Lien entre deux variables quantitatives:

individu $(x_i ; y_i)$

	X	Y	$\hat{Y}$	ε
1	x_1	y_1		
i	x_i	y_i		
⋮	⋮	⋮		
n	x_n	y_n		

avec $\hat{Y} = bx_i + a$

$\varepsilon = Y_i - \hat{Y}_i$

- **Les calculs:**

$\bar{x}$; $\bar{y}$ s^2x, s^2y

covariance : indicateur entre deux variables

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

si on calcule la covariance entre x et x on trouve la variance des x.

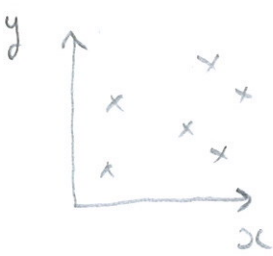
$$s_{xy} = \frac{1}{n} \sum x_i y_1 - \bar{x}\bar{y}$$

La covariance peut être négative

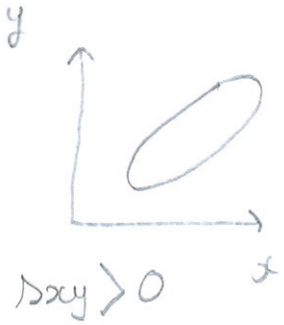
Le signe de la covariance indique la relation entre x et y.

- **graphique de dispersion:**

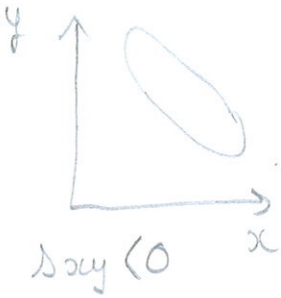
nuage de points.



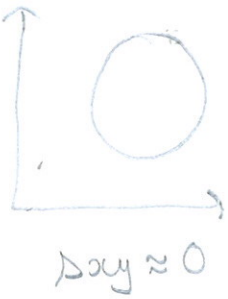
Le sens du nuage de point indique le lien entre x et y.



lorsque x et y ↑, liaison positive entre x,y.



quand $x \nearrow$ et $y \searrow$ → liaison négative entre x et y.

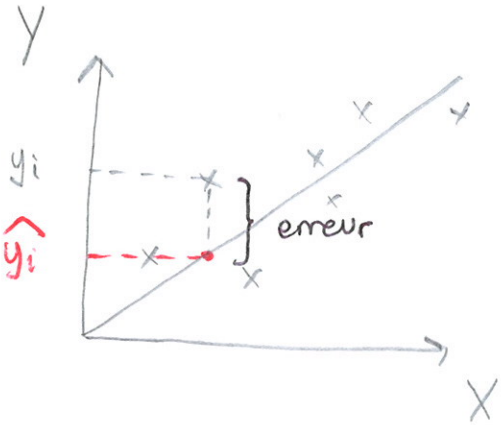


Si ce n'est pas linéaire il faut rechercher une équation (régression).

- <u>Régression linéaire :</u>

Y variable à expliquer

X variable explicative : variable fixe régresse.



il faut chercher une fonction telle que $Y = f(X)$

Il n'y aura pas une fonction qui passe par tout les points : on cherche une approximation qui passe par le maximum de points.

$Y = bx + a$: on prend en compte les distances verticales (car X fixe) entre les points et la courbe. C'est l'erreur : $\underline{\varepsilon_i = y_i - \hat{y_i}}$

$Y = bX + a$ telle que $\sum \varepsilon_i^2$ soit minimum (moindre carré ordinaire MCO)

$$S(b,a) = \sum (y_i - bx_i - a)^2$$

$$\begin{cases} \frac{dS}{da} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

$$b = \frac{s_{xy}}{s^2_x}$$

$$a = \bar{y} - b\bar{x}$$ → ordonnée à l'origine

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

$\hat{y_i} = bx_i + a$

coefficient de corrélation : $\boxed{r = \frac{s_{xy}}{s_x s_y}}$

$r \in [-1;1]$ donc $r \to 1, 0, -1$

r proche de 1 : corélation linéaire positive

_____ 0 : ∅ de corélation linéaire

_____ -1 : corélation linéaire négative

Signe de r = signe de la covariance

- modélisation statistique :

$y_i = b x_i + a + \varepsilon_i$ le modèle

on appelle ε_i les résidus ; $\varepsilon_i \sim N(0; \sigma_{res})$ et ils sont indépendants (on note ε_i = iid)

	$\hat{Y}$	ε
moyenne		0
variance	$s^2_{\hat{y}}$	s^2_ε

on estime ensuite :

$\sigma_{res} = \frac{n}{n-2} s^2_\varepsilon$

décomposition de la variance et R^2 :

$V_{totale} = V_{régression} + V_{résidus}$

$s^2_y = s^2_{\hat{y}} + s^2_\varepsilon$

coefficient de détermination : $\boxed{R^2 = \frac{s^2_{\hat{y}}}{s^2_y}}$

on a aussi $R^2 = r \times r$

R^2 est compris entre 0 et 1 car $r \in [0;1]$

- analyse des résidus :

pour vérifier la normalité des résidus :

- histogrammes
- graphique quantil/quantil.

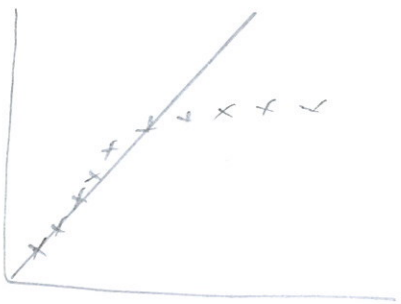
variance homogène : homoscédesticité.

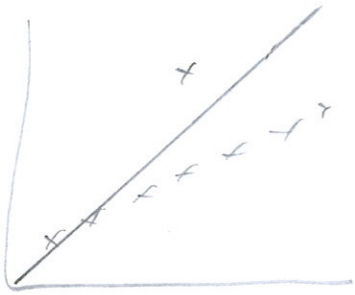
graphe des résidus: $\frac{\varepsilon_i}{\sigma_{res}}$

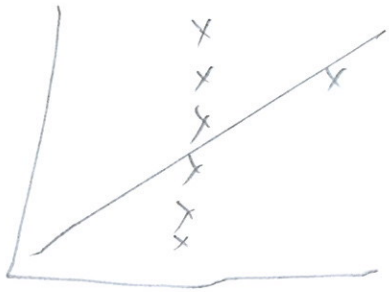
résidus standardisés

on va chercher les points externes aux limites

une corrélation n'est jamais la preuve d'un lien de causalité







coefficient de symétrie: γ_1 de Fisher

$$\gamma_1 = \frac{\mu_3}{\sigma_3}$$

<0 → étalement à gauche

$=0$ → symétrique

>0 → étalement à droite

coefficient de aplatissement: γ_2 de Fisher

$$\gamma_2 = \frac{\mu_4}{\sigma_4}$$

<0 + plate → platicurtique

$=0$ normale → mésocurtique

>0 + pointue → leptocurtique.