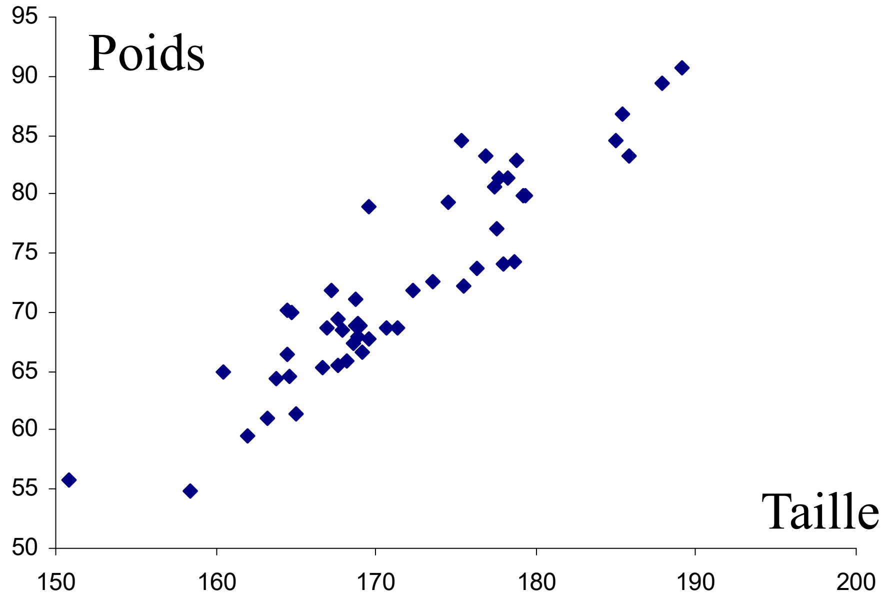
ETUDE DE 2 VARIABLES QUANTITATIVES

(1) MESURE DE LA LIAISON ENTRE 2 VARIABLES QUANTITATIVES

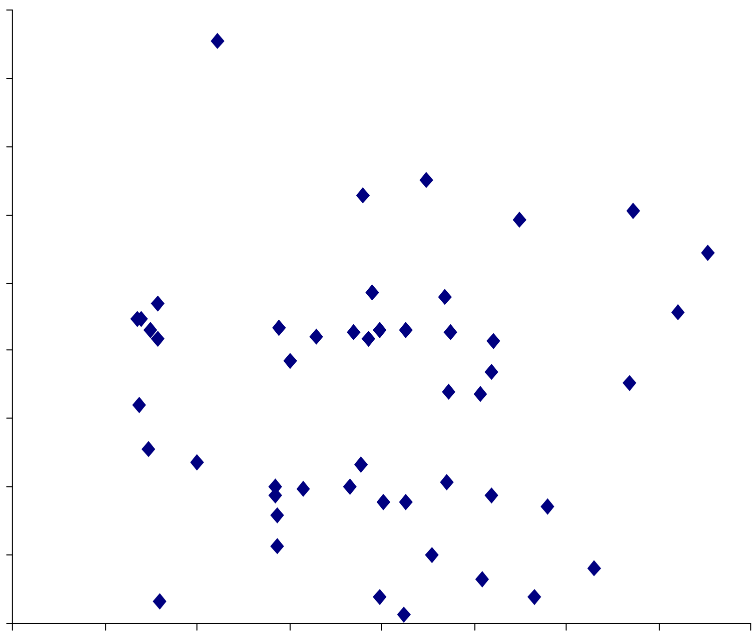
Nom	Taille x_i (cm)	Poids y_i (kg)
Pierre	175	73
Arantxa	168	56
…..	…..	…..
Martin	185	87



La connaissance de la taille x apporte une certaine information sur le poids y

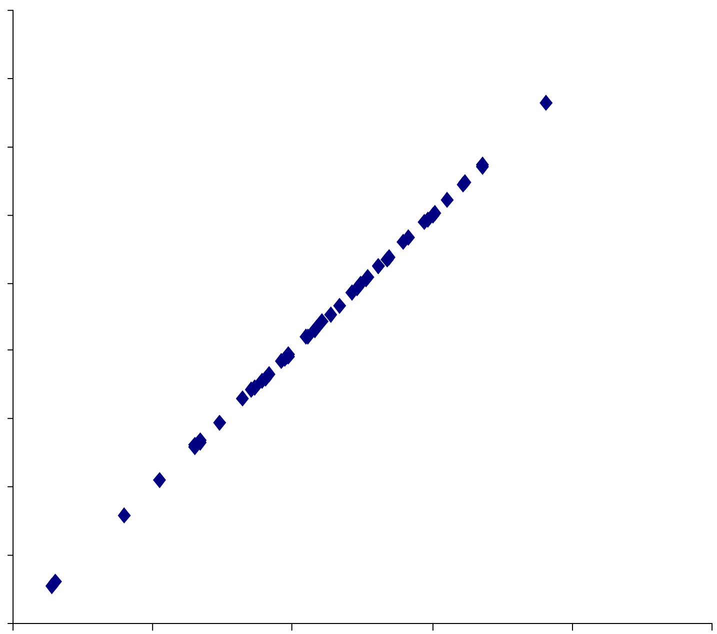
Il existe une relation de dépendance entre x et y

(2) MESURE DE LA LIAISON ENTRE 2 VARIABLES QUANTITATIVES



La connaissance de x n'apporte aucune certaine information sur y

x et y sont indépendantes



La connaissance de x permet de connaître exactement la valeur de y

Il existe une relation fonctionnelle entre x et y

(3) MESURE DE LA LIAISON ENTRE 2 VARIABLES QUANTITATIVES

Covariance : $$\mathrm{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Propriétés :

$\mathrm{Cov}(x,y) > 0 \Leftrightarrow$ x et y varient dans le même sens

$\mathrm{Cov}(x,y) < 0 \Leftrightarrow$ x et y varient en sens contraire

$$\mathrm{Cov}(x,y) = \mathrm{Cov}(y,x)$$

$$\mathrm{Cov}(x,x) = V(x)$$

$$\mathrm{Cov}(a\,x + b\,y\,,\,z) = a\,\mathrm{Cov}(x,z) + b\,\mathrm{Cov}(y,z)$$

(4) MESURE DE LA LIAISON ENTRE 2 VARIABLES QUANTITATIVES

Corrélation linéaire: $\rho = \dfrac{\mathrm{cov}(x,y)}{\sigma(x)\,\sigma(y)}$

Propriétés :

$$-1 \leq \rho \leq 1$$

$$y = a\,x + b \;\Leftrightarrow\; \begin{cases} \rho = 1 & \text{si} \quad a > 0 \\ \rho = -1 & \text{si} \quad a < 0 \end{cases}$$

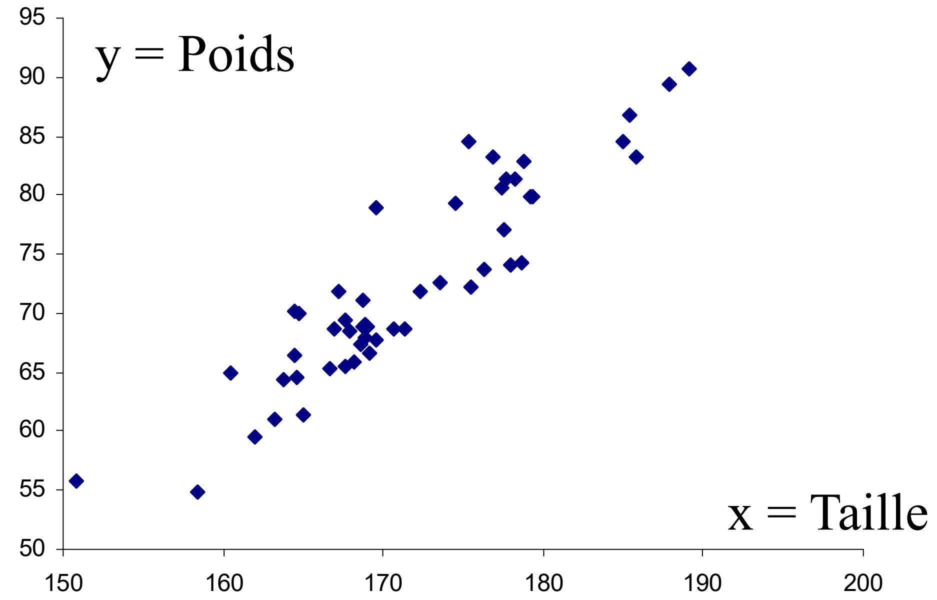
$|\rho| = 1 \;\Leftrightarrow\;$ Il existe une relation fonctionnelle entre x et y

$\rho = 0 \;\Leftrightarrow\;$ x et y sont indépendantes

$0 < |\rho| < 1 \;\Leftrightarrow\;$ Il existe une dépendance linéaire d'autant plus forte que $|\rho|$ est grand

Ne pas confondre causalité et corrélation

(1) AJUSTEMENT LINEAIRE

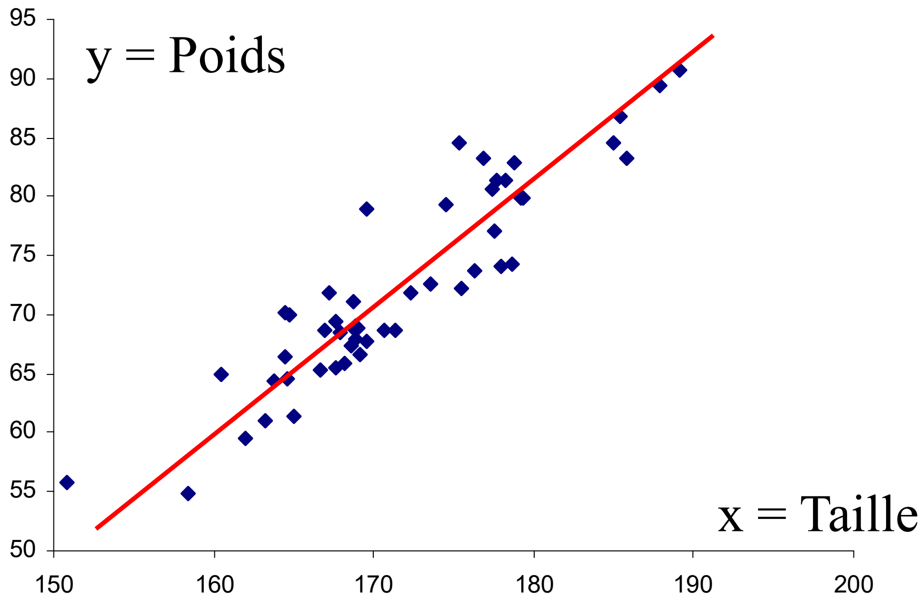


Est-il possible de trouver une fonction numérique f telle que y = f (x) ?

Si une telle fonction existe, on dit que f est un modèle du phénomène étudié.

x est la variable explicative.
y est la variable expliquée.

(2) AJUSTEMENT LINEAIRE



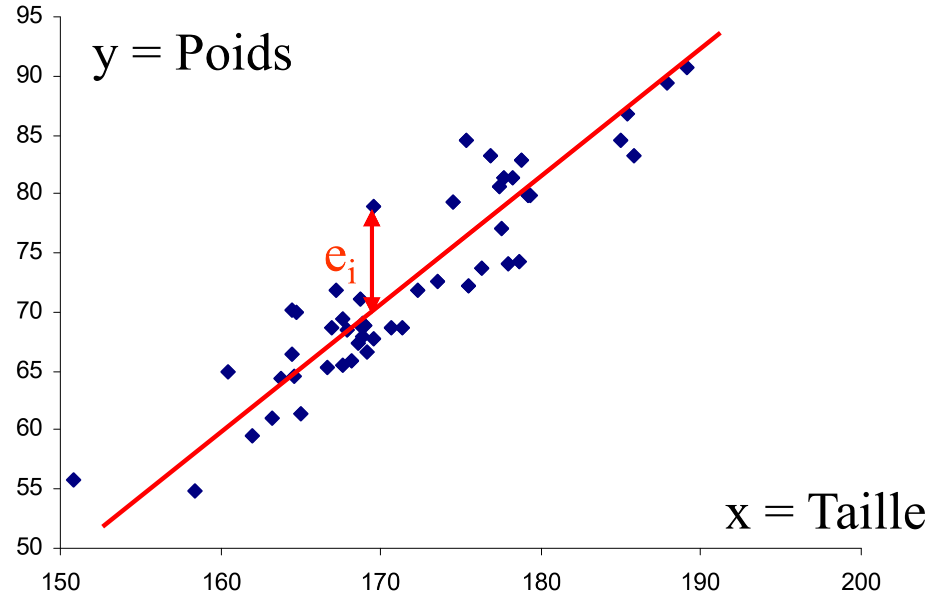
On désire trouver la droite qui passe « au mieux » à l'intérieur du nuage de points

(3) AJUSTEMENT LINEAIRE

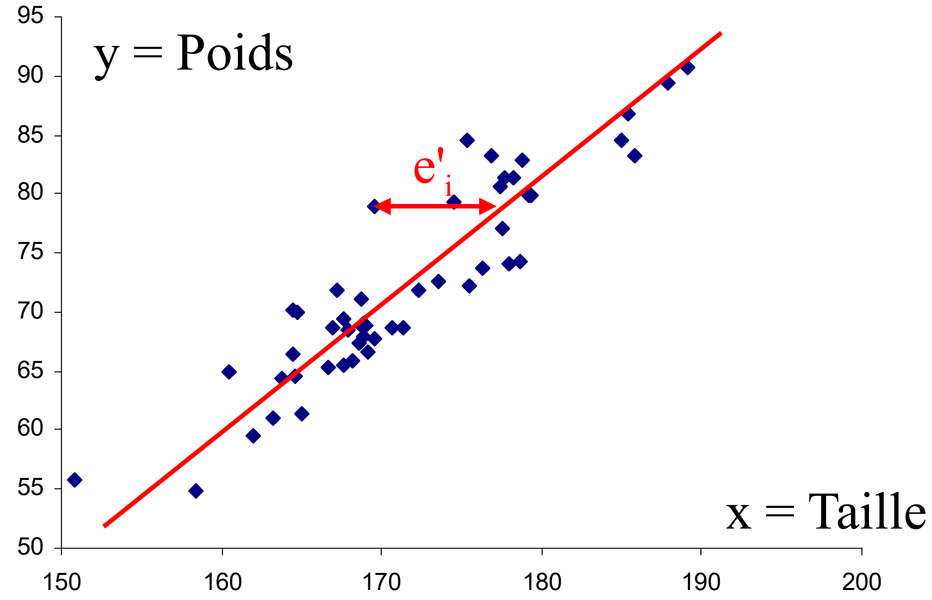
« au mieux »

Minimiser $S = \sum_{i=1}^{n} e_i^2$

Minimiser $S' = \sum_{i=1}^{n} e'^2_i$



Droite de régression de y en x

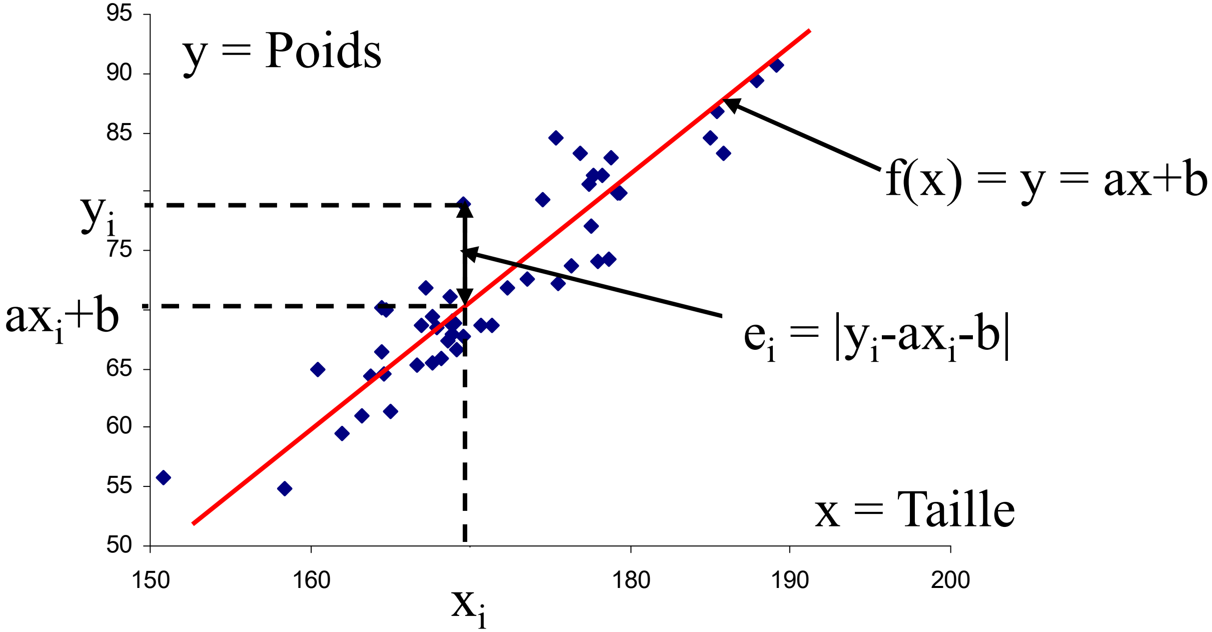


Droite de régression de x en y

(4) AJUSTEMENT LINEAIRE

REGRESSION LINEAIRE DE Y EN X

Droite de régression linéaire de y en x
$y = f(x) = ax + b$



La droite de régression linéaire de y en x, notée $D_{y/x}$, minimise $S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

$$a = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{Cov(x,y)}{V(x)}$$

$$b = \bar{y} - a\bar{x}$$

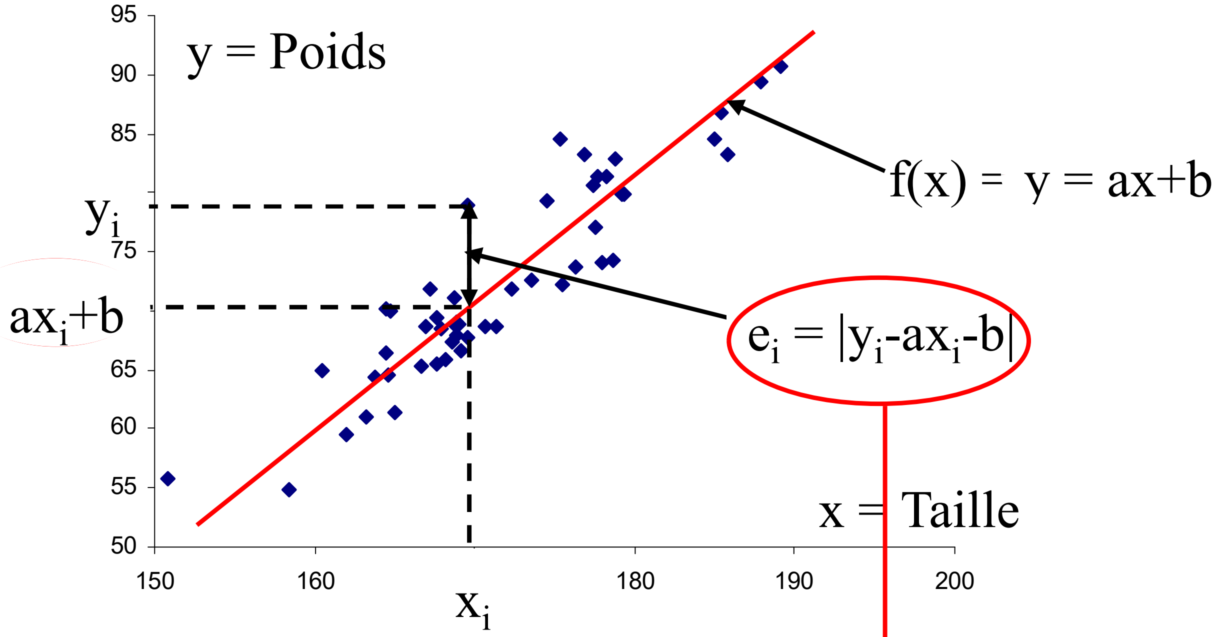
$D_{y/x}$ passe par le point moyen $(\bar{x}, \bar{y})$

(5) AJUSTEMENT LINEAIRE

REGRESSION LINEAIRE DE Y EN X

Droite de régression linéaire de y en x

$y = f(x) = ax + b$



$y = a\,x + b$ définit un modèle affine

$\hat{y}_i = a\,x_i + b$ = valeur de y_i prévue par le modèle

$r_i = y_i - \hat{y}_i$ = résidu de la ième observation

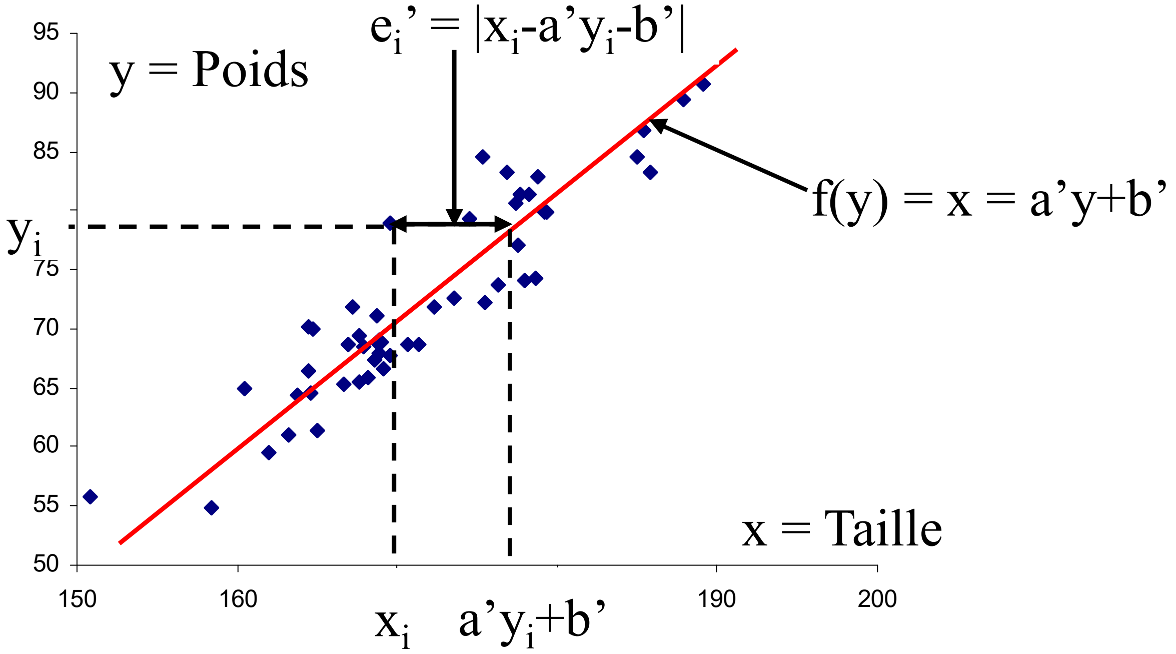
$e_i = |r_i| = |y_i - a\,x_i - b|$ = erreur due au modèle

(6) AJUSTEMENT LINEAIRE

REGRESSION LINEAIRE DE X EN Y

Droite de régression linéaire de x en y

$x = f(y) = a'y + b'$



La droite de régression linéaire de x en y, notée $D_{x/y}$, minimise $S' = \sum_{i=1}^{n} e'^2_i = \sum_{i=1}^{n}(x_i - a'y_i - b')^2$

$$a' = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{Cov(x,y)}{V(y)}$$

$$b' = \bar{x} - a'\,\bar{y}$$

$D_{x/y}$ passe par le point moyen $(\bar{x}\,,\,\bar{y})$

LIENS ENTRE CORRELATION ET DROITES DE REGRESSION

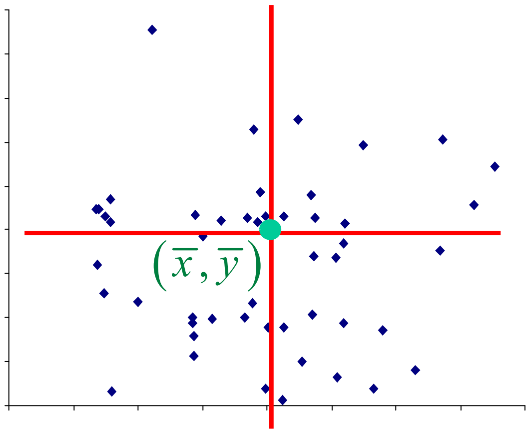
$D_{y/x}$: $y = ax + b$ $\quad a = \dfrac{\mathrm{Cov}(x,y)}{V(x)}$ $\quad b = \bar{y} - a\bar{x}$

$D_{x/y}$: $x = a'y + b'$ $\quad a' = \dfrac{\mathrm{Cov}(x,y)}{V(y)}$ $\quad b' = \bar{x} - a'\bar{y}$

$\Leftrightarrow y = \dfrac{1}{a'}x - \dfrac{b'}{a'}$

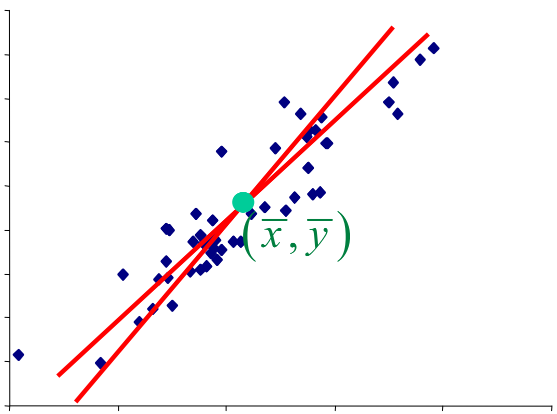
$\rho^2 = a\,a'$

$\rho = a\dfrac{\sigma(x)}{\sigma(y)} = a'\dfrac{\sigma(y)}{\sigma(x)}$



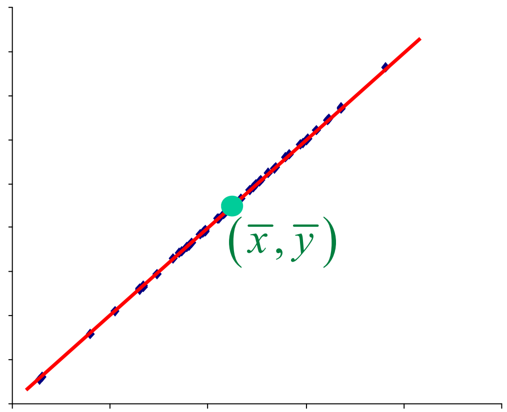
$\rho^2 = a\,a' = 0$

Indépendance linéaire



$0 < \rho^2 = a\,a' < 1$

Le degré de dépendance linéaire se mesure à la proximité des droites de régression

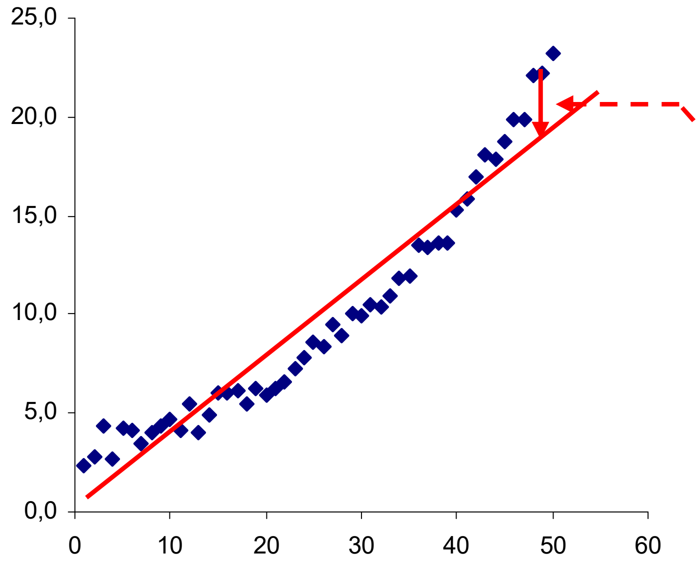


$\rho^2 = a\,a' = 1$

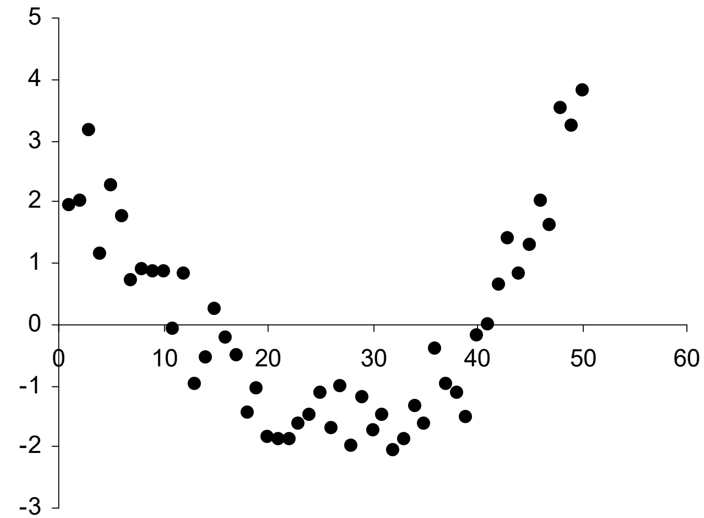
Liaison fonctionnelle linéaire

(1) AJUSTEMENT A UNE FONCTION EXPONENTIELLE

x_i	y_i
2,8	0,8
4,3	1,2
2,7	1,5
4,2	1,9
4,1	2,3
….	….
4,0	3,1



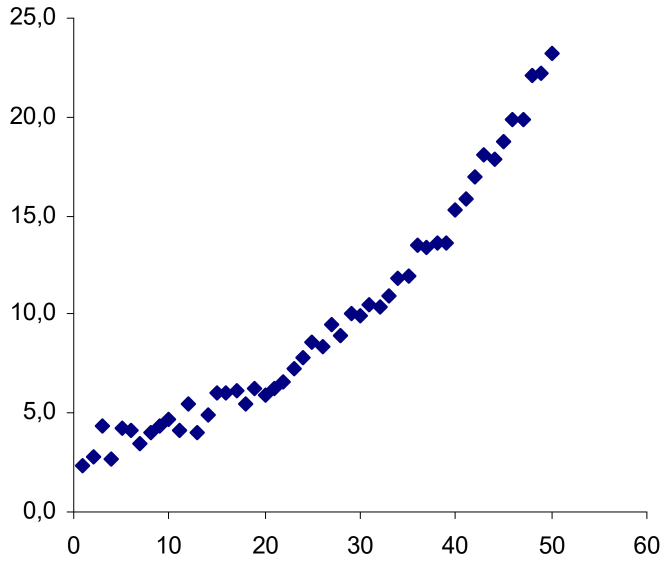
droite de régression linéaire
de y en x



Analyse des résidus

Les résidus devraient se répartir au hasard autour de l'axe des abscisses:
le modèle affine ne convient pas

(2) AJUSTEMENT A UNE FONCTION EXPONENTIELLE



Modèle exponentiel

$y = e^x$ exponentielle de base e

$y = a^x$ exponentielle de base a

$y = b\, a^x$ Forme exponentielle générale

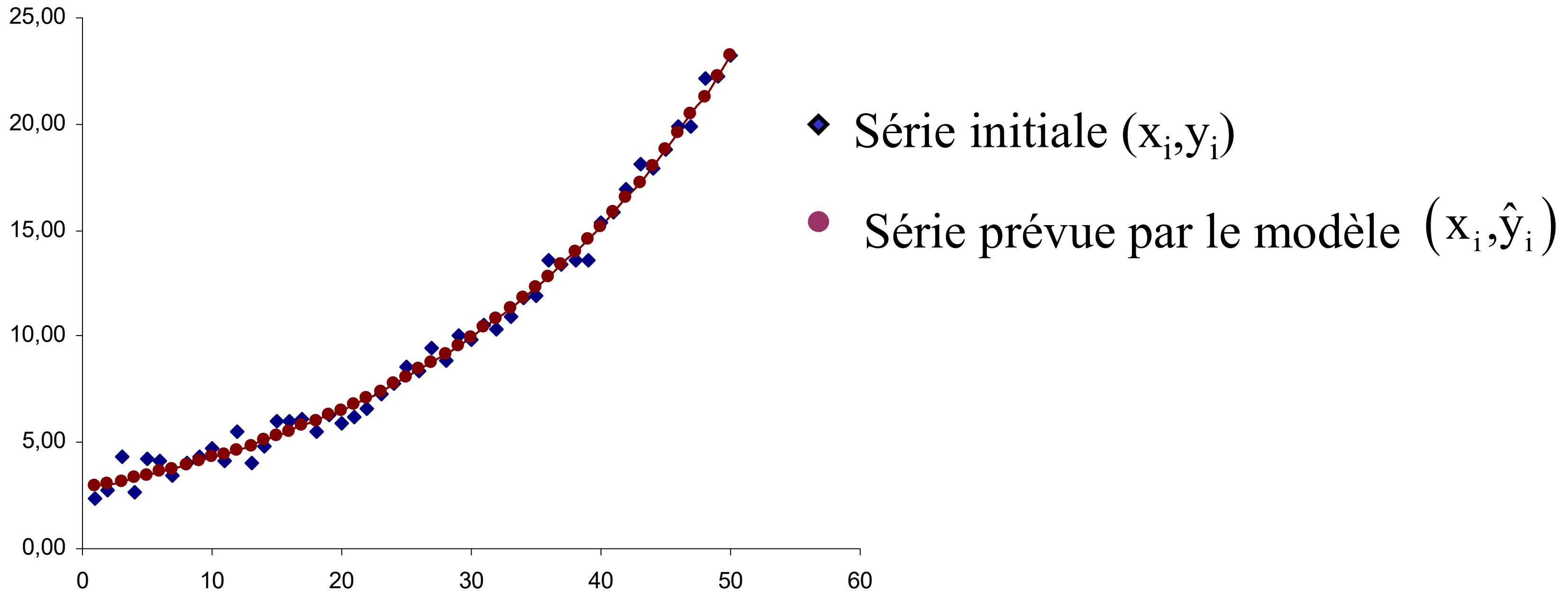
Changement de variable

$\ln y = \ln b + x \ln a$

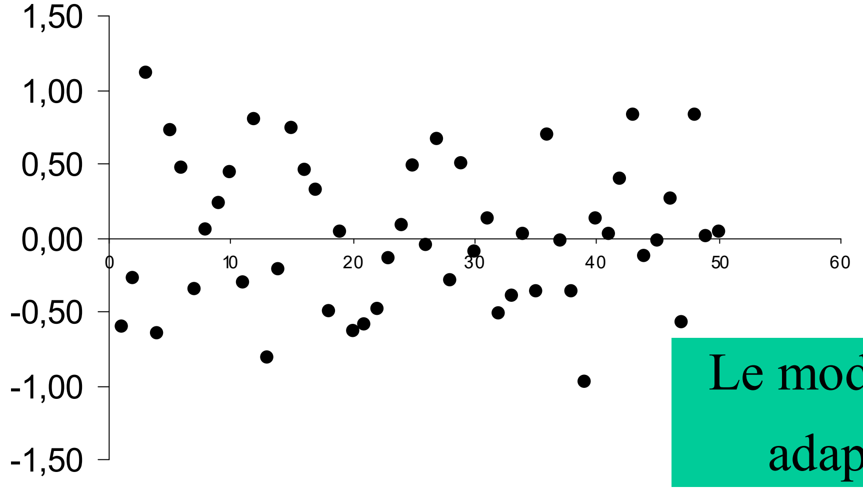
$Y = A X + B$ avec $Y = \ln y$
$X = x$
$A = \ln a$
$B = \ln b$

L'ajustement affine de Y en fonction de X donne A et B,
d 'où $a = e^A$, $b = e^B$, et le modèle $y = b\, a^x$

(3) AJUSTEMENT A UNE FONCTION EXPONENTIELLE

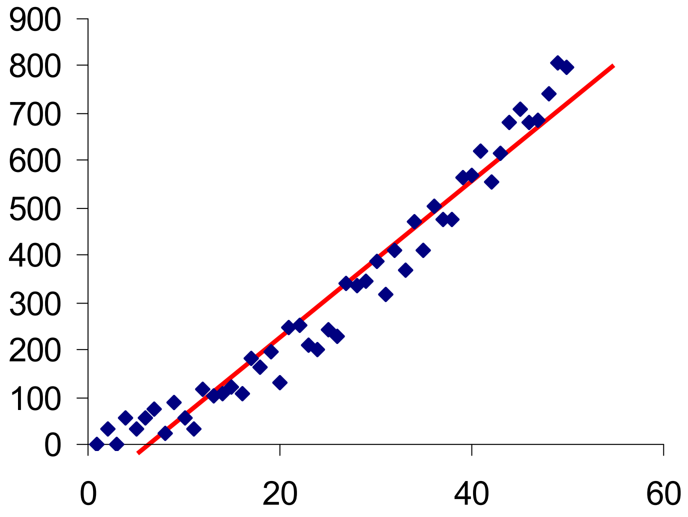


Analyse des résidus

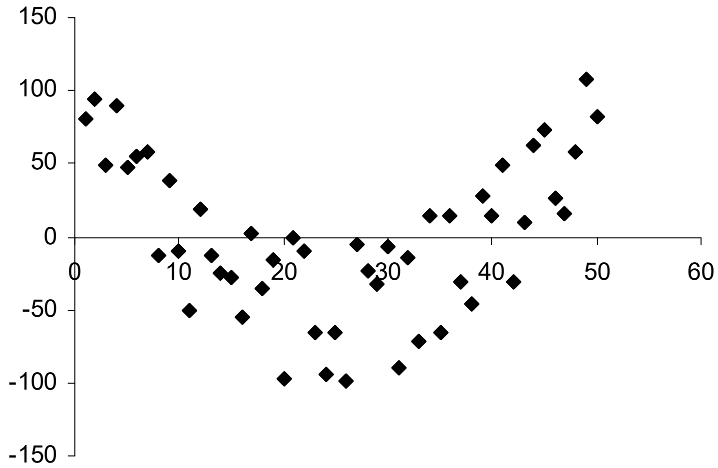


Le modèle exponentiel est mieux adapté que le modèle affine

(1) AJUSTEMENT A UNE FONCTION PUISSANCE



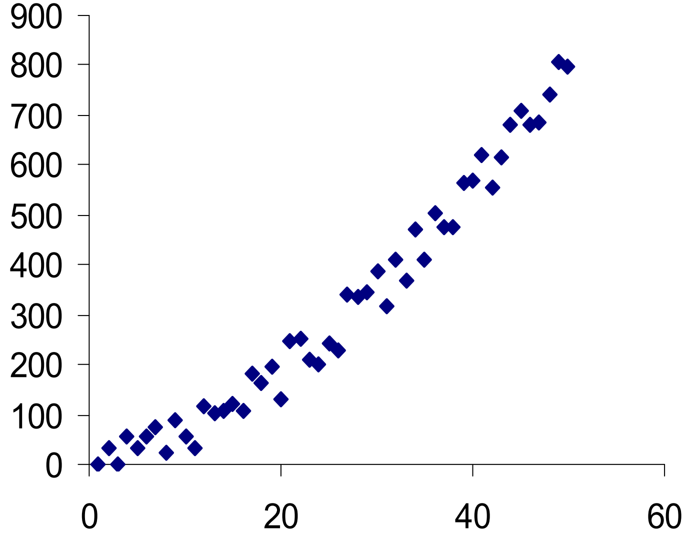
Droite de régression linéaire de y en x



Analyse des résidus

Le modèle affine ne convient pas

(2) AJUSTEMENT A UNE FONCTION PUISSANCE



Modèle puissance $y = b\,x^a$

Changement de variable

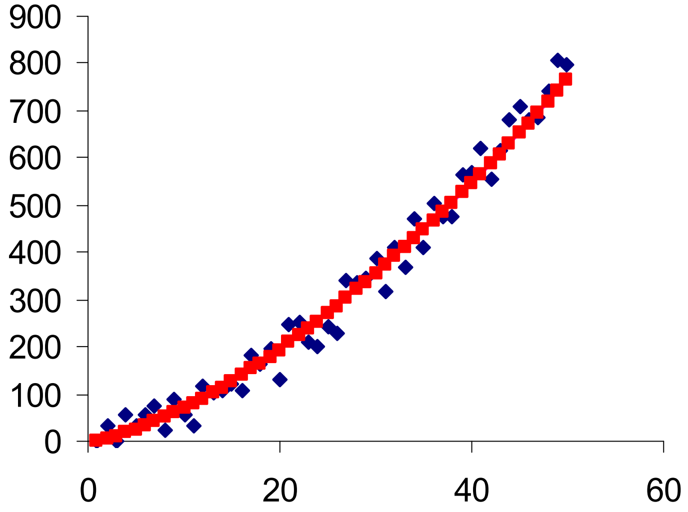
$$\ln y = \ln b + a \ln x$$

$Y = A\,X + B$ avec

$Y = \ln y$
$X = \ln x$
$A = a$
$B = \ln b$

L'ajustement affine de Y en fonction de X donne A et B,
d 'où $a = A$, $b = e^B$, et le modèle $y = b\,x^a$

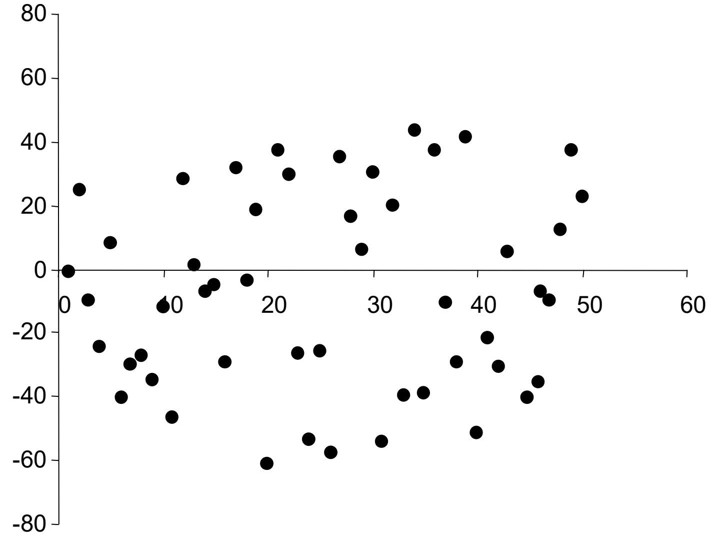
(3) AJUSTEMENT A UNE FONCTION PUISSANCE



◆ Série initiale (x_i,y_i)

● Série prévue par le modèle $\left(x_i,\hat{y}_i\right)$

Analyse des résidus



Le modèle puissance est mieux adapté que le modèle affine

QUALITE D'UN AJUSTEMENT

On montre que $$\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i-\bar{y})^2 + \sum(y_i-\hat{y}_i)^2$$

$$SCT = SCM + SCR \quad \Leftrightarrow \quad 1 = \frac{SCM}{SCT} + \frac{SCR}{SCT}$$

Somme des carrés des écarts à la moyenne = Somme des carrés des écarts du modèle + Somme des carrés des résidus

L'ajustement est d'autant meilleur que SCR est proche de 0, c.à.d. que SCR/SCT est proche de 0 ou SCM/SCT est proche de 1.

$$R = \frac{SCM}{SCT} = \text{Coefficient de détermination} = \rho^2 = (\text{coef. de corrélation})^2$$

= proportion de la variation totale due à l'ajustement

$$0 \leq R \leq 1$$