenne marginale de x sont données ci-dessus. On peut aussi calculer la variance marginale :

$$s_x^2 = \frac{18 \cdot 2^2 + 20 \cdot 3^2 + 24 \cdot 4^2 + 38 \cdot 5^2}{100} - 3.82^2 = 1.27.$$

De la même manière, la distribution marginale de y est donnée par

y_j	1	2	3	4	5
n_j	13	21	33	24	9

dont on déduit les moyenne et variance marginales :

$$\bar{y} = \frac{13 \cdot 1 + 21 \cdot 2 + 33 \cdot 3 + 24 \cdot 4 + 9 \cdot 5}{100} = 2.95$$
$$s_y^2 = \frac{13 \cdot 1^2 + 21 \cdot 2^2 + 33 \cdot 3^2 + 24 \cdot 4^2 + 9 \cdot 5^2}{100} - 2.95^2 = 1.33.$$

(d) On veut comparer les moyennes conditionnelles de y lorsque $x = 2$ et lorsque $x = 5$. On a

$$\bar{y}_{|x=2} = \frac{9 \cdot 1 + 5 \cdot 2 + 3 \cdot 3 + 1 \cdot 4 + 0 \cdot 5}{18} = 1.78$$
$$\bar{y}_{|x=5} = \frac{0 \cdot 1 + 5 \cdot 2 + 14 \cdot 3 + 14 \cdot 4 + 5 \cdot 5}{38} = 3.5.$$

Comme on pouvait s'y attendre, le chiffre d'affaire moyen est plus élevé lorsque le nombre d'appels est élevé.

(e) La covariance est donnée par

$$\begin{aligned} s_{xy} &= \frac{1}{n} \sum_{i,j} n_{ij} x_i x_j - \bar{x}\bar{y} \\ &= \frac{9 \cdot 2 \cdot 1 + \cdot + 0 \cdot 2 \cdot 5 + 4 \cdot 3 \cdot 1 + \ldots + 5 \cdot 5 \cdot 5}{100} - 3.82 \cdot 2.95 \\ &= 0.701. \end{aligned}$$

Le coefficient de corrélation est quant à lui donné par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0.701}{\sqrt{1.27 \cdot 1.33}} = 0.539.$$

Exercice 5. Une étude statistique a été réalisée auprès des 500 entreprises d'un même secteur industriel. Deux variables étaient considérées : la variable X correspond à la taille de l'entreprise (en nombre de salariés) et la variable Y au niveau de leur salaire mensuel (en unités monétaires). L'étude a permis d'obtenir les renseignements du tableau 5.

TABLE 5 – Répartition de 500 entreprises en fonction du nombre de salariés et du salaire mensuel

Nombre de salariés (X)	Niveau de salaire mensuel (Y)		
	5000	10 000	30 000
$[0, 500]$	13	5	2
$]500, 1500]$	60	25	15
$]1500, 3000]$	90	30	30
$]3000, 5000]$	137	65	28

(a) Déterminer les salaires mensuels moyen et médian des employés des entreprises considérées.

(b) Déterminer, pour chaque classe de taille d'entreprise, le salaire mensuel moyen.

(c) Quelle relation existe-t-il entre le salaire moyen global calculé en (a) et les salaires moyens calculés en (b).

Solution :

(a) La distribution marginale du salaire mensuel est donnée par

y_j	5000	10 000	30 000
n_j	300	125	75

Dès lors, le salaire mensuel moyen est donné par

$$\bar{y} = \frac{300 \cdot 5000 + 125 \cdot 10\ 000 + 75 \cdot 30\ 000}{500} = 10\ 000.$$

Le salaire mensuel médian est quant à lui donné par

$$\tilde{y} = \frac{y_{(250)} + y_{(251)}}{2} = \frac{5000 + 5000}{2} = 5000.$$

(b) Les salaires moyens conditionnels sont donnés par

$$\bar{y}_{|x_1} = \frac{13 \cdot 5000 + 5 \cdot 10\ 000 + 2 \cdot 30\ 000}{20} = 8750$$
$$\bar{y}_{|x_2} = \frac{60 \cdot 5000 + 25 \cdot 10\ 000 + 15 \cdot 30\ 000}{100} = 10\ 000$$
$$\bar{y}_{|x_2} = \frac{90 \cdot 5000 + 30 \cdot 10\ 000 + 30 \cdot 30\ 000}{150} = 11\ 000$$
$$\bar{y}_{|x_4} = \frac{137 \cdot 5000 + 65 \cdot 10\ 000 + 28 \cdot 30\ 000}{230} = 9456$$

où $x_1, \ldots, x_4$ représentent respectivement les classes $[0, 500], \ldots,]3000, 5000]$ de nombres de salariés.

(c) On a

$$\bar{y} = \frac{20 \cdot \bar{y}_{|x_1} + 100 \cdot \bar{y}_{|x_2} + 150 \cdot \bar{y}_{|x_3} + 230 \cdot \bar{y}_{|x_4}}{500}.$$

Autrement dit, le salaire moyen global n'est rien d'autre que la moyenne pondérée des salaires moyens conditionnels.

Exercice 6. Soit S la série double suivante, due à Anscombe (1973) :

x_i	10	8	13	9	11	14
y_i	8,04	6,95	7,58	8,81	8,33	9,96

x_i	6	4	12	7	5
y_i	7,24	4,26	10,84	4,82	5,68

(a) Représenter le nuage de points de cette série bivariée.

(b) Rechercher l'équation de la droite des moindres carrés. Calculer les résidus.

Solution :

(a) Le nuage de points se trouve en Figure 1.

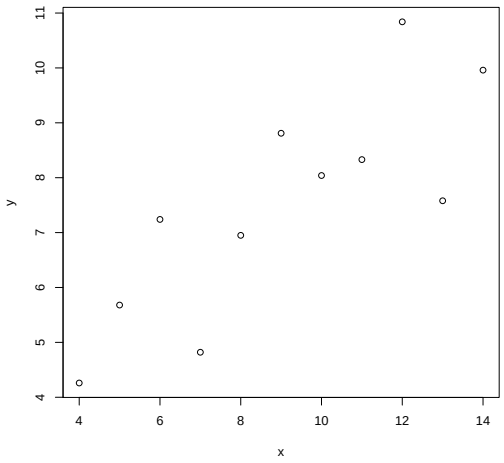

FIGURE 1 – Nuage de points de la série bivariée S

(b) On peut calculer les quantités suivantes :

$\sum_i x_i = 99$; $\sum_j y_j = 82.51$;
$\sum_i x_i^2 = 1001$; $\sum_j y_j^2 = 660.17$; $\sum_i x_i y_i = 797.6$

dont on déduit

$$\bar{x} = \frac{99}{11} = 9 \; ; \qquad \bar{y} = \frac{82.51}{11} = 7.5 \; ;$$

$$s_x^2 = \frac{1001}{11} - 9^2 = 10 \; ; \qquad s_y^2 = \frac{660.17}{11} - 7.5^2 = 3.77 \; ;$$

$$s_{xy} = \frac{797.6}{11} - 9 \cdot 7.5 = 5.$$

De là, on a

$$\hat{a} = \frac{s_{xy}}{s_x^2} = \frac{5}{10} = 0.5$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 7.5 - 0.5 \cdot 9 = 3.$$

L'équation de la droite des moindres carrés est ainsi donnée par $y = 0.5x + 3$.

Pour calculer les résidus, il faut commencer par calculer les valeurs ajustées : $\hat{y}_i = 0.5x_i + 3$. Ensuite, les résidus sont donnés par : $\delta_i = y_i - \hat{y}_i$. On obtient le tableau suivant :

x_i	10	8	13	9	11	14	6	4	12	7	5
$\hat{y}_i$	8	7	9.5	7.5	8.5	10	6	5	9	6.5	5.5
δ_i	0.04	-0.05	-1.92	1.31	-0.17	-0.04	1.24	-0.74	1.84	-1.68	0.18

Exercice 7. Au Grand-Duché de Luxembourg, on a relevé les taux de chômage suivants :

Année	Taux de chômage en %
75	4
76	4,4
77	5
78	5,2
79	5,9
80	6,3
81	7,4
82	8,1
83	8,3
84	9,7
85	10,2
86	10,4
87	10,5

(a) A partir de ces données, déterminer quel type de liaison fonctionnelle $y = f(x)$ existe entre le taux de chômage y et l'année x.

(b) Rechercher l'équation de la droite des moindres carrés exprimant y en fonction de x et calculer le coefficient de corrélation linéaire. Interpréter le coefficient de détermination.

(c) Estimer les taux de chômage livrés par la droite des moindres carrés et calculer les résidus. Commenter la situation des années 81, 86 et 87. Estimer le taux de chômage pour les années 2150 et 1965 ; commenter ces résultats et en tirer des conclusions pratiques.

Solution :

(a) Après avoir représenté le diagramme de dispersion de la Figure 2, on observe une relation linéaire entre l'année et le taux de chômage. On a donc $y = f(x)$ avec f une fonction linéaire croissante.

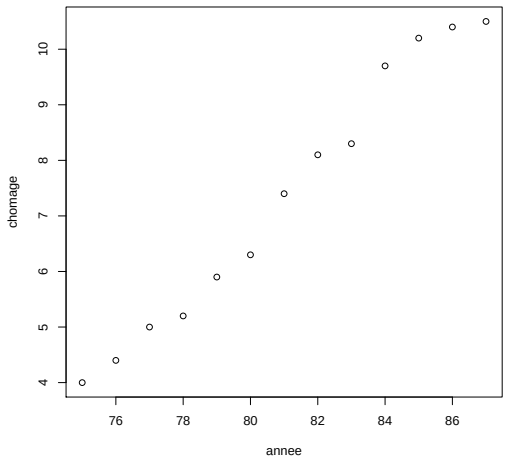

FIGURE 2 – Diagramme de dispersion du taux de chômage au Grand-Duché du Luxembourg en fonction des années.

(b) On peut, comme dans l'exercice précédent, calculer les quantités suivantes :

$\sum_i x_i = 1053$; $\sum_j y_j = 95.4$;
$\sum_i x_i^2 = 85475$; $\sum_j y_j^2 = 767.7$; $\sum_i x_i y_i = 7837.3$

dont on déduit

$\bar{x} = 81$; $\bar{y} = 7.34$;
$s_x^2 = 14$; $s_y^2 = 5.18$;
$s_{xy} = \dfrac{7837.3}{13} - 81 \cdot 7.34 = 8.33.$
De là, on a

$$\hat{a} = \frac{s_{xy}}{s_x^2} = \frac{8.33}{14} = 0.595$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 7.34 - 0.595 \cdot 81 = -40.855.$$

L'équation de la droite des moindres carrés est ainsi donnée par
$y = 0.595x - 40.855$.

Le coefficient de corrélation linéaire est donné par

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{8.33}{\sqrt{14 \cdot 5.18}} = 0.978,$$

qui est, comme on pouvait le prévoir au vu du diagramme de dispersion, positif et très proche de 1, traduisant une forte relation linéaire croissante entre les deux variables.

Le coefficient de détermination est quant à lui donné par

$$R^2 = r_{xy}^2 = 0.956.$$

Cela signifie que 96% de la variabilité du taux de chômage peut être expliquée par l'année.

(c) Le tableau suivant donne les valeurs estimées par la droite de régression ainsi que les résidus.

x_i (Année)	75	76	77	78	79	80
$\hat{y}_i = \hat{a}x_i + \hat{b}$	3.77	4.365	4.96	5.555	6.15	6.745
$\delta_i = y_i - \hat{y}_i$	0.23	0.035	0.04	-0.355	-0.25	-0.445

x_i (Année)	**81**	82	83	84	85	**86**	**87**
$\hat{y}_i = \hat{a}x_i + \hat{b}$	**7.34**	7.935	8.53	9.125	9.72	**10.315**	**10.9**
$\delta_i = y_i - \hat{y}_i$	**0.06**	0.165	-0.23	0.575	0.48	**0.085**	**-0.41**

Commentaires :
- En 81 et 86, le taux réel est *sous-estimé*.
- En 87, le taux réel est *sur-estimé*.

Estimations :
- 2150 : $\widehat{y}_{2150} = 0.595 \cdot 250 - 40.855 = 107.895 > 100$ % ! ! !
- 1965 : $\widehat{y}_{65} = 0.595 \cdot 65 - 40.855 = -2.18 < 0$ % ! ! !

$\Rightarrow$ Une régression n'est pas valable pour toutes les valeurs de x ; il ne faut pas trop s'éloigner des données récoltées.

Exercice 8. Dans les années 1950, beaucoup de recherches statistiques ont tenté d'établir un lien entre la consommation de cigarettes et le cancer du poumon. Une des enquêtes s'est concentrée sur 11 pays et sur les variables suivantes :

X est le nombre de cigarettes par habitant du pays en 1930

Y est le taux (calculé sur une base fixée) de mortalité due au cancer du poumon parmi la population du pays lors de l'année 1950.

Les données sont indiquées dans le tableau ci-dessous.

Pays	X	Y
Australie	455	170
Canada	510	150
Danemark	380	165
États-Unis	1280	190
Finlande	1115	350
Grande-Bretagne	1145	465
Hollande	460	245
Islande	220	58
Norvège	250	90
Suède	310	115
Suisse	530	250

(a) Considérer la série univariée S_Y. Construire une boîte à moustaches et commenter ses caractéristiques principales.

(b) Considérer la série double $S = \{(x_i, y_i); 1 \leq i \leq 11\}$. On donne les résumés numériques suivants :

$$\sum_{i=1}^{11} x_i = 6\ 655,\ \sum_{i=1}^{11} y_i = 2\ 248,\ \sum_{i=1}^{11} x_i^2 = 5\ 503\ 675,$$

$$\sum_{i=1}^{11} y_i^2 = 600\ 664, \sum_{i=1}^{11} x_i y_i = 1\ 698\ 535.$$

i. Une relation linéaire entre les deux variables semble-t-elle plausible ? Construire le nuage de points et justifier.

ii. Déterminer la droite de régression de Y en X obtenue par la méthode des moindres carrés. Représenter cette droite sur le nuage de points et commenter.

iii. Calculer le pourcentage de la variance de Y expliquée par la régression linéaire et commenter.

iv. Si, en Belgique, le nombre de cigarettes consommées par habitant en 1930 s'élève à 530, quel est le taux de mortalité due au cancer du poumon attendu en 1950 ?

v. Les États-Unis et la Grande-Bretagne se distinguent quelque peu des autres observations. Construire deux nouvelles séries de données S_{EU} et S_{GB} contenant les observations initiales sauf respectivement les États-Unis ou la Grande-Bretagne. Calculer, pour chaque série, les coefficients de la droite de régression de Y en X obtenue par la méthode des moindres carrés.

vi. Calculer le coefficient de détermination de chaque série et commenter.

vii. Comparer le résidu de la Grande-Bretagne calculé à partir de ces nouvelles droites avec le résidu obtenu pour la série complète.

Solution :

(a) Afin de construire la boîte à moustaches demandée, on peut commencer par construire la courbe cumulative des fréquences cumulées, donnée à gauche de la Figure 3. À partir de cette courbe cumulative, on retrouve les valeurs des trois quartiles :

$$Q_1 = 115\ ; \qquad \tilde{y} = 170\ ; \qquad Q_3 = 250.$$

Afin de déterminer les valeurs adjacentes, il faut d'abord déterminer l'écart interquartile et les valeurs pivots :

$$\begin{aligned} EIQ &= Q_3 - Q_1 = 135\ ; \\ a_1 &= Q_1 - 1.5 \cdot EIQ = 115 - 1.5 \cdot 135 = -87.5\ ; \\ a_2 &= Q_3 + 1.5 \cdot EIQ = 250 + 1.5 \cdot 135 = 452.5. \end{aligned}$$

Les valeurs adjacentes sont donc données par

$$y_{(g)} = 58 \; ; \qquad y_{(d)} = 350,$$

puisqu'il s'agit de la plus petite observation supérieure ou égale à a_1 et de la plus grande observation inférieure ou égale à a_2. On a donc une valeur extrême : 465.
Au final, la boîte à moustaches est représentée à droite de la Figure 3. On y observe une dissymétrie à gauche (étalement sur la droite) avec même une valeur extrême.

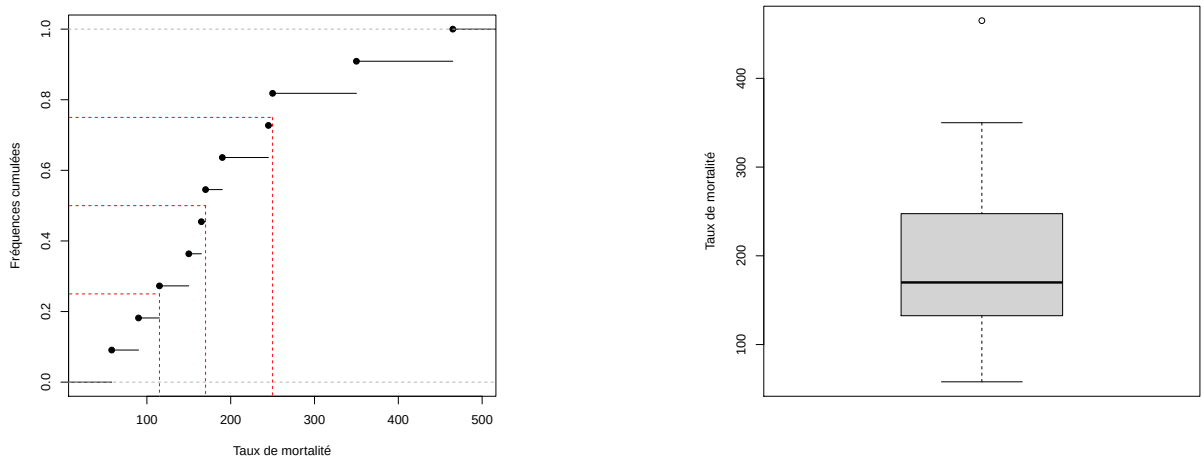

FIGURE 3 – Courbe cumulative des fréquences et boite à moustaches pour le taux de mortalité

(b) i. Le nuage de points est représenté sur la Figure 4. Ce graphique indique qu'une relation linéaire est plausible.

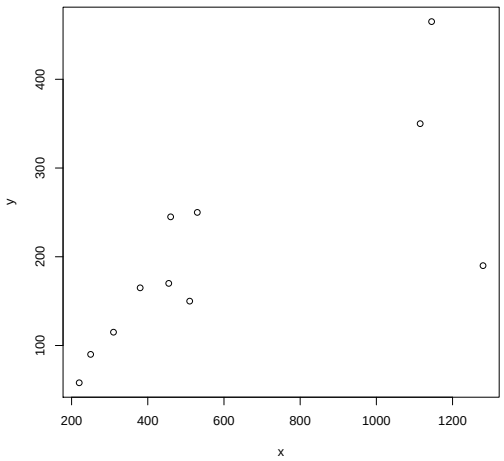

FIGURE 4 – Nuage de points de Y en fonction de X

ii. Les quantités suivantes

$$\bar{x} = \frac{6655}{11} = 605 \; ; \quad \bar{y} = \frac{2248}{11} = 204.38 \; ;$$
$$s_x^2 = \frac{5503675}{11} - 605^2 = 134309.1 \; ;$$
$$s_y^2 = \frac{600664}{11} - 204.36^2 = 12842.81 \; ;$$
$$s_{xy} = \frac{1698535}{11} - 605 \cdot 204.36 = 30774.5$$

permettent de calculer les paramètres de la droite de régression :

$$\hat{a} = \frac{30774.5}{134309.1} = 0.229$$
$$\hat{b} = 204.36 - 0.229 \cdot 605 = 65.815.$$

Ainsi, la droite de régression est donnée par $y = 0.229x + 65.815$. Elle est ajoutée sur le nuage de point à la Figure 5.

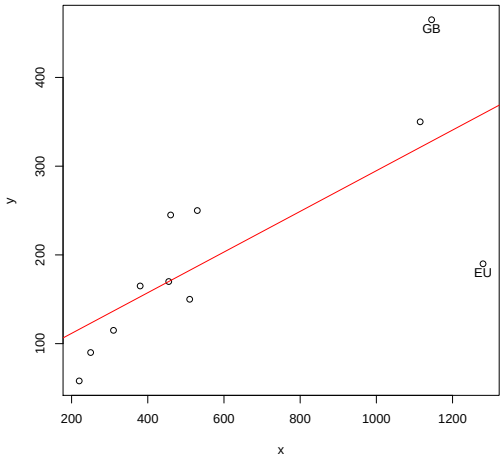

FIGURE 5 – Nuage de points de Y en fonction de X, avec la droite de régression des moindres carrés

iii. Le pourcentage de la variance de y expliquée par la régression linéaire est donné par le coefficient de détermination, i.e. le carré du coefficient de corrélation. On a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{30774.5}{\sqrt{134309.1 \cdot 12842.81}} = 0.741$$

et donc,

$$R^2 = 0.741^2 = 0.549.$$

Ainsi, seuls 55% de la variance de y sont expliqués par la régression. Ce faible pourcentage est notamment dû aux observations mises en évidence à la Figure 5, qui ne suivent pas vraiment la tendance linéaire générale.

iv. La valeur attendue est donnée par $\widehat{y} = 0.229 \cdot 530 + 65.815 = 187.185$.

v. On a, pour la série S_{EU},

$$\sum_i x_i = 6655 - 1280 = 5375 \ ;$$
$$\sum_i x_i^2 = 5503675 - 1280^2 = 3865275 \ ;$$
$$\sum_i y_i = 2248 - 190 = 2058 \ ;$$
$$\sum_i y_i^2 = 600664 - 190^2 = 564564 \ ;$$
$$\sum_i x_i y_i = 1698535 - 1280 \cdot 190 = 1455335$$

dont on déduit

$$\bar{x}_{EU} = \frac{5375}{10} = 537.5 \; ; \quad s^2_{x,EU} = \frac{3865275}{10} - 537.5^2 = 97621.25 \; ;$$
$$\bar{y}_{EU} = \frac{2058}{10} = 205.8 \; ; \quad s^2_{y,EU} = \frac{564564}{10} - 205.8^2 = 14102.76 \; ;$$
$$s_{xy,EU} = \frac{1455335}{10} - 537.5 \cdot 205.8 = 34916.$$

On obtient alors

$$a = \frac{34916}{97621.25} = 0.358$$
$$b = 205.8 - 0.358 \cdot 537.5 = 13.375$$

et la droite de régression est donnée par $y = 0.358x + 13.375$. Elle est ajoutée sur le nuage de point à la Figure 6.

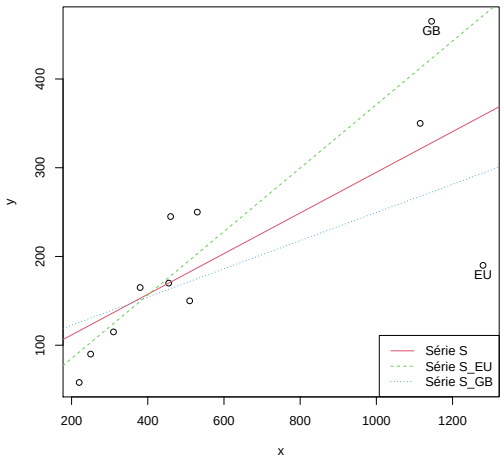

FIGURE 6 – Régression des moindres carrés pour les séries S, S_{EU} et S_{GB}

On remarque que la pente a augmenté car, sans l'observation des États-Unis, la droite est attirée par l'observation Grande-Bretagne.

Pour la série S_{GB}, on a

$$\sum_i x_i = 6655 - 1145 = 5510 \; ;$$
$$\sum_i x_i^2 = 5503675 - 1145^2 = 4192650 \; ;$$
$$\sum_i y_i = 2248 - 465 = 1783 \; ;$$
$$\sum_i y_i^2 = 600664 - 465^2 = 384439 \; ;$$
$$\sum_i x_i y_i = 1698535 - 1145 \cdot 465 = 1166110$$

dont on déduit

$$\bar{x}_{GB} = \frac{5510}{10} = 551 \; ; \quad s^2_{x,GB} = \frac{4192650}{10} - 551^2 = 115664 \; ;$$
$$\bar{y}_{GB} = \frac{1783}{10} = 178.3 \; ; \quad s^2_{y,GB} = \frac{384439}{10} - 178.3^2 = 6653.01 \; ;$$
$$s_{xy,GB} = \frac{1166110}{10} - 551 \cdot 178.3 = 18367.7.$$

On obtient alors

$$a = \frac{18367.7}{115664} = 0.159$$
$$b = 178.3 - 0.159 \cdot 551 = 90.691$$

et la droite de régression est donnée par $y = 0.159x + 90.691$. Elle est ajoutée sur le nuage de point à la Figure 6.
Cette fois-ci, la pente diminue vu l'attraction de l'observation des États-Unis.

vi. Les coefficients de détermination sont donnés par

$$R^2_{EU} = \frac{s^2_{xy,EU}}{s^2_{x,EU} s^2_{y,EU}} = \frac{34916^2}{97621.25 \cdot 14102.76} = 0.886 \; ;$$
$$R^2_{GB} = \frac{s^2_{xy,GB}}{s^2_{x,GB} s^2_{y,GB}} = \frac{18367.7^2}{115664 \cdot 6653.01} = 0.438.$$

On en conclut que, lorsque l'on retire l'observation concernant les États-Unis, le modèle est meilleur car la Grande-Bretagne se comporte plus comme le reste des autres pays.

vii. À partir de la première droite, on a

$$\hat{y}_{GB} = 0.229 \cdot 1145 + 65.815 = 328.02$$
$$\delta_{GB} = 465 - 328.02 = 136.98.$$

À partir de la droite estimée à partir de la série S_{EU}, on a

$$\hat{y}_{GB} = 0.358 \cdot 1145 + 13.375 = 423.285$$
$$\delta_{GB} = 465 - 423.285 = 41.715.$$

Enfin, à partir de la droite estimée à partir de la série S_{GB}, on a

$$\hat{y}_{GB} = 0.159 \cdot 1145 + 90.691 = 272.746$$
$$\delta_{GB} = 465 - 272.746 = 192.254.$$

Le résidu de la Grande-Bretagne est donc le plus petit (i.e. on a une estimation plus proche de la réalité) lorsque l'on ne considère pas les États-Unis, et le plus grand lorsque l'on ne considère pas la Grande-Bretagne.