République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER BISKRA

Faculté des Sciences Exactes et Sciences de la Nature et de la Vie
Département de Mathématiques

Première Année Master

Notes de Cours

Analyse de Données

Chapitre 1 : Régression linéaire simple (Séance 1)

Auteur des notes :

Dr. Sana BENAMEUR

Année universitaire : 2021-2022

Chapitre 1

Régression linéaire simple

La régression est une méthode statistique qui permet d'établir la relation entre une variable quantitative a expliquée (endogène ou dépendante) à une ou plusieurs autres variables quantitatives explicatives (exogènes ou indépendantes), sous la forme d'un modèle.

Elle est dite linéaire si elle impose une forme fonctionnelle linéaire dans les paramètres du modèle. Si on s'intéresse à la relation entre deux variables on parlera de la régression linéaire simple et si la relation porte entre une variable et plusieurs autres variables, on parlera de la régression linéaire multiple.

Cette méthode sert à prévoir les valeurs futures de l'une des variables en fonction de l'autre.

Dans la suite : $X, Y, Z, \ldots$ sont des variables réelles et leurs réalisations (valeurs observées) seront notées : $x, y, z, \ldots$

Exemple 1.1.

Considérons l'exemple suivant, représentant la relation entre le revenu, et la consommation mensuelle moyenne, de cinq familles (en $1000DA$*).*

R	18	25	27	35	45
C	17	19	30	32	35

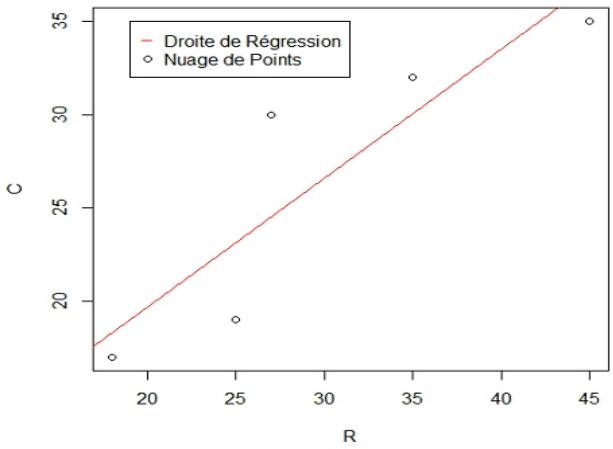

FIG. 1.1. Ajustement du nuage de point par une droite de régression

D'après cette figure, la relation entre R et C est de la forme linéaire :

$$c_i = b_0 + b_1 r_i + \varepsilon_i, \ \ i = 1, \ldots, n.$$

C : La consommation, R : le revenu, b_0 : consommation autonome, b_1 :propension marginale à consommer, ε : erreur ou consommation non liée au revenu, $n = 5$ la taille de l'échantillon.

1.1 Modèle

Un modèle linéaire simple est de la forme :

$$y_i = b_0 + b_1 x_i + \varepsilon_i, \ \ i = 1, \cdots, n :$$

où :

y_i : la i^{ème} observation de la variable aléatoire à expliquer Y,

x_i : la i^{ème} observation de la variable explicative X,

b_0 et b_1 : sont des constantes inconnues appelées paramètres du modèle,

ε_i : l'erreur (ou bruit) aléatoire du modèle.

n : la taille de l'échantillon.

Notons que le modèle de régression linéaire peut encore s'écrire sous forme matricielle, comme suit :

$$\begin{array}{rcl} Y & = & b_0 + b_1 X + \varepsilon. \\ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} & = & \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \end{array}$$

Les hypothèses relatives à ce modèle sont les suivantes :

(i) Les erreurs ε_i sont centrées, ont la même variance, et non corrélées entre elles :

$$\begin{aligned} &\mathbb{E}(\varepsilon_i) = 0,\ \mathbb{E}(\varepsilon_i^2) = \sigma_\varepsilon^2 < \infty,\ i = 1, \cdots, n \\ &Cov(\varepsilon_i, \varepsilon_j) = 0,\ \forall (i,j) \text{ tel que } i \neq j, \end{aligned}$$

(ii) L'erreur est indépendante de X :

$$cov(\varepsilon, X) = 0.$$

1.2 Estimation

On cherche les valeurs $\hat{b}_0$ et $\hat{b}_1$ des estimateurs de b_0 et b_1, définissant la droite de régression

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i,$$

tel que $\sum_{i=1}^{n} \varepsilon_i^2$ soit minimale. Cette méthode s'appelle la méthode des moindres carrés ordinaire (MCO).

Les estimateurs peuvent s'écrire sous la forme suivante :

$$\left(\hat{b}_0, \hat{b}_1\right) = \underset{(b_0,b_1)\in\mathbb{R}\times\mathbb{R}}{\arg\min}\ \Psi(b_0, b_1),$$

où

$$\Psi(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

1.2.1 Calcul des Estimateurs

Pour déterminé la valeur qui minimise Ψ, annulant les dérivées partielle par rapport à b_0 et b_1, nous obtenons le système d'équations suivant :

$$\left(\frac{\partial \Psi\left(b_0, b_1\right)}{\partial b_0}=0, \frac{\partial \Psi\left(b_0, b_1\right)}{\partial b_1}=0\right)_{\substack{b_0=\hat{b}_0 \\ b_1=\hat{b}_1}}$$

Posons :
$\overline{x}=\frac{1}{n} \sum_{i=1}^{n} x_i$ et $\overline{y}=\frac{1}{n} \sum_{i=1}^{n} y_i$ les moyennes empiriques des x_i et des y_i (respect).

$S_x=\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\sum_{i=1}^{n} x_i^2-n \overline{x}^2$.

$S_y=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=\sum_{i=1}^{n} y_i^2-n \overline{y}^2$.

$S_{xy}=\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)=\sum_{i=1}^{n} x_i y_{i-} n \overline{x y}$.

Les estimateurs de moindres carrés ordinaire sont alors :

$$\begin{cases} \hat{b}_0=\overline{y}-\hat{b}_1 \overline{x}, \\ \hat{b}_1=\dfrac{S_{xy}}{S_x}. \end{cases}$$