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

Chapitre 8 : Corrélation et régression linéaire

Introduction

Dans plusieurs études statistiques, que ce soit en sciences de la santé, en économie, en administration ou en d'autres domaines, la recherche de la relation entre deux ou plusieurs variables est fondamentale ; par exemple la relation entre les profits et les investissements d'une entreprise, entre les revenus et les dépenses en nourriture, entre la taille et le poids d'un groupe d'individus.

Pour saisir l'objet de l'étude de la régression et de la corrélation, nous sommes amenés à poser un certain nombre de questions jugées nécessaires. Ces questions sont les suivantes : Existe-t-il une relation ou une dépendance entre les variables statistiques ? Cette relation, si elle existe est-elle linéaire ou non ? Si une relation existe, comment peut-on la traduire par une équation mathématique ?

Ensuite, cette relation, est-elle importante ou faible ? Si l'équation mathématique de la relation est établie, comment prévoir les valeurs d'une variable à partir de la connaissance de valeurs de l'autre ou des autres variables ?

En fait, pour répondre à toutes ces questions, nous ferons appel à une théorie statistique appelée ***l'analyse de la régression*.**

- L'analyse de la régression est dite simple si elle permet de prédire les valeurs d'une variable dite dépendante à partir des valeurs prises par une autre variable dite indépendante.
- L'analyse de la régression est dite multiple si elle permet de déterminer les valeurs d'une variable dite dépendante à partir des valeurs prises par plusieurs autres variables dites indépendantes.

Dans ce chapitre, nous nous contenterons de l'étude de la régression simple. Pour en faciliter la compréhension, nous introduirons les concepts à partir d'une série d'exemples.

8.1. Régression linéaire simple

8.1.1 Définition

Nous appelons **Régression linéaire** l'ajustement d'une droite au nuage statistique d'une série de couple de données.

Cette régression linéaire va nous permettre de résumer, d'interpréter et de prévoir les fluctuations d'un caractère dit dépendant en fonction d'un autre caractère dit indépendant.

8.1.2. Détermination de la droite de régression

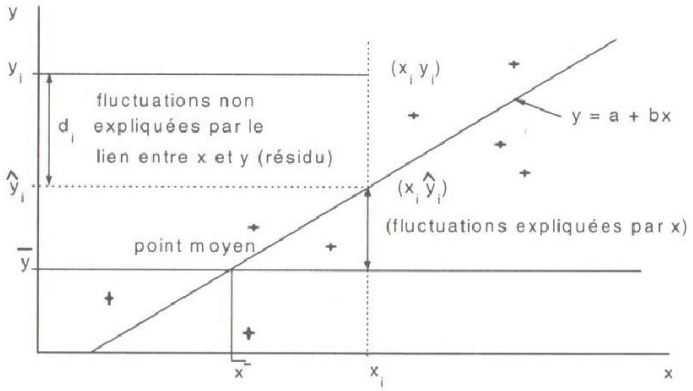
La méthode pour y arriver consiste à considérer n couples de données émanant de l'étude de deux variables statistiques X et Y.

X (variable indépendante)	X_1	X_2	…….	X_n
Y (variable dépendante)	Y_1	Y_2	…….	Y_n

Le problème consiste maintenant à trouver une **droite** qui traduit avec le plus de fidélité, la relation ou le lien entre X et Y. Cette droite est notée :

$$\hat{Y} = a + bx \text{ (Équation linéaire simple)}$$

La technique utilisée pour déterminer la droite la plus représentative [celle pouvant s'ajuster à un ensemble (X_i, Y_i)] est connue sous le nom de ***méthodes des moindres carrés.*** Le principe de cette méthode est de rendre minimale la somme du carré de la distance verticale (d_i) qui sépare chaque point (X_i, Y_i) de la droite elle-même.



Ce qui conduit à minimiser **G,** en calculant les dérivées partielles de l'expression par rapport à **(a)** et par rapport à **(b)**. **G** étant la somme du carré de la distance verticale (d_i). Ce qui entraîne :

$$G = \sum_{i=1}^{n} d_i^2 = d_1^2 + d_2^2 + \ldots\ldots\ldots + d_n^2$$

Avec $d_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i)$ puisque $\hat{Y} = a + bX_i$

La droite cherchée est encore celle qui minimise $\boldsymbol{G = \sum_{i=1}^{n}(Y_i - a - bX_i)^2}$

Nous appliquons ensuite les principes d'optimisation du calcul différentiel.

On obtient :

$G' = -2\sum_{i=1}^{n}(Y_i - a - bX_i) = 0$ (Par rapport à **a**)

et $G' = -2\sum_{i=1}^{n} X_i(Y_i - a - bX_i) = 0$ (Par rapport à **b**)

D'où le système d'équations :

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i$$

En divisant par n, l'équation fait ressortir une propriété importante de la droite des moindres carrées, qui passe par le point moyen. Ceci permet d'écrire :$\bar{Y} = a + b\bar{X}$

Et $\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2$

L'utilisation de la méthode des moindres carrés et la simplification des calculs permettent d'obtenir pour les valeurs (**a)** et (**b)** de la droite $\hat{Y} = \hat{a} + \hat{b}X$, les équations suivantes :

$\hat{b} = \frac{\frac{\sum X_i Y_i}{N} - \bar{X}\bar{Y}}{\frac{\sum X_i^2}{N} - \bar{X}^2} = \frac{Cov(X,Y)}{V(X)}$ ($\hat{b}$: pente).

Dans ce cas, nous avons fait intervenir l'expression de la **covariance** entre X et Y notée cov (X,Y) et $\hat{a} = \bar{Y} - \hat{b}\bar{X}$ ($\hat{a}$:constante).

Exemple :

Dans une étude de marché, la relation de régression intéressante dans une analyse économique est celle qui explique le prix Y_i (variable expliquée et dépendante) en fonction de l'offre X_i (variable explicative et autonome), ce qui permet d'écrire l'équation de la droite $\hat{Y} = \hat{a} + \hat{b}X$représente la droite retenue.

8.2. Corrélation linéaire

La notion de nuage statistique ou de diagramme de dispersion n'est rien d'autre que la représentation graphique dans le plan cartésien de l'ensemble de points ou de couples de données (X_i, Y_i).

A partir de cette représentation graphique, il est facile d'avoir une idée sur le degré de liaison entre les deux variables.

8.2.1. Définition

Après avoir souligné l'intérêt de l'ajustement d'une droite de régression à une série de paires de (X_i, Y_i), nous abordons maintenant le problème de la corrélation ou, le problème du degré de dépendance entre deux variables statistiques, **qui se mesurent par le degré de dispersion des points autour de la droite de régression.**

8.2.2. Détermination du coefficient de corrélation

Pour le calcul du coefficient de corrélation noté **r**, nous utilisons la formule la plus simple et la plus pratique qui est la suivante : $r = \frac{Cov(X,Y)}{\sigma(X).\sigma(Y)}$

Avec :$\sigma(X)$: représentant l'écart-type de la variable X

:$\sigma(Y)$: représentant l'écart-type de la variable Y

La définition du coefficient de corrélation nous permet d'établir que :

1. $-1 \leq r \leq +1$
2. $r < 0$ indique une **corrélation négative.** Ce qui signifie que la relation est négative entre X et Y lorsque X augmente la variable Y diminue.
3. $r > 0$ indique une **corrélation positive.** Ce qui signifie que la relation est positive lorsque la variable X augmente la variable Y augmente.

   Dans ces deux derniers cas [(2) et (3)], il s'agit d'une **relation relative.** En effet, lorsque la relation linéaire est positive et faible entre X et Y par exemple, cela signifie que les points du nuage sont plutôt dispersés par rapport à la droite de régression.
4. r proche de **-1** ou de **+1** indique une **dépendance très forte entre les deux variables.** Une corrélation positive et très forte ($r \approx +1$) entre X et Y par exemple, signifie que les fluctuations de la variable Y s'expliquent grandement par les variations de la variable X.
5. $r = \pm 1$ indique une **corrélation maximum**. Cela signifie que la droite de régression s'ajuste parfaitement aux données recueillies. Nous parlerons alors **de liaison fonctionnelle.**
6. $r = 0$ indique une **absence de corrélation linéaire** entre X et Y. Dans ce cas, les deux droites de régressions $D_{y/x} et D_{x/y}$ sont perpendiculaires et les pentes $b et b'$ sont nulles. Cela signifie que lorsque X augmente la variable Y ne va ni en augmentant, ni en diminuant. Les fluctuations de Y ne s'expliquent donc pas par les variations de la variable X.

Remarque 1 :

Par ailleurs, il peut exister une relation positive entre X et Y, mais cette relation **n'est pas linéaire.** Il s'agit alors, d'un nuage de points suggérant un **ajustement curviligne.**

Remarque 2 :

Il faut faire preuve d'une grande prudence dans l'interprétation et l'utilisation des résultats. Ainsi, un coefficient de corrélation de 0,99 indique une grande dépendance entre les deux variables (X Y). Cependant, cela ne signifie pas nécessairement qu'il y a une relation de cause à effet entre ces variables. En effet, ce coefficient **r** ne sert qu'à attirer l'attention sur la possibilité d'une relation entre les variables. C'est à l'économiste ou au statisticien de réfléchir ensuite sur la réalité de cette liaison.

La valeur $100r^2$ représente le **pourcentage de variation totale de Y s'expliquant par la liaison de Y par rapport X.** Il s'agit du **coefficient de détermination (R^2).**

Exemple1 : Pour un coefficient de corrélation **r** = 0,9, le coefficient de détermination ($\boldsymbol{R^2}$) donne$100r^2 = 81\%$, Ainsi, 81% de la variation de Y se trouve expliquée par le lien entre X et Y.

Exercice :

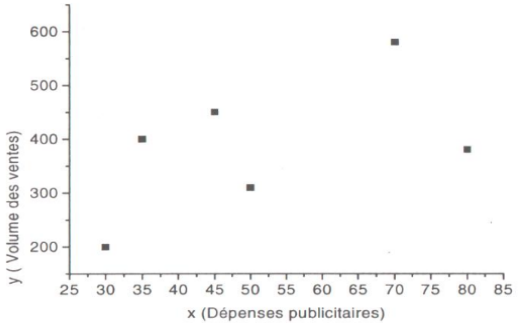
Une entreprise veut mener une étude sur la liaison entre les dépenses (hebdomadaires) mensuelles en publicité et le volume des ventes qu'elle réalise. Nous avons obtenu au cours des six derniers mois les données suivantes :

X_i **: Dépenses publicitaires (en MDA)**	70	80	30	50	35	45
Y_i **: Volume des ventes (en MDA)**	580	380	200	310	400	450

1- Tracer le nuage de points.
2- Ajuster la droite de régression.
3- Calculer le coefficient de corrélation.
4- Interpréter le coefficient de corrélation.

Corrigé :

1. **Le nuage de points**



A partir de cette représentation, nous soupçonnons l'existence d'une relation entre X et Y, mais la corrélation n'est pas très forte.

Nous le vérifierons avec la question (3).

2. Ajustement de la droite de régression

La réponse à cette question permet d'exprimer la relation à l'aide d'une équation mathématique : $\hat{Y} = \hat{a} + \hat{b}X$

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
70	580	40600	4900	336400
80	380	30400	6400	144400
30	200	6000	900	40000
50	310	15500	2500	96100
35	400	14000	1225	160000
45	450	20250	2025	202500
310	2320	126750	17950	979400

$\bar{Y} = \frac{\sum Y_i}{N} = \frac{2320}{6} = 386{,}66 \approx 387$;

$\bar{X} = \frac{\sum X_i}{N} = \frac{310}{6} = 51{,}6 \approx 52.$

$Var(X) = \frac{\sum X_i^2}{N} - \bar{X}^2 = \frac{17950}{6} - (52)^2 = 288$;

$\sigma(X) = \sqrt{288} \approx 17$

$Var(Y) = S^2(Y) = \frac{\sum Y_i^2}{N} - \bar{Y}^2 = \frac{979400}{6} - (387)^2 = 13464$;

$\sigma(Y) = \sqrt{13464} \approx 116.$

$$Cov(X,Y) = \frac{\sum XY}{N} - \bar{X}\bar{Y} = \frac{126750}{6} - (52) \times (387) = 1001.$$

Nous obtenons alors pour les valeurs $\hat{a}$ et $\hat{b}$ de la droite$\hat{Y} = \hat{a} + \hat{b}X$, les valeurs suivantes :

$$\hat{b} = \frac{Cov(X,Y)}{V(X)} = \frac{1001}{288} = 3{,}48$$

$\hat{a} = \bar{Y} - \hat{b}\bar{X} = 387 - (3{,}48) \times (52)$ =387-180,96≈387-181=206

D'où : $\hat{Y}$**=3,48X+206**

3. Calcul du coefficient de corrélation.

$$\boldsymbol{r = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} = \frac{1001}{17 \times 116} \approx 0{,}51.}$$

4. Interprétation du coefficient de corrélation.

Pour cette série, nous pouvons conclure qu'il y a une corrélation positive mais pas très forte entre le volume des ventes et les dépenses en publicité de cette entreprise. En effet, seulement : 26% ($100r^2 = 0{,}51^2 \times 100$) de la fluctuation totale de Y se trouve expliquée par le lien entre X et Y.