3 Statistiques à deux variables (séries doubles)

Dans certaines situations, on est amené à étudier deux caractères distincts d'une même population. On peut par exemple considérer la taille (x) et le poids (y) d'un ensemble d'individus. L'objectif principal de l'étude est de déterminer l'éventuel lien entre les deux variables x et y.

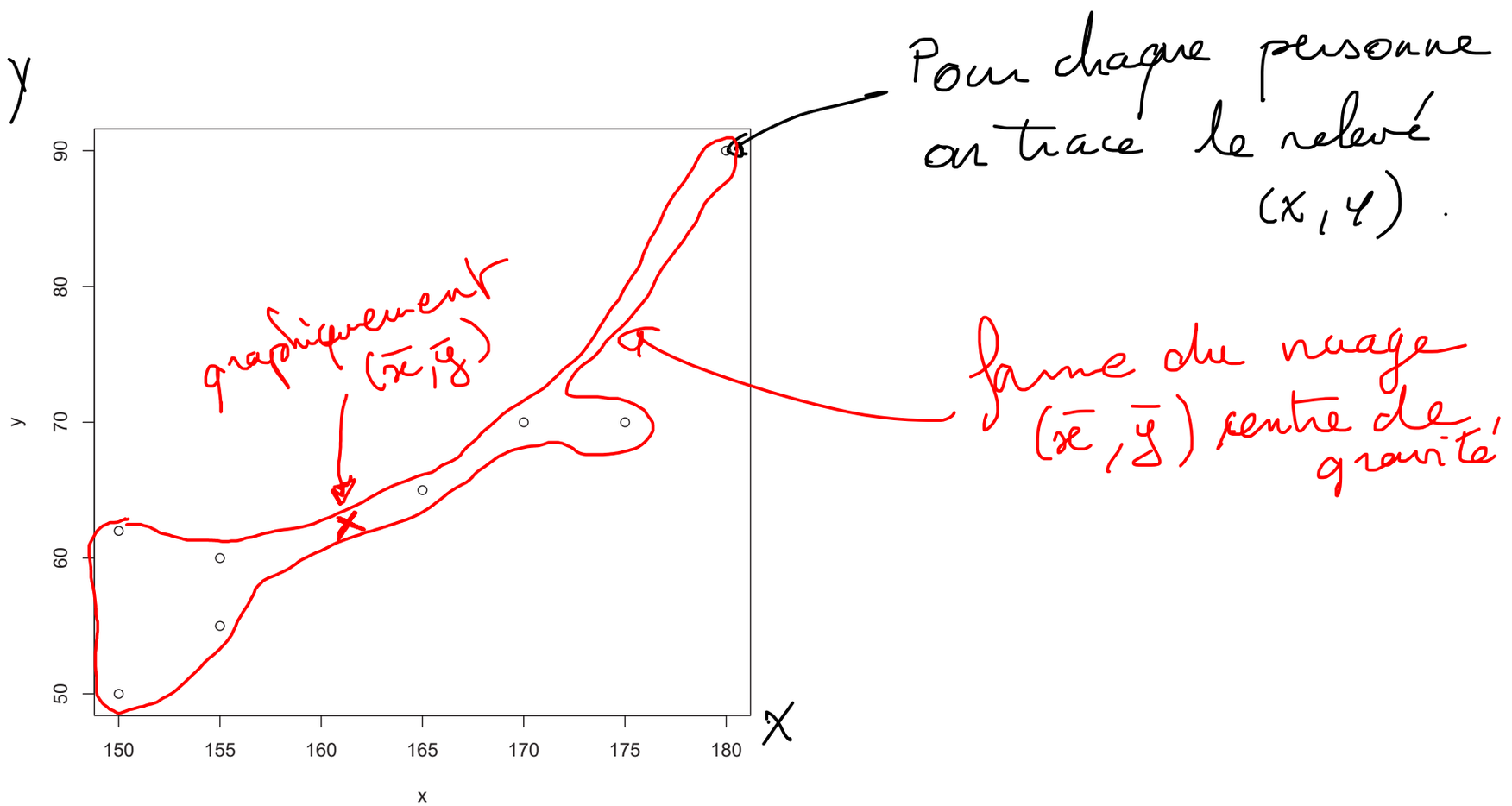
3.1 Nuage de points

On relève le couple (taille, poids) de 8 individus. On résume les données dans le tableau suivant.

taille	x	150	155	155	150	165	175	170	180
poids	y	50	55	60	62	65	70	70	90

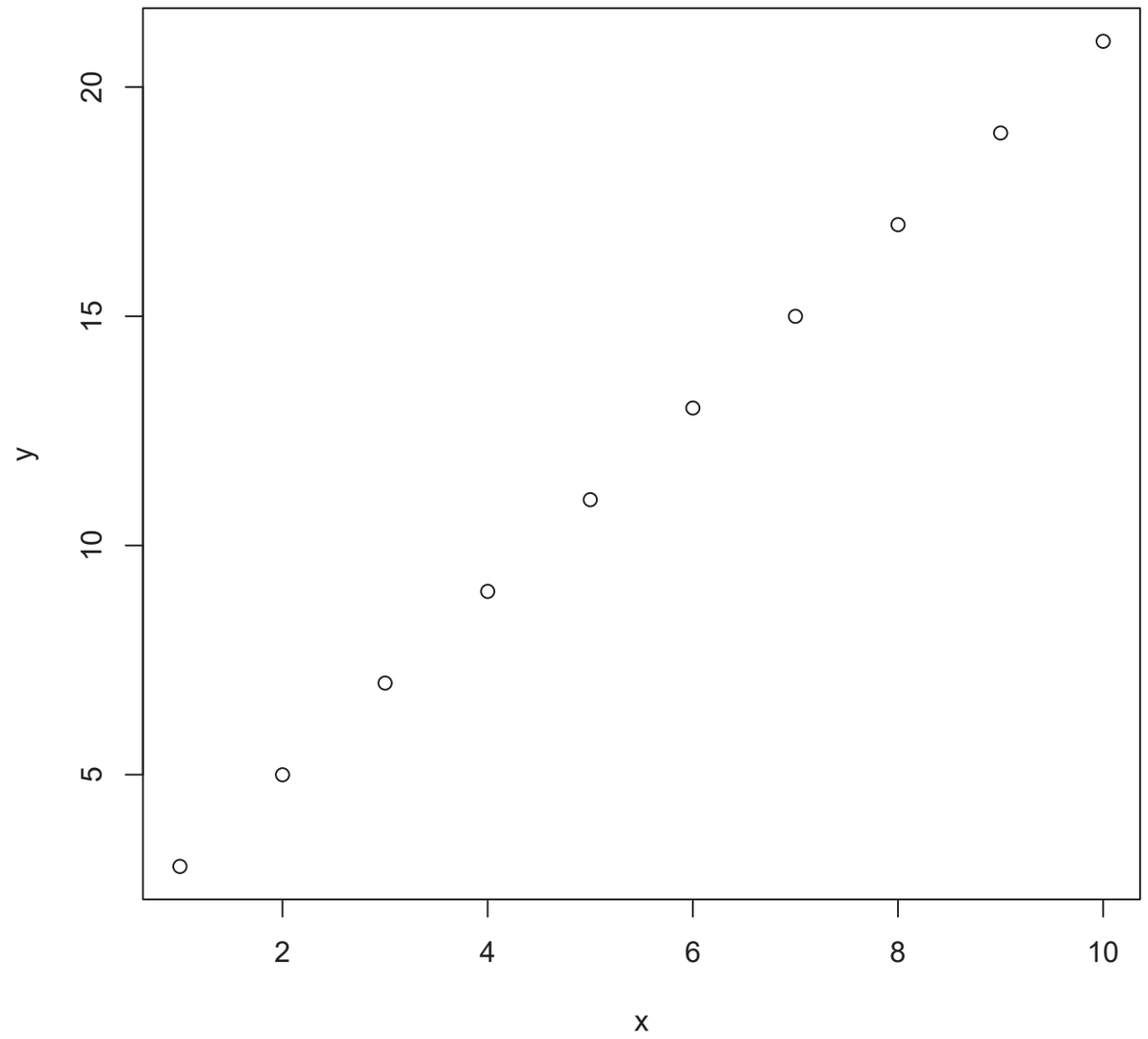
relevés sur population.

Definition 3.1 *Soit une population de N individus. Le graphe des N points (x_i, y_i) est appelé nuage de points de la série.*

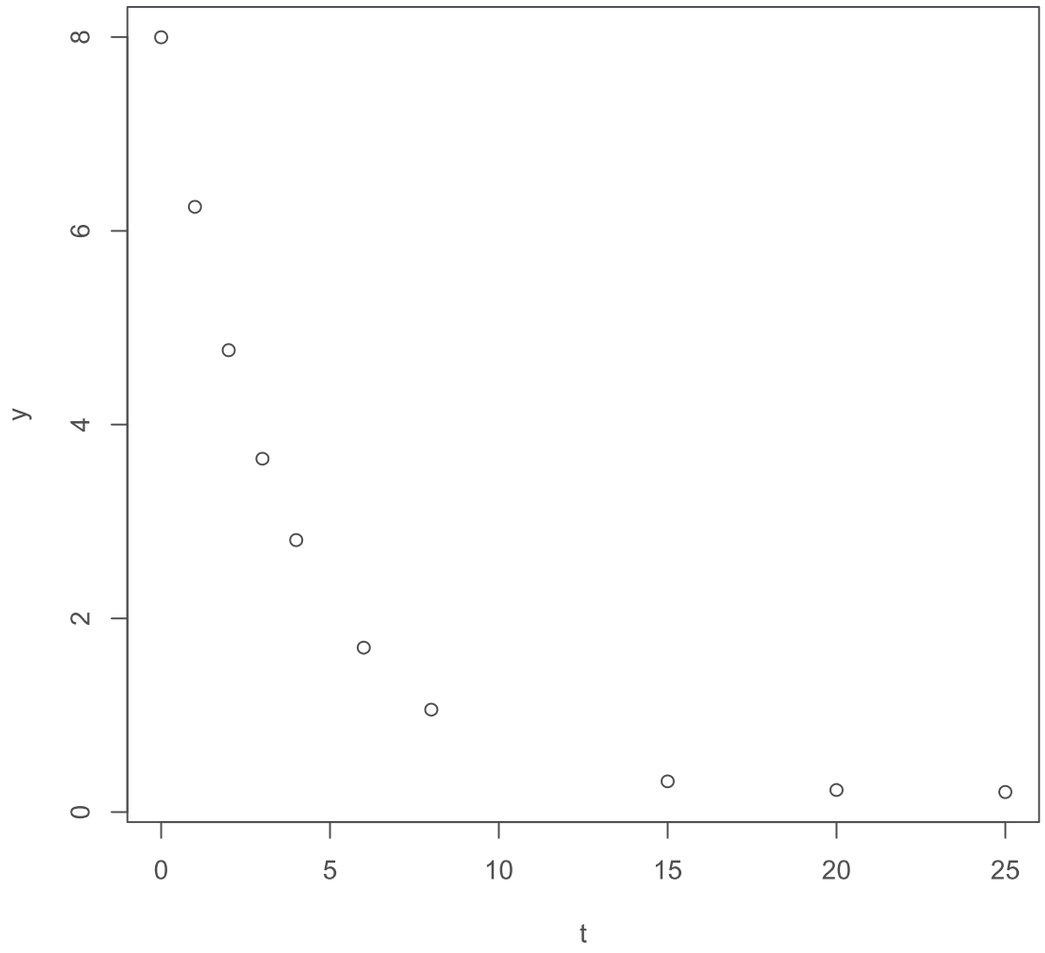


Definition 3.2 *Le point ayant pour coordonnées les moyennes $(\bar{x}, \bar{y})$ est appelé le point moyen.*

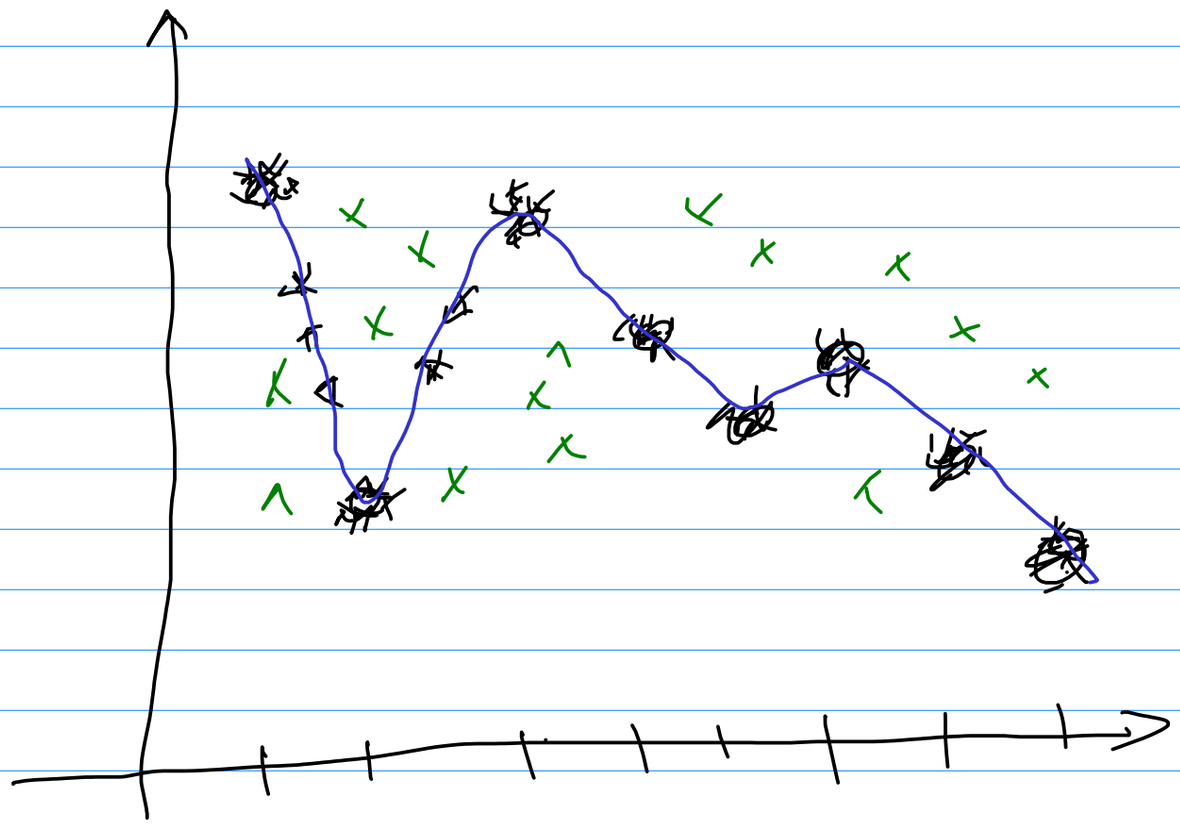
(i) forme allongée et rectiligne : les points sont plus ou moins alignés



(ii) forme allongée mais non rectiligne : les points ne sont pas alignés mais ont un profil ordonné

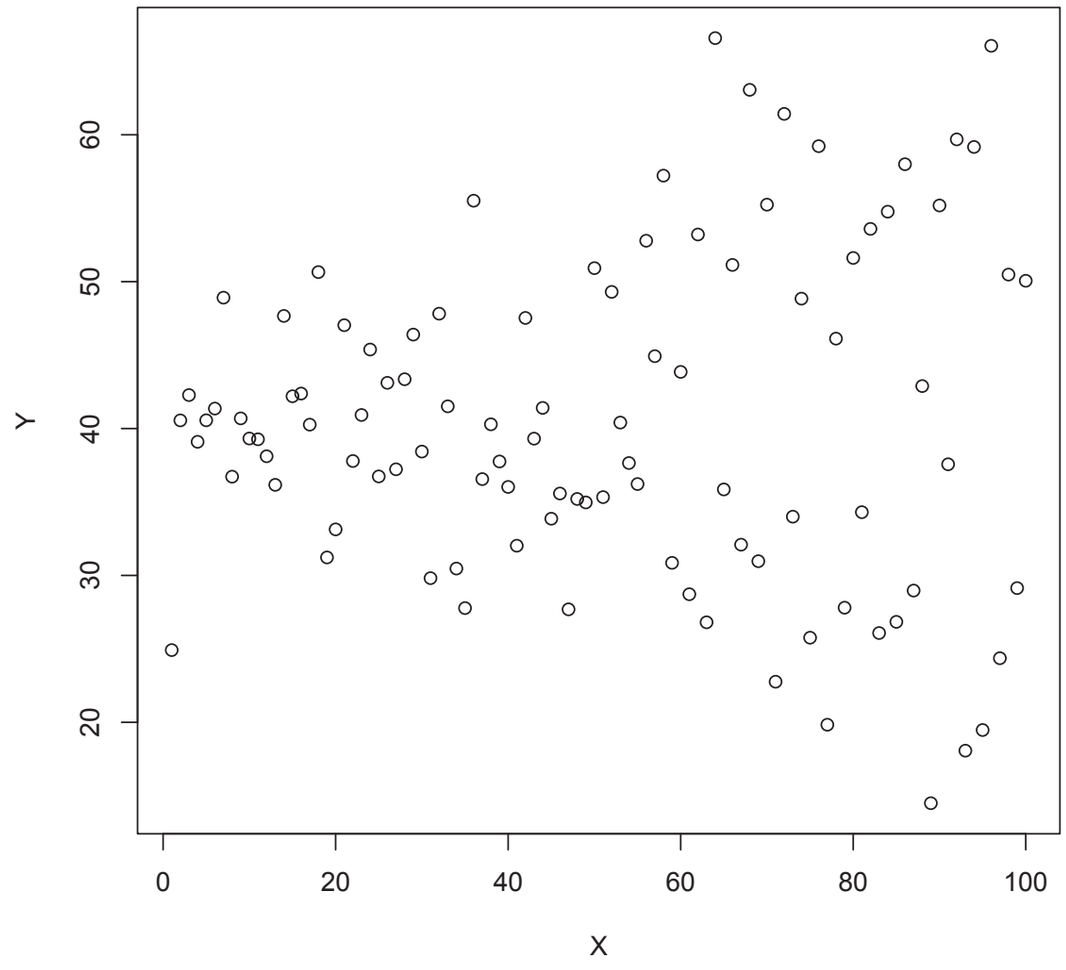


Note sur la dépendance (si on connaît une variable, on connaît "à peu près" l'autre).



(iii) forme quelconque

Indépendance des variables



3.3 Ajustement affine (droite de régression linéaire)

On s'intéresse plus particièrement au premier cas 3.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

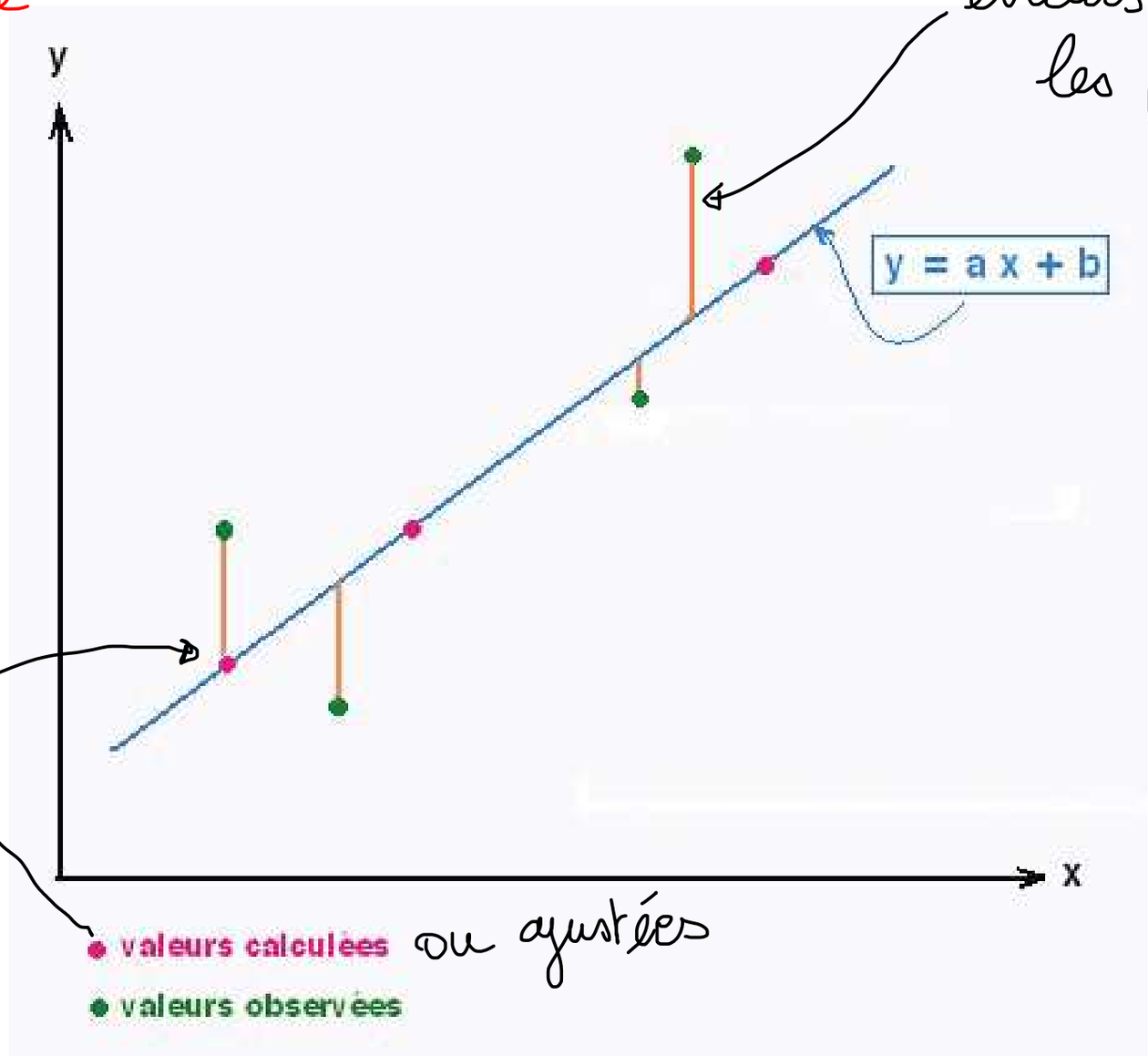
$$y = ax + b$$

qui passe au plus proche des points du nuage de points. Cette droite nous servira donc d'approximation. Bien évidemment, suivant la méthode utilisée pour la construire, on peut obtenir différentes droites. La méthode la plus utilisée car donnant la meilleure approximation est la méthode des moindres carrés.

3.3.1 La méthode des moindres carrés

L'idée de cette méthode est de chercher la droite qui minimise la somme des carrés des écarts verticaux entre la droite et les points du nuage, les *résidus*.

objectif de la méthode des moindres carrés: minimiser la somme des carrés des résidus -



"ecarts" entre la droite et les points du nuages: Résidus

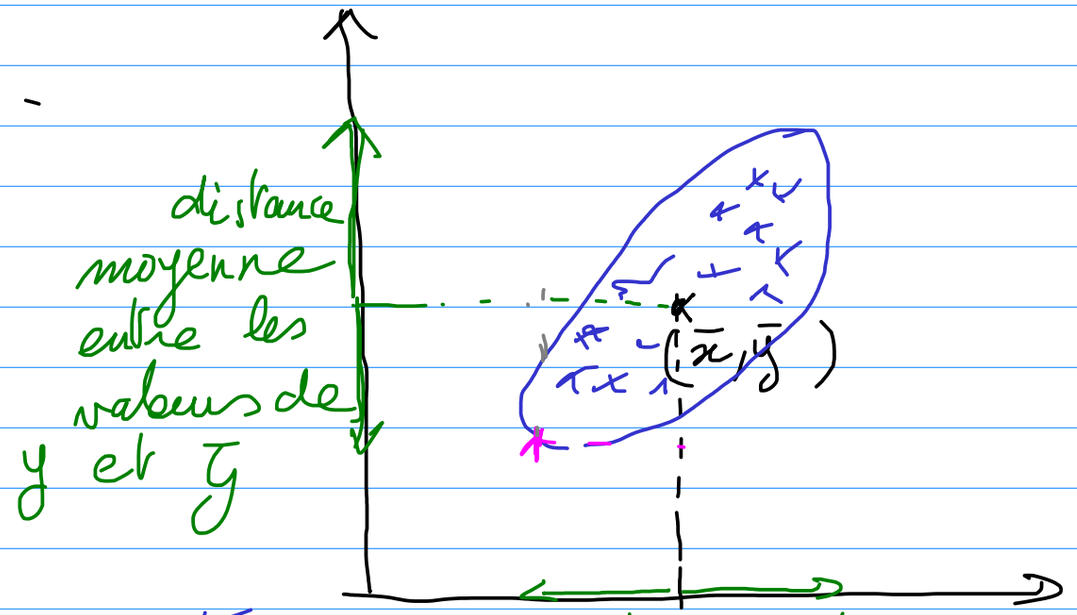
ou ajustées

Variances, écart-type, covariance

(moyennes = centre de gravité).

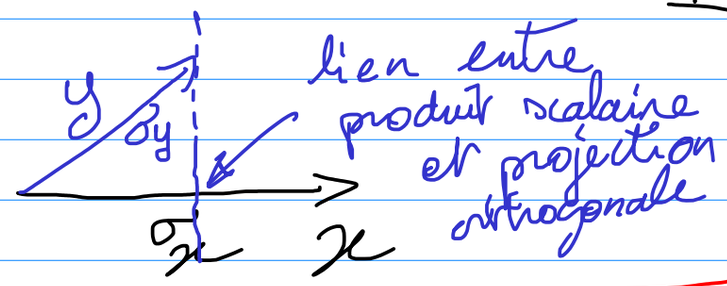
$$\sigma_x = \sqrt{\frac{1}{N}\sum_i (x_i - \bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{N}\sum_i (y_i - \bar{y})^2}$$

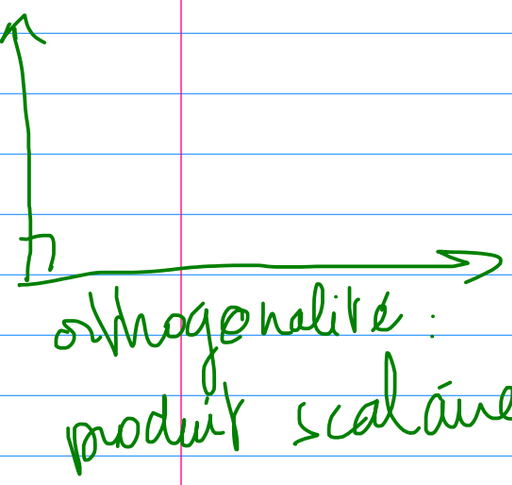


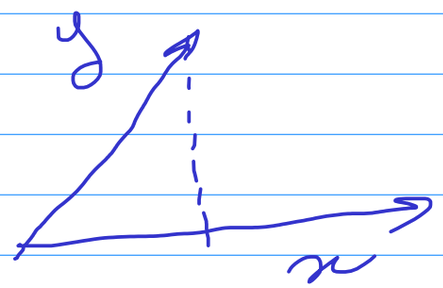
Abstraction :

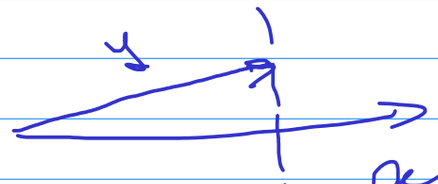
supposons : les séries x et y sont des vecteurs



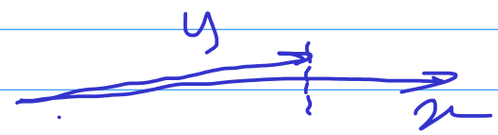
Cov(x,y) est un produit scalaire de la série x avec la série y.



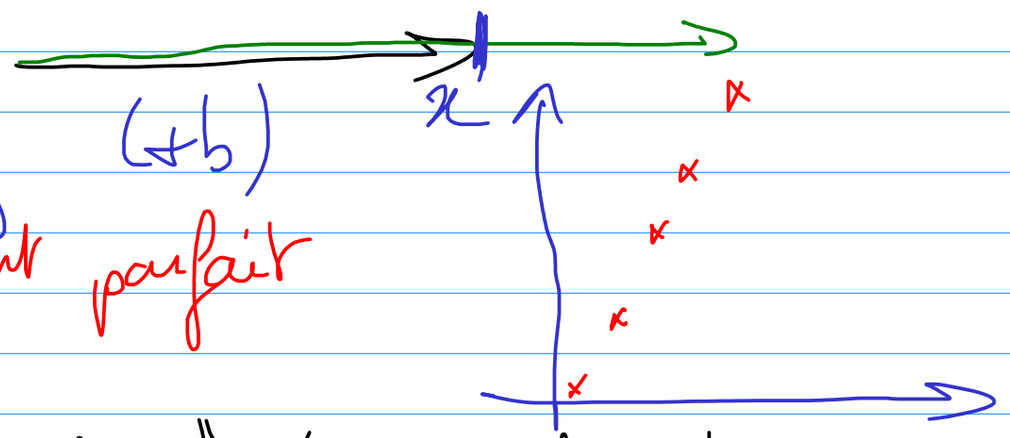




plus x et y vont dans la même direction, plus le produit scalaire est grand



Si on considère la covariance comme un produit scalaire : * Cov(x,y) est grande alors "x et y sont colinéaires".

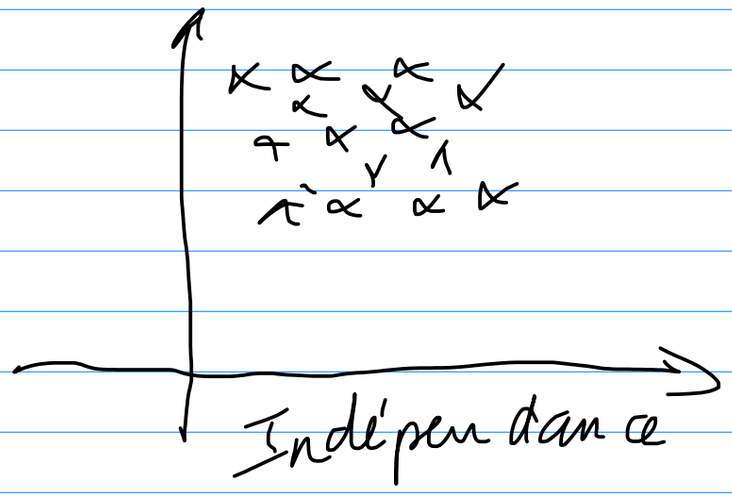


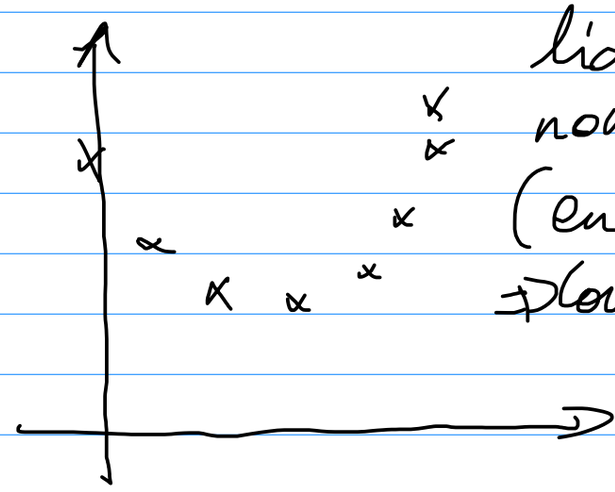
alors $x = \lambda y$ (+b)

colinéarité = alignement parfait
Cov(x,y) forte

Si Cov(x,y) est nulle ; il y "orthogonalité" entre x et y

Les variables ne fonctionnent pas linéairement ensemble.





On n'utilisera plus le vocabulaire de l'algèbre bilinéaire : pdt scalaire, colinéarité ...

En pratique, on détermine les coefficients de la droite $D : y = ax + b$ à l'aide d'un tableur. La droite ainsi obtenue est unique. Cette droite s'appelle la droite de régression linéaire de y en x par la méthode des moindres carrés. On note

relevés moyenne

$$Cov(x, y) = \sum (x_i - \bar{x})(y_i - \bar{y})/N.$$

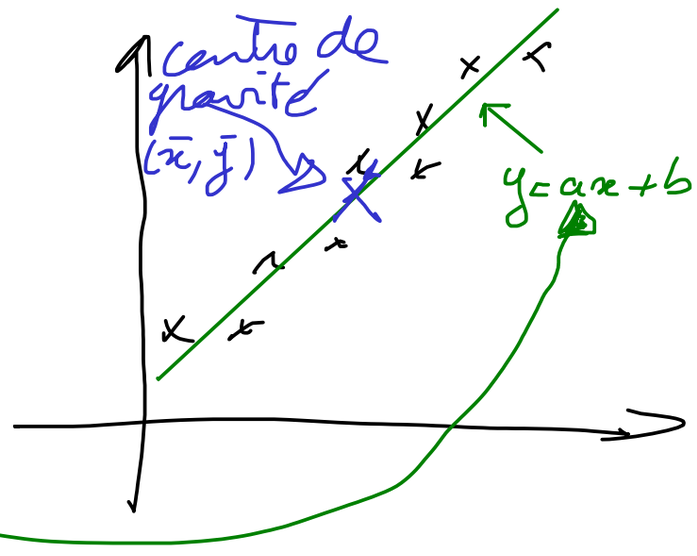
Cette quantité est nommée covariance de x et y. Si la quatité σ_x est une distance entre les valeurs de x est $\bar{x}$, on peut considérer la covariance comme un produit scalaire entre les variables x et y. Ainsi, si la covariance est proche de 0, on peut penser que les variables ont une dynamique qui n'ont rien de commun (penser à l'orthogonalité), c'est à dire le nuage 3.2.3.

On a

$$\begin{aligned} a &= cov(x, y)/\sigma_x^2, \\ b &= \bar{y} - a\bar{x}. \end{aligned}$$

la droite passe par $(\bar{x}, \bar{y})$

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



3.3.2 Coefficient de corrélation linéaire

Notons que la méthode des moindres carrés peut être utilisée pour n'importe quelle série double. On peut tout à fait obtenir une droite de régression dans le cas 3.2.3. Pour s'assurer de façon objective (et non purement visuelle) que l'ajustement est valide, on considère un autre paramètre de la série : le coefficient de corrélation r :

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

on normalise la covariance par les écarts-type pour obtenir le coefficient de corrélation linéaire.

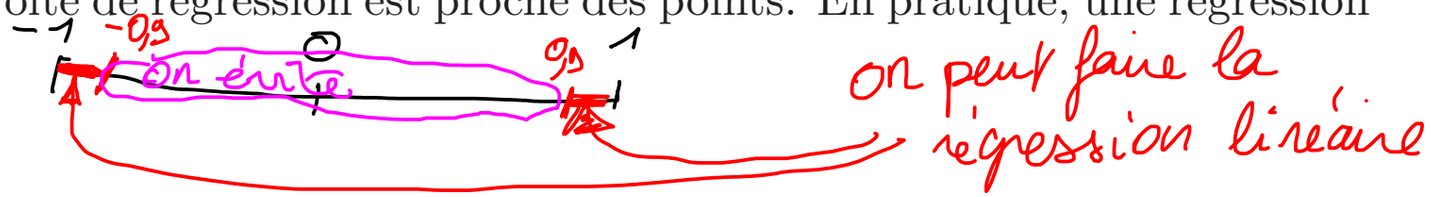
Proposition 3.3 *On a les propriétés suivantes :*

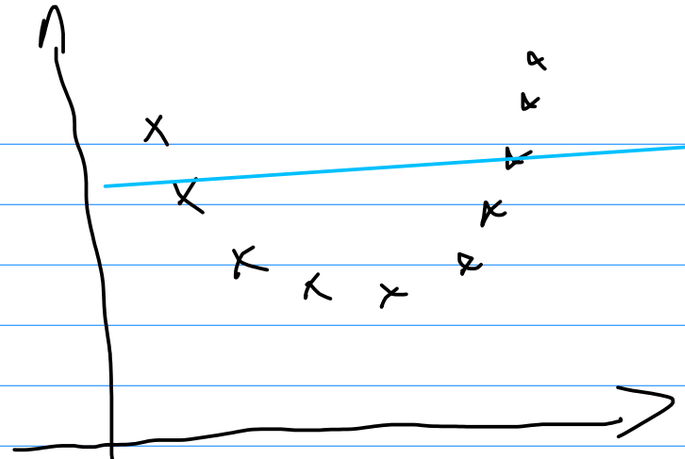
(i) on a toujours $-1 \leq r \leq 1$;

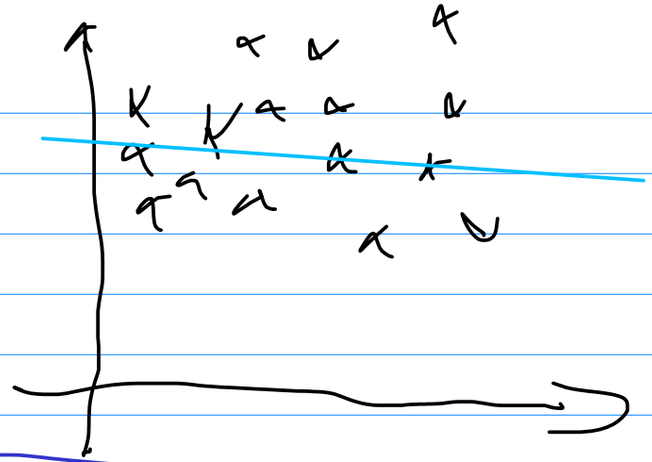
(ii) le coefficient directeur de la droite de régression et le coefficient de corrélation sont de même signe ;

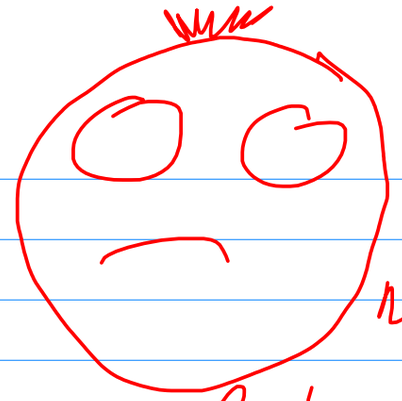
(iii) le degré de corrélation est d'autant plus fort que r *est proche de* 1 *ou* -1.

C'est l'assertion 3.*iii* qui nous permet de dire si la droite de régression est proche des points. En pratique, une régression linéaire est légitime si $r > 0.9$ ou si $r < -0.9$.



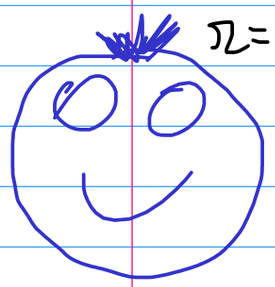




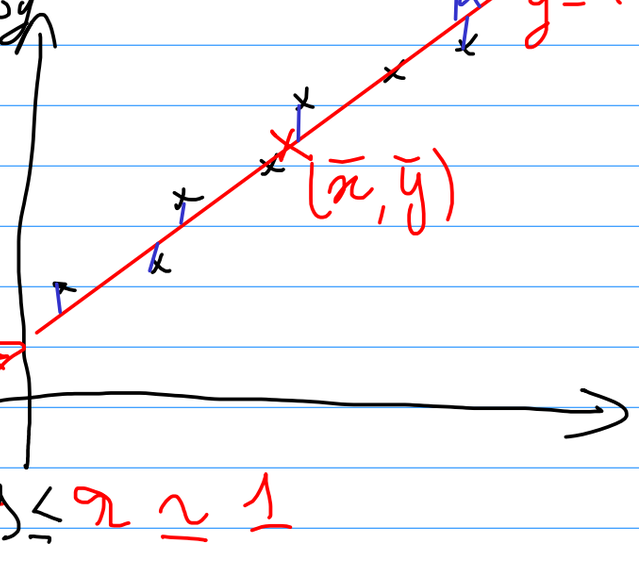


la droite n'est pas le bon outil

$r \simeq 0$

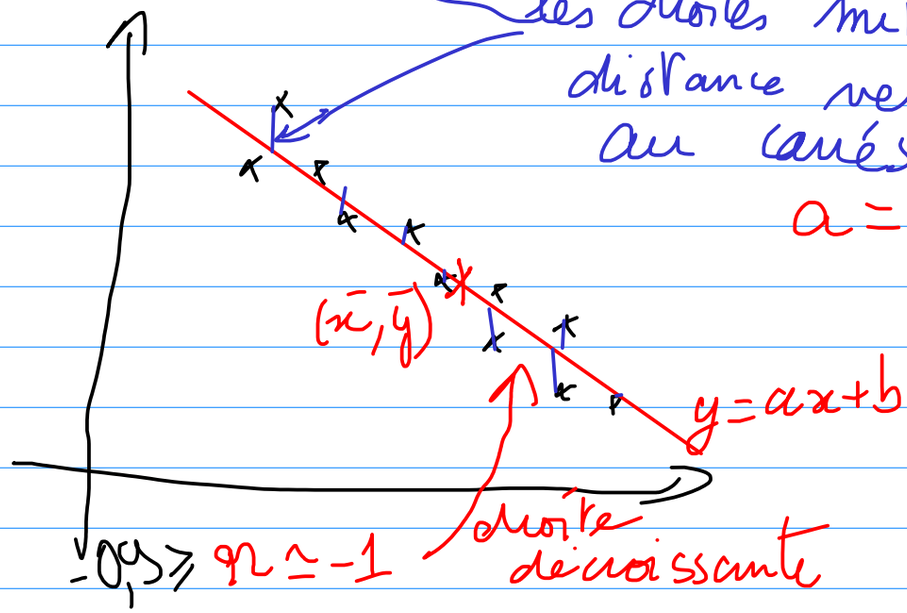


$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$



droite croissante

$0,9 \leq r \simeq 1$



$-0,9 \geq r \simeq -1$

droite décroissante

les droites minimisent la distance verticale (résidus) au carrés

$a = \frac{cov(x,y)}{\sigma_x^2}$ (signe, carré)

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

Nodes: plus r est proche de 1 ou -1 plus on considère qu'on peut faire des prévisions (à court terme, la régression linéaire n'est pas une boule de cristal).

3.3.3 Coefficient d'amélioration R^2

La variance est une bonne mesure de l'hétérogénéité d'une série (contrairement à la moyenne qui considère tous les individus comme semblables). La variance de la série x se décompose comme la variance expliquée par la droite de régression plus celle de l'erreur (résidus) :

$$Var(y) = Var(ax + b) + Var(e).$$

Le coefficient d'amélioration est le rapport de variance de y expliquée par la régression :

$$R^2 = \frac{Var(ax + b)}{Var(y)}. \tag{3.1}$$

Il se trouve que R^2 est le carré du coefficient de corrélation linéaire.

3.3.4 régression $x = my + p$

Le choix de représenter y en fonction de x est bien souvent arbitraire. Lorsque le caractère y dépend du caractère x par un lien de cause à effet clair (concentration y d'un composant lors d'une réaction chimique en fonction du temps x), on utlisera bien entendu la régression $y = ax + b$. Dans le cas d'une interdépendance (chiffre d'affaire x et budget publicité y), le choix de la régression $y = ax + b$ ou $x = my + p$ se pose. En fait, les deux régressions sont tout à fait valides (ainsi que toute droite située entre les deux).

Dans les faits, puisque nous choisissons d'effectuer des régressions linéaires uniquement dans le cas où le coefficient de corrélation r est proche de 1 ou -1, ces deux droites seront très proches et amènent aux mêmes conclusions.