Chapitre 02 :

Corrélation et régression linéaire

1. La régression linéaire :

Lorsqu'il existe une relation logique entre deux variables X et Y, il est intéressant de l'exprimer sous la forme d'un mobile mathématique qui sert à estimer la valeur Y correspondant à une valeur donnée de X. C'est ce qu'on appelle l'analyse de régression (ou théorie de la régression). Lorsque le nuage statistique (de points) indique qu'il existe une corrélation entre deux variables, on exprime mathématiquement cette relation par l'équation d'une droite

$$Y = aX + b$$

On appelle régression linéaire l'ajustement d'une droite au nuage statistique d'une série de couples de données (x_i, y_i). X (variable indépendante) tandis que Y (variable dépendante).

X = variable explicative / Y = variable expliquée

X = variable indépendante / Y = variable dépendante

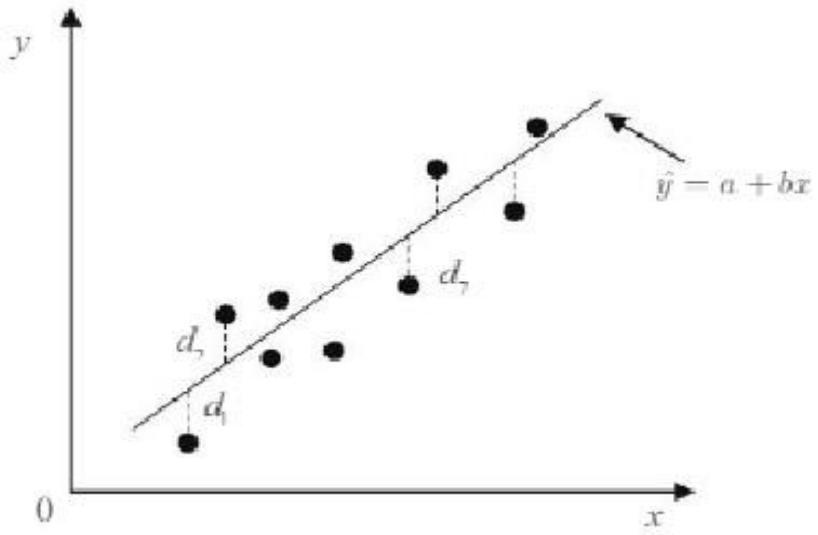
Le fait de trouver l'équation de la droite mettant en relation deux variables nous fournira un outil de prévision ou d'estimation. En effet, à partir de cette équation, on pourra estimer ou prévoir les valeurs d'une variable dite dépendante en fonction des valeurs prises par l'autre variable dite indépendante. La méthode pour y parvenir est la suivante : Considérons n couples de données provenant de l'étude de deux variables statistiques X et Y.

Considérons n couples de données provenant de l'étude de deux variables statistiques X et Y.

X (Variable indépendante)	x_1	x_2	x_3	$\cdots$	x_n
Y (Variable dépendante)	y_1	y_2	y_3	$\cdots$	y_n

On représente ces couples par un nuage statistique. Il s'agit maintenant de trouver une droite notée $Y = aX + b$, pouvant représenter convenablement la relation ou la tendance se manifestant entre la variable Y (variable dépendante) et la variable X (variable indépendante).

Le graphique suivant illustre la situation :

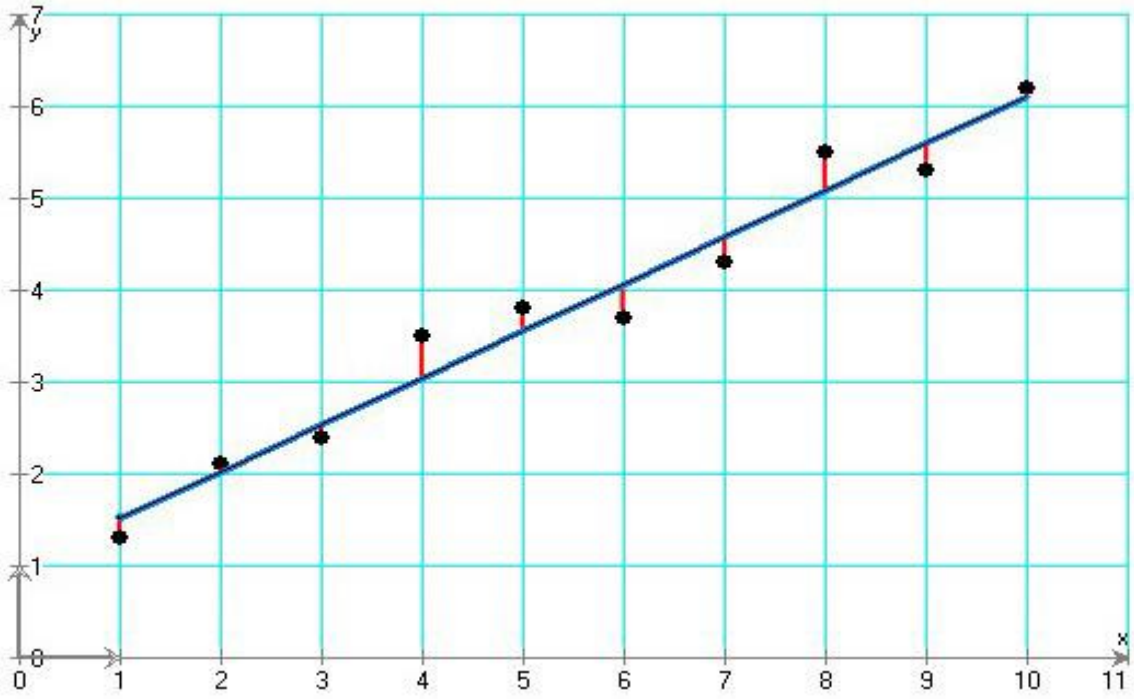


La méthode des moindres carrés :

Une situation courante en sciences d'avoir à sa disposition deux ensembles de données de taille n, $\{y_1, y_2, y_3, \dots, y_n\}$ et $\{x_1, x, x_3, \dots, x_n\}$, obtenus expérimentalement ou mesurés sur une population. Le problème de la régression consiste à rechercher une relation pouvant éventuellement exister entre les x et les y, par exemple de la forme $y = f(x)$. Lorsque la relation recherchée est affine, c'est-à-dire de la forme $y = ax + b$, on parle de régression linéaire.

Les données $\{(x_i, y_i), i = 1, \dots, n\}$ peuvent être représentées par un nuage de n points dans le plan (x, y), le diagramme de dispersion. Le centre de gravité de ce nuage peut se calculer facilement : il s'agit du point de coordonnées $(\bar{x}, \bar{y})$. Rechercher une relation affine entre les variables X et Y revient à rechercher une droite qui s'ajuste le mieux possible à ce nuage de points. Parmi toutes les droites possibles, on retient celle qui jouit d'une propriété remarquable : c'est celle qui rend minimale la somme des carrés des écarts des valeurs observées y_i à la droite $\hat{y_i} = ax_i + b$. Le principe des moindres carrés ordinaire (MCO) consiste à choisir les valeurs de a et de b qui minimisent

$$E = \sum_{i=0}^{n} \big(y_i - (ax_i + b)\big)^2.$$



Sur le dessin, chaque trait vertical rouge représente la valeur $y_i - \widehat{y_i}$

Les coefficients b et a de la droite de régression sont appelés coefficients de régression. Ils sont donnés par les formules ci-dessous :

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \; et \; b = \bar{Y} - a\bar{X} = \frac{\sum y_i}{n} - a\frac{\sum x_i}{n}.$$

2. Corrélation

On dira qu'il y a corrélation, ou dépendance, entre deux variables quantitatives $X \; et \; Y$ si elles ont généralement tendance à varier toutes deux dans le même sens ou en sens contraire. Les caractéristiques d'une corrélation entre deux variables $X \; et \; Y$ sont :

►La forme :

● Linéaire : les points du diagramme de dispersion ont tendance à se rapprocher d'une droite.

● Non linéaire : les points du diagramme de dispersion ont tendance à se rapprocher d'une courbe.

►le sens :

● Positif : les deux variables varient dans le même sens : quand les valeurs de la variable X augmentent, celles de la variable Y augmentent aussi.

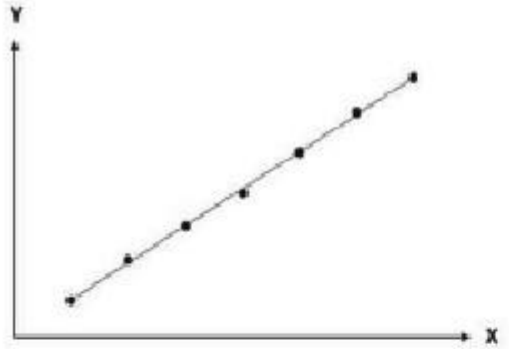
● Négatif : les deux variables varient en sens contraires : quand les valeurs de la variable X augmentent, celles de la variable Y diminuent.

►L'intensité :

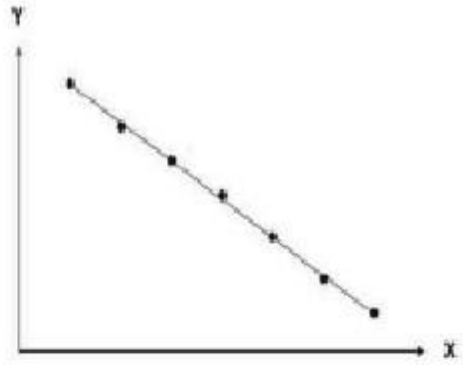
● Parfaite : les points du diagramme de dispersion sont parfaitement alignés, dans le cas d'une corrélation linéaire, ou tous situés sur la courbe dans le cas d'une corrélation non linéaire (Une dépendance parfaite permet de déterminer, pour chaque valeur de la variable X, la valeur exacte de la variable Y qui lui est associée, et vice-versa).

● Imparfaite (faible, forte, moyenne) : on constate une tendance moins forte des points du diagramme de dispersion à s'aligner ou à prendre la forme d'une courbe. Dans ce cas, on peut tout au plus estimer approximativement la valeur de la variable Y correspondant à une valeur donnée de la variable X.

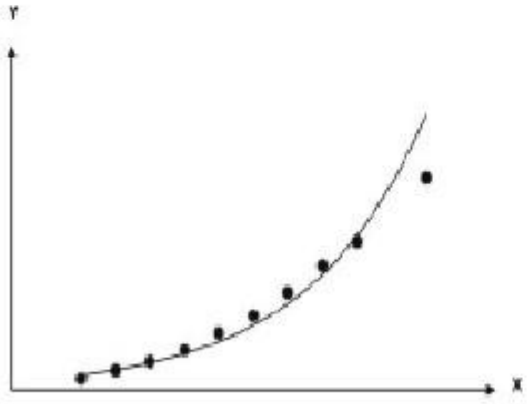
● Nulle ou inexistante : les points sont complètement éparpillés dans le plan et ne semblent suivre aucune orientation ni s'approcher d'une droite ou une courbe. On dit alors que les variables sont indépendantes. Il est alors impossible d'estimer la valeur de Y correspondant à une valeur donnée de X, et vice-versa.



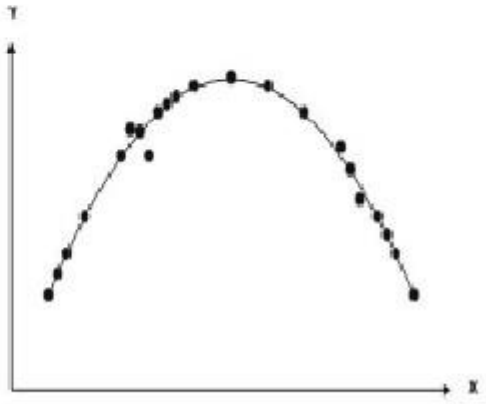
1. Corrélation linéaire positive parfaite



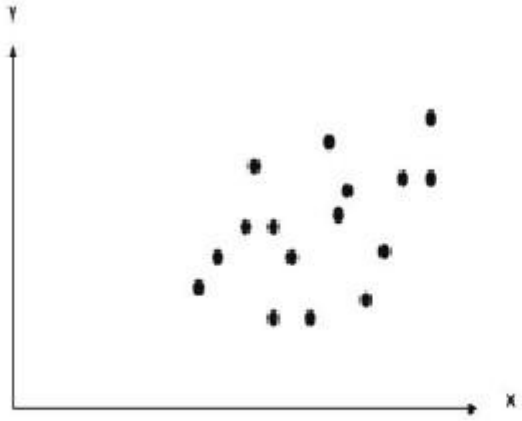
2. Corrélation linéaire négative parfaite



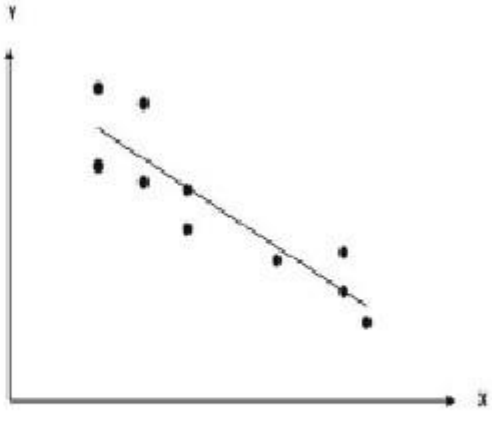
3. Corrélation non linéaire (exponentielle) positive forte



4. Corrélation non linéaire (parabolique ou quadratique) négative parfaite



5. Corrélation nulle



6. Corrélation linéaire négative faible

COEFFICIENT DE CORRÉLATIONS (OU DE PEARSON) :

Le nuage de points permet une analyse qualitative de la tendance à une relation linéaire entre les variables X et Y. Le coefficient de corrélation linéaire, ou coefficient de Pearson noté r, est un nombre sans dimension qui mesure quantitativement l'intensité de la corrélation (ou de la dépendance) linéaire entre les deux variables.

On le calcule à l'aide de la formule suivante :

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Interprétation du coefficient de corrélation :

Valeur r	Ce qu'elle dit sur la relation
+1	Corrélation directe parfaite
De 0.70 à 0.99	Corrélation directe forte
De 0.50 à 0.69	Corrélation directe moyenne
De 0.01 à 0.49	Corrélation directe faible
0	Aucune relation

Remarque : *Ce qui a été dit à propos de la corrélation directe s'applique également à la corrélation inverse avec un signe négatif* $(-)$.