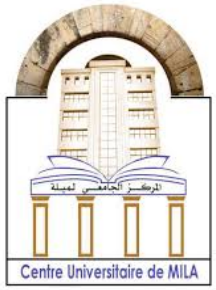
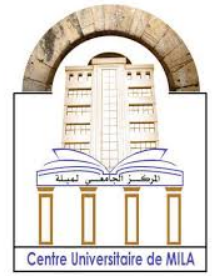
People's Democratic Republic Of Algeria
Ministry Of Higher Education And Scientific Research



Centtre Univercitaire Abdelhafid Boussouf Mila



Regression linéaire et corrélation linéaire

Octobre , 2022

En statistique, le terme de corrélation est utilisé afin de désigner la laison entre deux variables quantitatives (le plus souvent continues).

Coefficient de corrélation :

Le coefficient de corrélation permet de mesurer la dépendance linéaire entre deux variables quantitatives X et Y.

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1].$$

Coefficient de détermination : Le coefficient de détermination est le carré du coefficient de corrélation :

$$\rho^2_{XY} = \frac{\mathrm{Cov}^2(X, Y)}{V(X)V(Y)}.$$

Coefficient de corrélation :

Remarques :
- X et Y sont indépendantes, alors $\rho = 0$. La réciproque est fausse, sauf cas particulier ; si X et Y sont distribuées normalement.
- Si $\rho > 0$, les valeurs prises par Y ont tendance à croître quand les valeurs de X augmentent.
- Si $\rho < 0$, les valeurs prises par Y ont tendance à décroître quand les valeurs de X augmentent.
- $\mathrm{Si}|\rho| = 1$, alors il existe une relation linéaire parfaite entre X et Y.
- Un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux variables. Il peut exister une relation non linéaire entre elles.

Quelques exemples de corrélation (Ajustement linéaire :)

On considère n individus sur lesquels on mesure X et Y deux variables quantitatives. Pour chaque individu $i(1 \leq i \leq n)$, on dispose d'un couple d'observations (x_i, y_i) qui représente les valeurs prises par X et Y pour l'individu i.
Les données $(x_i, y_i), i = 1, \ldots, n$ peuvent être représentées par un nuage de n points dans le plan (x, y)

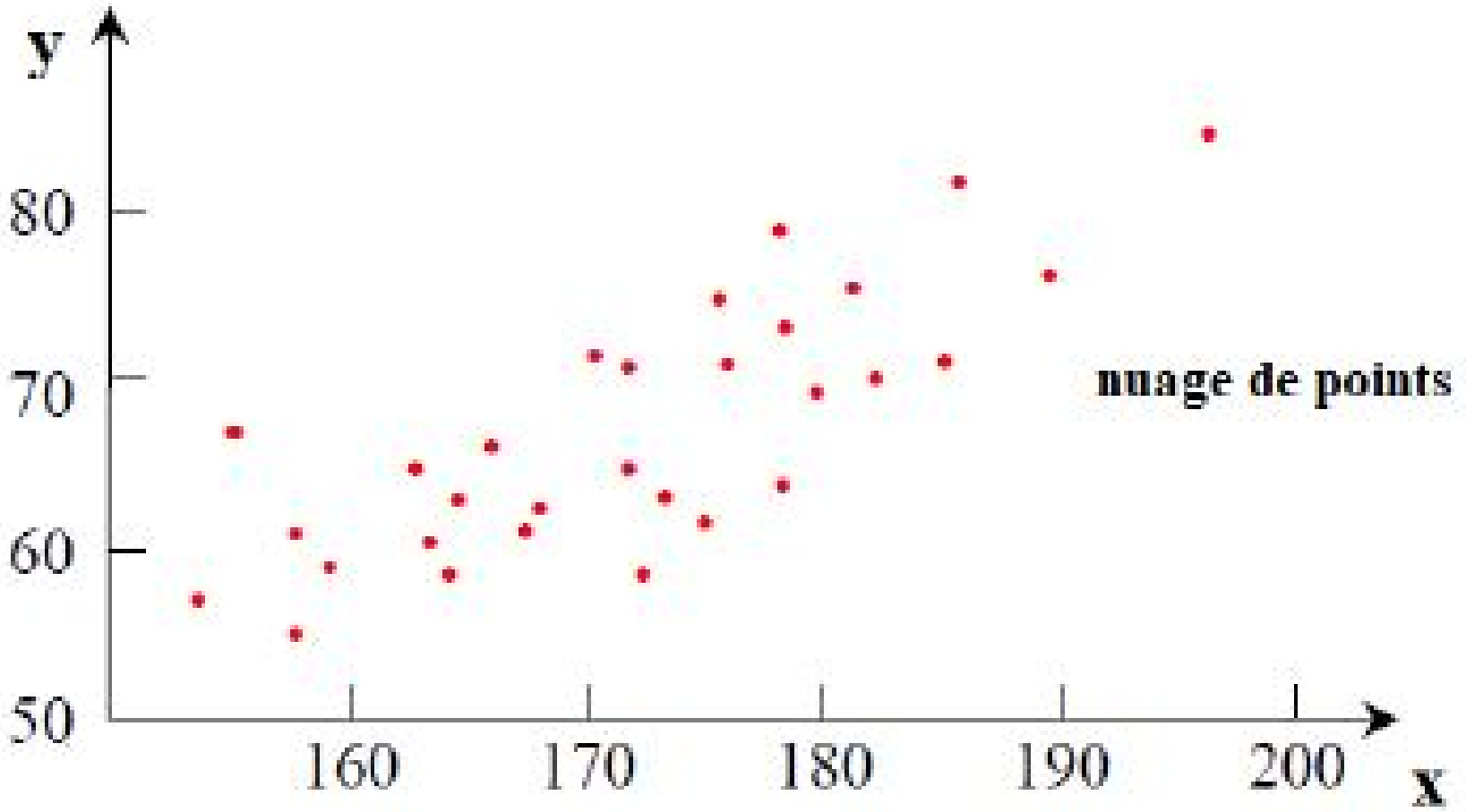


Figure – Nuage de points

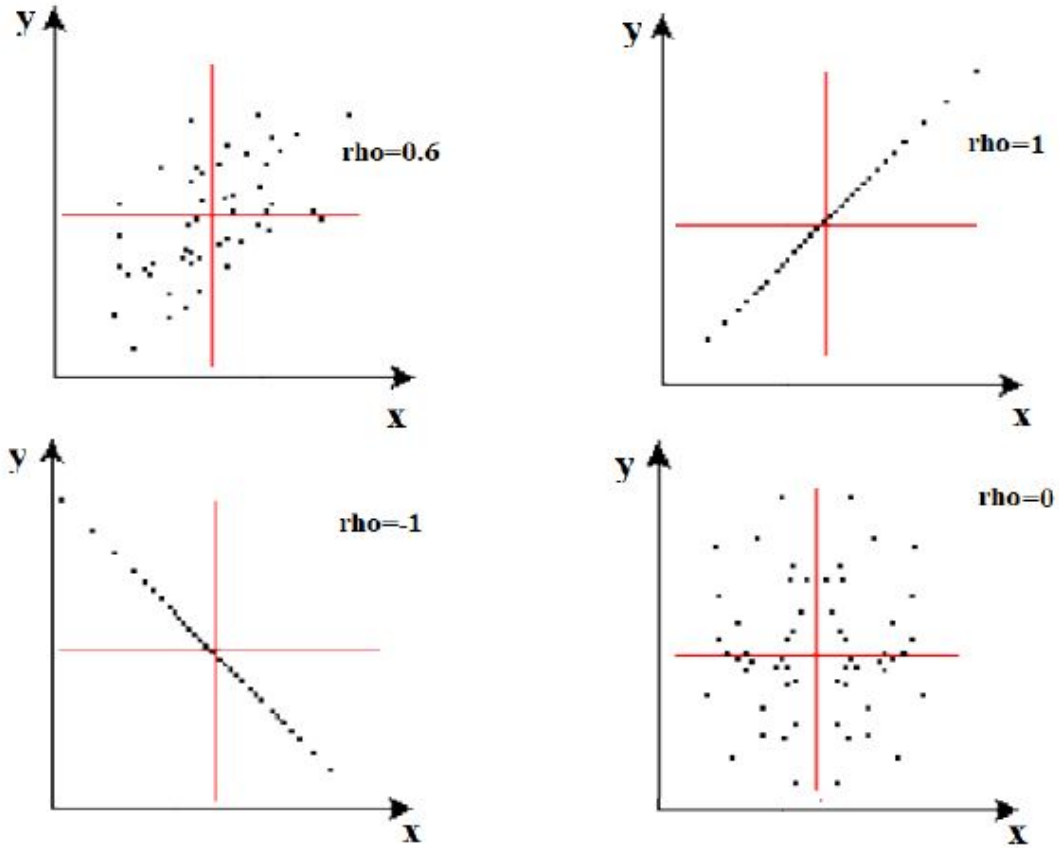


Figure – Nuage de points pour differentes valeure de rho

Ajustement linéaire au sens des moindres carrés :)

Selon l'allure du nuage, on a envie de remplacer ce nuage par le graphe d'une fonction f. Cette opération s'appelle un ajustement. La nature de l'ajustement dépend de la forme du nuage de points.
Nous étudions l'ajustement linéaire par la méthode des moindres carrés.
On considère les données $\{y_1, y_2, \ldots, y_n\}$ comme étant des réalisations d'une variable aléatoire Y et les données $\{x_1, x_2, \ldots, x_n\}$ les réalisations d'une variable aléatoire X.
- La variable X est une variable, aléatoire ou contrôlée, dite explicative.
- Y est une variable aléatoire dite à expliquer.

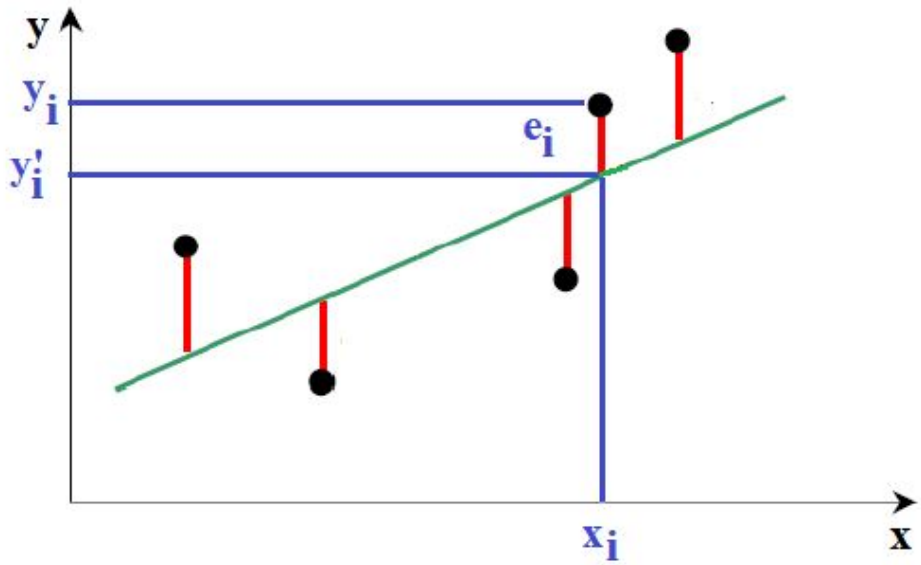


Figure – Nuage de points pour differentes valeure de rho

- Le problème de l'ajustement linéaire par la méthode des moindres carrés consiste à rechercher une relation affine entre les variables X et Y, ceci revient à trouver une droite qui s'ajuste le mieux possible à ce nuage de points.
Parmi toutes les droites possibles, on retient celle qui rend minimale la somme des carrés des écarts des valeurs observées y_i à la droite $y_i' = ax_i + b$.
Si les coefficients a et b étaient connus, on pourrait calculer les résidus de la régression définis par :

$$e_i = y_i - ax_i - b.$$

Le résidu e_i est l'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i. Les résidus peuvent être positifs ou négatifs.

Afin de déterminer la valeur des coefficients a et b nous utilisons le principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés des résidus :

$$\varepsilon(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 .$$

Calculons les coefficients a et b qui minimisent le critère des moindres carrés :
Pour déterminer le minimum la fonction ε, on doit avoir : $\frac{\partial \varepsilon}{\partial a} = 0$ et $\frac{\partial \varepsilon}{\partial b} = 0$.

On a $\frac{\partial \varepsilon}{\partial b} = 0 \Leftrightarrow \sum_{i=1}^{n}(-2)\left(y_i - ax_i - b\right) = 0 \Leftrightarrow \sum_{i=1}^{n}\left(y_i - ax_i - b\right) = 0$
$\Leftrightarrow \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = 0 \Leftrightarrow n\bar{y} - an\bar{x} - nb = 0 \Leftrightarrow \bar{y} - a\bar{x} - b = 0$
$\Leftrightarrow b = \bar{y} - a\bar{x}$.

Et $\frac{\partial \varepsilon}{\partial a} = 0 \Leftrightarrow \sum_{i=1}^{n} (-2x_i)(y_i - ax_i - b) = 0$
$\Leftrightarrow \sum_{i=1}^{n} x_i (y_i - ax_i - b) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} bx_i = 0$
$\Leftrightarrow n\overline{xy} - an\overline{x^2} - bn\bar{x} = 0 \Leftrightarrow \overline{xy} - a\overline{x^2} - (\bar{y} - a\bar{x})\bar{x} = 0$
$\Leftrightarrow \overline{xy} - a\overline{x^2} - \overline{yx} + a\bar{x}^2 = 0 \Leftrightarrow \overline{xy} - \overline{yx} = a\left(\overline{x^2} - \bar{x}^2\right) \Leftrightarrow a = \frac{\overline{xy} - \overline{yx}}{\left(\overline{x^2} - \bar{x}^2\right)}$
$\Leftrightarrow a = \frac{\mathrm{Cov}(X,Y)}{V(X)}$. D'ou la droite de régression de Y par rapport à X, est :

$$y = ax + b$$

avec

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

et.

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

Remarque La droite de régression de X par rapport à Y est :

$$x = a'y + b',$$

avec

$$a' = \frac{\mathrm{Cov}(X, Y)}{V(Y)} \text{ et } b' = \bar{x} - a'\bar{y}$$

En pratique, on estime que la régression est acceptable lorsque $|\rho_{XY}| \geq 0.85$.

Exemple : On s intéresse à la corrélation entre la taille X et le poids Y de quatre individus :

Individus	Taille (cm)	Poids (kg)
1	180	79
2	170	75
9	175	85
4	160	59

1. Calculer le coefficient de corrélation. 2. Trouver la droite de régression.

1. Les moyennes des variables X et Y.

$$\bar{x} = \frac{1}{4}(180 + 170 + 175 + 160) = 171.25; \bar{y} = \frac{1}{4}(79 + 75 + 85 + 59) = 74.5$$

2. Les variances des variables X et Y.

$$V(X) = \frac{1}{n}\sum_{i=1}^{4} x_i^2 - \bar{x}^2 = 54.68; V(Y) = \frac{1}{n}\sum_{i=1}^{4} y_i^2 - \bar{y}^2 = 92.75$$

3. La covariance de X et Y :

$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{4} x_i y_i - \overline{xy} = \overline{xy} - \overline{xy} = 63.12.$$

4. Le coefficient de corrélation :

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = 0.88.$$

5. La droite de régression de Y en X :

$$a = \frac{\mathrm{Cov}(X, Y)}{V(X)} = 1.15$$

$$b = \bar{y} - a\bar{x} = -122.43.$$

D'où

$$y = ax + b = 1.15x - 122.43$$

Fin du partie 2