Devoir Maison 9 - Régression linéaire - Correction

1 Préliminaires

Exercice 1 : Dans un repère orthonormé (O;I;J) du plan, la droite $\mathcal{D}$ est la représentation graphique de la fonction affine $f : x \mapsto 0,4x - 2,9$

a) Montrer que le point $A(9; 0,7)$ appartient à $\mathcal{D}$.

b) Le point B(2 015; 800) appartient-il à $\mathcal{D}$?

c) Peut-on trouver une valeur du réel t pour laquelle le point C(4;t) appartient à la droite $\mathcal{D}$? Si oui, peut-on en trouver un autre ?

d) Peut-on trouver une valeur du réel u pour laquelle le point D(u;4) appartient à la droite $\mathcal{D}$? Si oui, peut-on en trouver un autre ?

Proof. a) f(9)=0,7 donc $A \in \mathcal{D}$.

b) $f(2015 \neq 800$ donc B n'appartient pas à $\mathcal{D}$.

c)

$$\begin{aligned} C(4;t) \in \mathcal{D} &\iff f(4) = t && (1)\\ &\iff 0,4 \times 4 - 2,9 = t && (2)\\ &\iff 1,6 - 2,9 = t && (3)\\ &\iff t = -1,3 && (4) \end{aligned}$$

Donc t=-1,3 est la seule possibilité pour laquelle C(4;t) soit sur $\mathcal{D}$.

d)

$$\begin{aligned} D(u;4) \in \mathcal{D} &\iff f(u) = 4 && (5)\\ &\iff 0,4u - 2,9 = 4 && (6)\\ &\iff 0,4u = 6,9 && (7)\\ &\iff u = \frac{6,9}{0,4} = \frac{69}{4} && (8) \end{aligned}$$

Donc $u = \frac{69}{4}$ est la seule possibilité pour que D(u;4) soit sur $\mathcal{D}$. □

Exercice 2 : Soit $\mathcal{D}$ la droite d'équation $y = ax + b$. A quelle condition le point $M(x_M; y_M)$ appartient-il à $\mathcal{D}$?

Proof.

$$M(x_M; y_M) \in \mathcal{D} \iff y_M = f(x_M) \iff y_M = ax_M + b \quad (9)$$

□

Exercice 3 : Soit $M(x_M; y_M)$ et $N(x_N; y_N)$, avec $x_M \neq x_N$. On note $y = ax + b$ l'équation de la droite (MN). L'objectif de cette question est de trouver a et b, connaissant les coordonnées des deux points M et N.

a) Montrer que a et b sont solutions du système d'équation suivant :

$$\begin{aligned} y_M &= ax_M + b && (10)\\ y_N &= ax_N + b && (11) \end{aligned}$$

b) Résoudre le système. Montrer que l'on a $a = \frac{y_M - y_N}{x_M - x_N}$ et $b = \frac{y_N x_M - y_M x_N}{x_M - x_N}$. Laquelle de ces deux valeurs représente le coefficient directeur de (MN) ? Était-ce prévisible sans calculs ? Dans quel cas $b = 0$?

c) Application : soient M(1,2) et N(3;4). Calculer l'équation de la droite (MN).

c) (*) Pourquoi a-t-on supposé $x_M \neq x_N$? Que se passe-t-il si $x_M = x_N$, et quelle est l'équation correspondante ?

Proof. a) L'équation de la droite (MN) est $y = ax + b$. On sait que $M(x_M; y_M)$ est sur la droite (MN), donc ses coordonnées vérifient

$$y_M = ax_M + b \quad (12)$$

De même, $N(x_N; y_N)$ est sur (MN), donc ses coordonnées vérifient :

$$y_N = ax_N + b \quad (13)$$

Donc les deux équations 12 et 13 donnent le système que l'on voulait obtenir.

b) On donne la solution, donc une manière de procéder est de vérifier que les valeurs a et b données par l'énoncé conviennent.

Une autre manière est de résoudre le système.

En retranchant 12 à 13, on obtient :

$$y_M - y_N = ax_M + b - (ax_N + b) \tag{14}$$
$$\iff y_M - y_N = a(x_M - x_N) \tag{15}$$
$$\iff a = \frac{y_M - y_N}{x_M - x_N} \tag{16}$$

On remarque que a a la forme du taux de variation entre les points M et N.

Pour obtenir b, on remplace dans une des deux équations du début (par exemple 10) la nouvelle valeur de a que l'on vient de trouver.

Ainsi :

$$y_M = \frac{y_M - y_N}{x_M - x_N}x_M + b \tag{17}$$
$$\iff b = y_M - \frac{y_M - y_N}{x_M - x_N}x_M \tag{18}$$
$$\iff b = \frac{y_M(x_M - x_N) - (y_M - y_N)x_M}{x_M - x_N} \tag{19}$$
$$\iff b = \frac{y_M x_M - y_M x_N - y_M x_M + y_N x_M}{x_M - x_N} \tag{20}$$
$$\iff b = \frac{y_N x_M - y_M x_N}{x_M - x_N} \tag{21}$$

On a $b = 0$ si la droite est linéaire, c'est à dire passe par O(0;0).

c) Dans cet application numérique, on trouve $a = \frac{2-4}{1-3} = \frac{-2}{-2} = 1$ et $b = \frac{4\times1-2\times3}{1-3} = \frac{4-6}{-2} = 1$, donc la droite (MN) a pour équation y=x+1.

d) On a supposé que $x_M \neq x_N$, car sinon les points M et N auraient même abscisse. Dans ce cas, la droite (MN) serait verticale, et n'est pas représentable par une fonction (car pour un même x, il existerait plusieurs images (une infinité même!), ce qui n'est pas possible avec notre définition d'une fonction). □

2 Position du problème

On dispose de 3 points $M_1(x_1; y_1)$, $M_2(x_2; y_2)$ et $M_3(x_3; y_3)$. On veut tracer et trouver l'équation d'une droite $\mathcal{D}$ qui passe par ces trois points.

Question 2.1. *A quelle condition (géométrique) sur M_1, M_2 et M_3 la droite $\mathcal{D}$ va-t-elle passer exactement par ces trois points ? Traduire cette condition avec des vecteurs.*

Proof. $\mathcal{D}$ passe exactement par les 3 points si les points sont alignés, ou encore si les vecteurs $\overrightarrow{M_1M_2}$ et $\overrightarrow{M_2M_3}$ sont colinéaires. □

En général, $\mathcal{D}$ ne passera pas par les trois points. On va donc faire en sorte pour qu'elle passe à peu près par les trois points. On appelle $y = ax + b$ l'équation de la droite $\mathcal{D}$.

Question 2.2. *A quelle condition a-t-on $y_1 = ax_1 + b$?*

Proof. $y_1 = ax_1 + b \iff M_1 \in \mathcal{D}$. □

En général, $y_1 \neq ax_1 + b$; on introduit e_1 tel que $y_1 = ax_1 + b + e_1$; e_1 est l'erreur que l'on fait lorsque l'on dit que les 3 points forment une droite (si $e_1 = 0$, alors M_1 est bien sur la droite $\mathcal{D}$; sinon on fait une erreur en modélisant les trois points M_1, M_2 et M_3 par une droite).

De même, on introduit e_2 et e_3 tel que $y_2 = ax_2 + b + e_2$ et $y_2 = ax_2 + b + e_2$.

3 Résolution du problème

Définition 1. *La droite $\mathcal{D}$ qui passe "le mieux" est la droite des moindres carrés, c'est à dire celle qui minimise l'expression :*

$$S = e_1^2 + e_2^2 + e_3^2 \tag{22}$$

On **admet** que S est minimal si et seulement si les deux équations suivantes sont vérifiées :

$$x_1e_1 + x_2e_2 + x_3e_3 = 0 \tag{23}$$
$$e_1 + e_2 + e_3 = 0 \tag{24}$$

Dans la suite, on cherche les valeurs de a et de b pour lesquelles S est minimales (c'est à dire pour lesquelles les deux équations précédentes sont vérifiées).

Question 3.1. *(*) Montrer que l'on a* $b = \frac{1}{3}(y_1 + y_2 + y_3) - a\frac{1}{3}(x_1 + x_2 + x_3)$.
On note pour alléger les expressions $\bar{x} = \frac{1}{3}(x_1 + x_2 + x_3)$ *et* $\bar{y} = \frac{1}{3}(y_1 + y_2 + y_3)$. *Que représentent ces deux expressions ?*

Proof. On part de l'équation 24, que l'on réécrit en remplaçant les e_i par leur valeur ($e_i = y_i - ax_i - b$). Cela donne :

$$
\begin{aligned}
(24)\quad e_1 + e_2 + e_3 = 0 &\iff y_1 - ax_1 - b + y_2 - ax_2 - b + y_3 - ax_3 - b = 0 && (25)\\
&\iff y_1 + y_2 + y_3 - a(x_1 + x_2 + x_3) - 3b = 0 && (26)\\
&\iff b = \frac{1}{3}(y_1 + y_2 + y_3) - a\frac{1}{3}(x_1 + x_2 + x_3) && (27)\\
&\iff b = \bar{y} - a\bar{x} && (28)
\end{aligned}
$$

Cette formule pour b est finalement assez simple (et on peut la généraliser facilement pour n points). On peut l'interpréter en disant que la droite des moindres carré passe par le point moyen de coordonnée $(\bar{x}; \bar{y})$.

□

Question 3.2. *Montrer que l'équation* $y = ax + b$ *devient* $y - \bar{y} = a(x - \bar{x})$.

Proof. Comme $b = \bar{y} - a\bar{x}$, l'équation $y = ax + b$ devient $y = ax + \bar{y} - a\bar{x}$, c'est à dire $y - \bar{y} = a(x - \bar{x})$. □

Question 3.3. *(**) Montrer que* $a = \frac{(x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+(x_3-\bar{x})(y_3-\bar{y})}{(x_1-\bar{x})^2+(x_3-\bar{x})^2+(x_3-\bar{x})^2}$

Proof. Assez calculatoire, en partie du au fait que vous n'avez pas les bonnes notations à votre disposition pour simplifier les expressions.

On réécrit les deux équations 23 et 24 en remplaçant b par sa valeur. Puis on joue avec les équations (on multiplie 24 par $\bar{x}$ puis on retranche 23) pour obtenir la valeur de a.

$$
\begin{aligned}
&\begin{cases} x_1e_1 + x_2e_2 + x_3e_3 = 0 \\ e_1 + e_2 + e_3 = 0 \end{cases} && (29)\\
\iff &\begin{cases} x_1(y_1 - ax_1 - \bar{y} + a\bar{x}) + x_2(y_2 - ax_2 - \bar{y} + a\bar{x}) + x_3(y_3 - ax_3 - \bar{y} + a\bar{x}) = 0 \\ y_1 - ax_1 - b + y_2 - ax_2 - b + y_3 - ax_3 - b = 0 \end{cases} && (30)\\
\iff &\begin{cases} x_1(y_1 - \bar{y}) + x_2(y_2 - \bar{y}) + x_3(y_3 - \bar{y}) - a\Big(x_1(x_1 - \bar{x}) + x_2(x_2 - \bar{x}) + x_3(x_3 - \bar{x})\Big) = 0 \\ y_1 + y_2 + y_3 - a(x_1 + x_2 + x_3) - 3(\bar{y} - a\bar{x}) = 0 \end{cases} && (31)\\
\iff &\begin{cases} x_1(y_1 - \bar{y}) + x_2(y_2 - \bar{y}) + x_3(y_3 - \bar{y}) - a\Big(x_1(x_1 - \bar{x}) + x_2(x_2 - \bar{x}) + x_3(x_3 - \bar{x})\Big) = 0 \\ \Big((y_1 - \bar{y}) - a(x_1 - \bar{x})\Big) + \Big((y_2 - \bar{y}) - a(x_2 - \bar{x})\Big) + \Big((y_3 - \bar{y}) - a(x_3 - \bar{x})\Big) = 0 \end{cases} && (32)\\
\iff &\begin{cases} x_1(y_1 - \bar{y}) + x_2(y_2 - \bar{y}) + x_3(y_3 - \bar{y}) - a\Big(x_1(x_1 - \bar{x}) + x_2(x_2 - \bar{x}) + x_3(x_3 - \bar{x})\Big) = 0 \\ \Big((y_1 - \bar{y}) + (y_2 - \bar{y}) + (y_3 - \bar{y})\Big) - a\Big((x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x})\Big) = 0 \end{cases} && (33)\\
\iff &\begin{cases} x_1(y_1 - \bar{y}) + x_2(y_2 - \bar{y}) + x_3(y_3 - \bar{y}) - a\Big(x_1(x_1 - \bar{x}) + x_2(x_2 - \bar{x}) + x_3(x_3 - \bar{x})\Big) = 0 \\ (y_1 - \bar{y})\bar{x} + (y_2 - \bar{y})\bar{x} + (y_3 - \bar{y})\bar{x} - a\Big((x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x})\Big)\bar{x} = 0 \end{cases} && (34)\\
& && (35)
\end{aligned}
$$

On a presque fini : on soustrait membre à membre les deux dernières équations. Cela donne :

$$
\begin{aligned}
&(y_1 - \bar{y})(x_1 - \bar{x}) + (y_2 - \bar{y})(x_2 - \bar{x}) + (y_3 - \bar{y})(x_3 - \bar{x}) - a\Big((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\Big)\bar{x} - = 0 && (36)\\
\iff &a = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y})}{(x_1 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_3 - \bar{x})^2} && (37)
\end{aligned}
$$

On trouve bien le résultat de l'énoncé (ouf !).

□

Exemple Soient $M_1(0; 0)$ $M_2(1; 2)$ et $M_3(2; 3, 9)$. Calculer l'équation de la droite $\mathcal{D}$ qui passe le "mieux" par les trois points. Vérifier à la calculatrice si vous trouvez la même chose (à priori oui). Vérifiez parmi les trois points M_1, M_2 et M_3 lesquelles appartiennent à la droite $\mathcal{D}$. Cela pose-t-il problème si aucun des points n'appartient à la droite ?

Proof. Il était possible de traiter cet exemple sans avoir fait les questions précédentes. □

4 Analyse du résultat

L'objectif de cette partie est de trouver un critère simple pour pouvoir dire si la droite $\mathcal{D}$ passe presque par les points M_1, M_2 et M_3 ou non.

Question 4.1. *En partant de $(y_1 - \bar{y}) = a(x_1 - \bar{x}) + e_1$, montrer en mettant au carré les termes de droite et de gauche, que l'on a :*

$$(y_1 - \bar{y})^2 = a^2(x_1 - \bar{x})^2 + e_1^2 \tag{38}$$

On justifiera proprement que le double produit s'annule bien.

Proof. On fait ce que dit l'énoncé, et on met au carré l'expression.

$$\Big(y_1 - \bar{y}\Big)^2 = \Big(a(x_1 - \bar{x}) + e_1\Big)^2 \tag{39}$$

$$\iff \Big(y_1 - \bar{y}\Big)^2 = a^2(x_1 - \bar{x})^2 + e_1^2 + 2a(x_1 - \bar{x})e_1 \tag{40}$$

$$\iff \Big(y_1 - \bar{y}\Big)^2 = a^2(x_1 - \bar{x})^2 + e_1^2 + 2a(x_1 - \bar{x})(y_1 - ax_1 - b) \tag{41}$$

$$\iff \Big(y_1 - \bar{y}\Big)^2 = a^2(x_1 - \bar{x})^2 + e_1^2 + 2a(x_1 - \bar{x})(y_1 - ax_1 - \bar{y} - a\bar{x}) \tag{42}$$

$$\tag{43}$$

On veut faire annuler le double produit (le dernier terme de l'équation 42). On remarque que celui-ci est un produit de terme; il sera nul si un de ces terme est nul.

Dans $2a(x_1 - \bar{x})(y_1 - ax_1 - \bar{y} - a\bar{x})$, a n'est à priori pas nul (sauf gros cas particulier), $(x_1 - \bar{x})$ non plus. Il reste $(y_1 - ax_1 - \bar{y} - a\bar{x})$. Ce terme peut encore s'écrire comme $(y_1 - \bar{y} - a(x_1 - \bar{x}))$. On reconnaît l'équation de la question 3.2 ! □

Remarque Il y avait une erreur dans l'énoncé. Le double produit en général ne s'annule pas. Il faut en fait sommer sur les trois points pour avoir un terme qui s'annule. En effet, on obtiendrait le terme :

$$2a\Big((x_1 - \bar{x})(y_1 - ax_1 - \bar{y} - a\bar{x}) + (x_2 - \bar{x})(y_2 - ax_2 - \bar{y} - a\bar{x}) + (x_3 - \bar{x})(y_3 - ax_3 - \bar{y} - a\bar{x})\Big) \tag{44}$$

On reconnaît le terme de l'équation 36.

Donc dans ce cas, l'équation que l'on cherchait est :

$$(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 = a^2\Big((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2\Big) + e_1^2 + e_2^2 + e_3^2 = 0 \tag{45}$$

Question 4.2. *Écrire une équation similaire pour les points M_2 et M_3.*

Proof. Voir la remarque précédente □

Définition 2. *On définit les trois quantités suivantes :*

$$\textit{Total sum of squares} \quad SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 \tag{46}$$

$$\textit{Regression sum of squares} \quad SSR = a^2(x_1 - \bar{x})^2 + a^2(x_3 - \bar{x})^2 + a^2(x_3 - \bar{x})^2 \tag{47}$$

$$\textit{Error sum of squares} \quad SSE = e_1^2 + e_2^2 + e_3^2 \tag{48}$$

Question 4.3. *Montrer que SST = SSR + SSE.*

Proof. C'est l'équation 45 écrite de manière condensée. □

Définition 3. *On appelle le coefficient de régression linéaire (ou coefficient de corrélation), la quantité :*

$$r^2 = \frac{SSR}{SST} \tag{49}$$

Question 4.4. *(*) Quelles sont les valeurs pouvant être prises par r (ou r^2) ? Montrer que le cas $r^2 = 1$ correspond à SSE=0. Dans ce cas, que peut-on dire géométriquement sur les trois points ? De manière générale, à quelle condition la droite passe-t-elle presque par les trois points ?*

Proof. r^2 est compris entre 0 et 1, donc r est compris entre -1 et 1.

Le cas $r^2 = 1$ correspond à une corrélation parfaite, c'est à dire aux trois points M_1, M_2 et M_3 alignés sur la droite des moindres carré (qui passe donc par ces trois points). Plus r^2 s'éloigne de 1, plus la droite est une mauvaise approximation.

En pratique, lorsque l'on fait une régression linéaire, on doit avoir r^2 de l'ordre de 0,9999 (au moins trois ou quatre 9); en dessous c'est pas terrible, et soit la mesure expérimentale des points a mal été faite, soit les points ne correspondent pas à une droite, et il faut chercher autre chose. □

5 Généralisation

Question 5.1. *(**) Généraliser les résultats pour n points $M_1(x_1; y_1); \ldots; M_n(x_n; y_n)$.*