Régression linéaire - Applications aux séries chronologiques

Présentation de la droite de régression linéaire par les moindres carrés sur un exemple

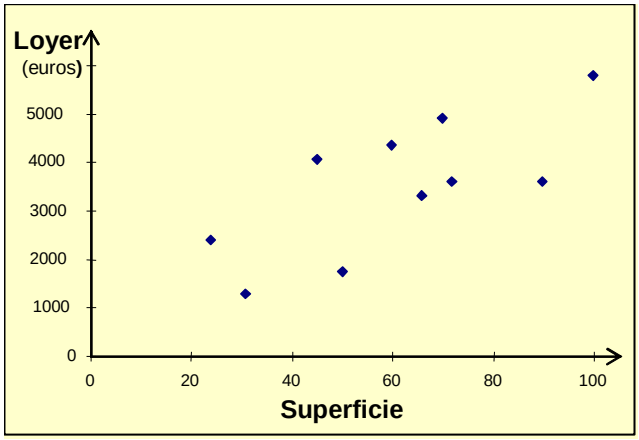
Loyer mensuel de locaux commerciaux (cf. en annexe jeu de données 1)

a) Les données

Population : 10 locaux commerciaux d'un quartier
Le loyer mensuel en euros : Y
La superficie en mètres carrés : X

Observations : $(X_i, Y_i) \quad i = 1, \ldots, n$

Loyer Y (en Euros)	**Superficie X (en m^2)**
2 395	24
1 265	31
4 050	45
1 730	50
4 350	60
3 300	66
3 600	72
4 900	70
3 600	90
5 760	100



b) Les résumés numériques

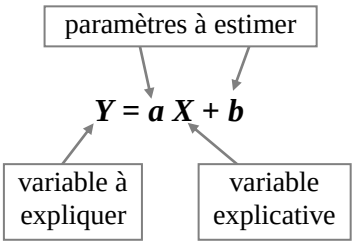
On obtient ici :

$\bar{X} = 60.8 \text{ m}^2$; $\bar{Y} = 3495 \text{ €}$; $\sigma(X) = 22.9 \text{ m}^2 \left(= \sqrt{\text{V}(X)}\right)$; $\sigma(Y) = 1322.9 \text{€} \quad \left(= \sqrt{\text{V}(Y)}\right)$

$\text{COV}(X,Y) = 22148.5 \text{ m}^2 \times \text{€}$; $\rho(X,Y) = 0.73$

On a une assez bonne corrélation linéaire positive. Il est possible d'utiliser un modèle linéaire pour expliquer le loyer mensuel par la superficie du local.

c) Le modèle algébrique



Estimation des paramètres par la méthode des moindres carrés des erreurs

Introduction des erreurs E_i

$$Y_i = aX_i + b + E_i \quad i = 1, \ldots, n$$

Les valeurs de a et de b qui rendent minimum la moyenne des carrés des erreurs :

$$f(a,b) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - aX_i - b)^2$$

sont : $\hat{a} = \dfrac{\text{COV}(X,Y)}{\text{V}(X)}$ et $\hat{b} = \bar{Y} - \hat{a}\bar{X}$

La droite d'équation $y = \hat{a}x + \hat{b}$ est appelée droite de régression linéaire de Y en X.

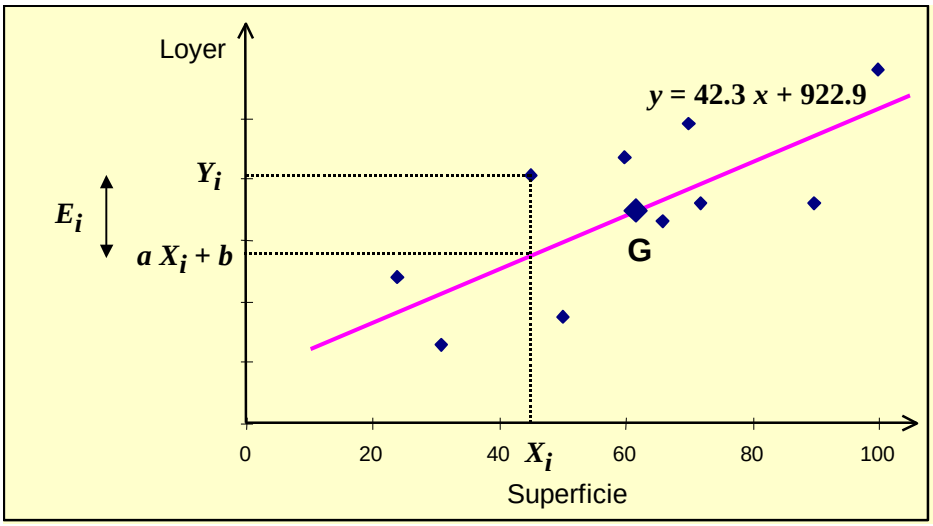
Ajustement du modèle et interprétation

La droite de régression linéaire de Y en X a pour équation :

$$y - 3495 = 42.3(x - 60.8) \text{ soit } y = 42.3x + 922.9$$

Le loyer mensuel moyen d'un local est de 3495 euros pour une superficie moyenne de 60.8 m^2 ;

- pour chaque m^2 en plus, le loyer augmente en moyenne de 42.3 euros
- pour chaque m^2 en moins, le loyer diminue en moyenne de 42.3 euros

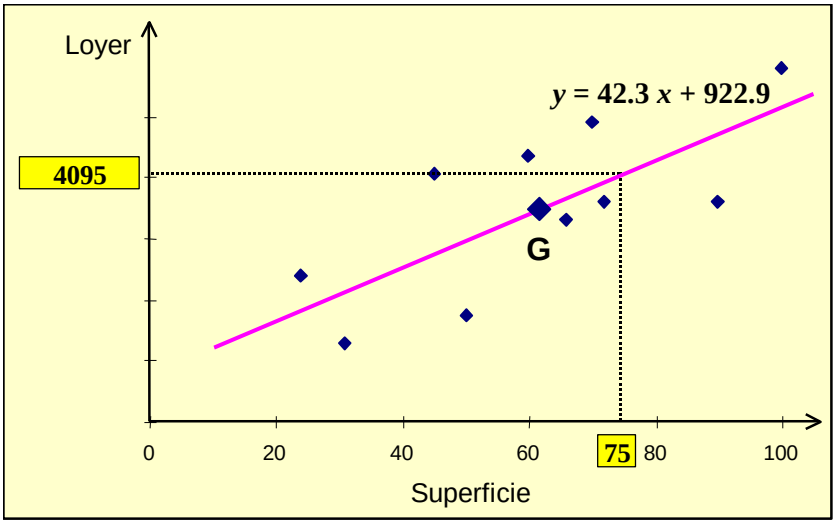


d) Prévision

Quel loyer prévoir pour un local de 75 m^2 ?

Le modèle ajusté est : $\hat{Y} = 42.3X + 922.9$

Pour X = 75 on a : $\hat{Y} = 42.3 \times 75 + 922.9 \cong 4095$ euros



e) Décomposition de la variance

On construit les deux variables :

$\hat{Y} = \hat{a}\,X + \hat{b}$ (le modèle) et

$\hat{E} = Y - \hat{Y}$ (l'erreur d'ajustement).

	***X* (Surface)**	***Y* (Prix)**	$\hat{Y} = \hat{a}\,X + \hat{b}$	$\hat{E} = Y - \hat{Y}$
	24	2 395	1 938.2	456.8
	31	1 265	2 234.3	-969.3
	45	4 050	2 826.6	1 223.4
	50	1 730	3 038.1	-1 308.1
	60	4 350	3 461.2	888.8
	66	3 300	3 715.0	-415.0
	72	3 600	3 968.8	-368.8
	70	4 900	3 884.2	1 015.8
	90	3 600	4 730.3	-1 130.3
	100	5 760	5 153.3	606.7
Moyenne	61	3 495	3 495	0
Variance	523.56	1 750 150	936 962.435	813 187 .565

On vérifie $\bar{\hat{Y}} = \bar{Y}$, $\bar{\hat{E}} = 0$, $\mathrm{COV}\left(\hat{Y}, \hat{E}\right) = 0$

et $$\boxed{\mathrm{V}(Y) = \mathrm{V}\left(\hat{Y}\right) + \mathrm{V}\left(\hat{E}\right)}$$

$R^2 = \dfrac{\mathrm{V}\left(\hat{Y}\right)}{\mathrm{V}(Y)}$ (= 0.535) est la part de variance expliquée par le modèle ;

$1 - R^2$ (= 0.465) est la part résiduelle.

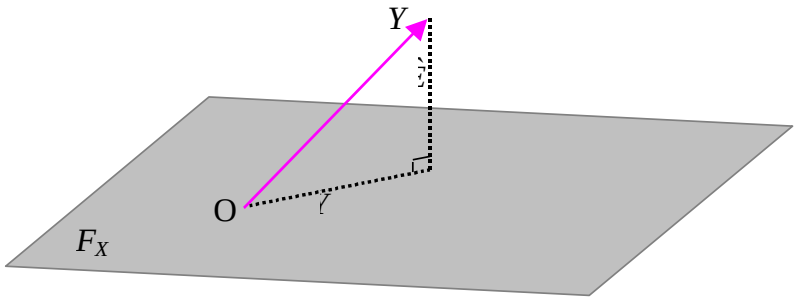
Interprétation géométrique

Dans $\mathbf{R}^n$ muni du produit-scalaire : $\langle x, y \rangle_n = \dfrac{1}{n}\sum_{i=1}^{n} x_i\, y_i$

$$f(a,b) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - a\,X_i - b\right)^2 = \left\| Y - a\,X - b\mathbf{1}_n \right\|_n^2$$

avec $Y = (Y_1, \ldots, Y_n)$, $X = (X_1, \ldots, X_n)$ et $\mathbf{1}_n = (1, \ldots, 1)$.

Soit $F_X = \left\{ a\,X + b\mathbf{1}_n \,\middle|\, (a,b) \in \mathbf{R}^2 \right\}$ le plan engendré par X et 1_n.



La recherche de a et b minimisant $f(a,b)$ est la recherche du vecteur de F_X le plus proche de Y.

La solution $\hat{Y} = \hat{a}\,X + \hat{b}\mathbf{1}_n$ est la projection orthogonale de Y sur F_X .

Généralisation

Plusieurs variables explicatives $\left(X_1,\ldots,X_p\right)$ → Modèle de régression linéaire multiple

Une variable explicative catégorielle à *p* modalités → Analyse de variance à un facteur

$\left(X_1,\ldots,X_p\right)$ les indicatrices des modalités

Plusieurs variables explicatives réelles ou catégorielles → Analyse de covariance

Coefficient de corrélation linéaire

$$\rho(X,Y)=\frac{\mathrm{COV}(X,Y)}{\sigma(X)\,\sigma(Y)} \qquad -1\le\rho\le 1$$

$$\left|\rho(X,Y)\right|=1 \iff \exists\, a\in\mathbf{R}_*^+ \text{ et } b\in\mathbf{R} \quad Y=aX+b$$

On a ici : $\rho(X,Y)=0.73 \quad \left(=\sqrt{R^2}\right)$

Origine historique du mot "régression"

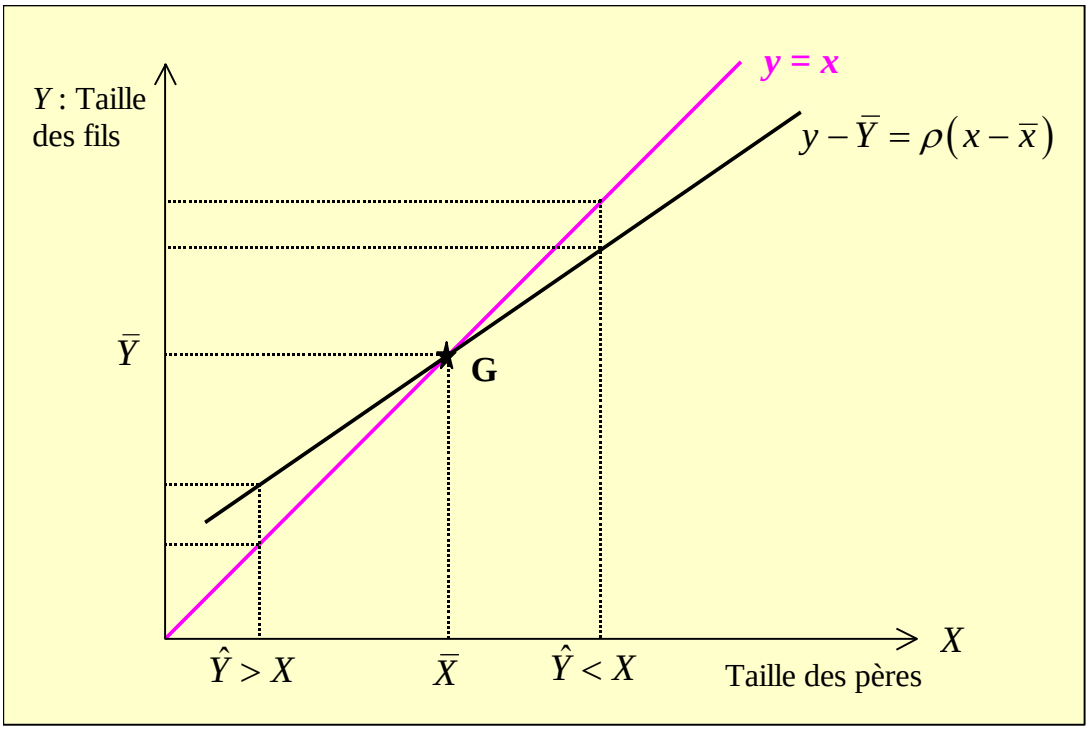
$$Y-\bar{Y}=\frac{\mathrm{COV}(X,Y)}{\mathrm{V}(X)}\left(X-\bar{X}\right)+E$$

Galton (1822, 1911) étudie la taille des fils (*Y*) en fonction de la taille des pères (*X*).

Bien que les moyennes (resp. les variances) des tailles des fils et des pères soient égales, il observe une *régression vers la moyenne*.

En effet, l'équation de la droite de régression linéaire de *Y* en *X* s'écrit alors :

$$y-\bar{Y}=\rho\left(x-\bar{X}\right) \text{ et } \rho<1$$



Applications aux séries chronologiques

Chiffre d'affaires trimestriel d'une entreprise (cf. annexe jeu de données 2)

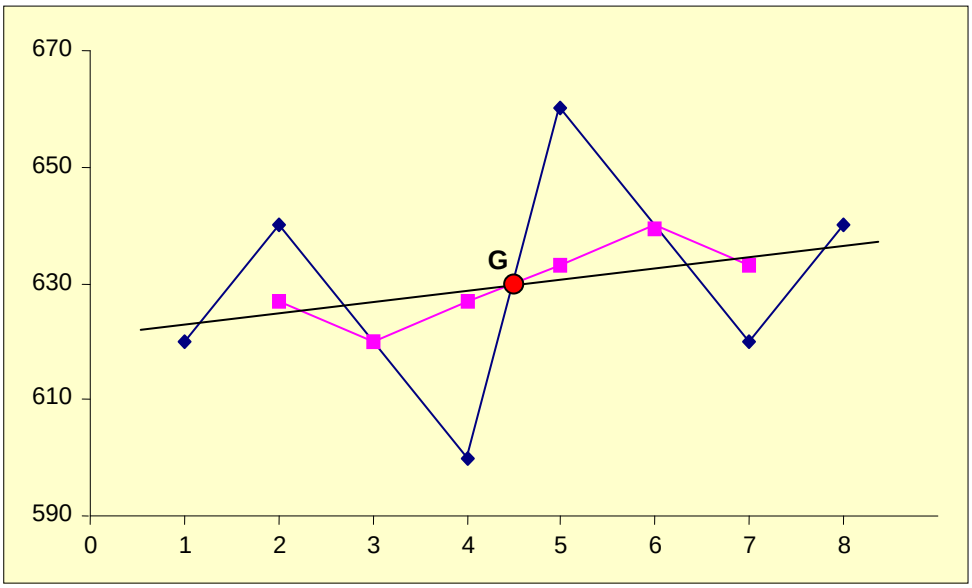
a) et b)

Le chiffre d'affaire (CA) trimestriel en KE d'une entreprise est observé sur 8 trimestres consécutifs. On a a priori une seule variable réelle, le CA, observée sur *n* individus ordonnés (les trimestres) mais pour étudier l'évolution de la variable en fonction du temps, on préfère considérer le temps comme une variable réelle explicative.

Soit *Y* le CA trimestriel en KE et *T* la variable temps (réindexé de 1 à 8). On représente les points (y_t, t) pour *t* = 1,…,8 dans un repère orthogonal.

Les moyennes mobiles d'ordre 3 sont données dans le tableau suivant et représentées sur le graphique.

Temps	1	2	3	4	5	6	7	8
CA	620	640	620	600	660	640	620	640
Moy. Mob.	/	627	620	627	633	640	633	/



Les moyennes mobiles sont une technique de lissage permettant de repérer la tendance d'une série chronologique ayant de fortes perturbations. Nous observons ici une légère croissance linéaire.

c) Soit $y = at + b$, *t* = 1,...,8, l'équation de la droite de régression du CA sur le temps.

Si *T* désigne la variable temps, on a $a = \frac{\mathrm{COV}(Y,T)}{\mathrm{V}(T)}$ et $b = \overline{Y} - a\overline{T}$.

La droite de régression linéaire du CA par rapport au temps passe par le point moyen (appelé aussi centre de gravité du nuage de points pondérés par les effectifs) $G(\overline{T}, \overline{Y})$. La droite est représentée sur le graphique.

Les calculs sont présentés dans le tableau suivant ; ils sont réalisés à partir du changement de variable $U = \frac{Y - 620}{20}$.

	t	y_t	u_t	t^2	u_t^2	$t\ u_t$
	1	620	0	1	0	0
	2	640	1	4	1	2
	3	620	0	9	0	0
	4	600	-1	16	1	-4
	5	660	2	25	4	10
	6	640	1	36	1	6
	7	620	0	49	0	0
	8	640	1	64	1	8
$\sum$	36		4	204	8	22
$(1/8)\sum$	4.5		0.5	25.5	1	2.75

On a : $\overline{T} = 4.5$, $\mathrm{V}(T) = 25.5 - (4.5)^2 = 5.25$, $\overline{U} = 0.5$, $\mathrm{V}(U) = 1 - (0.5)^2 = 0.75$, $\sigma(U) = 0.87$

$\mathrm{COV}(U,T) = 2.75 - (4.5)(0.5) = 0.5$

On en déduit : $\overline{Y} = 20\overline{U} + 620 = 630$, $\mathrm{V}(Y) = (20)^2 \mathrm{V}(U) = 300$, $\sigma(Y) = 17.32$

$\mathrm{COV}(Y,T) = 20\,\mathrm{COV}(U,T) = 10.$

Enfin on a : $a = \dfrac{\mathrm{COV}(Y,T)}{\mathrm{V}(T)} = \dfrac{10}{5.25} = 1.9$ et $b = \overline{Y} - a\,\overline{T} = 630 - 1.9 \times 4.5 = 621.43$

L'équation de la droite de régression linéaire du CA par rapport au temps est donc $y = 1.9t + 621.43$. La droite passe par le point moyen G de coordonnées (4.5, 630) et le coefficient directeur correspond à l'accroissement du CA pour une unité de temps, soit un accroissement de 1.9 KE en moyenne par trimestre.

d) La prévision du CA trimestriel pour le 3ème et le 4ème trimestre 2003 est obtenue à partir de la droite de régression linéaire du CA sur le temps, en posant $t = 9$ et $t = 10$.

Prévisions de Y: 3ème trimestre 03 $\hat{Y} = 1.9 \times 9 + 621.43 = 638.53$

4ème trimestre 03 $\hat{Y} = 1.9 \times 10 + 621.43 = 640.43$

Ces prévisions sont faites sans tenir compte des fortes fluctuations d'un trimestre à l'autre. L'augmentation de 1.9 KE en moyenne par trimestre est faible par rapport à ces fluctuations. Le coefficient de corrélation linéaire entre le CA et le temps et de 0.25 ce qui montre une corrélation positive mais bien faible. Les prévisions sont donc à prendre avec précaution.

Ventes trimestrielles d'un produit (cf. annexe jeu de données 3)

t	$Y(t)$	$Y(t)-Y(t-1)$	$Y(t) / Y(t-1)$
1	255	/	/
2	330	75	1.29
3	435	105	1.32
4	570	130	1.31
5	740	170	1.30
6	960	220	1.30

b/ Le modèle exponentiel est préférable au modèle linéaire car les rapports $Y(t) / Y(t-1)$ sont à peu près constants.

On pose donc : $Y(t) = b\,a^t$

On pose $X(t) = \ln Y(t)$, $A = \ln a$, $B = \ln b$

On a alors : $X(t) = At + B$ avec $A = \dfrac{\mathrm{COV}(X,T)}{\mathrm{V}(T)}$ et $B = \overline{X} - A\overline{T}$

t	$X(t)$	$\overline{X}'(t)=\ln X(t)$	$t\,X'(t)$
1	255	5.64	5.54
2	330	5.80	11.60
3	435	6.08	18.24
4	570	6.35	25.40
5	740	6.61	33.05
6	960	6.87	41.22
Σ		37.25	135.05
$(1/6)\,\Sigma$		6.21	22.51

On a :

$\overline{T}=\dfrac{6+1}{2}=3.5 \quad \mathrm{V}(T)=\dfrac{6^2-1}{12}=2.92$

$\overline{X}=6.21$

$COV(X,T)=22.51-(3.5)(6.21)=0.775$

$A=\dfrac{0.775}{2.92}=0.265$ d'où $a=\mathrm{e}^A=1.3$

$B=6.21-(0.265)(3.5)=5.28$ d'où $b=\mathrm{e}^B=197$

Il vient finalement : $Y(t)=197(1.3)^t$

Les prévisions pour les trimestres 7 et 8 sont :
$Y(7)=197(1.3)^7=1236$ unités vendues et $Y(8)=197(1.3)^8=1607$ unités vendues.

Quelques propriétés

Définitions de la covariance et du coefficient de corrélation linéaire.
Rappel des propriétés usuelles de ces indices.

- ***Deuxième formule de la covariance***

Données brutes

$$\mathrm{COV}(X,Y)=\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\overline{X}\right)\left(Y_k-\overline{Y}\right)=\frac{1}{n}\sum_{k=1}^{n}\left(X_kY_k-\overline{X}Y_k-X_k\overline{Y}+\overline{X}\,\overline{Y}\right)=\frac{1}{n}\sum_{k=1}^{n}X_kY_k-\overline{X}\,\overline{Y}$$

Distribution de fréquences

$$\mathrm{COV}(X,Y)=\sum_{i=1}^{I}f_i\left(x_i-\overline{X}\right)\left(y_i-\overline{Y}\right)=\sum_{i=1}^{I}f_i\left(x_iy_{i_k}-\overline{X}y_i-x_i\overline{Y}+\overline{X}\,\overline{Y}\right)=\sum_{i=1}^{I}f_ix_iy_i-\overline{X}\,\overline{Y}$$

- ***Propriétés de la covariance***

$\mathrm{COV}(X,Y)=\dfrac{1}{n}\sum_{k=1}^{n}\left(X_k-\overline{X}\right)\left(Y_k-\overline{Y}\right)=\mathrm{COV}(Y,X)$, la covariance est symétrique

$$\mathrm{COV}(aX+bY,Z)=\frac{1}{n}\sum_{k=1}^{n}\left(aX_k+bY_k-a\overline{X}-b\overline{Y}\right)\left(Z_k-\overline{Z}\right)$$

$$=a\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\overline{X}\right)\left(Z_k-\overline{Z}\right)+b\frac{1}{n}\sum_{k=1}^{n}\left(Y_k-\overline{Y}\right)\left(Z_k-\overline{Z}\right)=a\,\mathrm{COV}(X,Z)+b\,\mathrm{COV}(Y,Z)$$

la covariance est linéaire par rapport au premier argument ; étant symétrique la covariance est linéaire par rapport au deuxième argument et donc bilinéaire.
$\mathrm{COV}(X,X)=\mathrm{V}(X)\geq 0$ la covariance est positive et $V(X)=0$ si et seulement si X est constante.

- ***Propriétés du coefficient de corrélation linéaire***

1) On a $\rho(X,Y)=\dfrac{\mathrm{COV}(X,Y)}{\sigma(X)\sigma(Y)}=\dfrac{\mathrm{COV}(Y,X)}{\sigma(Y)\sigma(X)}=\rho(Y,X)$.

Le coefficient de corrélation linéaire est symétrique.

2) Pour tout réel a on a vu que l'on avait :

$\mathrm{V}(Y - aX) = a^2\ \mathrm{V}(X) - 2a\,\mathrm{COV}(X,Y) + \mathrm{V}(Y)$ (= $f(a)$)

Comme une variance est positive, pour tout a, le trinôme du second degré $f(a)$ est positif.

On en déduit que le discriminant $\mathrm{COV}^2(X,Y) - \mathrm{V}(X)\mathrm{V}(Y)$ est négatif ou nul,

c-à-d $\mathrm{COV}^2(X,Y) \le \mathrm{V}(X)\,\mathrm{V}(Y)$, $\dfrac{\mathrm{COV}^2(X,Y)}{\mathrm{V}(X)\mathrm{V}(Y)} \le 1$, $\left|\dfrac{\mathrm{COV}(X,Y)}{\sigma(X)\sigma(Y)}\right| \le 1$, $|\rho(X,Y)| \le 1$,

soit $-1 \le \rho(X,Y) \le 1$

Par ailleurs $f(a)$ atteint son minimum pour $a = \dfrac{\mathrm{COV}(X,Y)}{\mathrm{V}(X)}$ et, pour cette valeur de a,

on obtient : $f(a) = \dfrac{\mathrm{V}(X)\mathrm{V}(Y) - (\mathrm{COV}(X,Y))^2}{\mathrm{V}(X)}$; on en déduit :

$\mathrm{V}(Y - aX) = 0 \Leftrightarrow \exists b \in \mathbf{R}, Y = aX + b \Leftrightarrow \mathrm{V}(X)\mathrm{V}(Y) = [\mathrm{COV}(X,Y)]^2 \Leftrightarrow |\rho(X,Y)| = 1$

3) On peut préciser :

$\rho(X,Y) = 1 \Leftrightarrow Y = aX + b, \quad a > 0$

$\rho(X,Y) = -1 \Leftrightarrow Y = aX + b, \quad a < 0$

- ***Coefficients de la droite de régression linéaire de Y en X***

Montrer que la fonction f de deux variables réelles a et b définie ci-après est minimum pour $a = \dfrac{\mathrm{COV}(X,Y)}{\mathrm{V}(X)}$ et $b = \overline{Y} - a\overline{X}$

$$f(a,b) = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - aX_k - b\right)^2 = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y} - aX_k + a\overline{X} + \overline{Y} - a\overline{X} - b\right)^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y} - aX_k + a\overline{X}\right)^2 + \frac{2}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y} - aX_k + a\overline{X}\right)\left(\overline{Y} - a\overline{X} - b\right) + \frac{1}{n}\sum_{k=1}^{n}\left(\overline{Y} - a\overline{X} - b\right)^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \overline{Y} - aX_k + a\overline{X}\right)^2 + \left(\overline{Y} - a\overline{X} - b\right)^2$$ (car le double produit est nul).

C'est la somme de 2 termes positifs dont le second est minimum (car nul) pour $b = \overline{Y} - a\overline{X}$

Il s'agit donc de trouver a minimisant le premier terme ; il s'écrit :

$$\frac{1}{n}\sum_{k=1}^{n}[\left(Y_k - \overline{Y}\right)^2 - 2a\left(Y_k - \overline{Y}\right)\left(X_k - \overline{X}\right) + a^2\left(X_k - \overline{X}\right)^2]$$

c'est-à-dire : $a^2\ \mathrm{V}(X) - 2a\,\mathrm{COV}(X,Y) + \mathrm{V}(Y)$

Ce trinôme du second degré en a admet un minimum pour $a = \dfrac{\mathrm{COV}(X,Y)}{\mathrm{V}(X)}$ d'où la solution.

Annexe : jeu de données

1) Loyer mensuel de locaux commerciaux

Le loyer mensuel en euros et la superficie en m^2 de 10 locaux commerciaux d'un quartier sont donnés dans le tableau suivant :

Loyer mensuel (en €)	Superficie (en m^2)
2 395	24
1 265	31
4 050	45
1 730	50
4 350	60
3 300	66
3 600	72
4 900	70
3 600	90
5 760	100

a) Décrire les données et représenter les dans un repère cartésien, la superficie en abscisse, le loyer en ordonnée.

b) Calculer moyenne et écart-type de chacune des deux variables, covariance et coefficient de corrélation linéaire des deux variables. Commenter.

c) Calculer les coefficients *a* et *b* de la droite de régression linéaire, d'équation $y = a\,x + b$, du loyer *Y* par rapport à la superficie *X*. Représenter sur le graphique le point moyen et cette droite de régression linéaire. Interpréter le coefficient directeur de la droite.

d) Donner une prévision du loyer mensuel net en euro d'un local commercial du quartier de 75 m^2.

e) On note $\hat{Y} = aX + b$ le modèle et $\widehat{E} = X - \hat{Y}$ l'erreur.
Calculer les valeurs de ces variables sur les 10 observations.
Calculer moyenne, variance, écart-type de chacune de ces deux variables, covariance et coefficient de corrélation linéaire des deux variables.
Écrire le carré du coefficient de corrélation linéaire de *X* et *Y* en fonction des variances de $\hat{Y}$ et de *Y*. Commenter. (On renvoie à la correction pour des compléments sur la régression linéaire).

2) Chiffre d'affaires (CA) trimestriel d'une entreprise

Le CA en milliers d'euros d'une entreprise est observé sur deux années consécutives.

3° Trim 2001	4° Trim 2001	1° Trim 2002	2° Trim 2002	3° Trim 2002	4° Trim 2002	1° Trim 2003	2°Trim 2003
620	640	620	600	660	640	620	640

a) Décrire les données et représenter les dans un repère cartésien, le temps (noté *T*) en abscisse et le CA (noté *Y*) en ordonnée. Joignez les points consécutifs.

b) Calculer les moyennes mobiles d'ordre 3. Représenter les points correspondants sur le graphique. Commenter.

c) Calculer les coefficients *a* et *b* de la droite de régression linéaire, d'équation $y = a\,t + b$, du CA *Y* par rapport au temps *T* (on indexera de 1 à 8 les valeurs de *T*). Représenter sur le graphique le point moyen et cette droite de régression linéaire. Interpréter le coefficient directeur de la droite.

d) Donner une prévision du CA pour les $3^{ème}$ et $4^{ème}$ trimestres de l'année 2003. Commenter.

3) Ventes trimestrielles d'un produit

L'effectif des ventes d'un produit est donné dans le tableau suivant.

Trim.	1	2	3	4	5	6
Quantité vendue	255	330	435	570	740	960

a) Décrire les données et représenter les dans un repère cartésien, le temps (noté T) en abscisse et le nombre de produits vendus (noté Y) en ordonnée. Joignez les points consécutifs.

b) Pour t = 2 à 6, calculer $Y(t) - Y(t-1)$ et $Y(t)/Y(t-1)$. Entre le modèle linéaire $Y(t) = at + b$ et le modèle exponentiel $Y(t) = ba^t$, lequel pensez-vous le mieux adapté à cette série chronologique ?

c) Ajuster le modèle choisi en b) et prévoir les trimestres 7 et 8.