2. Ajustements

2.1. Un peu d'histoire

Adrien-Marie **Legendre**
(1752 - 1833)

De Legendre, on ne connaît que cette caricature. On sait depuis peu que le portrait « habituel » ci-contre est celui d'un autre Legendre.

Le problème de l'ajustement d'un ensemble de points représentés dans un système d'axes par une droite, ou plus généralement par une courbe, est essentiel dans le développement de la statistique.
Au 18ème siècle, Leonhard **Euler** et Tobias **Mayer** développent, indépendamment l'un de l'autre, la méthode des moyennes permettant d'ajuster des points par une droite.
Le premier texte paru faisant mention de la méthode des moindres carrés est dû à Adrien-Marie **Legendre** dans un article sur ses « nouvelles méthodes pour la détermination des orbites des comètes », publié en 1805. Un an plus tard, **Gauss** fait aussi allusion à cette méthode. C'est avec l'apparition de la loi normale que cette méthode va trouver sa justification et va devenir pour longtemps la méthode d'ajustement.
La paternité de la corrélation a donné lieu à une littérature abondante. Signalons simplement que **Galton** exprime le désir de construire un coefficient de réversion qui se mutera en régression et qu'en 1888 il utilise les termes de « partial co-relation » annonçant déjà la corrélation multiple. En 1896, Karl **Pearson** reprend les concepts de **Galton** pour leur donner leur forme actuelle. Au 20ème siècle, d'autres mesures d'association allaient naître comme, en 1904, le coefficient de corrélation de rang avec **Spearman** et la même année la statistique « classique » du chi-deux par **Pearson**.

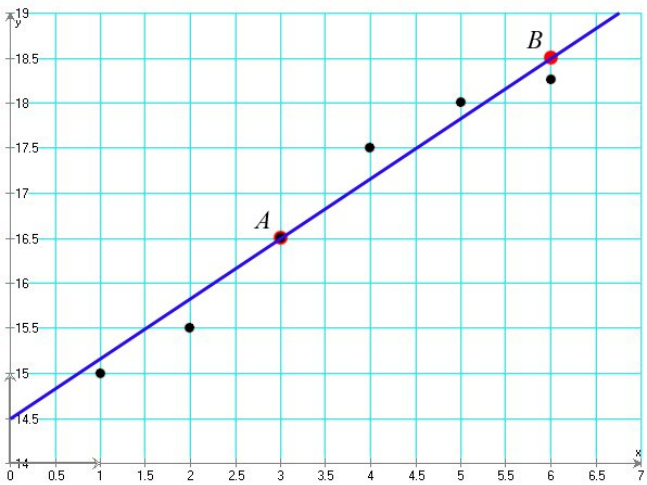
2.2. Ajustement affine graphique

Soient les n points du nuage représentant, dans un repère cartésien, la série des n valeurs (x_i, y_i) des variables x et y. Ajuster une droite d à ce nuage de points consiste à remplacer chaque point (x_i, y_i) par un point de même abscisse et d'ordonnée $\hat{y}_i$, les points $(x_i, \hat{y}_i)$ étant alignés sur la droite d.

Une fois l'équation de la droite d déterminée, on pourra l'utiliser pour faire des *interpolations* (calculs de valeurs intermédiaires) et des *extrapolations* (calculs de valeurs futures).

La méthode graphique consiste à tracer, **à l'œil**, à l'aide d'une règle **transparente**, une droite $y = m \cdot x + h$ s'ajustant le mieux possible sur le nuage de points.

Les points noirs représentent les données.
Les points rouges A et B sont les points choisis pour tracer la droite. Ils peuvent être choisis parmi les points noirs (A) ou pas (B).



Équation de la droite d'ajustement

Une fois la droite tracée, on choisit sur le dessin deux points A et B quelconques de la droite pour en déterminer l'équation. Ces points ne doivent pas obligatoirement faire partie du nuage de points.

Cette méthode est couramment employée, en raison de sa rapidité et de sa simplicité. Elle est empirique, mais donne de très bons résultats.

empirique : basé sur l'expérience

Rappel : L'équation de la droite passant par les points $A(x_A, y_A)$ et $B(x_B, y_B)$ est donnée par : $y - y_B = \frac{y_B - y_A}{x_B - x_A}(x - x_B)$

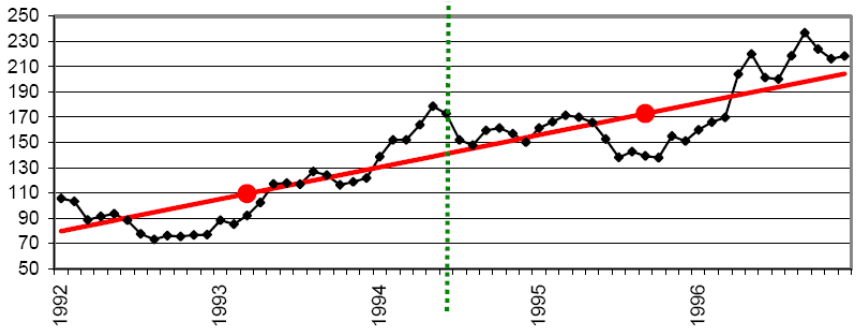
Les points A et B choisis dans notre exemple ont comme coordonnées (3, 16.5) et (6, 18.5). La droite passant par ces deux points est :

$$y - 18.5 = \frac{18.5 - 16.5}{6 - 3}(x - 6)$$

On obtient après simplification : $y = \frac{2}{3}x + 14.5$.

2.3. Droite de Mayer

Méthode On découpe le nuage de points en deux sous-ensembles de même effectif. Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et la moyenne des y_i. On obtient ainsi deux points $(\bar{x}_1, \bar{y}_1)$ et $(\bar{x}_2, \bar{y}_2)$, appelés **points moyens**. Il reste à tracer la droite passant par ces deux points.



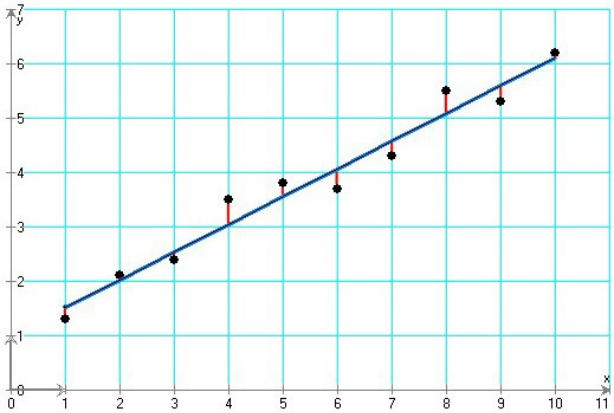
L'équation de cette droite s'obtient de la même façon que pour un ajustement affine graphique.

2.4. Ajustement analytique par la méthode des moindres carrés

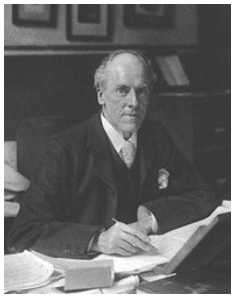
$\hat{y}_i$ est la coordonnée verticale du point **de la droite** d'abscisse x_i. Donc $\hat{y}_i = a x_i + b$.

L'ajustement linéaire par la méthode des moindres carrés consiste à déterminer la droite (que l'on appelle aussi **droite de régression**) telle que la somme des carrés des n valeurs $y_i - \hat{y}_i$ soit minimale (ce qui explique le nom de la méthode).

Sur le dessin, chaque trait vertical rouge représente la valeur $y_i - \hat{y}_i$



On veut donc minimiser la quantité $q=\sum (y_i-(a\,x_i+b))^2$.



Karl **Pearson**
(1857 - 1936)

Rappelons que la valeur minimale d'une fonction se calcule en posant sa dérivée égale à 0. Pour trouver *a* et *b*, calculons cette dérivée.

Calculons d'abord la dérivée de *q* par rapport à *a*.

$$\frac{dq}{da}=-2\sum ((y_i-a\,x_i-b)\,x_i)=0$$

$$\sum x_i\,y_i=a\sum x_i^2+b\sum x_i \qquad (1)$$

Calculons maintenant la dérivée de *q* par rapport à *b*.

$$\frac{dq}{db}=-2\sum (y_i-a\,x_i-b)=0$$

$$\sum y_i=\sum a\,x_i+\sum b$$

$$\sum y_i=\sum a\,x_i+n\,b \qquad \text{Divisons le tout par } n.$$

$$\bar{y}=a\,\bar{x}+b$$

$$b=\bar{y}-a\,\bar{x} \qquad (2)$$

Ce résultat indique que la droite passe par le point moyen $(\bar{x};\bar{y})$.
Introduisons le résultat de (2) dans (1) pour trouver *a* :

$$\sum x_i\,y_i=a\sum x_i^2+(\bar{y}-a\,\bar{x})\sum x_i$$

$$\sum x_i\,y_i=a\sum x_i^2+\bar{y}\sum x_i-a\,\bar{x}\sum x_i$$

$$a\sum x_i^2-a\,\bar{x}\sum x_i=\sum x_i\,y_i-\bar{y}\sum x_i$$

$$a=\frac{\sum x_i\,y_i-\bar{y}\sum x_i}{\sum x_i^2-\bar{x}\sum x_i}$$

$$a=\frac{\sum x_i\,y_i-n\,\bar{x}\,\bar{y}}{\sum x_i^2-n\,\bar{x}^2}=\frac{\frac{1}{n}\sum x_i\,y_i-\bar{x}\,\bar{y}}{\frac{1}{n}\sum x_i^2-\bar{x}^2}=\frac{\sigma_{xy}}{\sigma_x^{\,2}}$$

La droite des moindres carrés $y = ax + b$ a pour coefficients :

$$a=\frac{\frac{1}{n}\sum x_i\,y_i-\bar{x}\,\bar{y}}{\frac{1}{n}\sum x_i^2-\bar{x}^2} \qquad \text{et} \qquad b=\bar{y}-a\,\bar{x}$$

Remarque Certaines calculatrices ont des fonctions statistiques qui fournissent ces valeurs très rapidement.
Consultez le mode d'emploi de votre machine !

Exercice 2.1

Lors d'une expérience, on a relevé les valeurs suivantes :

x	1	2	3	4	5	6	7	8	9	10
y	1.1	3.1	4.7	7.3	9.2	11.1	12.9	15.4	17	18.8

Donnez l'équation d'une droite ajustant ces valeurs

a. à l'œil ;
b. par la méthode de Mayer ;
c. par la méthode des moindres carrés.
d. Dessinez les droites obtenues en **b** et **c.**
e. Interpolez la valeur de $\hat{y}$ pour $x = 6.3$ grâce aux droites obtenues en **b** et **c.**

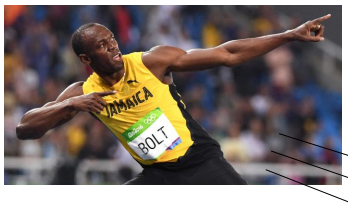
Exercice 2.2

Le tableau ci-dessous montre l'évolution des temps olympiques du 200 m plat, en secondes, pour les hommes et pour les femmes.

Vous remarquerez que les mesures au centième de seconde apparaissent en 1968 pour les hommes et en 1972 pour les femmes.

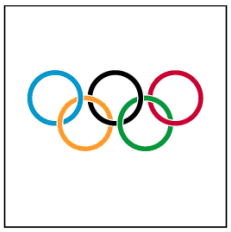
	200 m hommes	200 m femmes
Londres 1948	21.1	24.4
Helsinki 1952	20.7	23.7
Melbourne 1956	20.6	23.4
Rome 1960	20.5	24.0
Tokyo 1964	20.3	23.0
Mexico 1968	19.83	22.5
Munich 1972	20.00	22.40
Montréal 1976	20.23	22.37
Moscou 1980	20.19	22.03
Los Angeles 1984	19.80	21.81
Séoul 1988	19.75	21.34
Barcelone 1992	19.73	21.72
Atlanta 1996	19.32	22.12
Sydney 2000	20.09	21.84
Athènes 2004	19.79	22.05
Pékin 2008	19.30	21.74
Londres 2012	19.32	21.88
Rio de Janeiro 2016	(19.78)	(21.78)

Florence Griffith-Joyner



Usain Bolt

Comparez l'estimation et la réalité...



CITIUS • ALTIUS • FORTIUS

Donnez l'équation des droites (celle des performances des hommes et celle des femmes) ajustant ces valeurs, de 1948 à 2012,

a. à l'œil ;
b. par la méthode des moindres carrés.
c. Dessinez la droite obtenue en **b**.
d. Estimez les temps olympiques de 2016 et comparez-les avec les vrais résultats.
e. D'après la droite obtenue en **b**, quand les femmes courront-elles le 200 m plat aussi vite que les hommes ?
f. Ces ajustements affines sont-ils adéquats ?

2.4. Coefficient de corrélation linéaire

Définition On nomme **coefficient de corrélation linéaire** des variables x et y, le nombre réel :

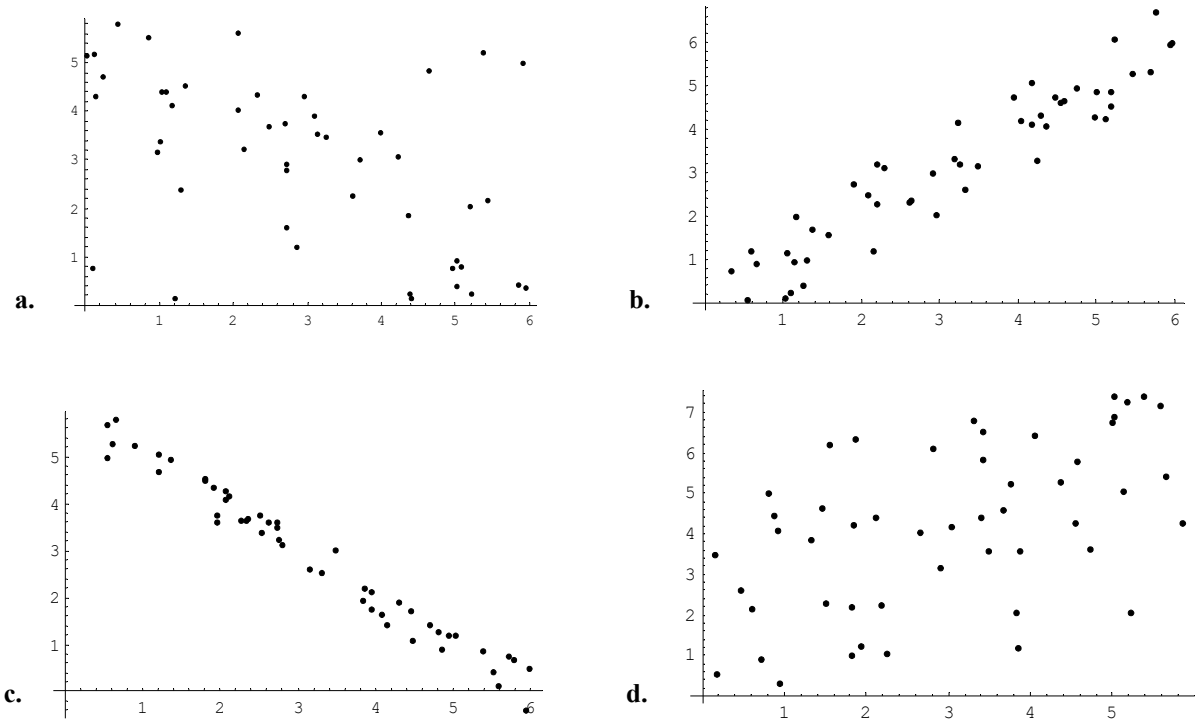
$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

avec $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$ $\quad \sigma_x = \sqrt{\frac{1}{n}\sum_{1}^{n} x_i^2 - \bar{x}^2}$ $\quad \sigma_y = \sqrt{\frac{1}{n}\sum_{1}^{n} y_i^2 - \bar{y}^2}$

Interprétation r est un nombre réel compris entre –1 et 1.
Quand | r | = 1, tous les points sont alignés.
Quand | r | est proche de 1, les variables x et y sont fortement corrélées.
Quand $r < 0$, la droite de régression a une pente négative.
Quand $r > 0$, la droite de régression a une pente positive.

Exercice 2.3

Rendez à chacun des nuages de points ci-dessous son coefficient de corrélation linéaire : –0.98, –0.50, 0.53, 0.94.



Plus d'exercices de ce type sur www.istics.net/Correlations/

Exercice 2.4

Les criquets ont un organe spécial sur leurs ailes avant qui produit un son lorsqu'ils frottent leurs ailes les unes contre les autres. En règle générale, plus la température de l'air est élevée, plus ils frottent leurs ailes rapidement. La relation entre la température et le nombre de pulsations par seconde est bien approchée par une droite de régression (chaque espèce a sa droite propre). On a relevé les mesures suivantes :

Température (°C) [x]	15°	17°	20°	21°	23°	24°	27°	28°	30°	32°	34°
# de pulsations par sec. [y]	13.5	14.1	14.5	14.4	16.3	15.5	17.1	17.8	18.2	20.2	20.1

a. Donnez la droite des moindres carrés ajustant ce nuage de points.
b. Calculez le coefficient de corrélation linéaire.
c. Si la température augmente de 3°C, de combien augmentera le nombre de pulsations ?

2.5. Ajustements non linéaires

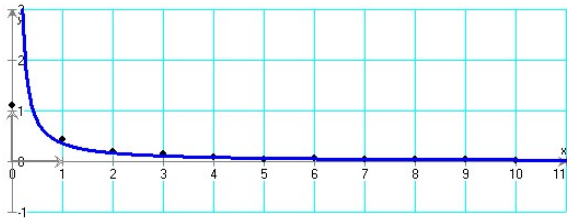
Dans certains cas, l'ajustement à une fonction linéaire n'est pas adéquat : un ajustement des données à une fonction non linéaire doit être envisagé.
Les cas que nous considérerons sont ceux où on peut se ramener par une simple transformation à un ajustement affine.

Ajustement par une hyperbole

droite de régression de z en x : z est la valeur à estimer et joue le rôle de l'ordonnée, x joue le rôle de l'abscisse.

Les points $(x_i\,;\,y_i)$ ne sont pas alignés, mais plutôt proches d'une certaine hyperbole de la forme $\hat{y}=\dfrac{1}{a\,x+b}$.

1. calculer $z_i=\dfrac{1}{y_i}$;
2. déterminer l'équation de la droite de régression de z en x avec la méthode des moindres carrés ;
3. de l'équation obtenue $z=a\,x+b$, on déduit immédiatement l'équation de l'hyperbole $\hat{y}=\dfrac{1}{a\,x+b}$.



Exercice 2.5

Ajustez ce nuage de points par une hyperbole $\hat{y}=\dfrac{1}{a\,x+b}$.

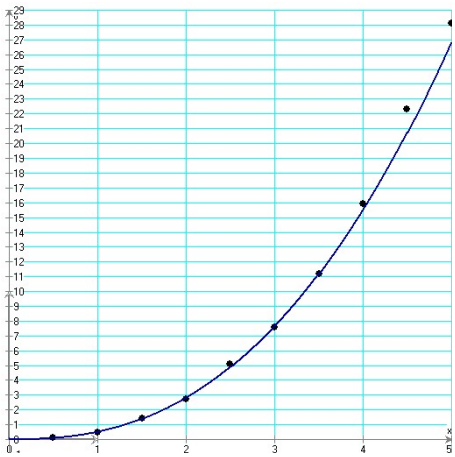
x	0	1	2	3	4	5	6	7	8	9	10
y	1.1	0.43	0.19	0.15	0.08	0.05	0.06	0.05	0.04	0.04	0.03

Ajustement par une fonction puissance

droite de régression de v en u : v est la valeur à estimer et joue le rôle de l'ordonnée, u joue le rôle de l'abscisse.

Les points $(x_i\,;\,y_i)$ sont proches d'une courbe de fonction puissance comme $\hat{y}=b\,x^a$. On remarque que $\ln(y) = a\ln(x) + \ln(b)$.

1. calculer $u_i=\ln(x_i)$ et $v_i=\ln(y_i)$;
2. déterminer l'équation de la droite de régression de v en u avec la méthode des moindres carrés ;
3. de l'équation obtenue $v = Au + B$, on déduit l'équation de la fonction puissance $\hat{y}=b\,x^a$, puisque $a = A$ et $b = e^B$.



Exercice 2.6

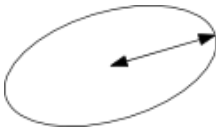
Ajustez ce nuage de points par une fonction puissance $\hat{y}=b\,x^a$.

x	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	0.1	0.5	1.4	2.7	5.1	7.6	11.2	15.9	22.3	28.1

Exercice 2.7

Dans le tableau ci-dessous, on a donné pour chaque planète du système solaire le demi-grand axe de l'orbite et la période de révolution. Le demi-grand axe est exprimé en unités astronomiques (UA). On appelle unité astronomique le demi-grand axe de l'orbite terrestre. Il vaut 149'600'000 km. La période de révolution est exprimée en années.

En géométrie, le **grand axe** d'une ellipse est un paramètre utilisé pour décrire la dimension de cette conique.
Le **demi-grand axe** est la moitié du grand axe.



Hale-Bopp a probablement été la comète la plus observée de l'histoire. Très brillante, elle a été visible 18 mois, avec son maximum d'éclat en 1997.

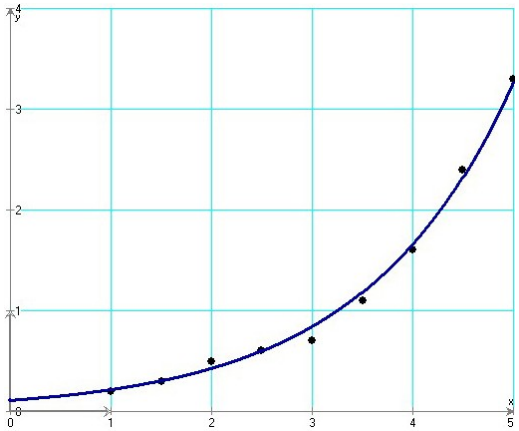
Planètes	Demi-grand axe (UA)	Période (années)
Mercure	0.38710	0.24084
Vénus	0.72333	0.61519
Terre	1	1
Mars	1.52369	1.88082
Jupiter	5.20280	11.8618
Saturne	9.53884	29.4567
Uranus	19.1819	84.0107
Neptune	30.0578	164.786

a. Trouvez la relation qui existe entre la période et le demi-grand axe à l'aide de la méthode des moindres carrés. Cette relation est la troisième loi de Kepler.
b. L'existence de cette relation est évidente lorsque l'on choisit une échelle logarithmique. Utilisez un papier millimétré log-log (http://customgraph.com) pour reporter les périodes en fonction du demi-grand axe.
c. Il y a un « trou » sur le graphe entre Mars et Jupiter. Comment peut-on l'expliquer ?
d. La comète de Hale-Bopp a une période de 2530 ans. Que vaut le demi-grand axe ?

Ajustement par une exponentielle

Les points $(x_i\,;\,y_i)$ sont proches d'une courbe d'une exponentielle de la forme $\hat{y}=b\,a^x$.
On remarque que $\ln(y) = x\ln(a) + \ln(b)$.

1. calculer $z_i=\ln(y_i)$;
2. déterminer l'équation de la droite de régression de z en x avec la méthode des moindres carrés ;
3. de l'équation obtenue $z = Ax + B$, on déduit l'équation de l'exponentielle $\hat{y}=b\,a^x$, puisque $a = e^A$ et $b = e^B$.



Exercice 2.8

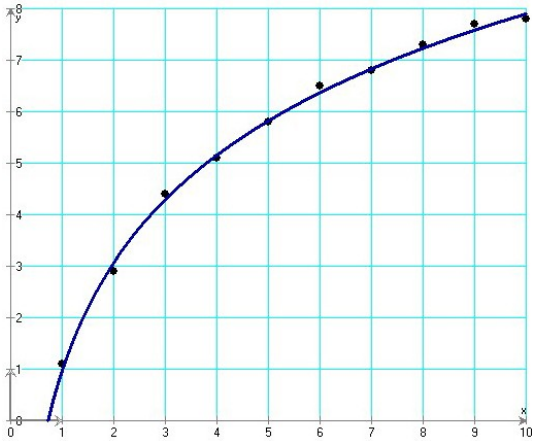
Ajustez ce nuage de points par une exponentielle de la forme $\hat{y}=b\,a^x$.

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
y	0.2	0.3	0.5	0.6	0.7	1.1	1.6	2.4	3.3

Ajustement par une fonction logarithmique

Les points $(x_i\,;y_i)$ sont proches d'une courbe logarithmique de la forme $\hat{y}=a\ln(x)+b$.

1. calculer $z_i=\ln(x_i)$;
2. déterminer l'équation de la droite de régression de y en z avec la méthode des moindres carrés ;
3. de l'équation obtenue $y = a\cdot z + b$, on déduit l'équation de la fonction logarithmique $\hat{y}=a\ln(x)+b$.



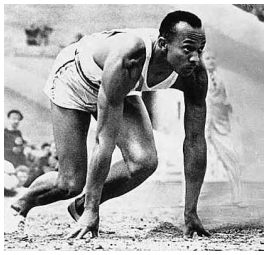
Exercice 2.9

Ajustez ce nuage de points par une fonction logarithmique $\hat{y}=a\ln(x)+b$.

x	1	2	3	4	5	6	7	8	9	10
y	1.1	2.9	4.4	5.1	5.8	6.5	6.8	7.3	7.7	7.8

Exercice 2.10

Étudions l'évolution des records de l'épreuve d'athlétisme du 100 mètres masculin. Pour cela, on cherche un ajustement des records pour en prévoir l'évolution. On donne dans le tableau suivant certains records, établis depuis 1900.



Jesse Owens aux J.O. de 1936 à Berlin

Année	1900	1912	1921	1930	1964	1983	1991	1999
Rang (x_i)	0	12	21	30	64	83	91	99
Temps en sec. (y_i)	10.80	10.60	10.40	10.30	10.06	9.93	9.86	9.79

1) Étude d'un modèle affine

a. Construisez le nuage de points $M(x_i\,;y_i)$, avec i compris entre 1 et 8, associé à cette série statistique double. Vous prendrez comme unité graphique 1 cm pour dix ans en abscisse et 1 cm pour un dixième de seconde en ordonnées.
On commencera les graduations au point de coordonnées (0; 9).

b. Peut-on envisager un ajustement affine à court terme ? Cet ajustement permet-il des prévisions pertinentes à long terme sur les records futurs ?

2) Étude d'un modèle exponentiel

Après étude, on choisit de modéliser la situation par une autre courbe. On effectue les changements de variables suivants : $X = e^{-0.00924x}$ et $Y = \ln(y)$. On obtient :

$X = e^{-0.00924x}$	1.000	0.895	0.824	0.758	0.554	0.464	0.431	0.401
$Y = \ln(y)$	2.380	2.361	2.342	2.332	2.309	2.296	2.288	2.281

a. Donnez une équation de la droite de régression de Y en X obtenue par la méthode des moindres carrés.

b. En déduire que l'on peut modéliser une expression de y en fonction de x sous la forme suivante :

$$y = \exp(a \cdot e^{-0.00924x} + b)$$ où a et b sont deux réels à déterminer.

c. À l'aide de cet ajustement, quel record du 100 mètres peut-on prévoir en 2010 ?

d. Calculez la limite en $+\infty$ de la fonction f définie sur $\mathbb{R}$ par l'expression suivante :

$$f(x) = \exp(0.154\, e^{-0.00924x} + 2.221)$$

e. Que peut-on en conclure, en utilisant ce modèle, quant aux records du cent mètres masculin, à très long terme ?

Note : $\exp(x) = e^x$

Record d'Usain Bolt en 2009 : 9.58 secondes

Exercice 2.11

Le 14 octobre 2012, l'autrichien Félix **Baumgartner** effectuait le plus spectaculaire des sauts en « chute libre », emmené par un ballon stratosphérique à l'altitude de 39'045 m. Ce plongeon dura au total 549 secondes. Un appareil a enregistré la vitesse verticale V_z en fonction du temps. La vitesse maximum de 373 m/s a été atteinte en t = 40 s. Le précédent record remontait à 1960.

Le tableau 1 contient les mesures faites pendant les 35 premières secondes de chute.

Tableau 1

T (s)	5	12	20	26	30	35
V_z (m/s)	49	115	195	254	287	340

Le tableau 2 contient les mesures réalisées entre 50 s et 260 s, une zone dans laquelle la force de traînée (la force aérodynamique de freinage) devient importante, jusqu'à compenser complètement le poids du sauteur.

Tableau 2

t (s)	50	70	100	130	180	230	260
$t' = t-50$ (s)	0	20	50	80	130	180	210
V_z (m/s)	352	254	158	102	69	51	51

a. Représentez la fonction $V_z(t)$, d'après les tableaux 1 et 2. En particulier, extrapolez la fonction entre les temps t = 35 s et t = 50 s.

b. Vérifiez que durant les 35 premières secondes la chute de F. Baumgartner est libre.

c. À l'aide du graphique du point **a**, estimez la distance parcourue par F. Baumgartner durant les 260 premières secondes du vol et sa vitesse limite V_{lim} avant l'ouverture du parachute.

La distance parcourue est l'aire sous la courbe de $V_z(t)$.

d. Représentez sur un graphique la fonction $f(t') = \ln(V_z(t') - V_{lim})$ pour $t' > 0$. Déduisez-en l'équation de la vitesse de F. Baumgartner en fonction de t'.

2.6. Corrélation ≠ causalité

Deux événements peuvent être corrélés sans pour autant avoir des rapports de cause à effet. Quelques exemples :

1. 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 sait les liens santé-pauvreté... À elle seule, cette étude ne permet pas de conclure.
2. Les assurances ont établi que 50 % des accidents arrivaient sur un trajet de moins de 30 km. On en a conclu – un peu vite – que l'habitude des courts trajets pour aller travailler favorisait le manque d'attention des conducteurs. Il

est possible que ce soit vrai, mais la « démonstration » est fausse : la plupart des trajets font moins de 30 km !

3. Le conseil de l'Ordre des médecins a publié une étude prouvant que ceux qui pratiquaient régulièrement le jogging à l'âge de 60 ans avaient une probabilité de se trouver en bonne santé à l'âge de 70 ans plus grande que la population normale. Conclusion de l'Ordre, le jogging est une bonne pratique. Il est encore possible que ce soit vrai, mais ce n'est pas une démonstration : la population qui pratique le jogging à 60 ans concentre ceux qui sont déjà en bonne santé. On a donc seulement prouvé que ceux qui sont en bonne santé à 60 ans ont plus de chance de l'être encore 10 ans plus tard.
4. Le très sérieux *New England Journal of Medecine* publiait en octobre 2012 un intéressant article du Dr F. Messerli. Partant de l'idée que les flavonoïdes, ces composés psychoactifs du chocolat, ont une excellente réputation d'anti-dépresseur et de stimulants intellectuels, ce bon docteur s'est posé la question d'un lien entre la consommation de chocolat et l'obtention de prix Nobel. Voici le graphique qu'il a obtenu :

On retrouve en tête la Suisse, avec 22 prix Nobel attribués pour 7 millions d'habitants et une consommation individuelle de 11.9 kg de chocolat par an, soit 30 grammes par jour. (chiffres de 2012).

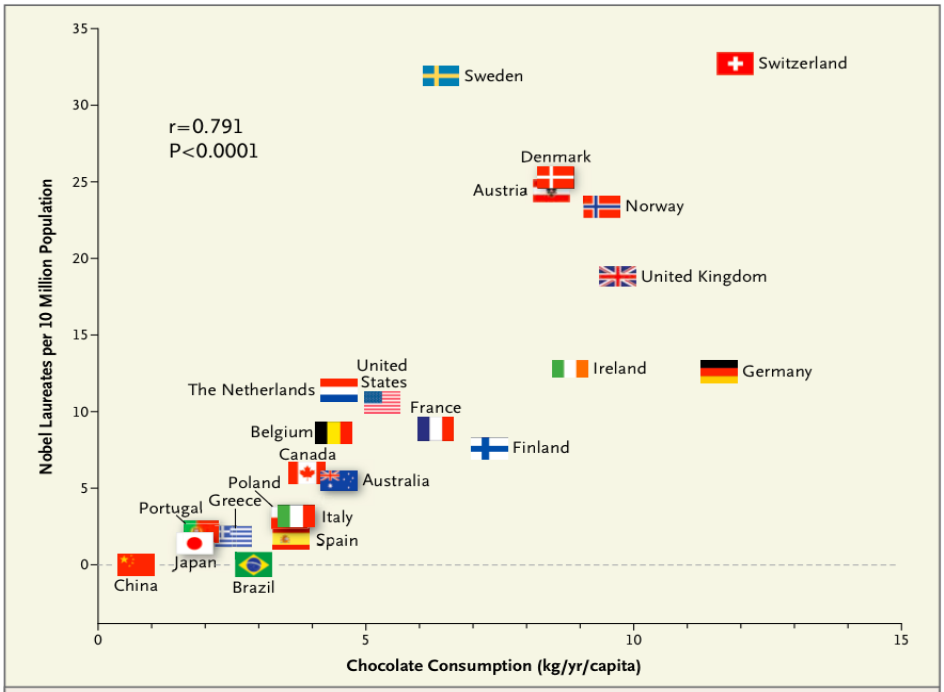


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

D'autres corrélations farfelues sur http://tylervigen.com/spurious-correlations.

2.7. Ce qu'il faut absolument savoir

Faire un ajustement affine graphique	❑ ok
Faire un ajustement affine par la méthode de Mayer	❑ ok
Faire un ajustement affine par la méthode des moindres carrés	❑ ok
Estimer et interpréter un coefficient de corrélation linéaire	❑ ok
Faire un ajustement par une hyperbole	❑ ok
Faire un ajustement par une fonction puissance	❑ ok
Faire un ajustement par une exponentielle	❑ ok
Faire un ajustement par une fonction logarithmique	❑ ok