Université A- mira de Bejaia
Faculté des sciences économiques, commerciales et des sciences de gestion
Département SEGC(LMD)
Module stat I
Enseignante : Dr. BERRAH

Chapitre 9 : Régression et corrélation

Introduction :

La régression fournit une expression de liaison entre deux variables sous la forme d'une fonction mathématique.
La corrélation renseigne sur l'intensité de la relation existante entre les deux variables.

1-Représentation graphique du nuage de points : L'étude simultanée sur deux variables quantitatives X et Y sur une population de n individus a donné les différents points de mesures : (x_1, y_1), (x_2, y_2), (x_3, y_3),....., (x_{n-1}, y_{n-1}), (x_n, y_n)

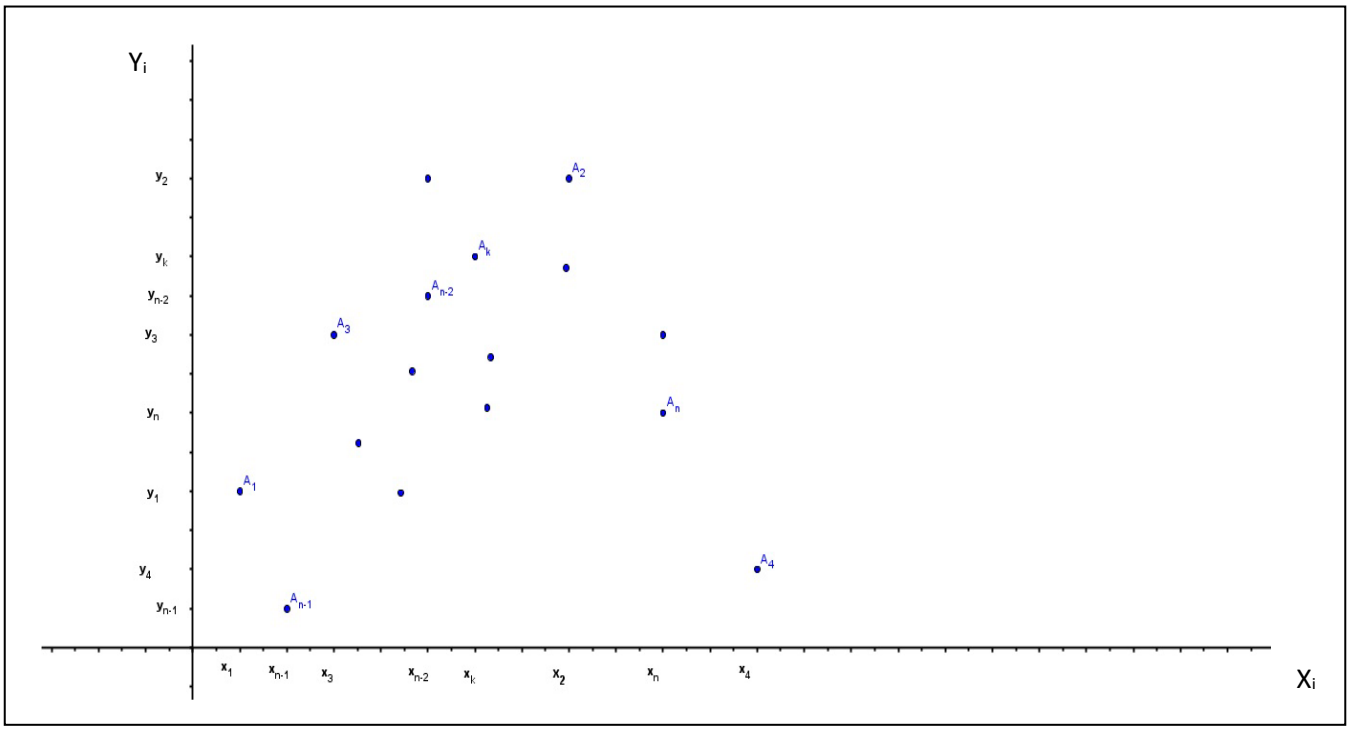
Ces données sont représentées par paires. le premier élément de la paire correspond à la valeur prise par la variable X et le second par Y . x_i et y_i , sachant k= i, · · · , n sont des valeurs observées.

On représente une distribution statistique à deux caractères quantitatifs par l'ensemble des points A_i, de coordonnées : (x_i, y_i). Chaque individu correspond à un point du plan.

On appelle nuage de points l'ensemble des points A_i, de coordonnées (x_i, y_i), i = 1, · · · , n. La représentation graphique du nuage de points est essentielle pour déterminer s'il existe ou non une relation entre les variables X et Y .

On représente sur l'axe des abscisses les mesures x_i, i = 1· · ·, n et sur l'axe des ordonnées les mesures y_i, i = 1· · ·, n , et le points A_i correspond à la paire (xi, y_i).

Figure 1 – Nuage de points



2- Mesure de l'intensité de la relation linéaire entre deux variables

L'objectif est de mettre en évidence l'existence d'une relation linéaire entre deux variables quantitatives (continues ou discrètes). On cherche un modèle de la forme : $Y = a + \beta X + \varepsilon$

où :

- Y est la variable à expliquer (dépendante).
- X est la variable explicative (indépendante).
- ε est l'erreur introduite par le modèle.
- a et b les paramètres du modèle avec b la pente de la droite d'ajustement et a l'ordonné à l'origine.

2-1-Covariance

La covariance (noté Cov(x, y)) des variables X et Y s'écrit :

$$Cov\,(X,\,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y}$$

La covariance dépend des unités de mesures dans lesquelles sont exprimées les variables.

2-2-Coefficient de corrélation linéaire

Le coefficient de corrélation est un nombre sans dimension destiné à mesurer l'intensité de la liaison entre les variations de la variable X et celles de Y , il s'écrit par la formule suivante:

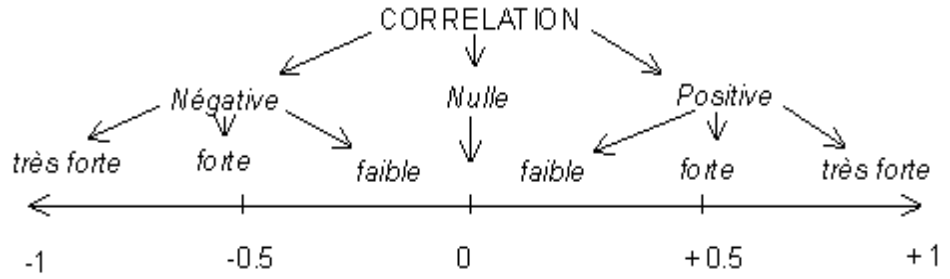
$$r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Sachant : σ_X : *écart type de X* , σ_y : *écart type de Y*

Le coefficient de corrélation est toujours compris entre -1 et 1 ($-1 \le r \le 1$)

- Si r = 1 ou -1 : la corrélation linéaire entre X et Y est parfaite.
- Si r = 0 : Alors il n'y a pas de relation linéaire entre X et Y (X et Y sont indépendants)
- Si r > 0 : la corrélation linéaire entre X et Y est positive.
- Si r < 0 : la corrélation linéaire entre X et Y est négative.
- Si r ≅ 1 : la corrélation linéaire entre X et Y est fortement positive.
- Si r ≅ -1 : la corrélation linéaire entre X et Y est fortement négative.

On peut résumer les propriétés par le schéma suivant :



2-3-Droite de régression

Si r_{xy} est proche de 1 et si l'examen du nuage de points indique qu'on peut supposer une relation de type linéaire entre X et Y , alors on cherche à déterminer les réels a et b de la

droite : $Y = a + \beta X$, telle que la distance entre cette droite et chaque point du nuage soit la plus petite possible.

La méthode des moindres carrés ordinaires (MCO) propose cette notion de proximité entre la droite et le nuage des points. Sous cette méthode nous avons obtenu les formules suivantes :

$$\hat{a} = \overline{Y} - \hat{\beta}\overline{X} \quad \text{Et} \quad \hat{\beta} = \frac{Cov(X, Y)}{\text{var}(X)}$$

Et donc, le modèle estimé (modèle ajusté) s'écrit : $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$

Exercices corrigés

Exercice 1 : Soit les 5 observations pour deux variables suivantes :

X	10	15	20	30	35
Y	400	700	800	900	950

1- Calculer la covariance.
2- Calculer le coefficient de corrélation et interpréter le résultat.
3- Ajuster la droite de régression.

Corrigé de l'exercice 1

1- La covariance : $Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y}$

Tout d'abord on calcule : $\overline{X}$, $\overline{Y}$ *et* $\sum_{i=1}^{n} x_i y_i$

$$\overline{X} = \frac{\sum xi}{N} = \frac{110}{5} = 22 \quad , \qquad \overline{Y} = \frac{\sum yi}{N} = \frac{3750}{5} = 750, \qquad \sum_{i=1}^{20} x_i y_i = 90750$$

Et donc : $Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y} = \frac{90750}{5} - 22 \times 750 = 1650$

2- Le coefficient de corrélation : $r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$

On doit calculer d'abord l'écart-type de X et l'écart-type de Y

$$\sigma(X) = \sqrt{v(x)} = \sqrt{\frac{\sum x_i^2}{N} - \overline{X}^2} = \sqrt{\frac{2850}{5} - (22)^2} = \sqrt{86}$$

$$\sigma(Y) = \sqrt{v(Y)} = \sqrt{\frac{\sum Y_i^2}{N} - \overline{Y}^2} = \sqrt{\frac{3002500}{5} - (750)^2} = \sqrt{38000}$$

Et donc $r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{1650}{\sqrt{86} \times \sqrt{38000}} = 0.91 \approx 1$

D'après ce résultat la corrélation linéaire entre X et Y est fortement positive.

3- La droite de régression : $\hat{Y}_i = \hat{a} + \hat{\beta}\ X_i$

On les relations suivantes : $\hat{a} = \overline{Y} - \hat{\beta}\overline{X}$ Et $\hat{\beta} = \frac{Cov(X, Y)}{\text{var}(X)}$

$$\hat{\beta} = \frac{Cov(X, Y)}{\text{var}(X)} = \frac{1650}{86} = 19{,}1866 \qquad \Rightarrow \hat{a} = \overline{Y} - \hat{\beta}\overline{X} = 750 - 19{,}186 \times 22 = 327{,}9$$

L'équation de la droite de régression s'écrit : $\hat{Y}_i = 327{,}9 + 19{,}186\hat{X}_i$

Exercice 2 : Soit le modèle linéaire simple suivant : $y_i = \alpha + \beta x_i + \varepsilon_i$ dont les valeurs de x_i et y_i sont présentées dans le tableau ci- dessus :

Y_i	X_i
128.2	32.1
106.1	31.1
107.8	29.8
124.7	43.8
78.6	20.4
157.1	45.1
215.6	66
137.1	56.6
171.1	50.3
137	36.8
141.5	35.2
105.4	42.6
139.5	55.7
104.3	32.4
88.6	24.6
109.5	33.2
79.9	26.4
124.1	35.9
122.4	40.5
131	39

1- Calculer le coefficient de corrélation et interpréter le résultat.
2- Estimer les paramètres du modèle.

Corrigé de l'exercice 2

1-Le coefficient de corrélation : $r_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$

$$Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y}$$

3$\overline{X} = \frac{\sum xi}{N} = \frac{777.5}{20} = 38.875, \qquad \overline{Y} = \frac{\sum yi}{N} = \frac{2509.5}{20} = 125.475, \qquad \sum_{i=1}^{20} x_i y_i = 103481.3$

Et donc : $Cov(X, Y) = \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y} = \frac{103481.3}{20} - 38.875 \times 125.475 = 296.2239$

$$\sigma(X)=\sqrt{v(x)}=\sqrt{\frac{\sum x_i^2}{N}-\overline{X}^2}=\sqrt{\frac{32758.63}{20}-(38.875)^2}=\sqrt{126.6659}=11.25459$$

$$\sigma(Y)=\sqrt{v(Y)}=\sqrt{\frac{\sum Y_i^2}{N}-\overline{Y}^2}=\sqrt{\frac{334283.7}{20}-(125.475)^2}=\sqrt{970.2079}=31.15$$

Et donc $r_{xy}=\frac{Cov(X,Y)}{\sigma_x\sigma_y}=\frac{296.2239}{11.25459\times 31.15}=0.845\approx 1$

D'après ce résultat la corrélation linéaire entre X et Y est fortement positive.

4-

2-Estimer les paramètres du modèle : $\hat{Y}_i=\hat{a}+\hat{\beta}\ X_i$

On a les relations suivantes : $\hat{a}=\overline{Y}-\hat{\beta}\overline{X}$ Et $\hat{\beta}=\frac{Cov(X,Y)}{\text{var}(X)}$

$$\hat{\beta}=\frac{Cov(X,Y)}{\text{var}(X)}=\frac{296.2239}{126.6659}=2.339 \qquad \Rightarrow \hat{a}=\overline{Y}-\hat{\beta}\overline{X}=125.475-\times 38.875=34.546$$

L'équation de la droite de régression s'écrit : $\hat{Y}_i=34.546+2.339\hat{X}_i$