le cnam

2025 - 2026

Statistiques à deux variables (.....)

Régression linéaire

Covariance et corrélation

Les différentes courbes de tendance.

Le coefficient de détermination.

Fiche n° 2

I. Régression linéaire

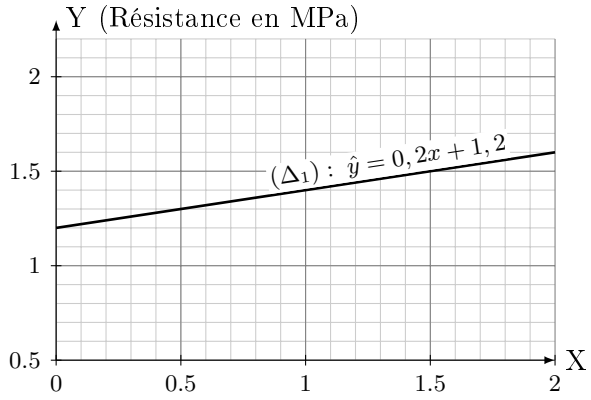
1. Nuage de points

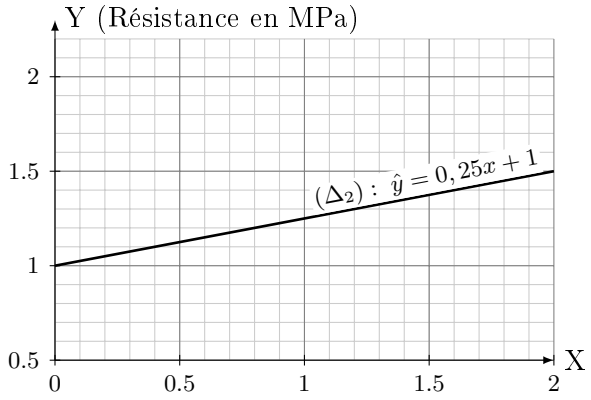
Exemple n° 1 : Avant de construire un ouvrage, on étudie le sol. On utilise souvent un pénétromètre (un cône que l'on enfonce dans le sol) pour mesurer la résistance du terrain à différentes profondeurs.

Profondeur X (m)	0,00	0,33	0,67	1,00	1,33	1,67	2,00
Résistance Y (MPa)	1,2	0,8	1,5	1,1	2,1	0,9	1,7

On dispose de deux échantillons ..... de mesures $(\ldots, \ldots)$.

Définition:
On considère deux séries statistiques (x_i) et (y_i) définies sur une même population.
On appelle ..... l'ensemble des points de coordonnées $(x_i\,,\,y_i)$





On cherche à déterminer, s'il existe, un lien entre la résistance Y, et la profondeur X. On peut par exemple chercher à exprimer les valeurs prises par la variable Y en fonction de celles prises par X sous la forme $Y = aX+b$. Graphiquement, on cherche donc une droite qui passe "au plus près" des points de coordonnées $(x_i\,,\,y_i)$.

Considérons les deux droites (Δ_1) et (Δ_2). Laquelle de ces deux droites passe "au plus près" de ces points ? Pour répondre à cette question, nous devons nous interroger sur ce qu'on entend par "au plus près". Pour ce faire, complète le tableau suivant :

Etude de l'approximation $(\Delta_1) : \hat{y} = 0,2x + 1,2$

x_i	0	0,3	0,7	1	1,3	1,7	2,00	
y_i	1,2	0,8	1,5	1,1	2,1	0,9	1,7	
$\hat{y}_i$	1,2		1,34	1,4		1,54	1,6	Total
$y_i - \hat{y}_i$	0		0,16	−0,3		−0,64	0,1	
	0		0,0256	0,09		0,4096	0,01	

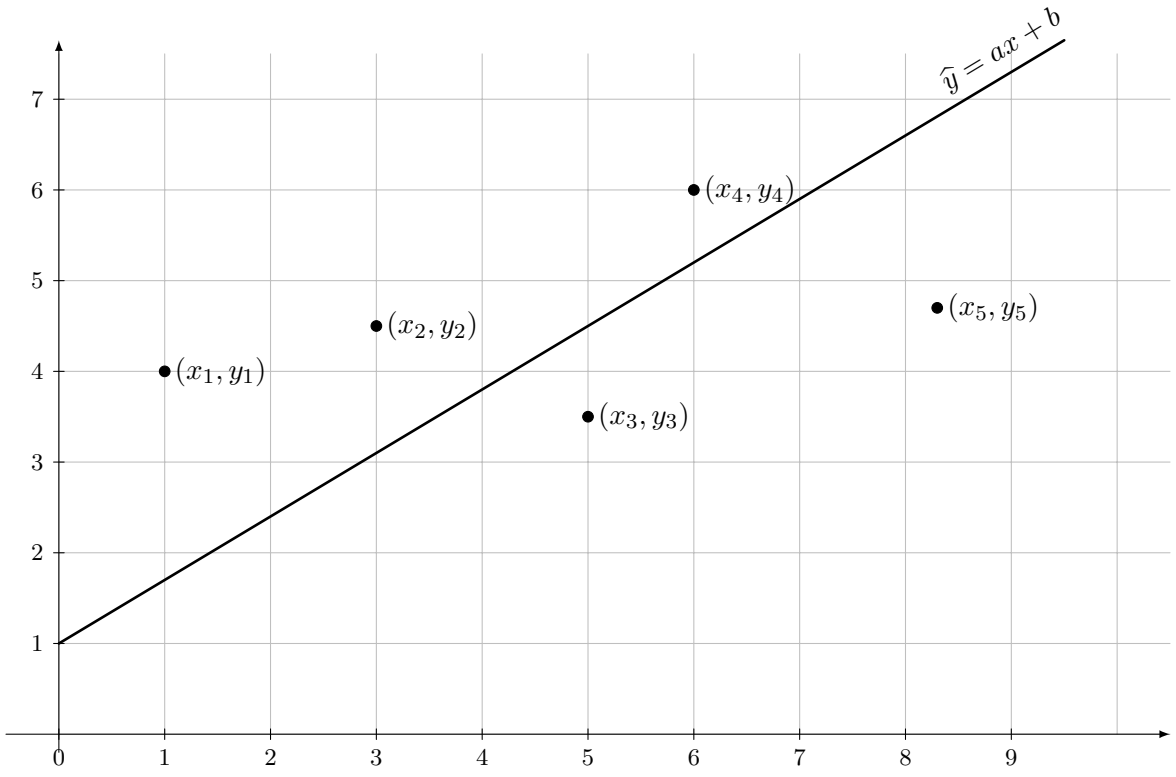
Etude de l'approximation $(\Delta_2) : \hat{y} = 0,25x + 1$

x_i	0	0,3	0,7	1	1,3	1,7	2,00	
y_i	1,2	0,8	1,5	1,1	2,1	0,9	1,7	
$\hat{y}_i$		$1,075$						Total
$y_i - \hat{y}_i$		-0.275						

Ainsi, pour (Δ_1) : $\sum_{i=1}^{7} \left(y_i - \hat{y}_i\right)^2 =$ ..... et pour (Δ_2) : $\sum_{i=1}^{7} \left(y_i - \hat{y}_i\right)^2 =$ .....
Au regard de la somme des carrés des écarts verticaux, ... , on peut dire que la droite (Δ_1) est plus ..... du nuage de points que la droite (Δ_2).

2. Méthode des moindres carrés

Etant donnée deux variables aléatoires X et Y associées à une même population. On recherche un lien affine entre X et Y. Plus précisément, on cherche à exprimer les valeurs prises par la variable Y en fonction de celles prises par X sous la forme : $Y = aX + b$. Cette relation s'appelle une régression linéaire.



On cherche donc, deux nombres a et b tels que la droite d'équation $\widehat{y} = ax + b$ soit "au plus près" du nuage de points $\left(x_i\,,\,y_i\right)$. On note $\widehat{y}_i$ les points d'abscisse x_i de la droite d'équation $\widehat{y} = ax + b$.

Pour cette recherche de lien, la méthode la plus utilisée est celle où l'on minimise la **S**omme des **C**arrés **R**ésiduels :

$$\text{SCR} = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2$$

Pour chaque individu i : $y_i - \widehat{y}_i$ est l' ..... commise lorsqu'on évalue y_i avec la relation $\widehat{y} = ax + b$. En fonction de la position de la droite par rapport au point $\left(x_i\,,\,y_i\right)$, cette quantité est positive si le point se situe au-dessus de la droite, négative sinon. Pour éviter que deux erreurs de signes opposés se neutralisent et, donc, masquent l'erreur, elles sont élevées au carré.

Autrement dit, on cherche la droite d'équation $\widehat{Y} = aX + b$ telle que $\sum_i d_i^2 = \sum_i \left[y_i - \underbrace{(ax_i + b)}_{\widehat{y}_i}\right]^2$ soit minimum,

d'où le nom : ..... .

Définition:

On appelle droite de régression linéaire la droite d'équation : $y = ax + b$ définie par la méthode des moindres carrés.

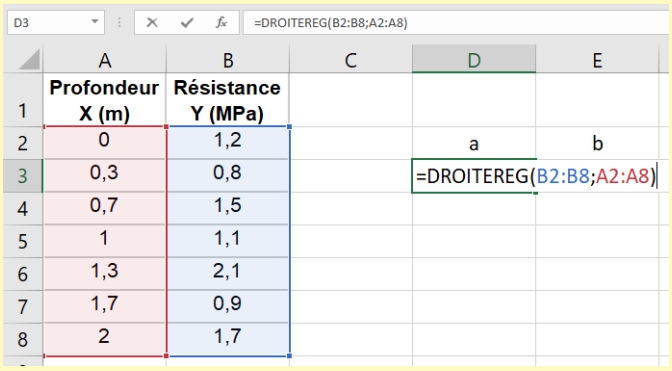
Propriété

Soit $(x_i\,,\,y_i)$ un nuage de points. Notons $\widehat{y} = ax + b$ l'équation réduite de la droite de régression linéaire.

- La **S**omme des **C**arrés **R**ésiduelle : $\sum_i \left(y_i - \widehat{y}\right)^2$ est minimum.
- Le point de coordonnées $\left(\overline{x}\,,\,\overline{y}\right)$, appelé le ..... du nuage de points, est sur cette droite.

Pour trouver a et b par la méthode des moindres carrés, on utilise la fonction

=DROITEREG(plage des y_i ; plage des x_i)



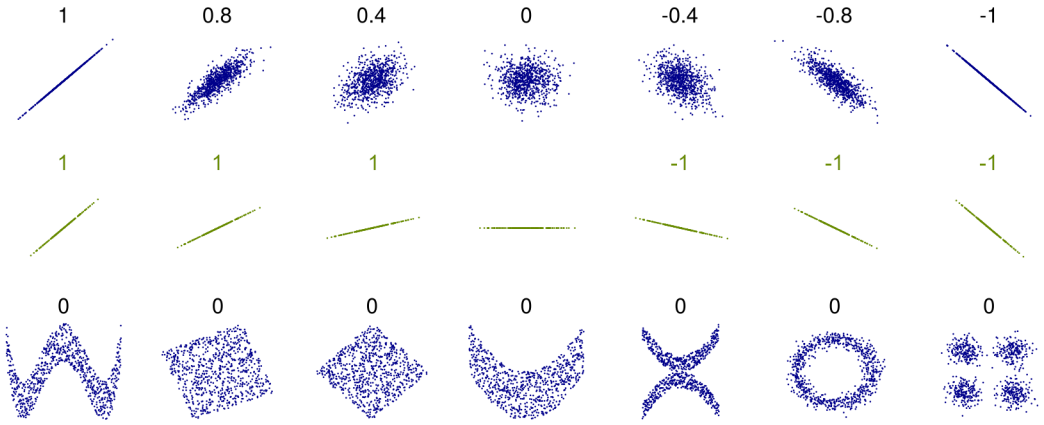
Définition:

On appelle coefficient de corrélation du nuage (X, Y), noté $\rho_{(X,Y)}$, le nombre $\dfrac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y}$

	A	B	C	D	E	F	G
1	Profondeur X (m)	Résistance Y (MPa)					
2	0	1,2		a	b		
3	0,3	0,8		0,23734177	1,09122966		
4	0,7	1,5					
5	1	1,1		σ_x	=ECARTYPE.PEARSON(A2:A8)		
6	1,3	2,1		σ_y	=ECARTYPE.PEARSON(B2:B8)		
7	1,7	0,9		cov(x,y)	=COVARIANCE.PEARSON(A2:A8;B2:B8)		
8	2	1,7		ρ(x,y)	=COEFFICIENT.CORRELATION(A2:A8;B2:B8)		

Exemple n° 1 (suite): Le coefficient de corrélation est ..... à 10^{-2} près.

Exemple : Coefficients de corrélations de différents nuages de points :



Propriété

Soit ρ le coefficient de corrélation d'un nuage de points.

- $\rho \in [-1\,;\,1]$
- Si $\rho = 1$ les points sont alignés suivant une droite ascendante.
- Si $\rho = -1$ les points sont alignés suivant une droite descendante.
- Plus ρ est proche de -1 ou de 1, plus la qualité de la prédiction par le modèle de régression linéaire est bonne, ce qui signifie que le nuage de points est resserré autour de la droite. A l'inverse, plus ρ est proche de 0, plus la qualité de la prédiction est mauvaise.

II. Covariance et corrélation par l'exemple.

La variance mesure la dispersion d'une seule variable par rapport à sa propre moyenne. Mathématiquement, c'est la moyenne des écarts au carré :

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

La variance répond à la question : « À quel point mes mesures s'écartent-elles de la valeur centrale : la moyenne ? ».

Si l'on veut étudier deux grandeurs simultanément (par exemple, la teneur en eau X et la résistance Y d'un sol), on ne peut plus se contenter de regarder comment chaque variable varie dans son coin. On veut savoir comment elles varient ensemble :
Pour cela, dans la formule de la variance écrite ci-dessus, on remplace le deuxième terme de l'écart de X par l'écart de Y. La formule devient :

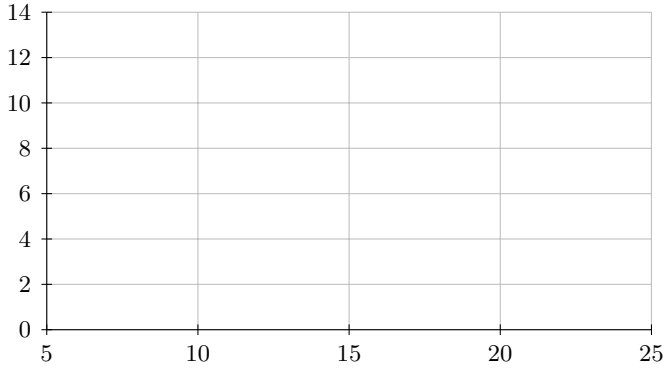
$$\mathrm{Cov}(X,Y) = \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

La variance n'est alors rien d'autre qu'un cas particulier de la covariance d'une variable avec elle-même :

$\mathrm{Cov}(X, X) =$ .....

Série n° 1 : relation linéaire parfaite.

	A	B	C	D	E	F	G	H
4	**Variable**	**Point 1**	**Point 2**	**Point 3**	**Point 4**	**Point 5**	**Métrique**	**Valeur**
5	**Teneur en eau x_i (%)**	5	10	15	20	25	Moyenne $\overline{x}$	
6	**Résistance y_i (MPa)**	12	10	8	6	4	Moyenne $\overline{y}$	
7							**Covariance**	
8							**Corrélation**	



Le point moyen est toujours ..... sur la droite de régression linéaire.

Ce graphique est parfait, il n'y a aucun doute : la résistance décroît de manière strictement ..... à la teneur en eau.

Série n° 2 : changement d'échelle.

Même relation physique que la série n° 1, mais exprimée en Pascals (Pa).

	A	B	C	D	E	F
11	**Variable**	**Point 1**	**Point 2**	**Point 3**	**Point 4**	**Point 5**
12	**Teneur en eau x_i (%)**	5	10	15	20	25
13	**Résistance y_i (MPa)**					

	G	H
14	**Covariance**	
15	**Corrélation**	

Conclusion :

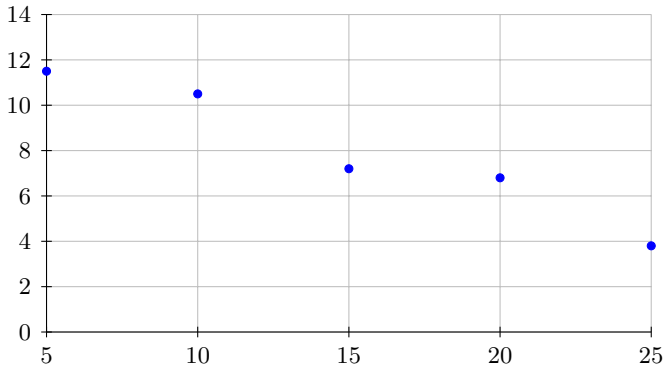
Le changement d'unité a fait exploser notre covariance. Sa valeur est difficile à interpréter, car elle est sensible aux changements d'échelle. D'où la nécessité de la normaliser, pour ce faire, on l'a divisée par les écarts-types, pour obtenir le ..... : $\rho = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$.

En normalisant la ..... , les séries n° 1 et 2 ont le même coefficient de corrélation $\rho = ...$ Ce qui est le peu de ce que l'on puisse attendre d'un paramètre qui étudie les mêmes données.

Série n° 3 : relation linéaire bruitée.

La tendance globale est identique à celle de la Série n° 1, mais intègre les incertitudes réelles de mesure et l'hétérogénéité du sol.

	A	B	C	D	E	F	G	H
18	**Variable**	**Point 1**	**Point 2**	**Point 3**	**Point 4**	**Point 5**	**Métrique**	**Valeur**
19	**Teneur en eau x_i (%)**	5	10	15	20	25	Moyenne $\overline{x}$	
20	**Résistance y_i (MPa)**	11,5	10,5	7,2	6,8	3,8	Moyenne $\overline{y}$	
21							**Covariance**	
22							**Corrélation**	



Conclusion :

En passant au coefficient ρ, la série n° 2 donne $\rho = -1$ (parfait) et la série n° 3 donne $\rho = -0.98$ (très fort mais bruité). La régression linéaire devient un outil indispensable pour donner une approximation statistique de la résistance du sol en fonction de la teneur en eau du sol.

Pour une teneur de 13% en eau, on estime la résistance à .....

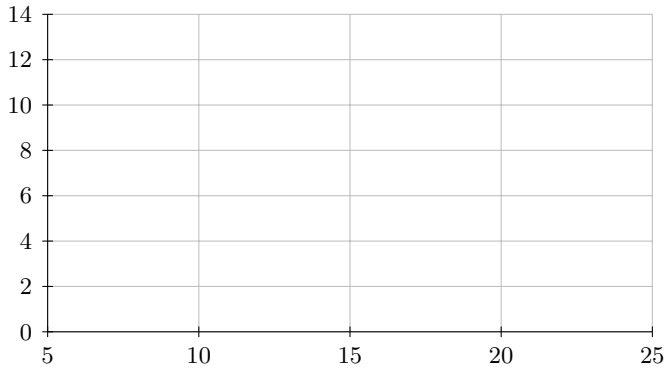
Série n° 4 : le piège de la non-linéarité (l'optimum de Proctor).

Relation quadratique parfaite (parabole).

	A	B	C	D	E	F	G	H
25	**Variable**	**Point 1**	**Point 2**	**Point 3**	**Point 4**	**Point 5**	**Métrique**	**Valeur**
26	**Teneur en eau x_i (%)**	5	10	15	20	25	Moyenne $\overline{x}$	
27	**Résistance y_i (MPa)**	6,0	10,0	12,0	10,0	6,0	Moyenne $\overline{y}$	
28	**Ecart aux moyennes**	**Point 1**	**Point 2**	**Point 3**	**Point 4**	**Point 5**	**Covariance**	
29	$(x_i - \overline{x})$						**Corrélation**	
30	$(y_i - \overline{y})$							
31	**Produit**							

Pourquoi la covariance est-elle ..... ?

- Entre 5% et 15% de teneur en eau : les écarts $(x_i - \bar{X})$ et $(y_i - \bar{Y})$ sont de même signe (soit tous deux positifs, soit tous deux négatifs). Leur produit est positif. Si le nuage de points est majoritairement dans ces zones, la covariance sera positive : quand X augmente, Y a tendance à augmenter.
- Entre 15% et 25% de teneur d'eau : Les écarts sont de signes opposés. Leur produit est négatif. Si le nuage s'aligne ici, la covariance sera négative : quand X augmente, Y diminue (ce qui est le cas de notre Série 1 où l'eau détruit la résistance).



En s'annulant, la covariance masque une relation ..... On dit aussi une relation .....

Pour une teneur de 13% en eau, on estime la résistance à

.....

Le point moyen n'a aucune raison d'être sur la parabole (ce n'est pas une régression linéaire).

Conclusion :

- La variance mesure la dispersion d'un phénomène sur un seul axe.
- La covariance mesure la tendance de deux phénomènes à s'écarter de leur moyenne dans le même sens (produit positif) ou en sens opposé (produit négatif). Elle capture la co-oscillation.

Remarque :

- Si on regarde uniquement les chiffres, on conclut : « pas de relation entre l'eau et la résistance ». Pourtant, si on trace le nuage de points, la relation est parfaite mais .....

- La covariance et le coefficient de corrélation ne mesurent que la liaison linéaire. Un ingénieur qui n'affiche pas ses graphiques passera à côté d'un phénomène physique majeur.

Conséquences :

- Sans dimension : les unités du numérateur et du dénominateur se simplifient. Le coefficient ρ est un nombre pur, sans unité.
- Borné (Inégalité de Cauchy-Schwarz) : Le dénominateur $\sigma_X\sigma_Y$ représente la valeur maximale absolue que peut atteindre la covariance. Il s'ensuit la double inégalité : $-1 \leqslant \rho \leqslant 1$.
- Interprétable au premier coup d'œil : $\rho = 1$: Relation linéaire positive parfaite.
  - $\rho = \ldots$ ou $\rho = \ldots$: relation linéaire parfaite.
  - $\rho = \ldots$: Absence totale de relation linéaire.

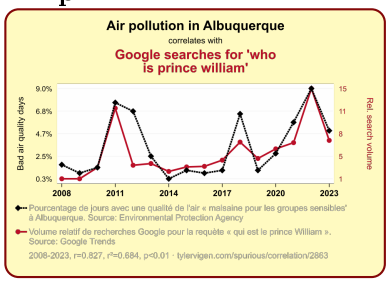
Le risque de surinterprétation :

Il faut veiller à ne pas donner un sens excessif, abusif ou non justifié au coefficient de corrélation ρ :

- Un bon ajustement linéaire se traduit par un ρ^2 proche de 1.
- A contrario, un ρ^2 proche de 1 ne traduit pas forcément un lien linéaire.
- Un ρ^2 proche de 0 traduit un mauvais ajustement linéaire, mais n'implique pas qu'aucune relation ne puisse être établie entre les variables.

Il ne faut pas confondre corrélation et relation ..... . Une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.

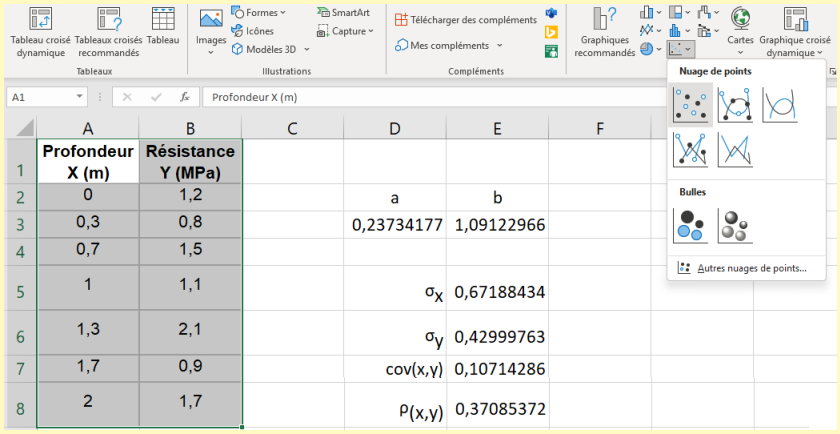
Exemple :



sur le site `https://www.tylervigen.com/spurious-correlations`, on trouve de nombreuses corrélations non causales

1. Comment déterminer l'équation de la droite de régression.

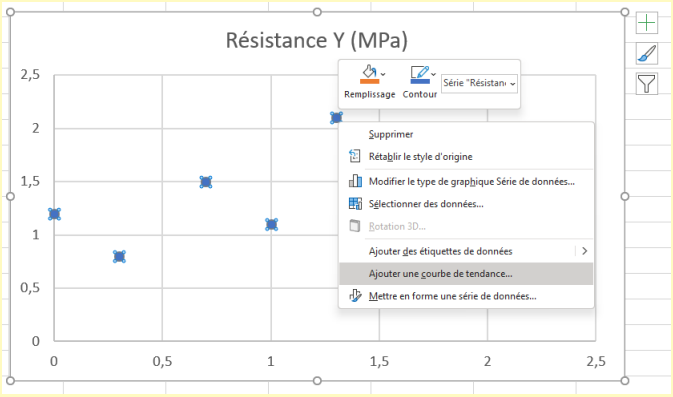
(1) On sélectionne les deux séries et on insère un graphique **nuage de points** :



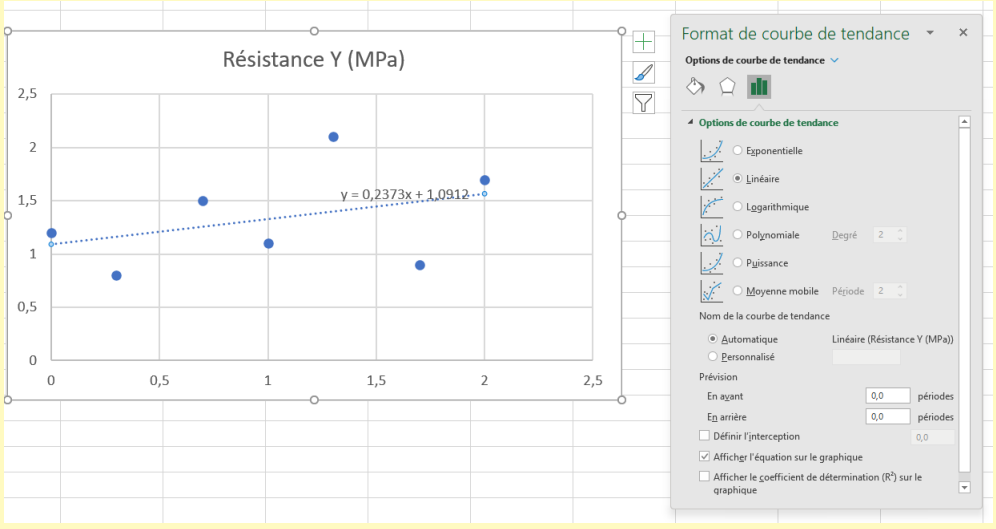
	A	B	C	D	E	F
1	Profondeur X (m)	Résistance Y (MPa)				
2	0	1,2		a	b	
3	0,3	0,8		0,23734177	1,09122966	
4	0,7	1,5				
5	1	1,1		σ_x	0,67188434	
6	1,3	2,1		σ_y	0,42999763	
7	1,7	0,9		cov(x,y)	0,10714286	
8	2	1,7		ρ(x,y)	0,37085372	

(2) On sélectionne le nuage de points, on coche **Courbe de tendance**,

puis Autres options...



(3) Dans les options on coche **Afficher l'équation sur le graphique**



Exemple n° 1 (suite): Reprenons notre étude de sol.

- Cette droite permet-elle de faire une bonne approximation ? .....

  .....

- Calcule une approximation de la pression à une profondeur de 1,5 m : .....

  .....

Exercice n° 1: On a les données suivantes :

X (m)	Y (MPa)	X (m)	Y (MPa)	X (m)	Y (MPa)
0,00	1,2	5,33	4,2	10,67	12,1
0,33	0,8	5,67	4,4	11,00	14,2
0,67	1,5	6,00	4,8	11,33	13,9
1,00	1,1	6,33	4,9	11,67	15,1
1,33	2,1	6,67	5,3	12,00	14,8
1,67	0,9	7,00	5,5	12,33	16,3
2,00	1,7	7,33	5,8	12,67	15,5
2,33	1,4	7,67	6,1	13,00	17,2
2,67	2,3	8,00	6,3	13,33	16,8
3,00	2,6	8,33	7,5	13,67	18,5
3,33	2,9	8,67	8,9	14,00	17,9
3,67	3,1	9,00	9,2	14,33	19,2
4,00	3,2	9,33	11,1	14,67	18,7
4,33	3,6	9,67	10,4	15,00	20,1
4,67	3,7	10,00	11,8		
5,00	3,9	10,33	13,5		

1. Quel est le coefficient de corrélation à 10^{-4} près ? .....
2. Au vu de ce coefficient de corrélation, est-il judicieux de faire un ajustement linéaire ? Pourquoi ? .....

   .....

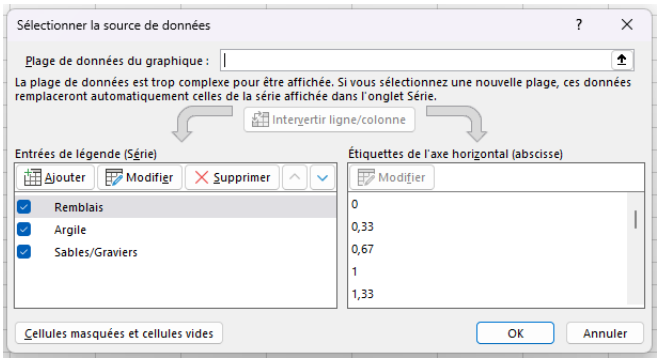
3. Complète les profondeurs en mètres :

- 0 à .... m (Remblai) : Faible résistance, très dispersée : le bruit est fort car
  - L'hétérogénéité des composants : Un remblai de chantier contient un mélange imprévisible de terre, de morceaux de béton, de briques, de gravats, de racines et parfois de cavités ou de poches d'air.
  - L'effet "obstacle" sur la pointe : Si la pointe rencontre un morceau de brique ou un galet, la résistance mesurée (q_c) va grimper en flèche d'un coup (ex : $2,1$ MPa), alors que si quelques centimètres plus bas, si la pointe glisse dans une poche de terre meuble ou un vide, la résistance va chuter brutalement (ex : $0,8$ MPa).
- .... à .... m (Zone Argileuse) : Comportement parfaitement linéaire avec une pente douce et régulière.
- .... à .... m (Sable dense / Graviers) : Transition brutale, la résistance grimpe en flèche avec une forte dispersion.
  - Le blocage des grains :Les graviers se coincent les uns contre les autres (phénomène d'imbrication). Pour avancer, la pointe doit forcer pour écarter ou briser ces grains compactés sous le poids des 8 mètres de terre au-dessus. La résistance grimpe d'un coup très haut.

– La rupture locale : Dès qu'un groupe de grains se déplace ou glisse, la résistance rechute brièvement avant le blocage suivant.

- Ajouter les trois courbes de tendance :

1. Crée un tableau avec les trois séries en colonnes les unes à côté des autres.
2. Sélectionne l'ensemble de ton tableau (les 6 colonnes : Remblais, Argile, Sable).
3. Insère un nuage de points
4. Réorganise tes données en faisant un clic droit au milieu de ton graphique et choisis Sélectionner des données :



4. Complète :

	Régression linéaire	Coefficient de corrélation
Remblais		
Argile		
Sable/graviers		

5. **Applications :** détermine une approximation de la résistance du sol à :

(a) 4,20 m : .....

.....

(b) 1,50 m : .....

.....

(c) 17 m : .....

.....

ÉVITE LES ERREURS D'ARRONDIS

	A	B	C	D	E	F	G
26	Coefficient de corrélation			a	b	Profondeur en m	Estimation en MPa
27	Remblais			0,18186887	1,12539543	4,20	
28	Argile			0,74196223	0,31331088	1,50	
29	Sable/Graviers			1,73543066	-5,8324053	17,00	

Exercice n° 2: Le tableau suivant donne l'évolution du nombre de clients d'une entreprise de vente par internet pendant cinq années consécutives.

Rang de l'année	1	2	3	4	5
Nombre de clients	2463	5817	11210	20620	34900

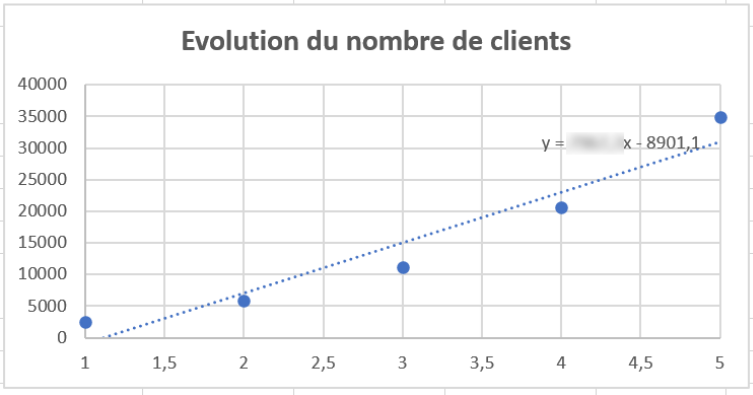
1. Quel est le coefficient de corrélation à 10^{-3} près ? .....

2. Le coefficient de corrélation permet-il de faire une régression linéaire ? Pourquoi ?

.....

.....

	A	B	C	D	E	F	G
1	**Rang de l'année**	1	2	3	4	5	
2	**Nombre de clients**	2463	5817	11210	20620	34900	
3							
4	Coefficient de corrélation :	[illegible]5051					
5							
6	**a**	**b**					
7	[illegible]	-8901,1					
8							
9	La sixième année :	[illegible]					
10							
11							



3. La formule saisie dans la cellule A7 est : ☐

4. Combien de clients peut-on prévoir la sixième année ? .....

5. La formule saisie dans la cellule B9 est : ☐

6. Copie cette feuille sur une autre feuille, et ajoute dans le tableau la sixième année avec son estimation.

	A	B	C	D	E	F	G
1	**Rang de l'année**	1	2	3	4	5	6
2	**Nombre de clients**	2463	5817	11210	20620	34900	

(a) La formule saisie dans la cellule G2 est : ☐

(b) Modifie les plages de données de la droite de régression linéaire pour y ajouter le rang 6.
La droite a-t-elle changé d'équation ? Pourquoi ? .....

.....

.....

. . . . . .

. . . . . .

. . . . . .

. . . . . .

(c) Le coefficient de corrélation a-t-il changé ? . . . . . .

. . . . . .

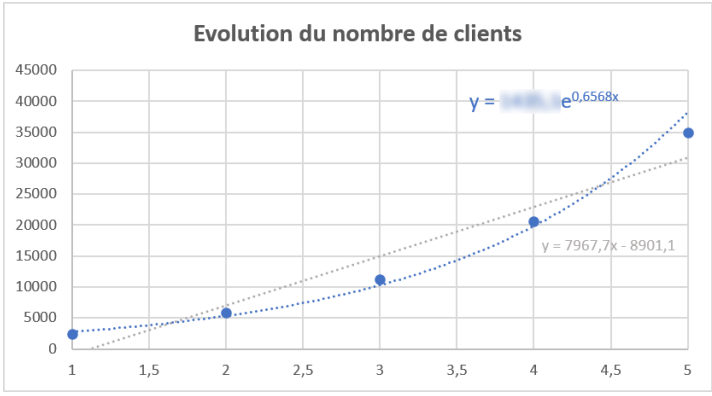
(d) On souhaite estimer le nombre de clients prévisible pour les quatre années suivantes, quelle formule saisir dans la cellule G2 ?

	A	B	C	D	E	F	G	H	I	J	K
1	**Rang de l'année**	1	2	3	4	5	6	7	8	9	10
2	**Nombre de clients**	2463	5817	11210	20620	34900	[illegible]	46873	54841	62808	70776

III. Les différentes courbes de tendance.

Exercice n° 3: Reprenons la nuage de point d'un exercice précédent présentant l'évolution du nombre de clients d'une entreprise de vente par internet pendant cinq années consécutives.

Rang de l'année	1	2	3	4	5
Nombre de clients	2463	5817	11210	20620	34900



Il semble que la courbe de tendance exponentielle approxime mieux les données.

1. Quelle est l'équation de cette courbe ? . . . . . .

2. Pour faire des prévisions, quelle formule saisir dans la cellule G2 ?

	A	B	C	D	E	F	G	H	I	J
1	**Rang de l'année**	1	2	3	4	5	6	7	8	9
2	**Nombre de clients**	2463	5817	11210	20620	34900	[illegible]	[illegible]	[illegible]	[illegible]

3. Y a-t-il une grande différence avec une prévision linéaire ? . . . . . .

. . . . . .

4 Détermination des coefficients du modèle exponentiel.

(a) Ajoute la ligne $\ln(y_i)$:

	A	B	C	D	E	F
1	**Rang de l'année**	1	2	3	4	5
2	**Nombre de clients**	2463	5817	11210	20620	34900
3	**ln(Nombre de clients)**	7,8091	8,6685	9,3246	9,9340	10,4602

(b) Calcule le logarithme népérien de l'équation trouvée à la question 1 :

.....

(c) Comment retrouver les deux paramètres de la courbe exponentielle ? .....

.....

(d) Déduis-en :

	H	I	J
1	**a**	**b**	[illegible]
2	0,6568	7,2690	1435,1

(e) La formule saisie dans la cellule H2 est

(f) La formule saisie dans la cellule J2 est

IV. Le modèle linéaire.

DROITEREG cherche à estimer les paramètres d'un modèle de régression linéaire multiple.
Soit Y la variable endogène (le vecteur des observations de dimension $n \times 1$) et X la matrice des variables exogènes (de dimension $n \times (p+1)$ si l'on inclut la constante). Le modèle s'écrit :

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon} \text{ où } \widehat{Y} = \begin{pmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{pmatrix} = X\beta$$

et $\beta = \begin{pmatrix} \beta_0 & \beta_1 & \cdots & \beta_p \end{pmatrix}^T$ est le vecteur des coefficients à estimer, et ε est le vecteur des résidus.

On cherche à minimiser la norme euclidienne au carré du vecteur des résidus (critère des moindres carrés) :

$$\min_{\beta} S(\beta) = \min_{\beta} \|Y - X\beta\|_2^2 = \min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

Théorème
D'un point de vue purement théorique, si X est de plein rang-colonne ($rg(X) = p+1$), la solution unique est donnée par les célèbres équations normales : $\hat{\beta} = (X^T X)^{-1} X^T Y$

Nous allons démontrer ce théorème dans le cas du premier degré. Pour ce faire, nous aurons besoin de ne pas oublier que si $\overrightarrow{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$ et $\overrightarrow{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ dans une base orthonormée, alors

$$a^T b = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = a_1 b_1 + \dots + a_n b_n = \quad \dots\dots\dots$$

1. Dans le cas du premier degré.

On cherche le modèle $y = \beta_0 + \beta_1 x + \varepsilon$.

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

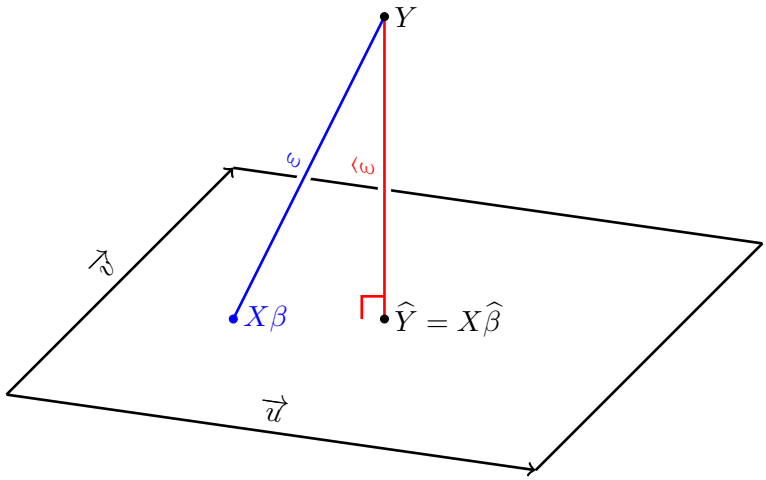
Le théorème donne comme solution unique : le vecteur des coefficients de la droite affine :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^T X)^{-1} X^T Y$$

Démonstration

Posons $\overrightarrow{u} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ et $\overrightarrow{v} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, on $X = \begin{pmatrix} & \\ & \\ & \end{pmatrix}$ et $X\beta = \begin{pmatrix} & \\ & \\ & \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} & \\ & \\ & \end{pmatrix}$

soit $X\beta = \beta_0 \overrightarrow{u} + \beta_1 \overrightarrow{v}$: quand β décrit $\mathbb{R}^2$, $X\beta$ décrit le plan engendré par $\overrightarrow{u}$ et $\overrightarrow{v}$.



$\varepsilon = Y - X\beta$ est minimum lorsque le vecteur ε est orthogonal au plan engendré par $\overrightarrow{u}$ et $\overrightarrow{v}$.

Si un vecteur est orthogonal à un plan, son produit scalaire avec n'importe quel vecteur de ce plan est nul. En particulier, il est orthogonal aux colonnes de la matrice X qui forment la base de ce plan. Donc :

$$X^T \widehat{\varepsilon} = 0$$
$$X^T(\quad \dots\dots\dots \quad) = 0$$
$$X^T Y - X^T X \widehat{\beta} = 0$$
$$X^T X \widehat{\beta} = \quad \dots\dots$$

Si la matrice $X^T X$ est inversible alors

$$\widehat{\beta} = \quad \dots\dots\dots\dots\dots$$

Remarque :

En posant $\overline{Y} = \begin{pmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} = \overline{y} \overbrace{\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}}^{\overrightarrow{u}}$, on a $\widehat{Y} - \overline{Y} = \left(X\widehat{\beta} - \overline{y}\,\overrightarrow{u}\right) \in \left\langle \overrightarrow{u}, \overrightarrow{v} \right\rangle$. Donc, $\overrightarrow{\varepsilon} \cdot \left(\widehat{Y} - \overline{Y}\right) = 0$.

Autrement dit, l'orthogonalité de $\overrightarrow{\varepsilon}$ au plan $\left\langle \overrightarrow{u}, \overrightarrow{v} \right\rangle$ entraîne .....

Remarque : dans le cas du premier degré, on trouve après calcul :

On trouve la pente $\hat{\beta}_1 = \dfrac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \dfrac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$, et l'ordonnée à l'origine $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

2. Dans la cas du second degré.

On cherche le modèle $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\varepsilon}$$

Le théorème donne la solution unique : le vecteur des coefficients de la parabole :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X^T X)^{-1} X^T Y$$

Remarque :

Le modèle parabolique est bien un modèle ..... , alors que la régression associée n'est pas une droite. C'est un modèle linéaire, car du point de vue mathématique, l'approximation est un vecteur $\widehat{\beta}$, la projection orthogonale sur le sous-espace vectoriel engendré par les colonnes de X (sous-espace vectoriel de dimension ...). Cette projection $\widehat{\beta}$ a pour coordonnées les trois coefficients de la parabole.

3. Le cas exponentiel.

On cherche le modèle exponentiel $y = A \cdot e^{Bx}$ où A et B sont les paramètres à estimer, et $y > 0$. Si on applique le logarithme népérien (ln) des deux côtés de l'égalité, on obtient :

$$\ln(y) = \ln(A \cdot e^{Bx}) = \ln(A) + Bx$$

On pose alors un changement de variables $Y^* = \ln(y)$, $\beta_0 = \ln(A)$, et $\beta_1 = B$ qui nous ramène à un modèle linéaire du premier degré :

$$Y^* = \beta_0 + \beta_1 x + \varepsilon^*$$

En transformant le problème, la fonction DROITEREG ne minimise plus l'écart quadratique sur les données réelles, mais sur les données logarithmiques. Le modèle sous-jacent avec son terme d'erreur est :

$$\ln(y_i) = \ln(A) + Bx_i + \varepsilon_i \implies y_i = A \cdot e^{Bx_i} \cdot e^{\varepsilon_i}$$

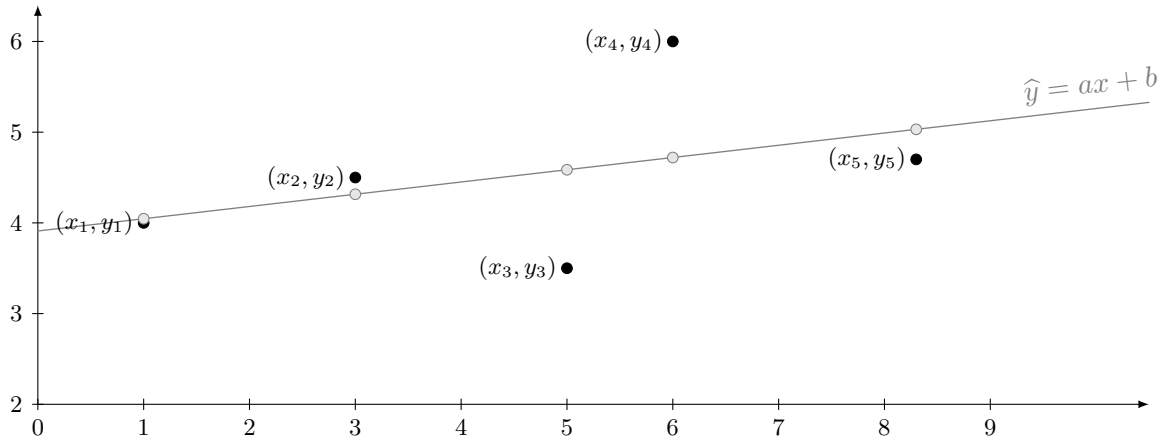
L'erreur (e^{ε_i}) n'est plus additive mais multiplicative.

Remarque :

Le modèle exponentiel n'est pas un modèle ..... en soi, dans la mesure où il faut modifier (ou transformer) les données pour se ramener à un modèle linéaire.

V. Le coefficient de détermination

Une fois la régression linéaire effectuée, on a un modèle qui mathématise la réalité, celui formé par le nuage de points ..... :



On appelle **S**omme des **C**arrés **T**otale la somme $SCT = \sum_{i=1}^{n} (y_i - \overline{y}_i)^2$. On a :

$$\begin{aligned} \text{SCT} &= \sum_{i=1}^{n} (y_i - \overline{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{y} + \widehat{y} - \overline{y}_i)^2 = \sum_{i=1}^{n} \left[(y_i - \widehat{y}) + (\widehat{y} - \overline{y}_i) \right]^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{y})^2 + 2 \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y})(\widehat{y} - \overline{y}_i)}_{\vec{\varepsilon} \cdot (\widehat{Y} - \overline{Y}) = 0} + \sum_{i=1}^{n} (\widehat{y} - \overline{y}_i)^2 \end{aligned}$$

Il s'ensuit une propriété remarquable de la méthode des moindres carrés :

$$\underbrace{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}_{\text{SCE}} + \underbrace{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}_{\text{SCR}}$$

L'égalité n'est vraie que pour les modèles linéaires (droites, paraboles, polynômes).

- SCT (**S**omme des **C**arrés **T**otale) traduit la variation totale de Y.
- SCE (**S**omme des **C**arrés **E**xpliquée) traduit la variation expliquée par le modèle.
- SCR (**S**omme des **C**arrés **R**ésiduels) traduit la variation inexpliquée par le modèle.

Définition:

On appelle ..... la quantité suivante : $R^2 = \dfrac{\text{SCE}}{\text{SCT}}$

Ce coefficient R^2 est dans $[0, 1]$, puisque : $0 \leqslant \text{SCE} \leqslant \text{SCT}$

☞ Si $R^2 = 1$, on a alors SCE = SCT : toute la variation est expliquée par le modèle.

☞ Si $R^2 = 0$, on a alors SCE $= 0$ donc SCE = SCT : aucune variation n'est expliquée par le modèle.

☞ Plus R^2 est proche de 1, plus la qualité de la prédiction par le modèle de régression linéaire est bonne, ce qui signifie que le nuage de points est resserré autour de la droite. A l'inverse, plus R^2 est proche de 0, plus la qualité de la prédiction est mauvaise.

Le coefficient de détermination R^2 est la ..... de la variation totale expliquée par le modèle.

$R^2 = \dfrac{\text{SCE}}{\text{SCT}} = \dfrac{\text{SCT} - \text{SCR}}{\text{SCT}} =$ ..... . Or la méthode des moindres carrés est la meilleure méthode pour minimiser la Somme des Carrés Résiduels, donc elle est la meilleure méthode pour maximiser le coefficient de détermination R^2.

Remarque

- Dans le cadre d'une régression linéaire simple (avec une seule variable explicative x), $R^2 = \rho^2$.
- Le coefficient de détermination est compris entre 0 et 1. Plus il est proche de 1 plus le modèle explique les variations des données.
- Le coefficient de corrélation classique est rigoureusement limité à la mesure d'une régression linéaire (relation en ligne droite).
- Si par exemple, on l'applique la formule classique $\rho = \dfrac{\text{cov}(x, y)}{\sigma_x \sigma_y}$ à des données qui forment une parfaite parabole, on pourrait trouver un coefficient de corrélation ρ proche de 0, laissant croire à tort qu'il n'y a aucune relation algébrique entre les variables.

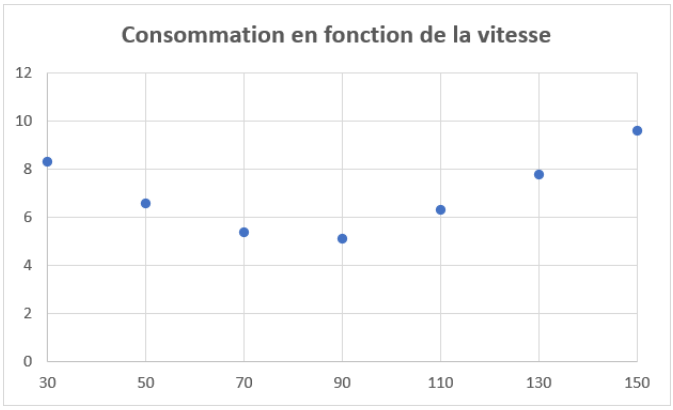
1. Etude d'un exemple linéaire.

En automobile, la courbe de consommation est typiquement en forme de U (une parabole) : la consommation est élevée à basse vitesse (rapports de vitesse courts, arrêts fréquents), atteint un minimum optimal (souvent entre 70 et 90 km/h), puis grimpe en flèche à haute vitesse à cause de la résistance de l'air (force aérodynamique).

Vitesse x_i (km/h)	30	50	70	90	110	130	150
Consommation réelle y_i (L/100 km)	8,3	6,6	5,4	5,1	6,3	7,8	9,6

Exemple n° 2 : On a relevé la consommation en carburant d'une automobile en fonction de sa vitesse :

Vitesse x_i (km/h)	**Consommation réelle y_i** (L/100 km)
30	8,3
50	6,6
70	5,4
90	5,1
110	6,3
130	7,8
150	9,6



1. Quelle courbe de tendance semble être la plus adaptée par rapport à celles proposées par Excel ?

   .....

2. L'équation de la courbe de tendance est de la forme ..... , le tableur affiche sur le graphique :

   $y =$ .....

3. Complète le tableau suivant à 10^{-2} près :

	A	B	C	D	E
27	**Vitesses**	85	120	180	1000
28	**Consommation**				

4. La formule écrite dans la cellule B28 est ☐

5. Dans la cellule D3, saisis : =DROITEREG(plage des y_i ; plage des x_i ^SEQUENCE(1 ;2))

   Complète à 10^{-7} près

	D	E	F
2	**a**	**b**	**c**
3	0,0010119		

6. Complète le tableau suivant à 10^{-2} près :

	A	B	C	D	E
27	**Vitesses**	85	120	180	1000
28	**Consommation**	5,27			
29	**Consommation calculée avec plus de décimales**				

7. La formule saisie dans la cellule B29 est ☐

8. Construis et complète la feuille suivante :

	A	B	C	D	E	F
1	**Etude de la consommation de carburant en fonction de la vitesse**					
2						
3				**SCT**	**SCE**	**SCR**
4	**Vitesse x_i (km/h)**	**Consommation réelle y_i (L/100km)**	**$\hat{y}_i$**	**$(y_i-\bar{y})^2$**	**$(\hat{y}_i-\bar{y})^2$**	**$(y_i-\hat{y}_i)^2$**
5	30	8,3				
6	50	6,6				
7	70	5,4				
8	90	5,1				
9	110	6,3				
10	130	7,8				
11	150	9,6				
12			**Total**			
13						
14	**a**	**b**	**c**			
15	0,001011905	-0,169285714	12,4345238			
16						
17	$\bar{y}$					
18						

9 La formule saisie dans la cellule A15 est

10 La formule saisie dans la cellule :

A18 est E5 est

D5 est F5 est

11 Complète à 10^{-4} près : SCT= ..... SCE= ..... SCR= .....

12 On aurait pu les calculer directement : SCT :

SCE : SCR :

13 L'égalité SCT = SCE + SCR est-elle bien vérifiée ? .....

14 Calcule le coefficient de détermination : R^2 = ...

15 Le modèle parabolique est-il adapté pour expliquer la relation entre la consommation et la vitesse ?

.....

Sur LibreOffice, la fonction SOMME ne fait que des additions. Pour passer en mode matriciel et comprendre des formules de la forme (B5 :B11-A$18)^2, elle doit passer en mode matriciel. Pour cela, il faut remplacer la fonction SOMME par SOMMEPROD.

La fonction SOMMEPROD (Somme de Produits) est conçue dès le départ pour forcer le mode matriciel sur les plages de cellules qu'elle contient. Elle va faire exactement le même calcul (soustraction, mise au carré, puis somme), mais elle est comprise nativement par LibreOffice (et restera compatible avec Excel si jamais vous devez réouvrir le fichier avec ce tableur).

On peut aussi passer en mode matriciel, en écrivant la formule entre accolades.

2. La fonction SEQUENCE.

Construis dans deux nouvelles feuilles les tableaux suivants :

Feuille n° 1 :

	A	B
1	Vitesse x_i en km/h	Consommation réelle y_i
2	30	8,3
3	50	6,6
4	70	5,4
5	90	5,1
6	110	6,3
7	130	7,8
8	150	9,6

Feuille n° 2 :

	A	B	C	D	E	F	G	H
1	Vitesse x_i en km/h	30	50	70	90	110	130	150
2	Consommation réelle y_i	8,3	6,6	5,4	5,1	6,3	7,8	9,6

Pour obtenir les coefficient a, b, et c de la régression quadratique

- de la feuille n° 1, on écrit :
- de la feuille n° 2, on écrit :

Dans la feuille n° 1, ajoute le tableau suivant :

	A	B	C
10	Vitesse x_i en km/h	x_i^2	Consommation réelle y_i
11			8,3
12			6,6
13			5,4
14			5,1
15			6,3
16			7,8
17			9,6

(a) Pour compléter les colonnes A et B, dans la cellule A11 on écrit :

=A2:A8^SEQUENCE(1;2)

(b) Pour obtenir les coefficient a, b, et c de la régression quadratique, on écrit

La fonction SEQUENCE(Ligne ; Colonne ; debut ; pas) :

- La fonction SEQUENCE(1;4;5;2) créé le tableau suivant | 5 | 7 | 9 | 11 |
- La fonction SEQUENCE(1;4) créé le tableau suivant | 1 | 2 | 3 | 4 |
- La fonction SEQUENCE(1;2) créé le tableau suivant | 1 | 2 |
- La formule A2:A8^SEQUENCE(1;2) créé le tableau a deux colonnes où dans la première, on a la colonne A2:A8 élevée à la puissance ..., et dans la deuxième, on a la colonne A2:A8 élevée à la puissance ...

Dans la feuille n° 2, ajoute le tableau suivant :

	A	B	C	D	E	F	G	H
5	Vitesse x_i en km/h							
6	x_i^2 en km/h							
7	Consommation réelle y_i	8,3	6,6	5,4	5,1	6,3	7,8	9,6

(a) Pour compléter les lignes 5 et 6, dans la cellule B5 on écrit : ____

(b) Pour obtenir les coefficient a, b, et c de la régression quadratique,
on écrit ____

3. Interprétation des Sommes des Carrés.

Le « E » signifie bien Expliquée. Ce terme est très imagé et prend tout son sens quand on comprend l'objectif d'un modèle statistique, comme le modèle parabolique que l'on vient d'étudier.

Pour comprendre pourquoi on dit « Expliquée », il faut voir nos données sous l'angle de la variabilité. Imaginons que nos points y montent et descendent : ils varient. On cherche à comprendre pourquoi ils varient.

La variabilité Totale (SCT) est séparée en deux morceaux :

$$\text{Variabilité Totale (SCT)} = \text{Variabilité Expliquée (SCE)} + \text{Variabilité Résiduelle (SCR)}$$

Voici concrètement ce que cela signifie :

(a) La part « Expliquée » par la parabole (SCE)

C'est la variation de y qui est directement causée par l'évolution de x, selon la logique de notre équation. Dans notre étude, la parabole montre qu'en accélérant, la consommation baisse d'abord puis remonte. La SCE mesure toute la variation de consommation que notre modèle arrive à anticiper et à justifier uniquement grâce à la vitesse. La parabole dit : "Je sais pourquoi ce point est plus haut que la moyenne : c'est parce que la vitesse x a augmenté". Le modèle explique cette part de la réalité.

(b) La part « Résiduelle » ou Inexpliquée (SCR)

C'est tout le reste. Ce sont les écarts (les résidus) entre vos points réels et votre courbe. Pour reprendre l'exemple de la voiture : le vent, la pression des pneus ou le style de conduite du pilote font que la consommation réelle dévie un peu de la courbe théorique. Votre parabole ne peut pas deviner ces facteurs. Cette part de variabilité lui échappe : elle est donc inexpliquée par le modèle.

En résumé : La Somme des Carrés Expliquée quantifie mathématiquement la part de réalité que notre parabole a réussi à décoder du nuage de points. Plus la SCE est proche de la SCT (et donc plus le ratio $\frac{SCE}{SCT}$ est proche de 1), plus notre modèle est performant !

4. Etude d'un exemple linéarisable.

La relation $SCT = SCE + SCR$ repose entièrement sur la méthode des moindres carrés linéaires. Une parabole ($y = ax^2 + bx + c$) est un modèle dit "linéaire par rapport à ses paramètres" (a, b, c).
Un modèle exponentiel classique, de la forme :

$$y = a \cdot e^{bx}$$

est un modèle non linéaire. À cause de cette courbure exponentielle, l'orthogonalité de $\overrightarrow{\varepsilon}$ et $\widehat{Y}-\overline{Y}$ qui annulaient le produit scalaire

$$\sum_{i=1}^{n}(y_i-\widehat{y})(\widehat{y}-\overline{y}_i)$$

ne fonctionnent plus. Les résidus ne sont plus orthogonaux aux prédictions. En conséquence :

$$SCT \neq SCE + SCR$$

C'est d'ailleurs pour cela que le coefficient de détermination R^2 perd de son sens ou devient très difficile à interpréter dans les régressions purement non linéaires (il peut même parfois devenir négatif si l'on utilise la formule $1-\frac{SCR}{SCT}$).

La ruse des statisticiens : la ..... pour pouvoir utiliser à nouveau nos outils classiques (et retrouver notre égalité parfaite), on applique un logarithme népérien (ln) pour "aplatir" l'exponentielle et la transformer en droite.

Ainsi, en appliquant le logarithme des deux côtés de l'équation $y = ae^{bx}$, on obtient :

$$\ln(y) = \ln\left(a \times e^{bx}\right) = \ln(a) + bx$$

En posant $Y = \ln(y)$ et $A = \ln(a)$, l'équation devient :

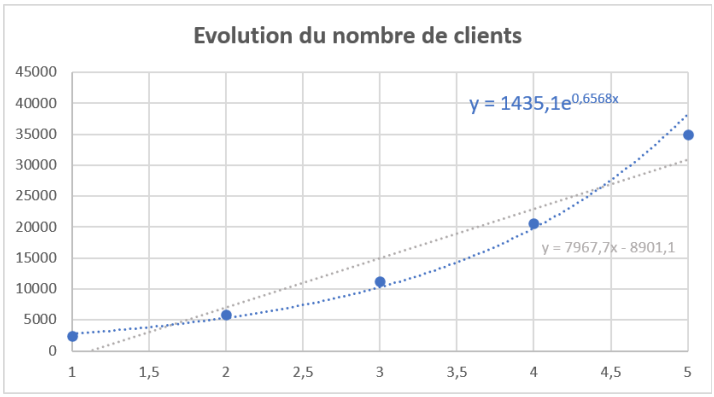
$$Y = bx + A$$

En prenant les ..... de Y, le modèle est devenu une simple droite. La méthode des moindres carrés fonctionne à nouveau parfaitement. L'égalité est donc vraie pour les logarithmes :

$$SCT_{\ln(y)} = SCE_{\ln(y)} + SCR_{\ln(y)}$$

En résumé : l'égalité n'est vraie que pour les modèles linéaires ..... .
Pour un modèle exponentiel, elle n'est vraie que si vous travaillez avec le logarithme de vos ordonnées.

Exemple n° 3 : Reprenons l'étude de l'évolution du nombre de clients d'une entreprise de vente par internet pendant cinq années consécutives.



	A	B	C	D	E	F
1	Rang de l'année	1	2	3	4	5
2	Nombre de clients	2463	5817	11210	20620	34900
3	ln(Nombre de clients)	7,8091	8,6685	9,3246	9,9340	10,4602
4	$\hat{y}_i=ax_i+b$					
5	$\hat{y}_i=\exp(ax_i+b)$					

	H	I	J	K	L
1	a	b	e^b	Moyenne des	
2	0,6568	7,2690	1435,1	$\ln(y_i)$	y_i
3					
4					
5	Modèle linéarisé			Modèle non linéarisé	
6	SCT			SCT	
7	SCE			SCE	
8	R^2			R^2	

Complète en écrivant les formules saisies dans les cellules :

(1) K3 : ..... (2) L3 : .....

3 B4 :

4 B5 :

5 I6 :

6 I7 :

7 I8 :

8 L6 :

9 L7 :

10 L8 :

Dans le cas linéaire, le coefficient de détermination est .....

.....

Dans le cas non linéaire, le coefficient de détermination est .....

.....

VI. Le quartet d'Anscombe

Le quartet d'Anscombe est constitué de quatre nuages de 11 points qui ont les mêmes propriétés statistiques similaires alors qu'en réalité, ils sont très différents, ce qui se voit facilement lorsqu'on les représente sous forme de graphiques. Ils ont été construits en 1973 par le statisticien Francis Anscombe. Ces nuages montrent l'importance de tracer des graphiques avant d'analyser des données, ce qui permet notamment d'estimer l'incidence des données aberrantes sur les différents indices statistiques que l'on pourrait calculer [1].

1 Construis la feuille suivante :

1. Wikipedia : Quartet d'Anscombe

	A	B	C	D	E	F	G	H	I
1		Le quartet d'Anscombe							
2		Nuage n°1		Nuage n°2		Nuage n°3		Nuage n°4	
3		X	Y	X	Y	X	Y	X	Y
4		10	8,04	10	9,14	10	7,46	8	6,58
5		8	6,95	8	8,14	8	6,77	8	5,76
6		13	7,58	13	8,74	13	12,7	8	7,71
7		9	8,81	9	8,77	9	7,11	8	8,84
8		11	8,33	11	9,26	11	7,81	8	8,47
9		14	9,96	14	8,1	14	8,84	8	7,04
10		6	7,24	6	6,13	6	6,08	8	5,25
11		4	4,26	4	3,1	4	5,39	19	12,5
12		12	10,8	12	9,13	12	8,15	8	5,56
13		7	4,82	7	7,26	7	6,42	8	7,91
14		5	5,68	5	4,74	5	5,73	8	6,89
15	Moyenne								
16	Ecart-type								
17	Covariance								
18	Coefficient de correlation								
19	a								
20	b								
21	Régression linéaire								

(2) La formule saisie dans la cellule :

☞ B15 est ☞ B16 est

☞ B17 est

☞ B18 est

☞ B19 est ☞ B20 est

☞ B21 est

(3) Construis pour chaque nuage le graphique correspondant avec la droite de régression linéaire.