Devoir Maison 9 - Régression linéaire

A rendre pour le 19 Janvier

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 ?

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 $\mathcal{D}$?

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 :

$$y_M = ax_M + b \tag{1}$$
$$y_N = ax_N + b \tag{2}$$

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 ?

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.*

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 $y_1 = ax_1 + b$?*

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 :*

$$S = e_1^2 + e_2^2 + e_3^2 \tag{3}$$

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

$$x_1e_1 + x_2e_2 + x_3e_3 = 0 \tag{4}$$
$$e_1 + e_2 + e_3 = 0 \tag{5}$$

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 ?

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

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

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 ?

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 :*

$$(y_1 - \bar{y})^2 = a^2(x_1 - \bar{x})^2 + e_1^2 \tag{6}$$

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

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

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

$$\textit{Total sum of squares} \quad SST = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 \tag{7}$$

$$\textit{Regression sum of squares} \quad SSR = a^2(x_1 - \bar{x})^2 + a^2(x_3 - \bar{x})^2 + a^2(x_3 - \bar{x})^2 \tag{8}$$

$$\textit{Error sum of squares} \quad SSE = e_1^2 + e_2^2 + e_3^2 \tag{9}$$

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

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

$$r^2 = \frac{SSR}{SST} \tag{10}$$

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 ?*

5 Généralisation

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