جامعة الإخوة منتوري قسنطينة 1
Frères Mentouri Constantin 1 University
Université Frères Mentouri Constantine 1

DEPARTEMENT BIOCHIMIE

1ere année master biochimie applique et biochimie

COURS BIOSTATISTIQUE

RESPONSABLE DU MODULE: DR.ZEGHBID NASSIM LOTFI

Année Universitaire : 2020

Corrélation et régression linéaire

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 (xi, yi). 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.

La méthode pour y parvenir est la suivante :

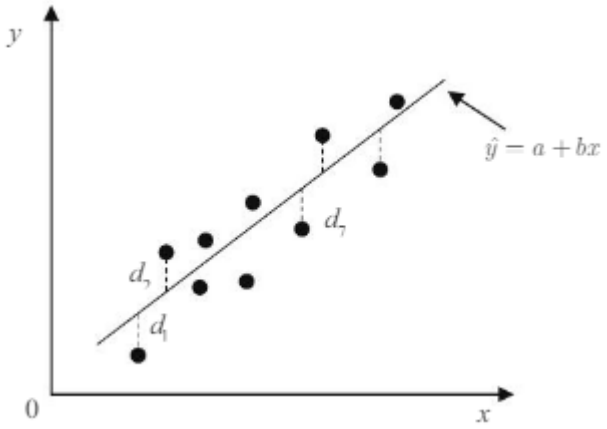
Considérons n couples de données provenant de l'étude de deux variables statistiques X et Y.

X (variable indépendante)	x_1	x_2	x_3	...	x_n
Y (variable dépendante)	y_1	y_2	y_3	...	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 dépendant) et la variable X (variable indépendante).

Le graphique suivant illustre la situation :



Exemple :

Etude de la relation entre la tension artérielle et l'âge d'un individu

Objectif On souhaite savoir si, de façon générale, l'âge a une influence sur la tension artérielle et sous quelle forme cette influence peut être exprimée. Le but est d'expliquer au mieux comment la tension artérielle varie en fonction de l'âge et éventuellement de prédire la tension à partir de l'âge.

Population et variables étudiées

Population générale d'individus. Sur cette population, on définit deux variables.

La variable Y : variable tension ; c'est la variable à expliquer, appelée encore variable à régresser, variable réponse, variable dépendante (VD).

La variable X : variable âge ; c'est la variable explicative, appelée également régresseur, variable indépendante (VI).

La méthode de calcule de la régression linaire et la méthode des moindres carrés

Tout d'abord on cherche l'equation $Y = aX + b$

$$\hat{a} = \frac{n\sum X_i y_i - \sum X_i \sum y_i}{n\sum X_1^2 - (\sum X_i)^2}$$

$$\widehat{b} = \overline{y} - \hat{a}\overline{X} \quad = \quad \frac{\sum y_i}{n} - \hat{a}\frac{\sum X_i}{n} = \widehat{b}$$

Pour x et y c la moyenne arithmétique avec simple donné

N c le nombre de population statistique

Exemple l : les données suivante exprime le relation entre la consommation du sucre et le taux de glycémie

y taux de glycemie	**10**	**15**	**5**	**4**	**3**	**13**
X sucre	7	10	4	3	2	8

1- Trouvé l équation de la régression linaire de y sur x.
2- Quelle sera le niveau de la glycémie lorsque la consommation du sucre 20 unité. **(X=20)**
   Solution

y	**x**	**X y**	x^2
10	7	70	49
15	10	150	100
5	4	20	16
4	3	12	9
3	2	6	4
13	9	117	81
50	**35**	**375**	**259**

On calcule

Y = aX + b

$$\hat{a} = \frac{n\sum X_i y_i - \sum X_i \sum y_i}{n\sum X_1^2 - (\sum X_i)^2}$$

$$= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{6.375 - (35.50)}{6.259 - (35)^2} = \frac{83.33}{54.83} = 1.52$$

$$\hat{b} = \bar{y} - \hat{a}\bar{X} \quad = \quad \frac{\sum y_i}{n} - \hat{a}\frac{\sum X_i}{n} = \hat{b}$$

$$= \frac{50}{6} - 1.52\frac{35}{6} = -32$$

Alors l'équation de Y = aX + b est de

$$\hat{y} = -3.2 + 1.52x$$

Estimation du taux de glycémie lorsque la consommation du sucre augmente de **(X=20) elle devient 27.2 .**

$$\hat{y} = -3.2 + 1.52.20 = 27.2$$

Exemple 2 : soit les données suivantes qui designe un relation entre deux variable

$$\sum y = 144 \quad \sum x = 103 \quad \sum xy = 2093$$

$$\sum y^2 = 3012 \quad \sum x^2 = 1531 \quad n=7$$

1- Trouvé l équation de la régression linaire de y sur x.
2- Estimation du taux du gras lorsque la consommation du sucre augmente de **(X=10)**

Solution :

On calcule :

Y = aX + b

$$\hat{a} = \frac{n\sum X_i y_i - \sum X_i \sum y_i}{n\sum X_1^2 - (\sum X_i)^2}$$

$$= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{7.2093 - (103.144)}{7.1531 - (103)^2}$$

$$= \frac{-181}{108} = -1.68$$

$$\hat{b} = \bar{y} - \hat{a}\bar{X} \quad = \quad \frac{\sum y_i}{n} - \hat{a}\frac{\sum X_i}{n} = \hat{b}$$

$$= \frac{144}{7} - (-1.68)\frac{103}{7} = 20.57 + 24.72 = 45.29$$

Alors l'équation de Y = aX + b est de

$$\hat{y} = 45.29 - 1.68x$$

Estimation du taux de glycémie lorsque la consommation du sucre augmente de **(X=10) elle devient 28.49 .**

$$\hat{y} = 45.29 - 1.68.(10) = 28.49$$

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. C'est ce type de corrélation que nous étudions (exemples : graphiques 1, 2 et 6).

Non linéaire : les points du diagramme de dispersion ont tendance à se rapprocher d'une courbe (exemples : les graphiques 3 et 4).

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 (exemples : les graphiques 1 et 3).

- négatif : les deux variables varient en sens contraires : quand les valeurs de la variable X augmentent, celles de la variable Y diminuent (exemples : les graphiques 2, 4 et 6).

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.

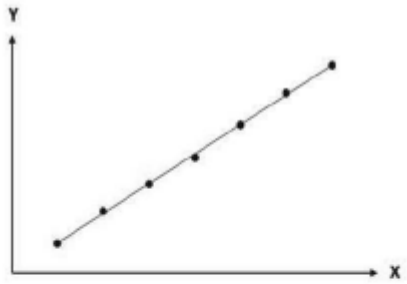
-Une dépendance parfaite permet de déterminer, pour chaque valeur de la variable X,

la valeur exacte de la variable Y qui lui est associée, et vice-versa

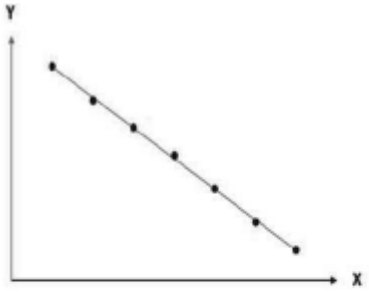
(exemples : les graphiques 1, 2 et 4).

- Une dépendance 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 (exemples : le graphique 3 et 6).

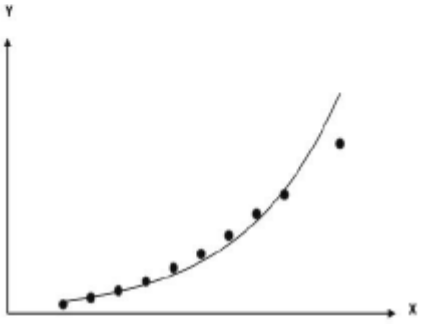
- nulle ou inexistante : les points sont complètement éparpillés dans le plan et ne semblent suivre aucune orientation ni s'approcher d'une droite ou une courbe. On dit alors que les variables sont indépendantes. Il est alors impossible d'estimer la valeur de Y correspondant à une valeur donnée de X, et vice-versa (exemples : le graphique 5).



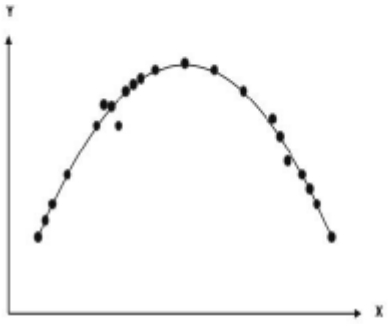
1. Corrélation linéaire positive parfaite



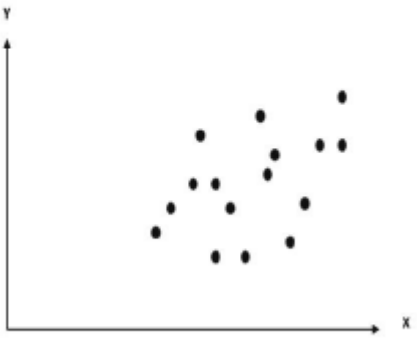
2. Corrélation linéaire négative parfaite



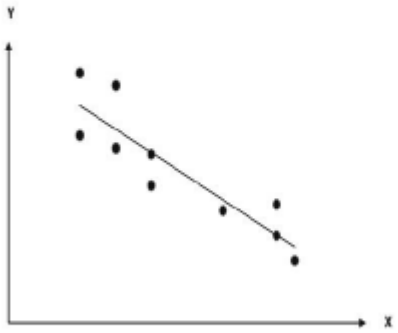
3. Corrélation non linéaire (exponentielle) positive forte



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 :

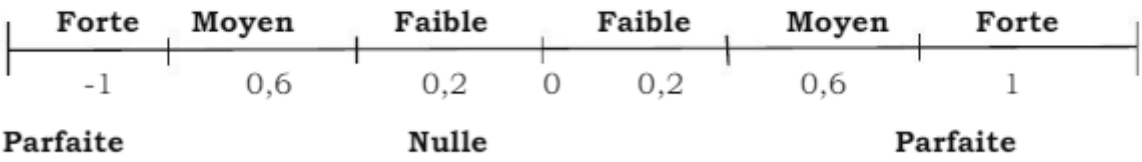
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

Interprétation du coefficient de corrélation :

P1. $-1 \leq r \leq 1$.

P2. La corrélation linéaire et parfaite et positive de $r = 1$, parfaite et négative si $r = -1$ et nulle si $r = 0$,

P3. Dans le cas d'une corrélation linéaire positive ($r \geq 0$), plus la valeur de r est pris de 1, plus la corrélation entre X et Y est forte. Il en est de même pour une corrélation linéaire négative : plus la valeur est près de -1, plus la corrélation entre X et Y est forte. Le schéma suivant donne une idée de la force d'une corrélation en sciences humaines.



P4. La valeur $100\,r^2$ donne le pourcentage de variation totale de Y qui s'explique par la dépendance de Y par rapport à X.

exemples : soit les données suivant qui montre une relation linaire en deux variable quantitative

10	7	6	2	X
4	8	10	12	Y

1- Trouvé l équation de la régression linaire de y sur x.
2- Calculez le coefficient de correlation Pearson.

Solution :

On calcule :

Y = aX + b

y^2	x_i^2	x_iy_i	y	x
144	4	24	12	2
100	36	60	10	6
64	49	56	8	7
16	100	40	4	10
324	**189**	**180**	**34**	**25**

$$\hat{a} = \frac{n\sum X_i y_i - \sum X_i y_i}{n\sum X_1^2 - (\sum X_i)^2}$$

$$\hat{a} = \frac{720 - 850}{756 - 625} = \frac{130}{131} = -0.99$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = \frac{34}{4} + 0.99\frac{25}{4} = 8.5 \times 6.55$$

$$\hat{b} = 14.7$$

$$\hat{\gamma} = 14.7 - 0.99X_i$$

coefficient de corrélation Pearson

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

$$\gamma = \frac{n\sum X_i y_i - \sum X_i \sum y_i}{\sqrt{n\sum X_1^2 - (\sum X_i)^2}\sqrt{n\sum y_1^2 - (\sum y_i)^2}}$$

$$\gamma = \frac{-130}{\sqrt{(756 - 625)(1296 - 1156)}} = \frac{130}{\sqrt{131 \times 140}}$$

$$\gamma = -0.962$$

Interprétation Corrélation fort parfaite

3- Corrélation de Spearman**:**

Définition et propriété:

Le coefficient de corrélation de Spearman, symbolisé par , mesure le degré de liaison existant entre le classement des éléments selon la variable X et le classement selon Y .il

s’agit en fait d’un coefficient de corrélation de Pearson calculé non pas sur les valeurs de X et Y , mais sur les rangs des valeurs de X et Y .

On peut le calculez avec l’équation

$$\gamma_s = 1 - \frac{6\sum_{i=1}^{n} d_1^2}{n(n^2-1)}$$

Avec :

Avant de trouve RX et RY on doit ordonné x et y soit de l’inferieure au supérieur / soit de supérieur a l’inferieure pour les deux l’important ces qu’ils soient ordonnés de la même façon

RX correspond au rang de la variable x soit de l’inferieure au supérieur / soit de supérieur a l’inferieure

RY correspond au rang de la variable y soit de l’inferieure au supérieur / soit de supérieur a l’inferieure

A noter que pour les deux variable x ou y le rang doit être de la même façon établie soit du supérieur a l’inferieure soit de l’inferieure au supérieur.

Di Correspond à l’ écart = **RX-RY.**

Exemple : les données suivante définie une relation linaire entre deux variable x y

9	6	13	7	16	8	11	10	**(X)**
4	2	9	1	12	2	7	3	**(y)**

1-Calculez le coefficient de Spearman.

Solution

D_{i2}	d_i	RY	RX	Y	X
1	**1**	**4**	**5**	**1**	**6**
0	**0**	**6**	**6**	2	7
0.25	**0.5**	**2.5**	**3**	2	8
0	**0**	**8**	**8**	3	9
1	**1**	**1**	**2**	4	10
0	**0**	**7**	**7**	7	11
2.25	**-1.5**	**2.5**	**1**	9	13
1	**-1**	**5**	**4**	12	16
5.5					

$$\gamma_s = 1 - \frac{6\sum_{i=1}^{n} d_1^2}{n(n^2-1)}$$

$$\gamma_s = 1 - \frac{6 \times 5.5}{8(64-1)} = 0.9345$$

L'Interprétation du coefficient spearman c la même interprétation du coefficient de corrélation pearson

Exemple : les données suivante définie une relation linaire entre un traitement x et une maladie y

moyen	faible	bien	moyen	moyen	faible	moyen	bien	bien	bien	**module (x)**
faible	moyen	bien	bien	faible	moyen	bien	bien	bien	faible	**module (y)**

1-Calculez le coefficient de Spearman.

di^2	di	Ry	Rx	y	x	
42.25	−6.5	9	2.5	b	b	
0.25	−0.5	3	2.5	b	b	
0.25	3.5	3	2.5	b	b	
12.25	3	3	6.5	b	b	
9	−2.5	−6.5	9.5	b	m	
6.25	3.5	9	6.5	M	m	
12.25	−0.5	3	6.5	m	m	
0.25	3	3	2.5	f	m	
9	−2.5	6.5	9.5	f	f	
6.25		9	6.5	f	f	
98						

$$\gamma_s = 1 - \frac{6\sum_{i=1}^{n} d_1^2}{n(n^2-1)}$$

$$= -1\frac{6 \times 98}{10 \times 99}$$

$$= 1 - 0.594$$

$$= 0.406$$