Statistique descriptive
Solutions des exercices

Chapitre 4 : Séries statistiques bivariées

1. (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 $f_{54} = 0.024$.

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

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

— $\bar{y} = 1.89$;
— $\tilde{y} = 2$;
— $s_y^2 = 1.13$;
— $y_M = 2$.

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

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

2. (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 $10 \times 35 + ... + 25 \times 29 = 2680$.

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

(d) $\bar{x}_{|y=10} = \dfrac{1}{35}(11 \times 50 + 14 \times 150 + ... + 2 \times 450) = 167.14$;
$\bar{x}_{|y=15} = 181.37$;
$\bar{x}_{|y=20} = 233.33$;
$\bar{x}_{|y=25} = 267.24$.

3. (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 $\bar{x} = 27.24$ et $\bar{y} = 24.12$

(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 $12/38 = 31.6\%$.

4. (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 $\bar{x} = \dfrac{1}{100}\left(18 \times 2 + \ldots + 38 \times 5\right) = 3.82$.

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

$$s_x^2 = \frac{1}{100}\left(18 \times 2^2 + \ldots + 38 \times 5^2\right) - 3.82^2 = 1.27.$$

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 :

$$\bar{y} = \frac{1}{100}\left(13 \times 1 + \ldots + 9 \times 5\right) = 2.95$$
$$s_y^2 = \frac{1}{100}\left(13 \times 1^2 + \ldots + 9 \times 5^2\right) - 2.95^2 = 1.33.$$

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

$$\bar{y}_{|x=2} = \frac{1}{18}\left(9 \times 1 + 5 \times 2 + \ldots + 0 \times 5\right) = 1.78$$
$$\bar{y}_{|x=5} = \frac{1}{38}\left(0 \times 1 + 5 \times 2 + \ldots + 5 \times 5\right) = 3.5.$$

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

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

Le coefficient de corrélation est quant à lui donné par

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

5. (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

$$\bar{y} = \frac{1}{500}\left(300 \times 5000 + 125 \times 10\,000 + 75 \times 30\,000\right) = 10\,000.$$

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

$$\bar{y}_{|x_1} = \frac{1}{20}\left(13 \times 5000 + 5 \times 10\ 000 + 2 \times 30\ 000\right) = 8750$$
$$\bar{y}_{|x_2} = \frac{1}{100}\left(60 \times 5000 + 25 \times 10\ 000 + 15 \times 30\ 000\right) = 10\ 000$$
$$\bar{y}_{|x_2} = \frac{1}{150}\left(90 \times 5000 + 30 \times 10\ 000 + 30 \times 30\ 000\right) = 11\ 000$$
$$\bar{y}_{|x_4} = \frac{1}{230}\left(137 \times 5000 + 65 \times 10\ 000 + 28 \times 30\ 000\right) = 9456$$

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

(c) On a

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

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

6. Cet exercice peut être réalisé à la main ou directement à l'aide du logiciel `R`. Le code utilisé pour la réalisation avec le logiciel sera donné dans un fichier séparé. (Notez que des différences entre les valeurs données par le logiciel et les valeurs calculées à la main peuvent apparaître. Celles-ci sont dûes aux arrondis dans les quantités utilisées pour arriver aux résultats finaux.)

(a) Voir Figure 13.

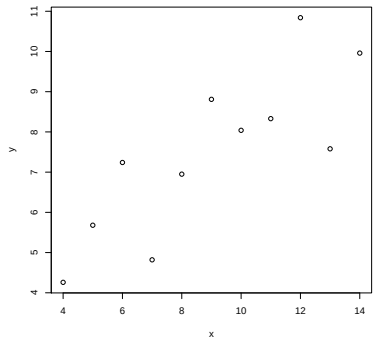

FIGURE 13 – Exercice 4.6(a)

(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

$\bar{x} = 9$; $\bar{y} = 7.5$;

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

$s_{xy} = \dfrac{797.6}{11} - 9 \times 7.5 = 5.$

De là, on a

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

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

7. Cet exercice peut être réalisé à la main ou directement à l'aide du logiciel `R`. Le code utilisé pour la réalisation avec le logiciel sera donné dans un fichier séparé. (Notez que des différences entre les valeurs données par le logiciel et les valeurs calculées à la main peuvent apparaître. Celles-ci sont dûes aux arrondis dans les quantités utilisées pour arriver aux résultats finaux.)
   (a) Après avoir représenté le diagramme de dispersion de la Figure 14, 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.
   (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

   $\bar{x} = 81$; $\bar{y} = 7.34$;
   $s_x^2 = 14$; $s_y^2 = 5.18$;
   $s_{xy} = \dfrac{7837.3}{13} - 81 \times 7.34 = 8.33.$

   De là, on a

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

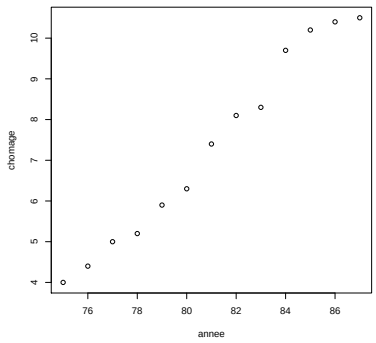

FIGURE 14 – Exercice 4.7(a)

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

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

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

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

8. Cet exercice peut être réalisé à la main ou directement à l'aide du logiciel `R`. Le code utilisé pour la réalisation avec le logiciel sera donné dans un fichier séparé. (Notez que des différences entre les valeurs données par le logiciel et les valeurs calculées à la main peuvent apparaître. Celles-ci sont dûes aux arrondis dans les quantités utilisées pour arriver aux résultats finaux.)

   (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 15. 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 :

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

   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 15. On y observe une dissymétrie à gauche (étalement sur la droite) avec même une valeur extrême.

   (b) i. Le nuage de points est représenté sur la Figure 16. Ce graphique indique qu'une relation linéaire est plausible.

   ii. Les quantités suivantes

   $$\bar{x} = \frac{6655}{11} = 605\ ; \quad \bar{y} = \frac{2248}{11} = 204.38\ ;$$

   $$s_x^2 = \frac{5503675}{11} - 605^2 = 134309.1\ ;$$

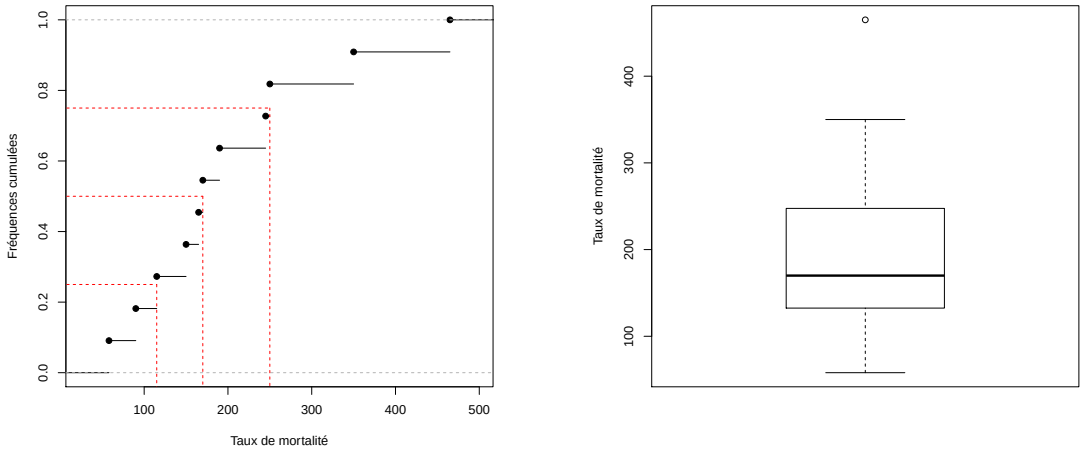

FIGURE 15 – Exercice 4.8(a)

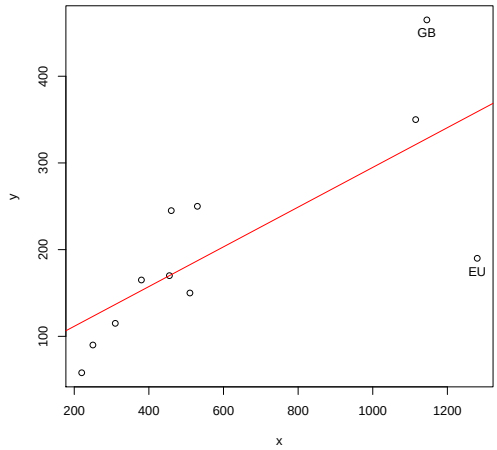

FIGURE 16 – Exercice 4.8(b)

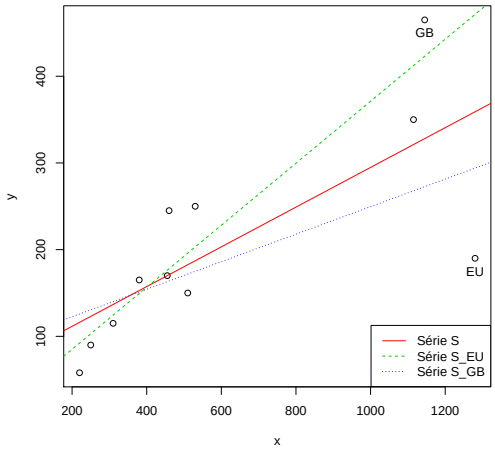

FIGURE 17 – Exercice 4.8(b)

$$s_y^2 = \frac{600664}{11} - 204.36^2 = 12842.81 \ ;$$
$$s_{xy} = \frac{1698535}{11} - 605 \times 204.36 = 30774.5$$

permettent de calculer les paramètres de la droite de régression :

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

Ainsi, la droite de régression est donnée par $y = 0.229x + 65.815$. Elle est ajoutée sur le nuage de point à la Figure 16.

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

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

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 misees en évidence à la Figure 16, qui ne suivent pas vraiment la tendance linéaire générale.

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

v. On a, pour la série S_{EU},

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

dont on déduit

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

On obtient alors

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

et la droite de régression est donnée par $y = 0.358x + 13.375$. Elle est ajoutée sur le nuage de point à la Figure 17.

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

Pour la série S_{GB}, on a

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

dont on déduit

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

On obtient alors

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

et la droite de régression est donnée par $y = 0.159x + 90.691$. Elle est ajoutée sur le nuage de point à la Figure 17.
Cette fois-ci, la pente diminue vu l'attraction de l'observation des Etats-Unis.

vi. Les coefficients de détermination sont donnés par

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

On en conclut que, lorsque l'on retire l'observation concernant les Etats-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

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

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

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

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

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

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 Etats-Unis, et le plus grand lorsque l'on ne considère pas la Grande-Bretagne.