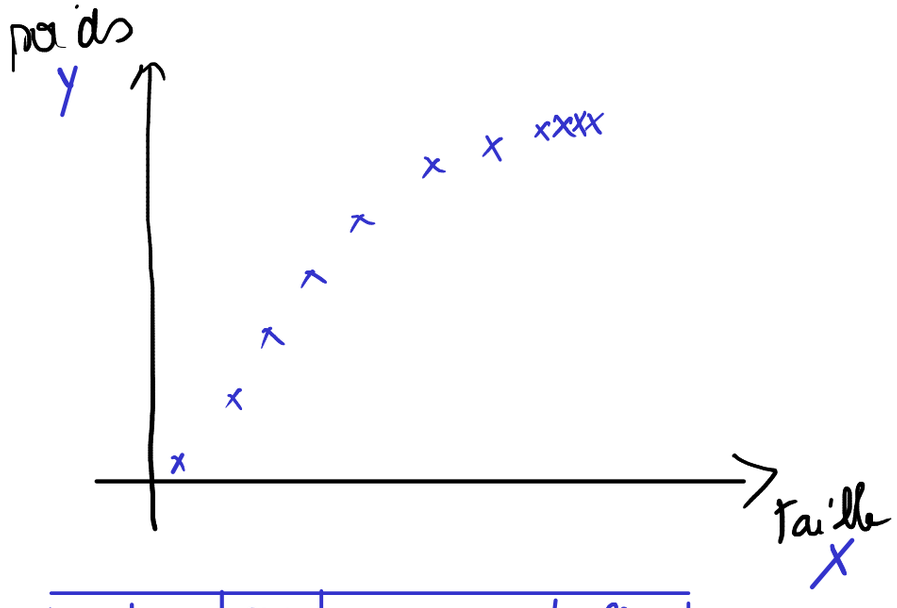
Part III

Régressions linéaires

Carnet de santé :



x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

⚠ on ne regroupe pas par modalité

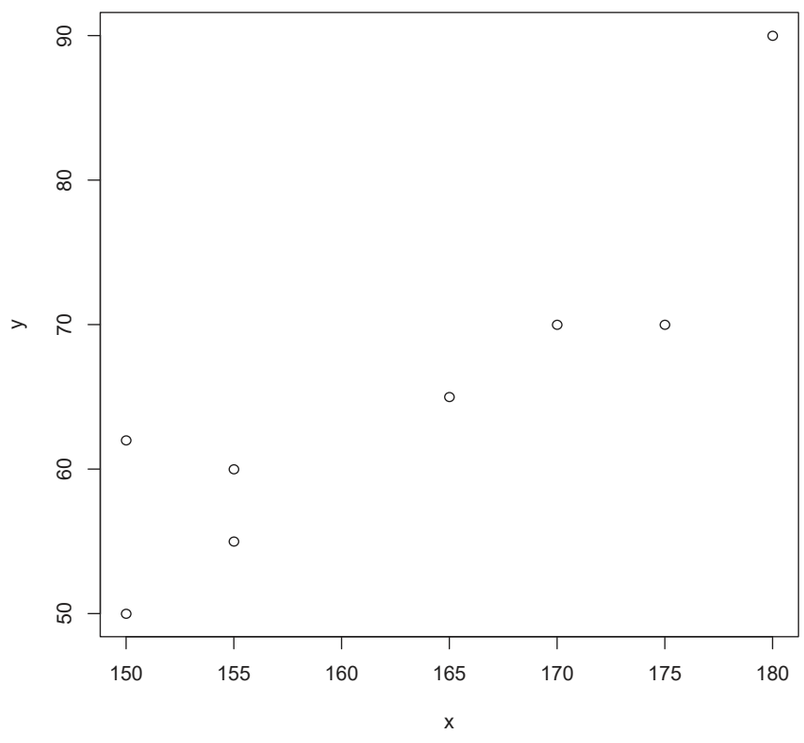
l'effectif de chaque observation est 1

4.1 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

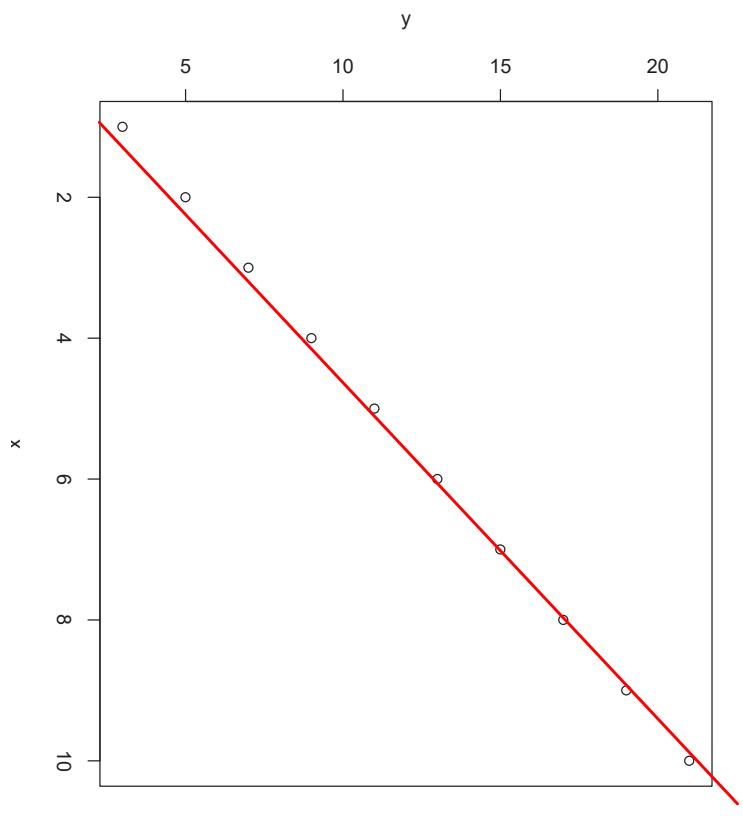
Definition 4.1 *Soit une population de N individus. Le graphe des N points (x_i, y_i) est appelé nuage de points de la série.*



Nuage de points

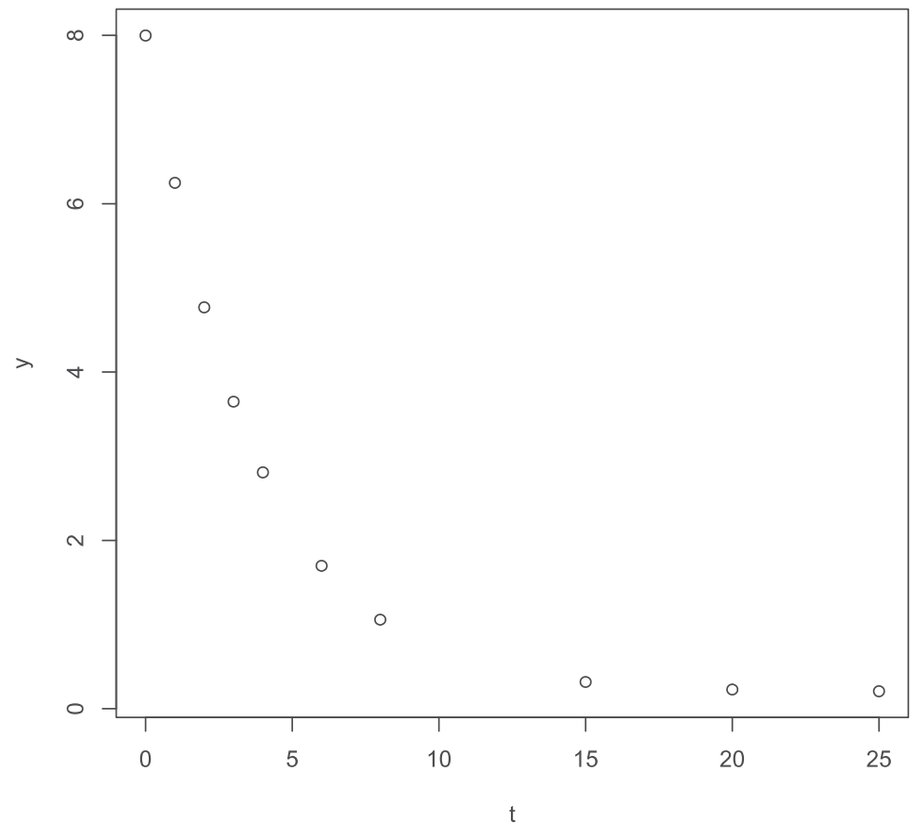
Definition 4.2 *Le point ayant pour coordonnées les moyennes $(\bar{x}, \bar{y})$ est appelé le point moyen.*

Nuage de point rectiligne.

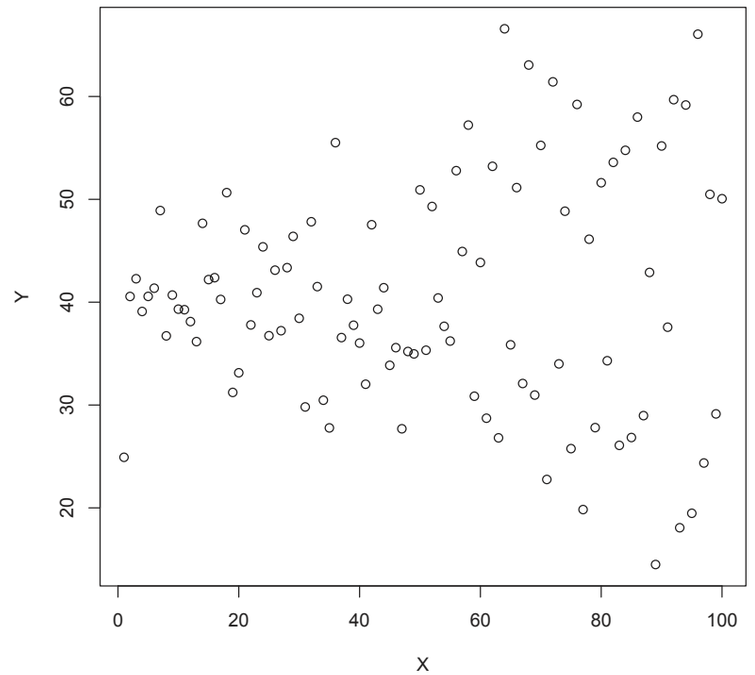


Idée : Cherchons la droite qui se trouve au plus proche des points !

Nuage ordonné mais non rectiligne



(iii) forme quelconque

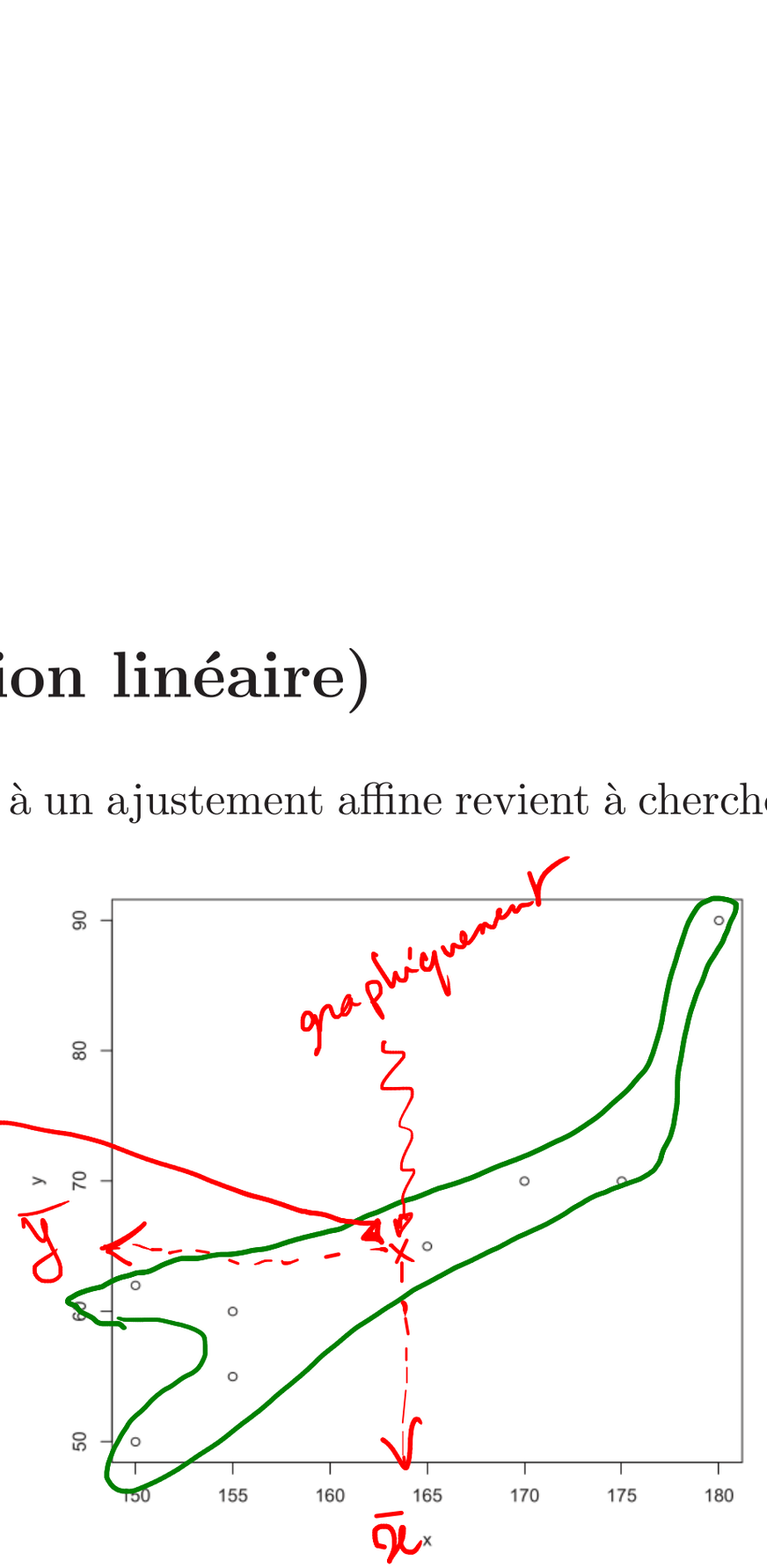


5 Ajustement affine (droite de régression linéaire)

On s'intéresse plus particièrement au premier cas 4.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

$$y = ax + b$$

Centre de gravité $(\bar{x}, \bar{y})$



Toute droite de régression passe par le centre de gravité.

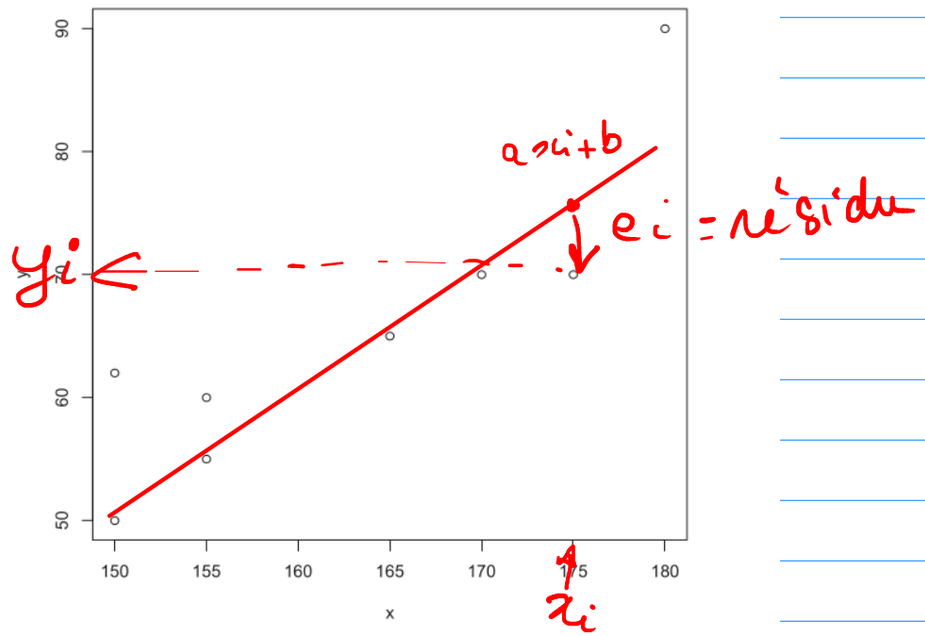
Note : Toute droite de régression passe par le centre de gravité.

$$y_i = a x_i + b + e_i$$

e_i : erreur entre la droite et l'observation

D'où $\bar{y} = \overline{ax + b + e}$

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



On ne veut pas d'erreur systématique donc $\bar{e} = 0$. (Sinon, on retire $\bar{e}$ à b pour obtenir un modèle sans erreur systématique).

D'où $\bar{e} = 0 \iff \bar{y} = a\bar{x} + b$

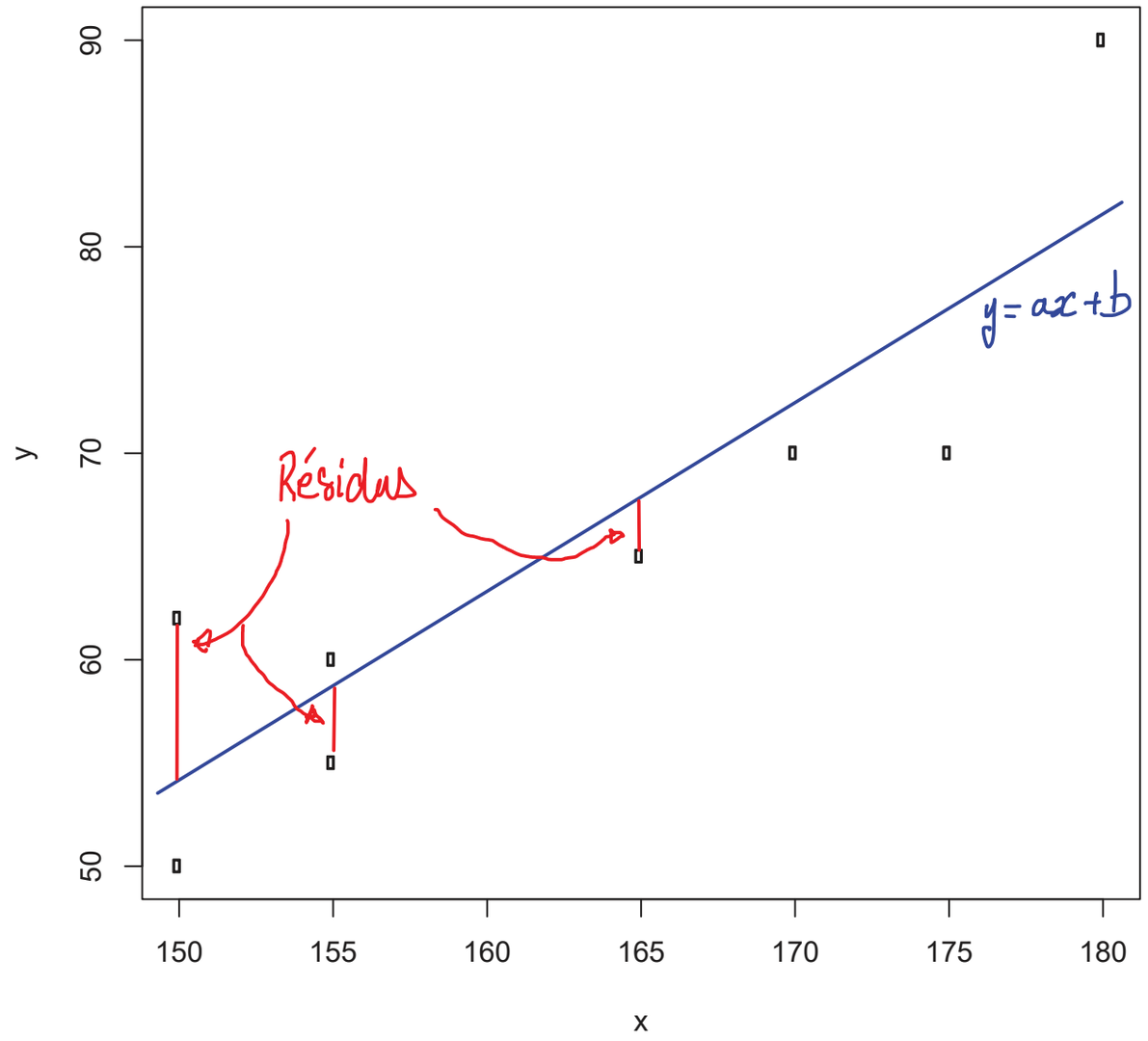
Donc $(\bar{x}, \bar{y})$ vérifie l'équation de la droite.

Toute droite de régression passe par le centre de gravité.

Pb : Trouver a dans $y = ax + b$!

5.1 La méthode des moindres carrés

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



On cherche à minimiser la distance verticale entre les points et la droite :

$\sum_i e_i^2$

résidus

On cherche donc le "a" qui minimise la somme des carrés des résidus $\sum_i e_i^2$

Rappel: $\bar{x} = \frac{1}{N} \sum_i x_i$ $\qquad \bar{y} = \frac{1}{N} \sum_i y_i$

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

Var y ...

$$= m(x^2) - \bar{x}^2$$

$$= \frac{1}{N} \sum_i x_i^2 - \bar{x}^2$$

$\sigma_x = \sqrt{\text{Var } x}$ $\qquad \sigma_y = \sqrt{\text{Var } y}$

$$\text{Cov}(x, y) = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) = m(xy) - \bar{x}\,\bar{y}$$

Note: $\text{Cov}(x, x) = \frac{1}{N} \sum_i (x_i - \bar{x})(x_i - \bar{x})$

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

démontré en S4.

Thm : $a = \dfrac{\mathrm{Cov}(x,y)}{\mathrm{Var}\,x}$

et on trouve b en vérifiant $\bar{y} = a\bar{x} + b$.

Exemple :

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

$$\bar{x} = \frac{150+155+155+\dots+170+180}{8} = 162,5$$

$$\bar{y} = \frac{50+55+\dots+70+90}{8} = 65,25$$

$$\text{Var } x = m(x^2) - \bar{x}^2 = \frac{150^2+155^2+\dots+180^2}{8} - 162,5^2$$

$$= 26\,525 - 26\,406,25$$

$$= 118,75$$

$$\text{Var } y = m(y^2) - \bar{y}^2 = 4386,75 - 4257,562 \simeq 129,2 .$$

Calculons $\text{cov}(x,y)$.

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

$$Cov(x,y) = m(x \times y) - \bar{x}\,\bar{y}$$

$$= \frac{50\times150 + 55\times155 + 60\times155 + \dots + 90\times180}{8} - 162{,}5 \times 65{,}25$$

$$= 10\,712{,}\overline{5} - 10\,603{,}12$$

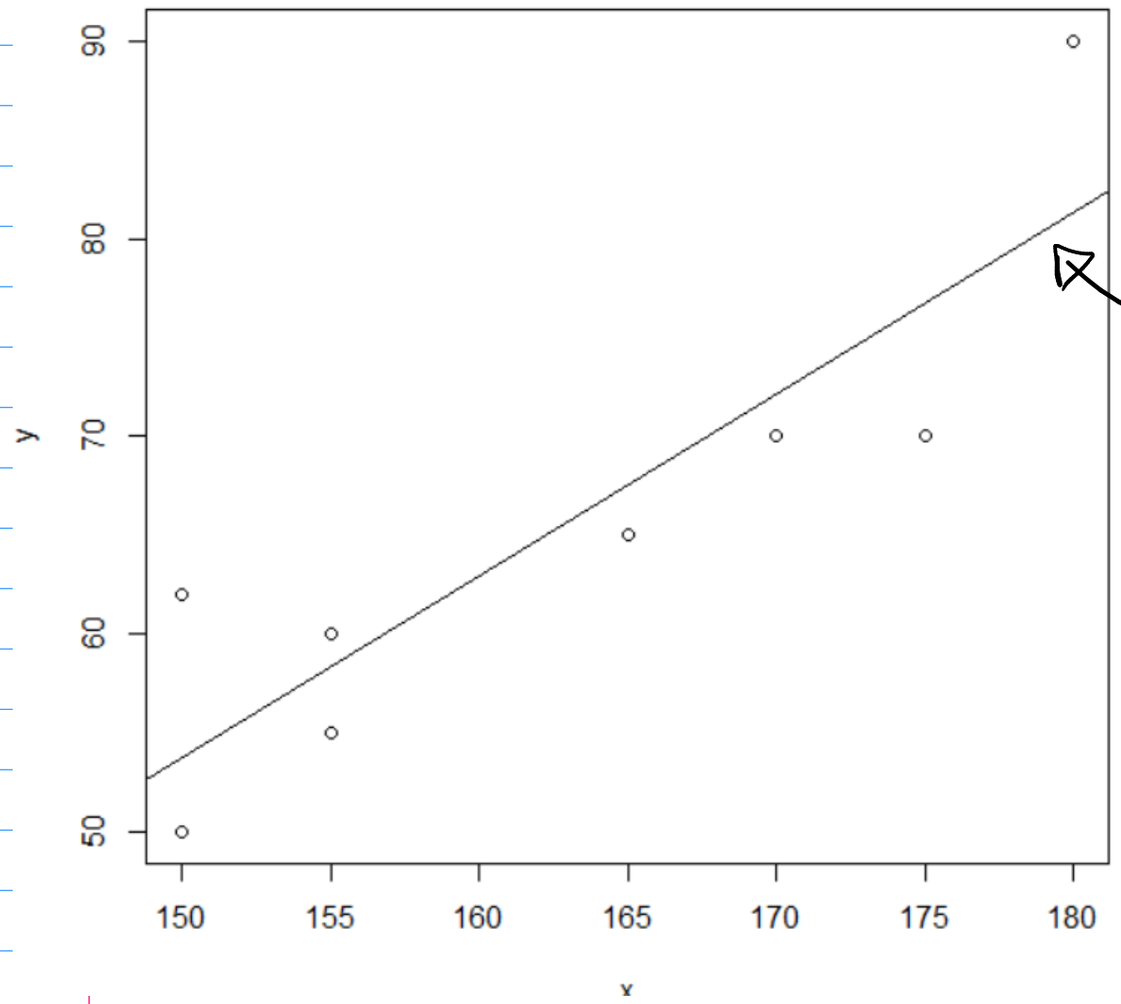
$$= 109{,}4$$

Équation de la droite des moindres carrés : $y = ax + b$.

$$a = \frac{Cov(x,y)}{Var(x)} = \frac{109{,}4}{118{,}75} = 0{,}92$$

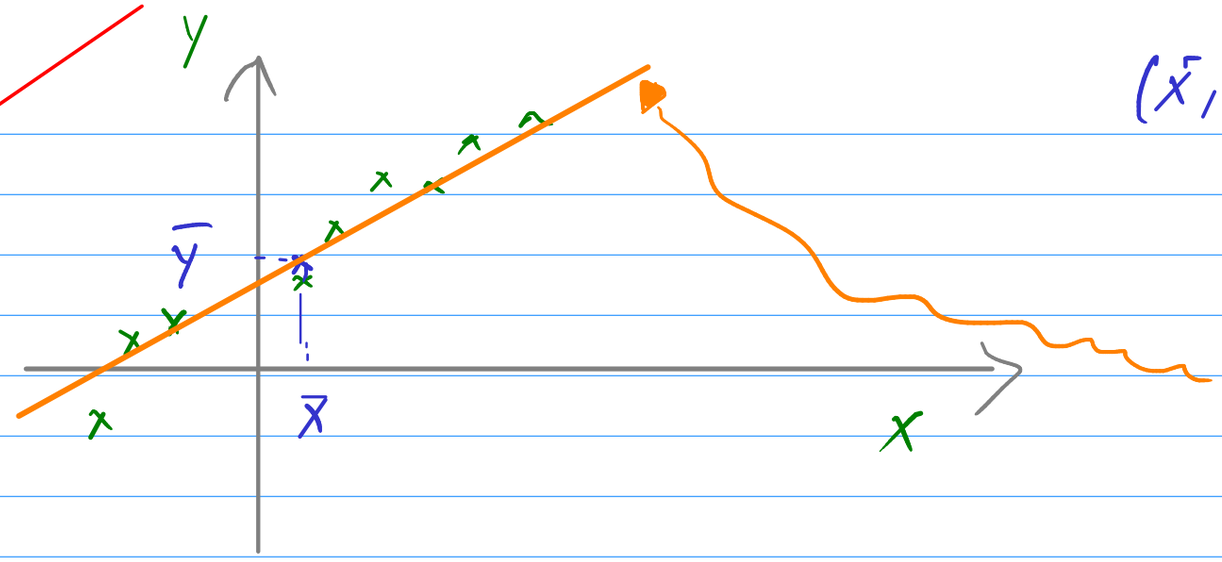
et $\bar{y} = a\bar{x} + b \Leftrightarrow \quad b = \bar{y} - a\bar{x} = 65{,}25 - 0{,}92 \times 162{,}5$

$$= -84{,}25$$



$y = 0{,}92x - 84{,}25$.

Rappel



$(\bar{x}, \bar{y})$ centre de gravité du nuage de points

droite qui passe par le centre de gravité et au plus proche des points

$y = ax + b$.

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

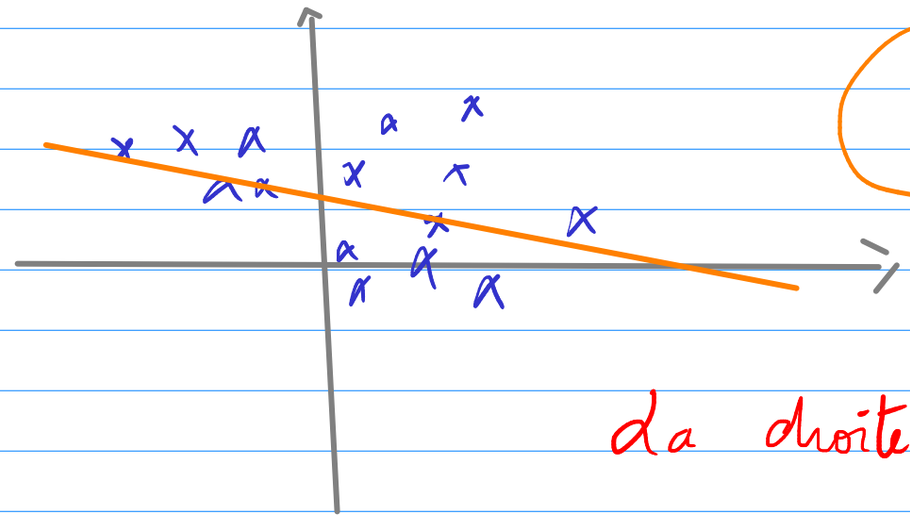
$$\mathrm{Cov}(x,y) = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$
$$= m(xy) - \bar{x}\,\bar{y}$$

$$\mathrm{Var}\, x = \frac{1}{N} \sum_i (x_i - \bar{x})^2$$
$$= m(x^2) - \bar{x}^2 .$$

$\bar{y} = a\bar{x} + b$ car $(\bar{x}, \bar{y})$ est sur la droite, donc

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

Note: Ainsi définis, on peut toujour calculer a et b et ainsi trouver une droite qui passe au plus proche des points -



la droite n'est pourtant pas un modèle satisfaisant.

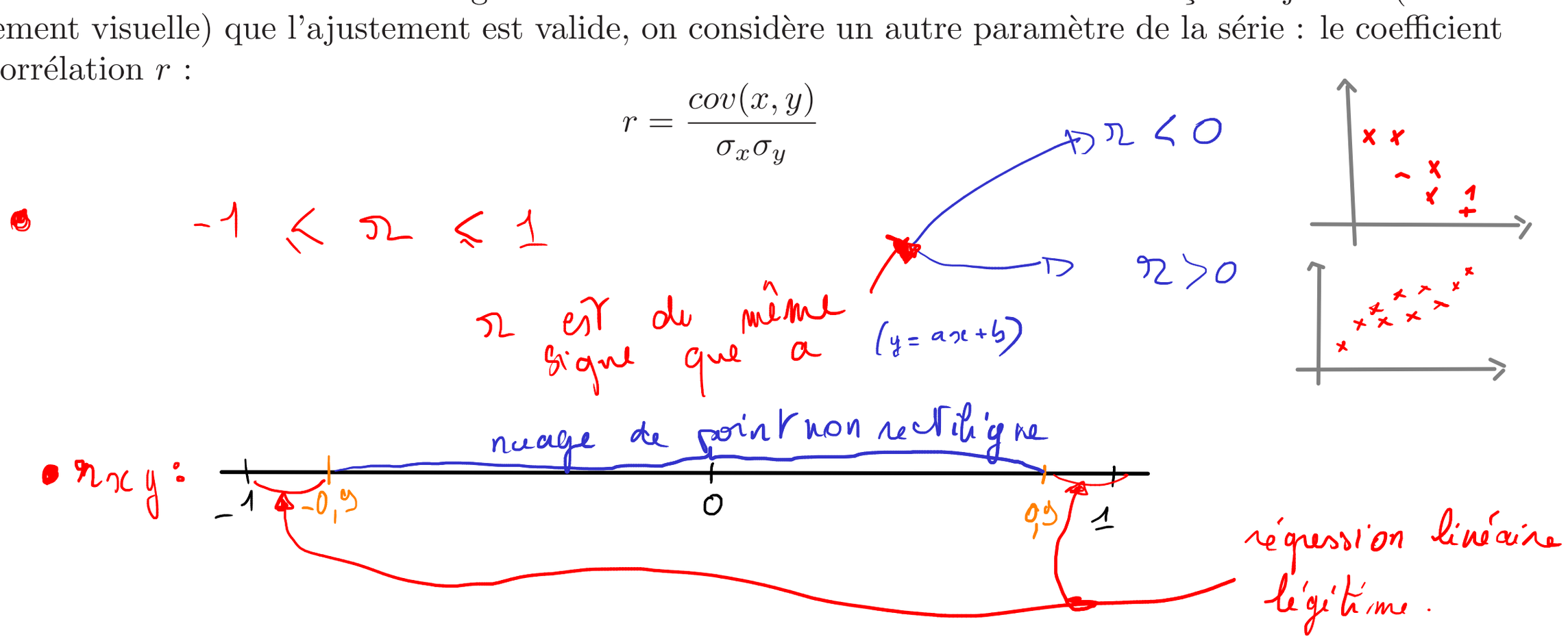
* Gardons le modèle de droite de régression pour le cas où le nuage est rectiligne -

⇒ C'est le coefficients r_{xy} qui nous l'indique ...

5.2 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 4.2.3. 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 :

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



Dans l'exemple :

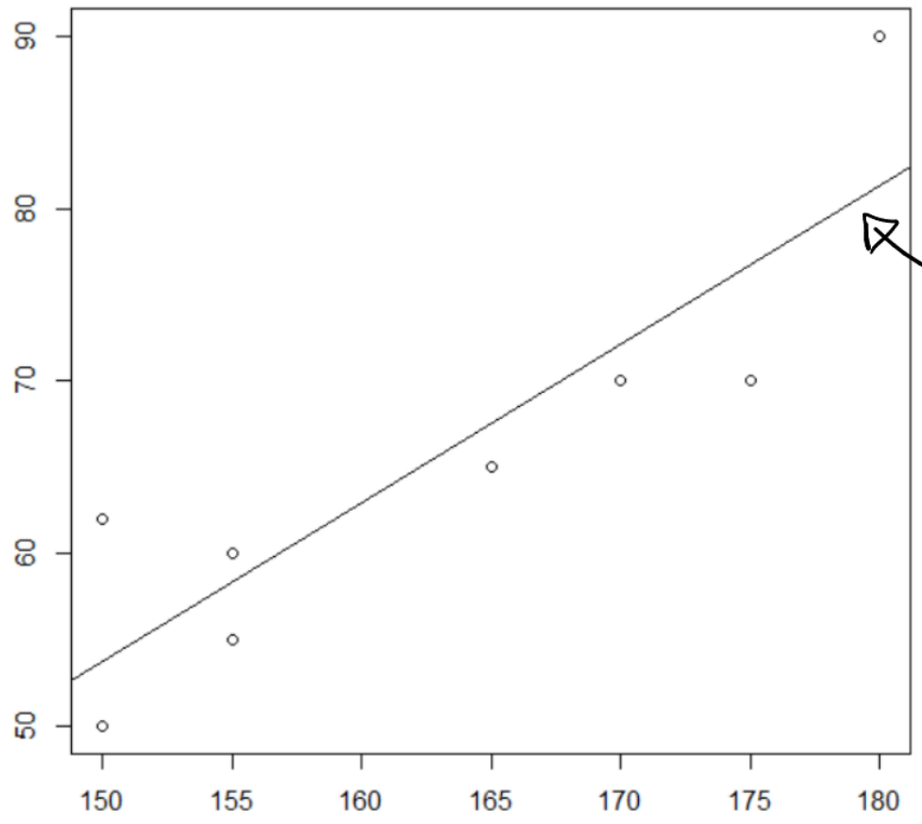
taille	x	150	155	155	150	165	175	170	180
poids	y	50	55	60	62	65	70	70	90

$$Cov(x,y) = 109,4$$

$$Var\ x = 118,75$$

$$Var\ y = 129,2$$

$$r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{Cov(x,y)}{\sqrt{Var\ x}\sqrt{Var\ y}} = \frac{109,4}{\sqrt{118,75}\sqrt{129,2}} = 0,88 .$$



$$y = 0,92x - 84,25$$

$-0,9 < r_{xy} < 0,9$ La droite n'est pas très proche des points.

On devrait éviter d'utiliser ce modèle.

$R^2 = r_{xy}^2 = 0{,}88^2 = 0{,}78.$

78 % de la variance des observation est expliquée par la droite.

5.3 Coefficient de détermination 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 de détermination est le rapport de variance de y expliquée par la régression :

$$R^2 = \frac{Var(ax + b)}{Var(y)}. \tag{5.1}$$

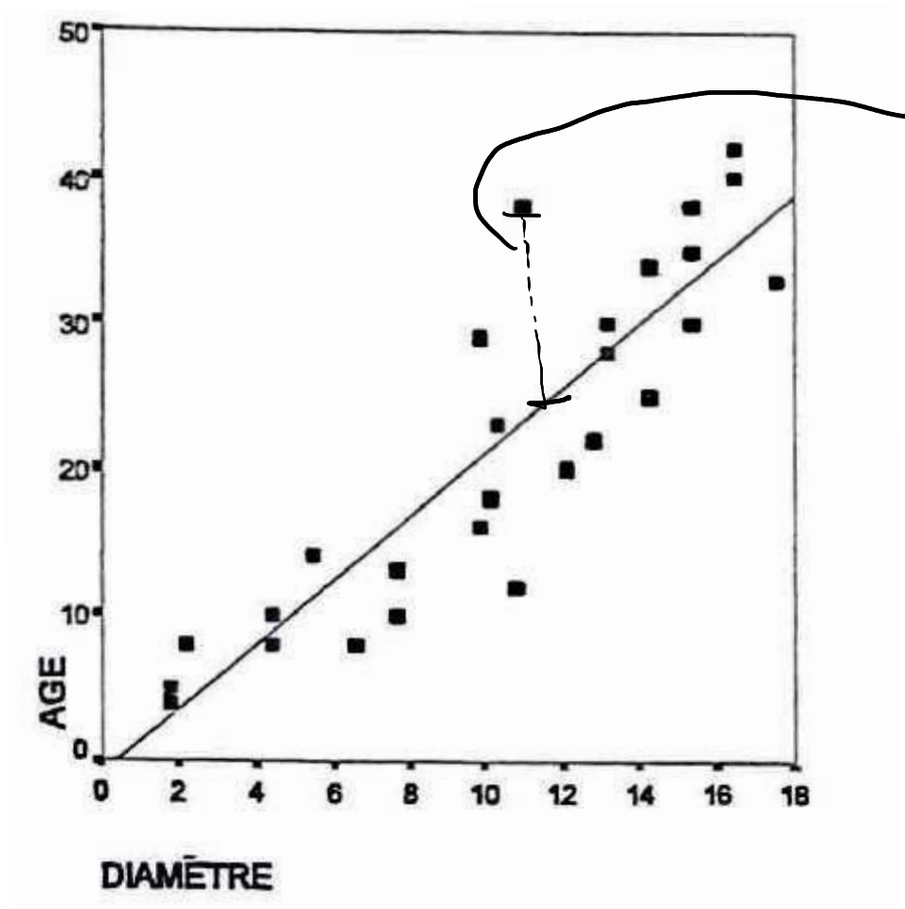
Il se trouve que R^2 est le carré du coefficient de corrélation linéaire r_{xy}.

$r_{xy} > 0{,}9$ ou $r_{xy} < -0{,}9 \Leftrightarrow R^2 > 0{,}8$

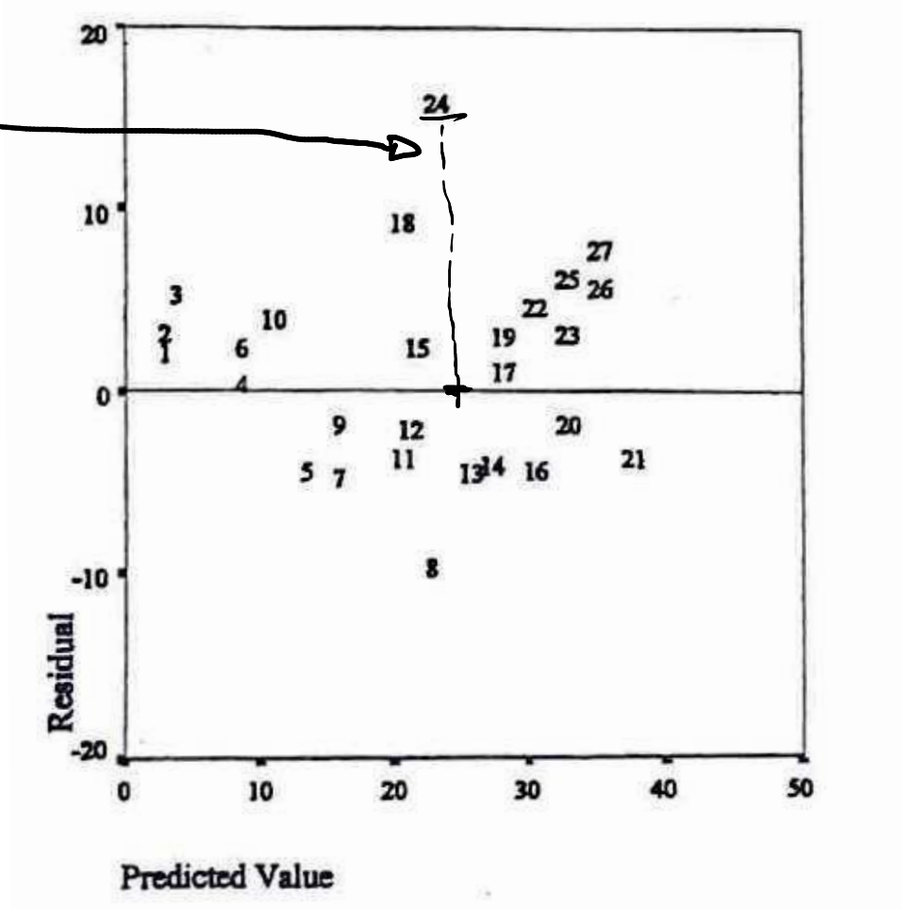
$\uparrow$ 80% de variance expliquée.

Sur logiciel : Age et diamètre de 27 arbres.

Nuage de points



Résidus



Résumé de la régression

Aller chercher les infos importantes ...

	Somme des carrés	ddl	Carré moyen	F	Signification
Régression	2905,549	1	2905,55	93,44	,000
Résidu	777,414	25	31,097		
Total	3682,963	26			

		Coefficients non standardisés		Coefficients standardisés		
		B	Erreur standard	Bêta	t	Signification
b →	(constante)	-,974	2,604		-,374	,711
a →	DIAMETRE	2,206	,228	,888	9,67	,000

coefficients

$r_{xy} \simeq 0,888$.

Age = 2,206 diamètre − 0,974.

$|r_{xy}| < 0,9$.

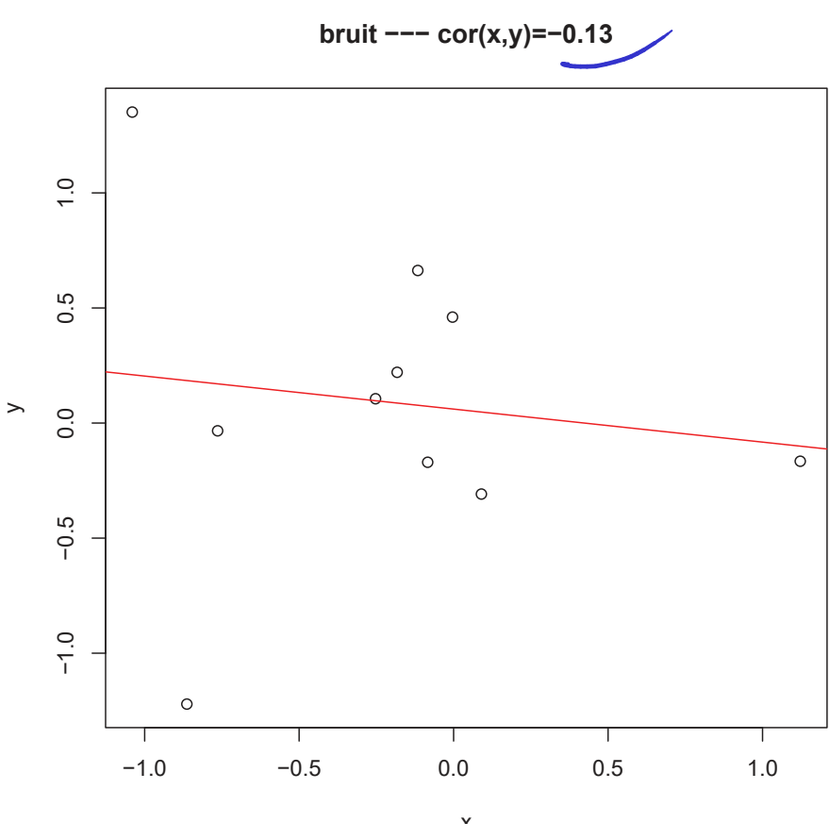
La droite donne une tendance mais pas une prévision fiable.

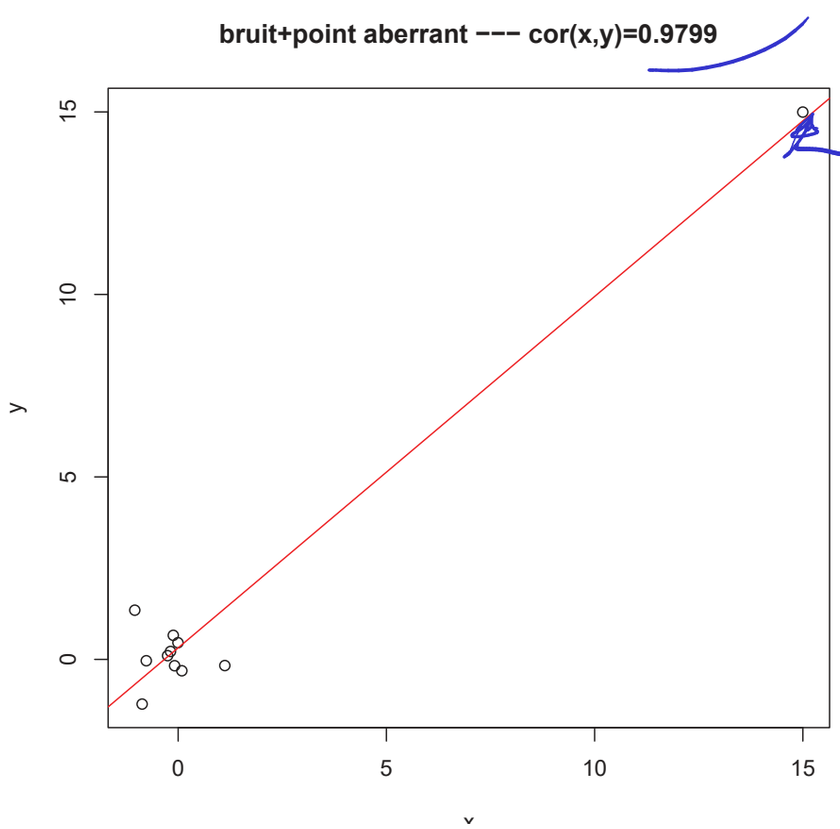
6 Discussions

6.1 Manipulation du coefficient r

Le coefficient de corrélation linéaire r ou le coefficient de détermination R^2 mesurent le caractère rectiligne du nuage. Il ne suffira pas à décrire, seul, la pertinence d'un ajustement.

Il faudra, en particulier, faire attention aux points aberrants.

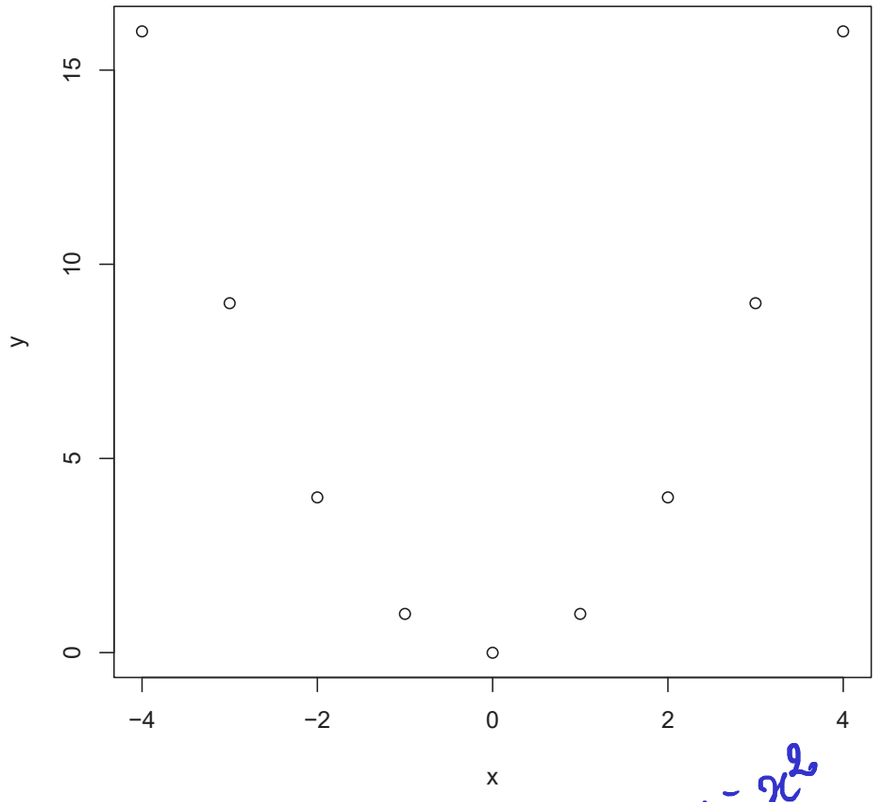




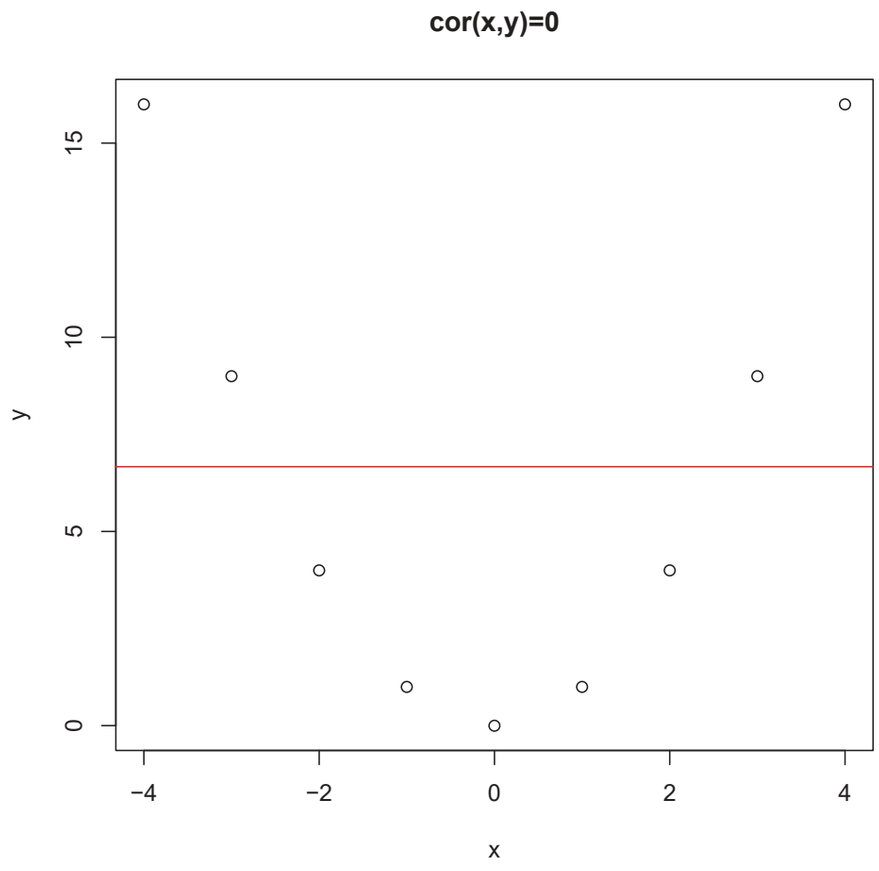
point aberrant

Toujours tracer le nuage de points.

Notons également qu'ils passent pleinement à coté d'une forme en cloche. Bien qu'il y ait une liaison entre les variables, la nature de cette liaison n'est pas linéaire.



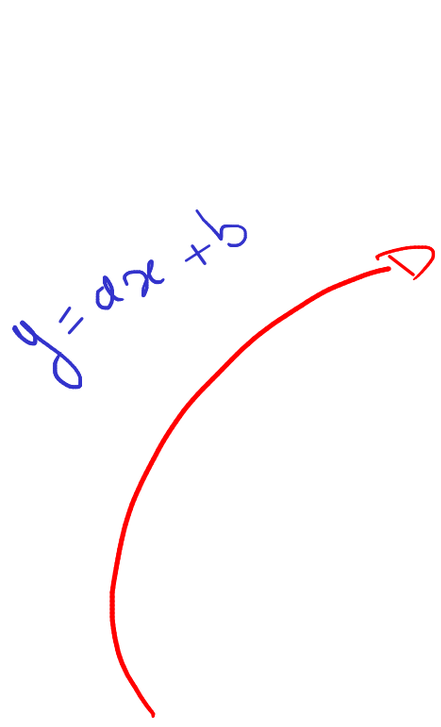
$y = x^2$

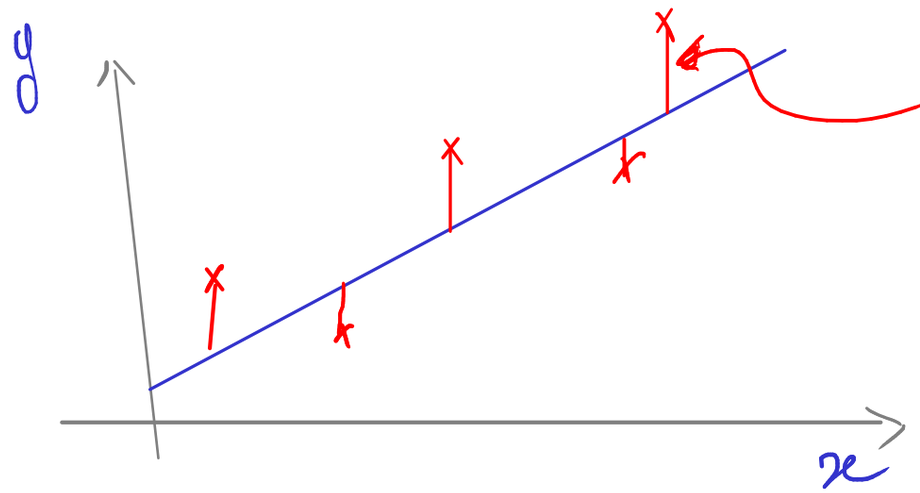


changement de variable ?

La forte corrélation mais non linéaire.

$\Rightarrow r_{xy} \simeq 0$





On a minisé ces distances distances verticales.

⇒ Pas d'erreur sur x

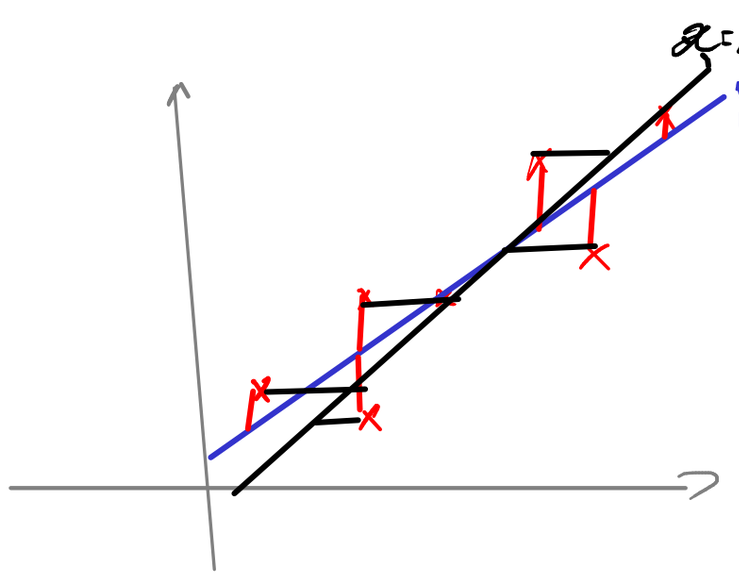
Pas de souci si y dépend de x

(pr une évolution temporelle par exemple)

6.2 Régression $x = my + p$

x chiffre d'affaires
y budget pub

→ y dépend de x et x dépend de y.))) il y a de l'erreur verticale (y) et horizontale x



Laquelle choisir ?

Si R^2 est proche de 1, le nuage est rectiligne et les droites sont proches des points donc les droites sont très proches.

Faux problème.

6.3 Changement de variable

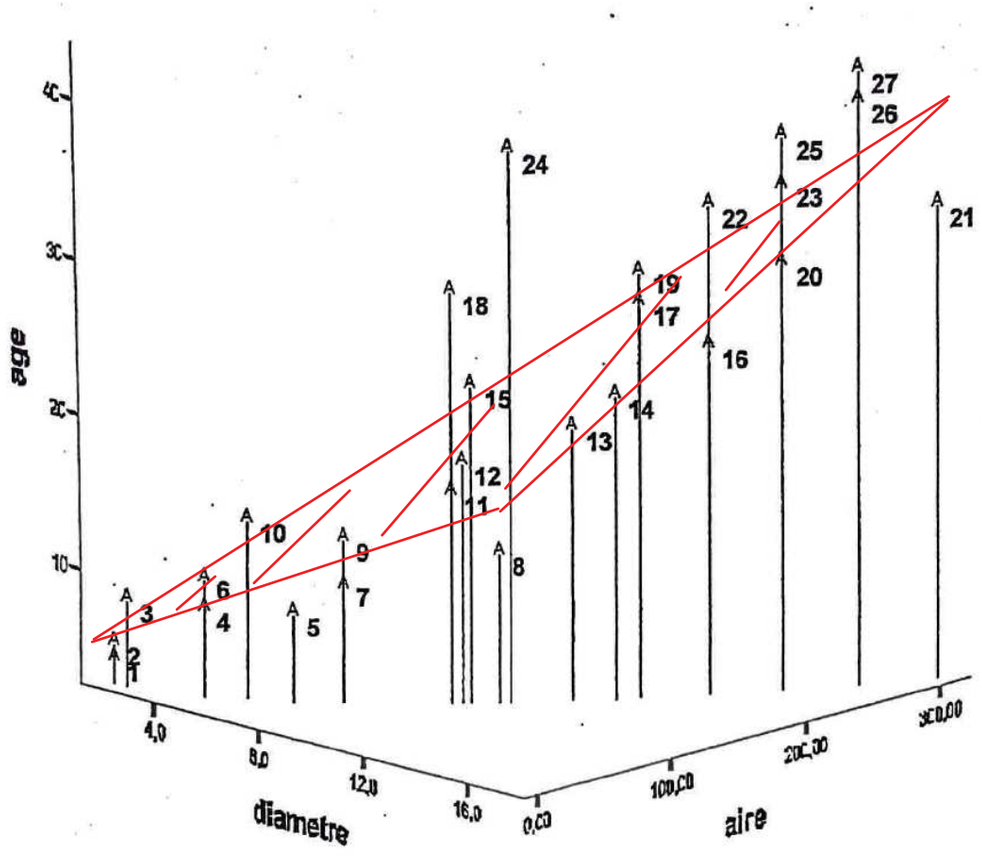
6.4 Régressions linéaire multiple

On peut souhaiter expliquer une variable réponse y par plusieurs variables explicatives $x_1, \cdots x_d$ en utilisant une équation cartésienne :

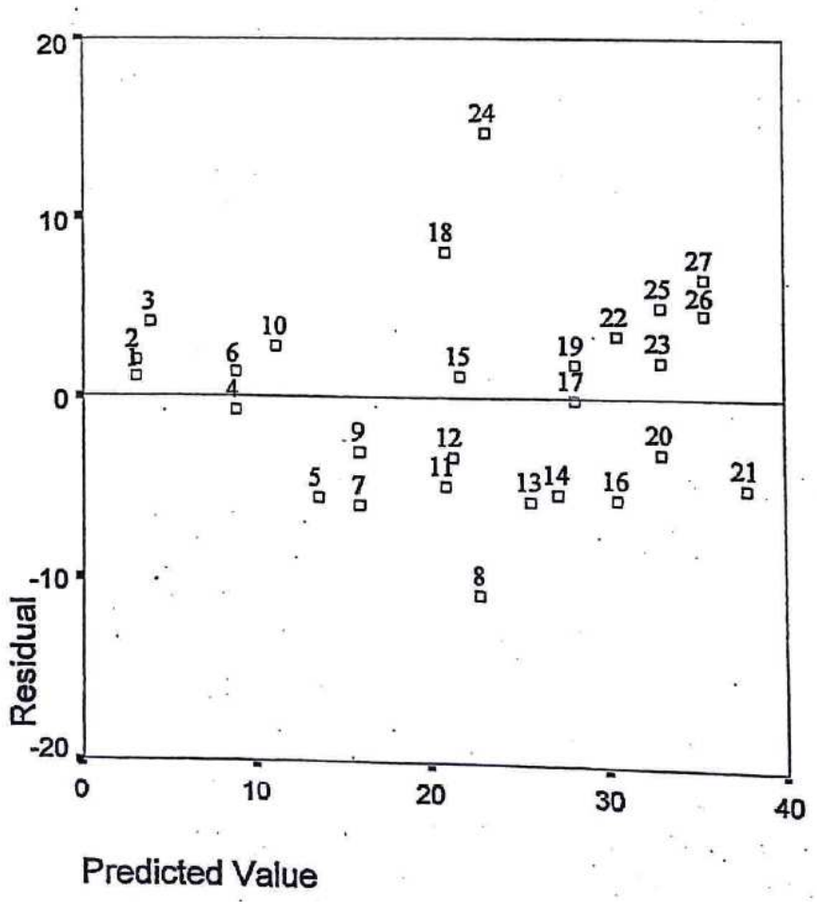
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d,$$

Reprenons notre exemple d'âge des arbres, et intéressons nous au diamètre et à la surface de la section (les arbres ne sont pas des cylindres parfaits, le diamètre n'explique pas totalement la surface). On obtient les graphes suivants

Nuage de points



Résidus



Résumé de la régression

	Coefficients non standardisés			
	B	Erreur standard	R	R-deux
(constante)	2,720	4,317		
DIAMETRE	1,161	1,001		
AIRE	,055	,051	,894[a]	,799