Part IV

Régression linéaire.

On rappelle la régression linéaire entre deux variables d'un point de vue descriptif.

9 Régression linéaire – Statistiques descriptives

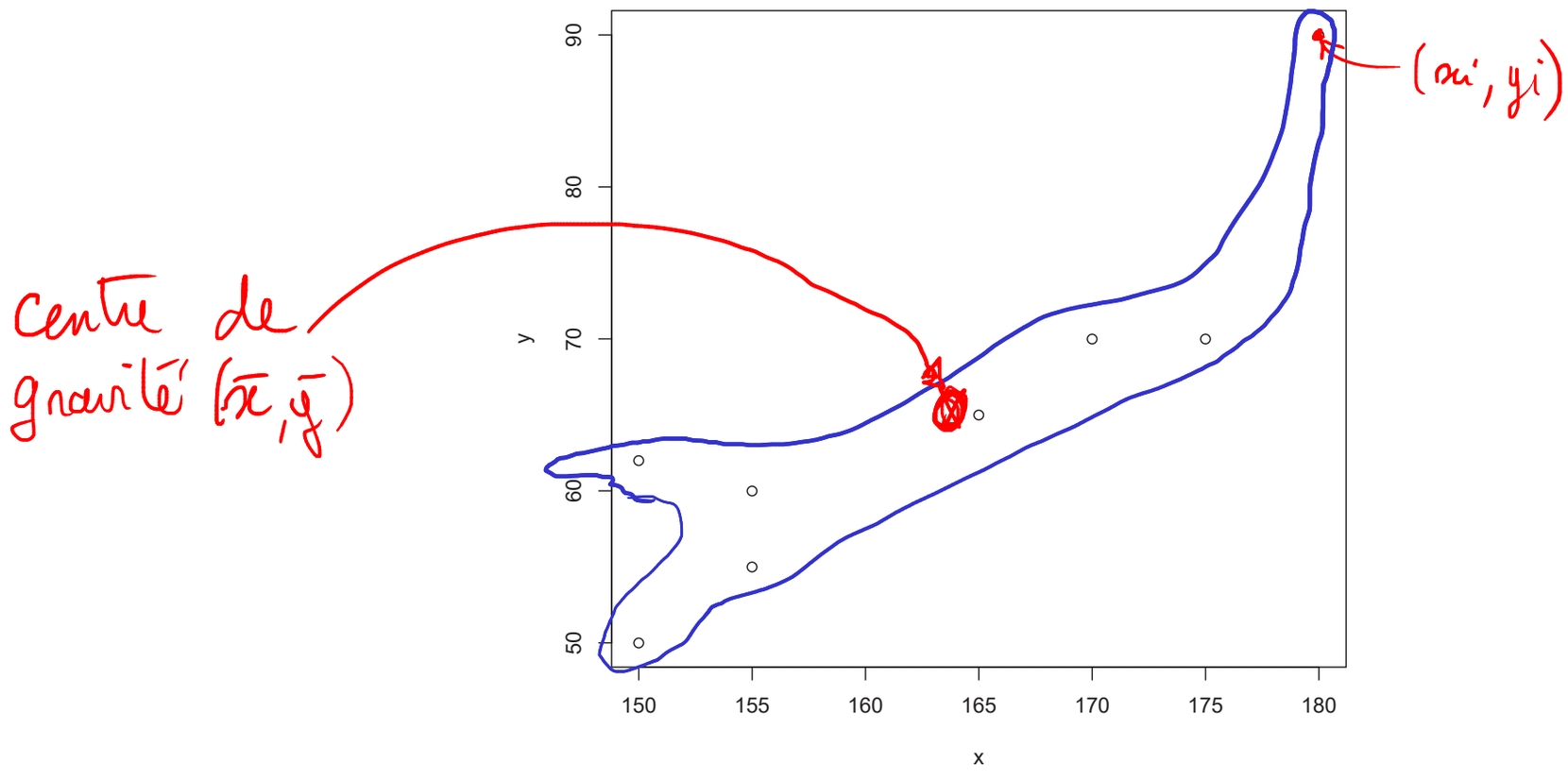
Dans certaines situations, on est amené à étudier deux caractères distincts d'une même population. On peut par exemple considérer la taille (x) et le poids (y) d'un ensemble d'individus. L'objectif principal de l'étude est de déterminer l'éventuel lien entre les deux variables x et y.

9.1 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 9.1 *Soit une population de N individus. Le graphe des N points (x_i, y_i) est appelé nuage de points de la série.*

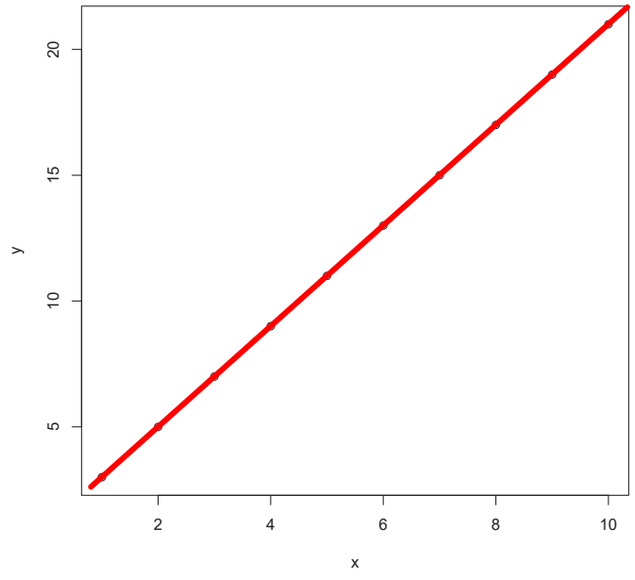


Definition 9.2 *Le point ayant pour coordonnées les moyennes $(\bar{x}, \bar{y})$ est appelé le point moyen.*

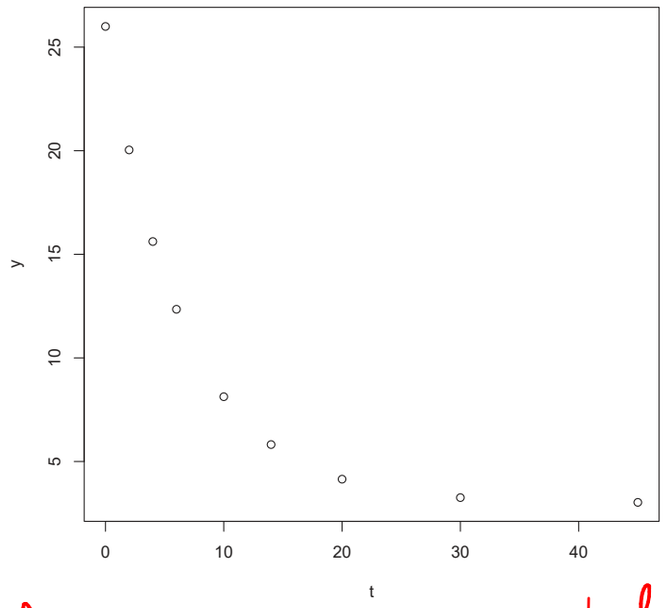
9.2 Forme du nuage de points

D'une manière générale, trois cas peuvent se présenter en ce qui concerne le profil du nuage :

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

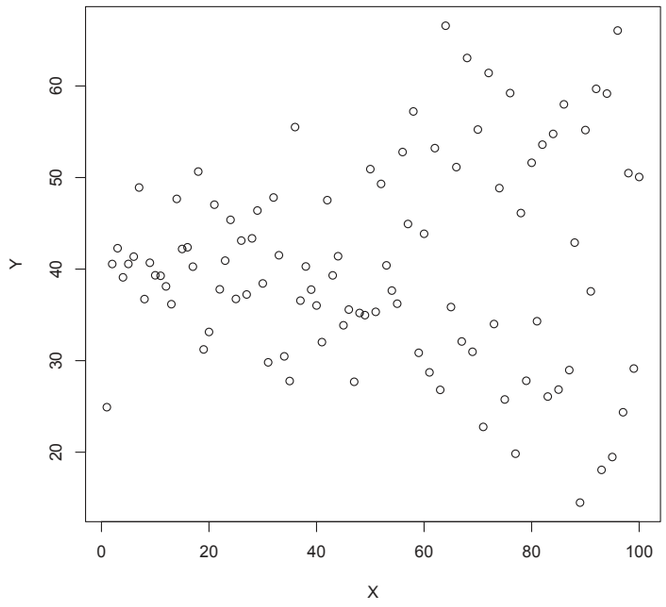


(ii) forme allongée mais non rectiligne : les points ne sont pas alignés mais ont un profil ordonné



changement de variable ?
(seulement si c'est imposé par le contexte).

(iii) forme quelconque

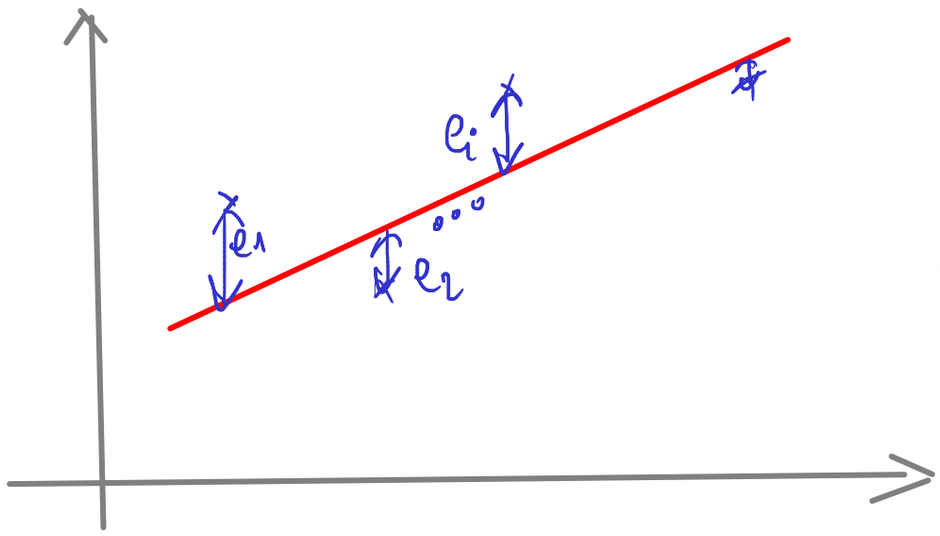


9.3 Ajustement affine (droite de régression linéaire)

On s'intéresse plus particièrement au premier cas 9.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

$$y = \beta_1 x + \beta_0$$

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.



But : minimiser la somme des écarts verticaux au carré (penser distance associée à une norme 2, euclidienne).

9.3.1 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*.

⚠ cache stat-descriptives sur population entière (pas d'estimation).

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

$$Var(y) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

En pratique, on détermine les coefficients de la droite $D : y = \beta_1 x + \beta_0$ à l'aide de `R` ou 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

$$\sigma_{xy} = Cov(x, y) = \frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y}).$$

$\sigma_x = \sqrt{Var\, x}$ $\sigma_y = \sqrt{Var\, y}$.
analogues à des normes

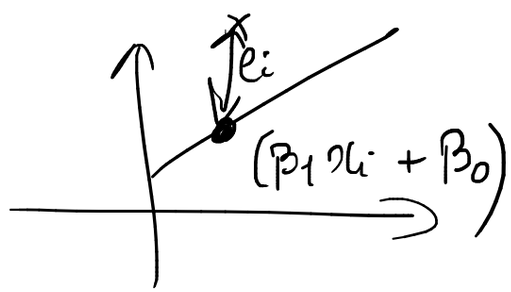
σ_{xy} analogue à un produit scalaire.

thm :

On a

$$\begin{aligned} \beta_1 &= cov(x,y)/\sigma_x^2, \\ \beta_0 &= \bar{y} - \beta_1 \bar{x}. \end{aligned}$$

La droite de régression passe par le centre de gravité.



$$e_i = y_i - (\beta_1 x_i + \beta_0)$$

$$e_i = y_i - \beta_1 x_i - \beta_0$$

Preuve. On pose la somme des carrés des résidus :

$$M(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2 = \sum_{i=1}^{n} e_i^2$$

linéaire en β_1 et β_0

Le minimum de $M(\beta_1, \beta_0)$ s'obtient en annulant les dérivées partielles par rapport à β_1 et β_0.

On cherche β_1, β_0 qui minimisent cette quantité.

$$\begin{cases} \dfrac{\partial M}{\partial \beta_1}(\beta_1, \beta_0) = \sum_{i=1}^{n} -2 x_i (y_i - \beta_1 x_i - \beta_0) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i - \beta_0) = 0 \\ \dfrac{\partial M}{\partial \beta_0}(\beta_1, \beta_0) = \sum_{i=1}^{n} -2 (y_i - \beta_1 x_i - \beta_0) = -2 \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0) = 0 \end{cases}$$

on les annule pour minimiser

On simplifie par -2

$$\begin{cases} \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i - \beta_0) = 0 \\ \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i - \bar{y} + \beta_1 \bar{x}) = 0 \\ \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \beta_0 \right) = 0 \end{cases}$$

$\times \frac{1}{n}$

$\times \left(\frac{1}{n}\right)$ $\Leftrightarrow$

$$\begin{cases} \frac{1}{n}\left(\sum_{i=1}^{n} x_i (y_i - \bar{y}) - \beta_1 \sum_{i=1}^{n} x_i (x_i - \bar{x}) \right) = 0 \\ \bar{y} - \beta_1 \bar{x} - \frac{n}{n}\beta_0 = 0 \Leftrightarrow \beta_0 = \bar{y} - \beta_1 \bar{x} \end{cases}$$

Notons que : $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i (y_i - \bar{y}) - \frac{1}{n} \bar{x} \sum_{i=1}^{n} (y_i - \bar{y})$

$\text{Cov}(x,y)$ ⟶

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

$\sum_{i=1}^{n} y_i = \bar{y}$

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

$\text{Var}(x)$ ⟶ $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} x_i (x_i - \bar{x})$.

La première ligne du système revient à

$$\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i-\bar{y}\right)-\beta_1\frac{1}{n}\sum_{i=1}^{n} x_i\left(x_i-\bar{x}\right)=0$$

$$Cov(x,y)-\beta_1 Var(x)=0$$

$$\Leftrightarrow \beta_1=\frac{Cov(x,y)}{Var(x)}$$

Pour montrer que c'est un minimum, on peut calculer la Hessienne et son déterminant > 0 (cf preuve du cours)

On intuitivement, on minimise une somme de carré, on ne peut avoir qu'un minimum.

9.3.2 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 9.2.3. 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

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

Si (x_i, y_i) forme une droite parfaite :

$|cov(x,y)| = \sigma_x \sigma_y$

$\Leftrightarrow r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \pm 1$.

) "produit scalaire/produit de norme"

Inégalité de Cauchy-Schwarz :

$|cov(x,y)| \leq \sigma_x \sigma_y$

Cas d'égalité si x et y sont colinéaires.

$\Leftrightarrow$ (x_i, y_i) forme une droite.

Proposition 9.3 *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 $r < -0.9$.