

Régression linéaire

1. Nuage de points

Exemple :

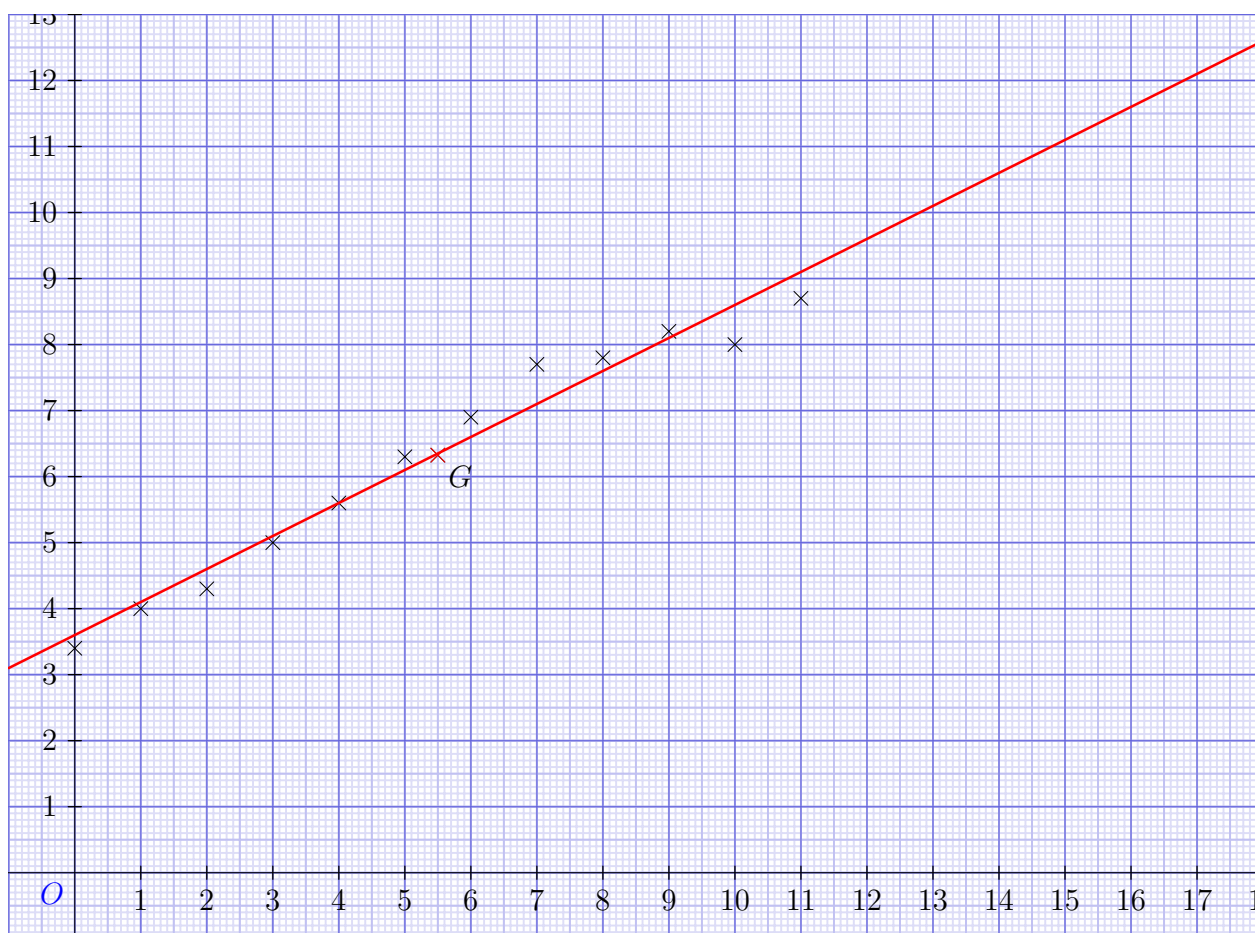
On donne ci-dessous l'évolution de la part de soins et de biens médicaux dans le PIB de la France de 1960 à 2004.

Remarque : La consommation de soins et de biens médicaux (CSBM) comprend les soins hospitaliers, les soins ambulatoires (médecins, dentistes, auxiliaires médicaux, laboratoires d'analyse, thermalisme), les transports sanitaires, les médicaments et les autres biens médicaux (optique, prothèses, petit matériel et pansements). Seules les dépenses qui concourent au traitement d'une perturbation provisoire de l'état de santé sont prises en compte. Ainsi, les dépenses de soins aux personnes handicapées et aux personnes âgées en institution sont exclues.

Année	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000	2004
Rang de l'année : x	0	1	2	3	4	5	6	7	8	9	10	11
Part dans le PIB (%) : y	3,4	4	4,3	5	5,6	6,3	6,9	7,7	7,8	8,2	8	8,7

Représenter sur le graphique ci-dessous le nuage de points associé à cette série.

Placer le point moyen G du nuage qui est par définition le point de coordonnée $(\bar{x} ; \bar{y}) = (5,5 ; 6,325)$.



■ DÉFINITION :

Soit $(x_i ; y_i)_{1 \leq i \leq N}$ une série statistique double de N couples.

Dans un repère orthogonal $(O ; I, J)$, l'ensemble des points A_i de coordonnées $(x_i ; y_i)_{1 \leq i \leq N}$ s'appelle le **nuage de points** associé à cette série double.

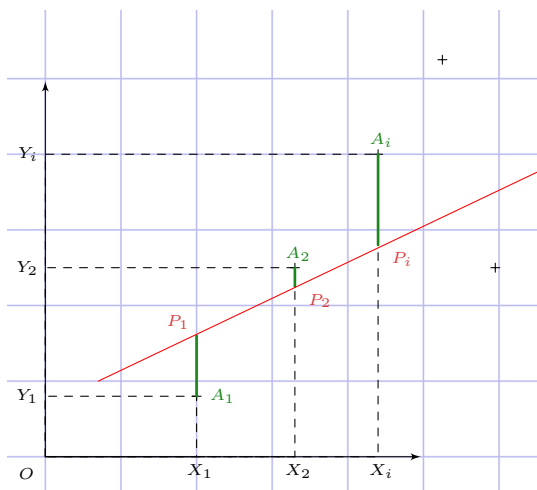
Le point G de coordonnées $(\bar{x} ; \bar{y})$ s'appelle le **point moyen du nuage**.

2. Modélisation : ajustement affine

Lorsque, comme dans notre exemple, le nuage de points présente une « forme allongée », on peut envisager de l'ajuster (c'est-à-dire de le modéliser) par une droite, dans le but de faire des prévisions (qui seront des estimations) : on dit alors qu'on a procédé à un **ajustement affine du nuage**, c'est-à-dire par une droite.

On peut procéder de plusieurs manières pour obtenir une équation de droite de la forme $y = ax + b$.

On va s'intéresser ici à la **méthode dite des « moindres carrés »**.



On considère une droite (d) ajustant un nuage de points $(A_i)_{1 \leq i \leq N}$, et on appelle P_1 le point de la droite (d) dont l'abscisse est la même que celle de A_1 , c'est-à-dire x_1 , puis de même les points $P_2, \dots, P_N$.

On calcule alors le nombre S défini par :

$$S = A_1 P_1^2 + A_2 P_2^2 + \dots + A_N P_N^2$$

- Quel est le signe du nombre S ?

Le nombre S est positif car c'est une somme de carrés

- Si $S = 0$, que peut-on dire du nuage de points ?

Si $S = 0$, on peut dire que chaque point du nuage se trouve exactement sur la droite

Le nombre S mesure, d'une certaine manière (celle des moindres carrés), la distance entre une droite ajustant le nuage et le nuage lui-même : au minimum, lorsque les points du nuage sont parfaitement alignés, $S = 0$; plus la droite est « éloignée du nuage », plus S est grand.

La théorie mathématique montre que parmi toutes les droites (d) ajustant le nuage, il en existe une et une seule qui minimise le nombre S (bien sûr, si les points du nuage ne sont pas alignés, cette valeur minimale est strictement positive) : c'est cette droite qu'on appelle **droite de régression de y en x par la méthode des moindres carrés**.

3. Droite de régression de y en x par la méthode des moindres carrés

A. Notion de covariance

■ DÉFINITION :

Soit $(x_i ; y_i)_{1 \leq i \leq N}$ une série statistique double de N couples.

La **covariance** de cette série est le nombre noté $\text{Cov}(x; y)$ ou σ_{xy} et définie par :

$$\text{Cov}(x; y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

Remarque : Lorsque les séries x et y sont les mêmes,

$$\begin{aligned} \text{Cov}(x ; y) &= \text{Cov}(x ; x) \\ &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x}) \\ &= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \\ &= V(x) \end{aligned}$$

On retrouve la variance de la série x .

Remarque : Pour faciliter le calcul de la covariance, on peut aussi utiliser une extension de la formule de König-Huygens :

$$\text{Cov}(x ; y) = \overline{x \times y} - \bar{x} \times \bar{y}$$

Exemple :

Calculer manuellement la covariance $\text{Cov}(x ; y)$ des séries x et y de l'exemple de la page 1, puis la variance $V(x)$ de la série x , et enfin le nombre

$$a = \frac{\text{Cov}(x ; y)}{V(x)}$$

On sait que $\bar{x} = 5,5$, $\bar{y} = 6,325$ et $\overline{x \times y} = 40,725$.
Ainsi

$$\text{Cov}(x ; y) = 40,725 - 5,5 \times 6,325 = 5,9375$$

Puis,

$$V(x) = \overline{x^2} - \bar{x}^2 = \frac{506}{12} - (5,5)^2 \approx 11,917$$

D'où

$$a \approx \frac{5,9375}{11,917} \approx 0,498$$

B. Déterminer une équation de la droite de régression

■ PROPRIÉTÉ :

Soient $(x_i ; y_i)_{1 \leq i \leq N}$ une série statistique double de N couples et son nuage de points $A_i(x_i ; y_i)_{1 \leq i \leq N}$ associé dans un repère orthogonal $(O ; I, J)$.

La droite de régression linéaire de y en x (par la méthode des moindres carrés associée) à ce nuage est la droite (d) d'équation cartésienne réduite $y = ax + b$, de coefficient directeur $a = \frac{\text{Cov}(x ; y)}{V(x)}$ et qui passe par le point moyen G .

Donc (d) a pour équation : $y - \bar{y} = a(x - \bar{x})$.

Exemple :

Donner une équation de la droite de régression de y en x par la méthode des moindres carrés associée au nuage de la page 1 (on arrondira a et l'ordonnée à l'origine au dixième).

On a calculé a dans l'exemple précédent : $a \approx 0,498$ donc $a \approx 0,5$ car on nous demande d'arrondir au dixième.

On obtient $y - 6,325 = 0,5(x - 5,5)$

c'est-à-dire $y = 0,5(x - 5,5) + 6,325$

ou encore $y = 0,5x - 2,75 + 6,325$

c'est-à-dire $y = 0,5 + 3,575$

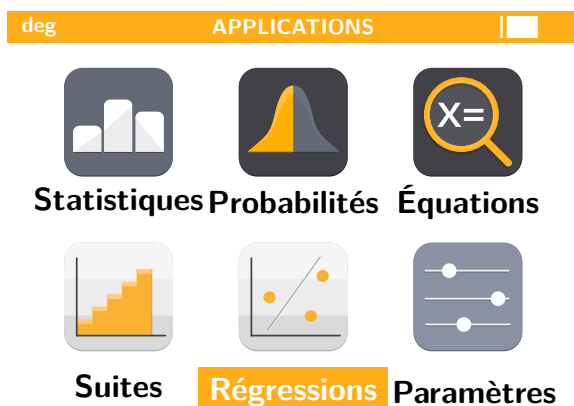
L'énoncé demande d'arrondir 3,575 au dixième, ainsi la l'équation réduite de la droite de régression attendue est

$$y = 0,5 + 3,6$$

Construire sur le graphique la droite de régression de y en x .

Remarque : En pratique, au baccalauréat franco-allemand, aucun détail de calcul n'est exigé pour les ajustements affines. Il faut donc savoir extraire l'équation demandée de la calculatrice.

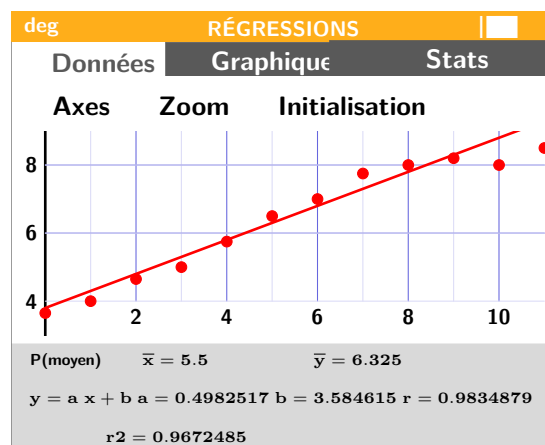
Tout d'abord, choisir le menu Régressions.



Ensuite rentrer les Données dans les colonnes X1 pour x et Y1 pour y .

deg	REGRESSIONS		
Données	Graphique	Stats	
X1	Y1	X2	
0	3.4		
1	4		
2	4.3		
3	5		
4	5.6		
5	6.3		
6	6.9		
7	7.7		

Choisir ensuite l'onglet Graphique : on retrouve le nuage de points avec le point moyen G et ses coordonnées calculées en bas de l'écran. Une équation approchée de la droite de régression est donnée et elle est construite sur le graphique. Comparer avec les résultats calculés à la main.



Exemple :

En supposant que l'évolution de la part de soins et de bien médicaux dans le PIB de la France se poursuive selon ce modèle :

- Quelle devrait être sa valeur en 2012? Elle devrait être environ $y = 0,5 \times 12 + 3,6 = 9,6$
- À partir de quelle année devrait-elle dépasser 12 % du PIB? On cherche x tel que $y > 12$. On résout donc $0,5x + 3,6 > 12$, c'est-à-dire $0,5x > 8,4$ ou encore $x > 16,8$. Donc la part du PIB devrait dépasser les 12 % à partir de la dix-septième année.
- En réalité, en 2012, sa valeur a été de 8,8 % du PIB. Que penser de l'estimation? L'estimation est trop loin de la réalité.

Le modèle n'est donc pas pertinent.

C. Pertinence d'un ajustement affine

Le coefficient r fourni par la calculatrice en même temps que l'équation de la droite de régression de y en x s'appelle le coefficient de corrélation linéaire entre les séries x et y .

■ DÉFINITION :

Le **coefficient de corrélation linéaire r entre les séries x et y** est le nombre défini par :

$$r = \frac{\text{Cov}(x ; y)}{\sigma_x \times \sigma_y}$$

■ PROPRIÉTÉ :

Le nombre r vérifie : $-1 \leq r \leq 1$.

La corrélation linéaire entre les séries x et y est très forte si $r \approx -1$ ou $r \approx 1$.

■ DÉFINITION :

L'ajustement affine est considéré comme **raisonnable** si : $r \geq 0,87$ ou $r \leq -0,87$.

Remarque : Au baccalauréat, il n'est pas demandé de savoir calculer le coefficient de corrélation linéaire.

Exemple :

Calculer le coefficient de corrélation linéaire r des séries x et y .

On a $V(x) \approx 11,917$, $V(y) \approx 3,059$ et $\text{Cov}(x ; y) = 5,9375$

Donc

$$r \approx \frac{5,9375}{\sqrt{11,917} \times \sqrt{3,059}} \approx 0,983$$

Comparer avec la valeur donnée par la calculatrice.

La calculatrice donne une valeur très proche : $r \approx 0,9834879$

La valeur de r trouvée justifie-t-elle (*a posteriori*) l'ajustement affine effectué ?

$r \approx 1$, donc la corrélation linéaire entre les séries x et y est très forte.

Remarque : Le terme de droite de « régression » vient d'une étude historique de Francis Galton (1822-1911) entre la taille y d'un individu et celle de son père x : le coefficient directeur a qu'il a trouvé est positif mais inférieur à 1, c'est-à-dire que la taille des individus est en général inférieure à celle de leur père dans cette étude : d'où régression de la taille.

Remarque : Il ne faut pas confondre corrélation et causalité : *Cum hoc ergo propter hoc* (phrase latine signifiant *avec ceci, donc à cause de ceci*) est un sophisme qui consiste à prétendre que si deux événements sont corrélés alors il y a un lien de cause à effet entre les deux.

Exemple :

Il y a une corrélation entre la peinture et le niveau en mathématiques chez les collégiens (un troisième paramètre, l'âge des adolescents, est la cause commune des deux événements).

Exemple :

Il existe une corrélation entre la vente de crème glacée et le nombre de morts par noyade.

Exemple :

Une étude anglaise a prouvé que les gens habitant près de pylônes à haute tension étaient significativement plus souvent malades que le reste de la population. Est-ce la faute du courant électrique ? Ce n'est pas évident parce qu'une autre étude a révélé que les habitants sous les pylônes étaient en moyenne plus pauvres ; et on connaît les liens santé-pauvreté... À elle seule, cette étude ne permet pas de conclure.

Exemple :

Le réchauffement climatique est une conséquence directe du déclin de la population de pirates en s'appuyant sur une corrélation inverse entre la population de pirates et la température moyenne sur Terre.

Exemple :

Le fait de dormir avec des chaussures est corrélé au fait de se réveiller avec un mal de tête. Peut-on en conclure que dormir avec des chaussures fait mal à la tête ? Une explication plus vraisemblable est que ces deux événements font suite à des soirées trop arrosées...

D. Droite de régression de x en y

C'est une droite « proche » de la droite de régression de y en x puisque ces deux droites ajustent le même nuage mais en général, elles sont distinctes. Pour obtenir une équation de la droite de régression de x en y , il suffit d'inverser les rôles de x et y .

Revenir dans l'onglet « Données » du menu Régressions et sélectionner la colonne X2.

deg	REGRESSIONS		
	Données	Graphique	Stats
	X1	Y1	X2
	0	3.4	
	1	4	
	2	4.3	
	3	5	
	4	5.6	
	5	6.3	
	6	6.9	
	7	7.7	

Cliquer. Il apparaît :

deg	RÉGRESSIONS		
	Données	Graphique	Stats
	Options de la colonne		
	Effacer la colonne		
	Remplir avec une formule		
	Régression $a * X + t$		

Sélectionner « Remplir avec une formule ».
Puis remplir la colonne X2 avec les données de la colonne Y1.



Attention

Il faut respecter la casse : Y doit être tapé en majuscule (donc shift + alpha + y)

deg	REGRESSIONS		
	Données	Graphique	Stats
	X1	Y1	X2
	0	3.4	3.4
	1	4	4
	2	4.3	4.3
	3	5	5
	4	5.6	5.6
	5	6.3	6.3
	6	6.9	6.9
	7	7.7	7.7

Exemple :

Trouver ainsi une équation de la droite de régression de x en y pour l'exemple de la page 1.

Grâce à la calculatrice, on obtient : à finir

On peut retrouver ce résultat par le calcul « en inversant les rôles de x et y dans les calculs » :

On sait que $\bar{x} = 5,5$, $\bar{y} = 6,325$ et $\overline{x \times y} = 40,725$.

Ainsi

$$\text{Cov}(x ; y) = 5,9375$$

Puis,

$$V(y) = \overline{y^2} - \bar{y}^2 = \frac{516,77}{12} - 5,5^2 \approx 6,325^2 \approx 3,059$$

D'où

$$a \approx \frac{5,9375}{3,059} \approx 1,94$$

Ainsi l'équation de la droite de régression de x en y est (attention ! x et y dans l'équation ne désignent pas ceux de la série statistique !):

$$y - \bar{y} = a(x - \bar{x})$$

$$y - 5,5 = 1,94(x - 6,325)$$

$$y = 1,94(x - 6,325) + 5,5$$

$$y = 1,94x - 12,2705 + 5,5$$

$$y = 1,94x - 6,7705$$



En allemand...

- La régression linéaire : die lineare Regression
- Un nuage de points : eine Punktwolke
- Le point moyen : Mittelwert
- La covariance : Kovarianz
- La droite de régression : die Regressionsgerade
- Le coefficient de corrélation linéaire : der lineare Korrelationskoeffizient
- Ajustement affine : das Trendmodell

