République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
ECOLE SUPÉRIEURE EN INFORMATIQUE
8 Mai 1945 - Sidi-Bel-Abbès

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
المدرسة العليا للإعلام الآلي
8 ماي 1945 - سيدي بلعباس

Régression et Modèle linéaire Simple

Présenté par Pr.Nabil KESKES.

Année 2021-2022

.

PLAN

- Introduction
- Nature de données étudiées
- Démarches de la méthode
- Algorithme General
- Conclusion

1. Introduction

1.1 Definition

La régression linéaire est une méthode statistique relève a la fois de l'analyse de données descriptive et de la statistique inferentielle pour évaluer la part du hasard dans les résultats. Elle reste une des méthodes les plus utilisées en pratique. Elle permet d'analyser en détail les liaisons entre une variable quantitative et plusieurs autre, en recherchant une liaison linéaire approximative entre elle,

1.2 Objectifs

Objectif : Exprimer le lien entre Y et X.

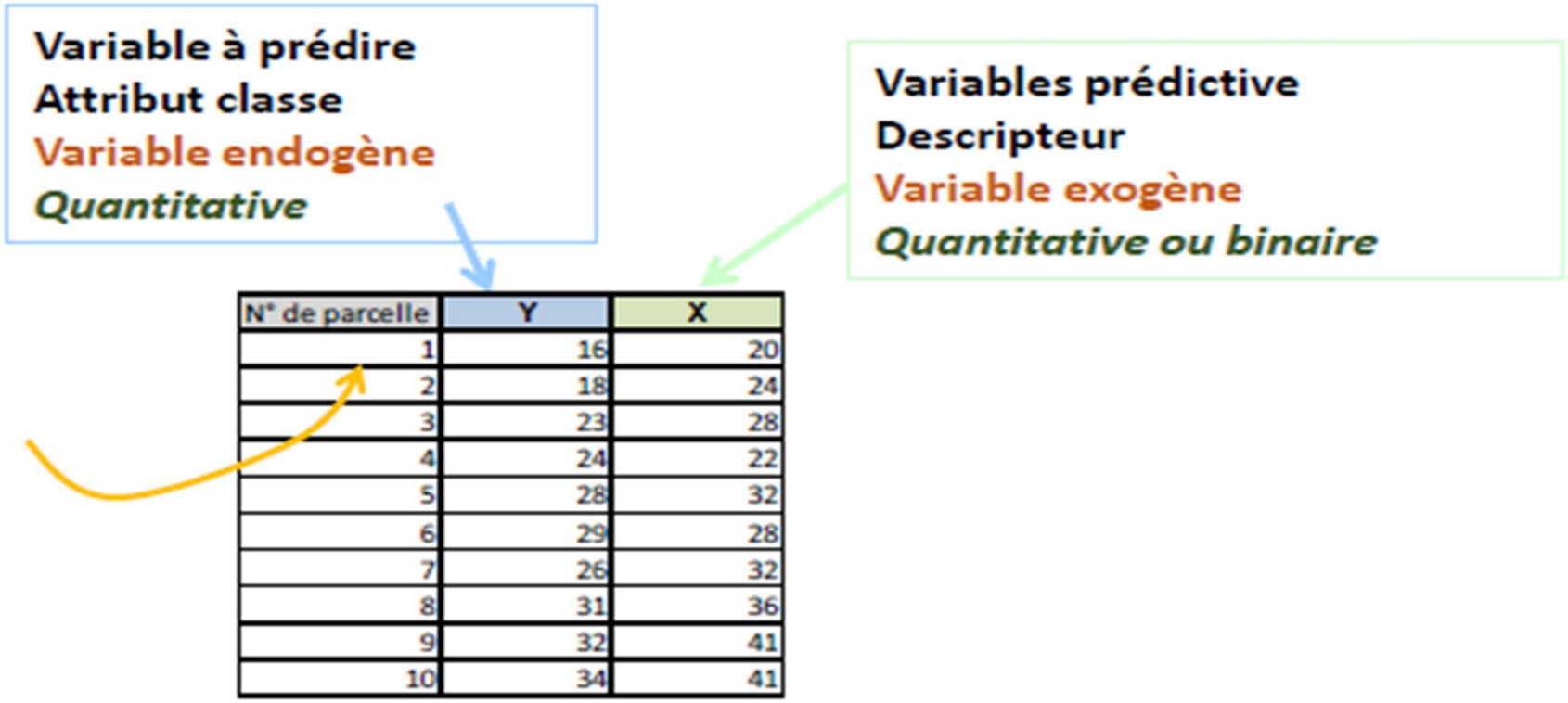
$$Y = f(X) + \epsilon$$

Il existe une infinité de liaisons fonctionnelles $\longrightarrow$ la plus simple est linéaire

2. Nature de données

2.1 Données initiales

Il est fréquent de présenter les données sous la forme d'un tableau comme ci-dessous :



N° de parcelle	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Figure 1.1 : ***Tableau de données***

Modèle de régression linéaire

$\forall i \in I$, y_i est la réalisation de la v.a.r. Y_i telle que

$$Y_i = \beta_1 x_i + \beta_0 + \epsilon_i$$

Avec

- ϵ_i : erreur du modèle (v.a.r.) (part de variabilité de Y qui n'est pas expliquée par le lien fonctionnel linéaire)
- β_0, β_1 : coefficients du modèle, constantes (valeurs fixes dans la population).

Hypothèses du modèle

- $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{V}[\epsilon_i] = \sigma^2$ (hypothèse d'*homoscédasticité*)
- L'erreur est indépendantes des $X_j \rightarrow \text{COV}(x_{ij}, \epsilon_i) = 0$
- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (normalité des résidus) $\rightarrow$ tests dans le modèle
- Les ϵ_i, $1 \leq i \leq n$, sont mutuellement indépendantes (absence d'autocorrélation des résidus) $\rightarrow \text{Cov}(\epsilon_i, \epsilon_j) = 0$ si $i \neq j$.

Droite de régression au sens des moindres carrés

Objectif : estimer β_0 et β_1 grâce à leur estimateurs B_0 et B_1 et leur réalisations b_0 et b_1 sur un échantillon d'observations i.i.d. de taille n.

Trouver b_0 et b_1 qui minimisent un critère d'ajustement.

$\Rightarrow$ Méthode des moindres carrés ordinaires

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2$$

$$\rightarrow \min S(\beta_0, \beta_1)$$

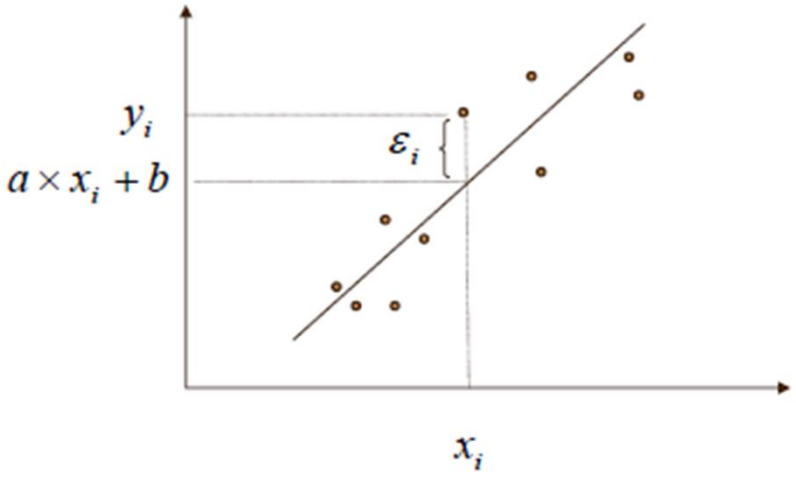
Estimation des paramètres

Dérivées partielles → Systèmes aux équations normales

$$\text{Solutions :} \quad b_1 = \frac{s_{xy}}{s_x^2} \text{ et } b_0 = \bar{y} - b_1\bar{x}$$

Estimation des paramètres

Critère numérique



Critère des moindres carrés : trouver les valeurs de a et b qui minimise la somme des carrés des écarts entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction.

$$S = \sum_{i=1}^{n} \varepsilon_i^{\,2}$$

$$S = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2$$

$$S = \sum_{i=1}^{n} [y_i - ax_i - b]^2$$

Remarque : Pourquoi pas la somme des erreurs ? Ou la somme des écarts absolus ?

SOLUTION

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \sum_i x_i y_i - a \sum_i x_i^{\,2} - b \sum_i x_i = 0 \\ \bar{y} - a\bar{x} - b = 0 \end{cases}$$

Equations normales

$$\begin{cases} \hat{a} = \dfrac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$$

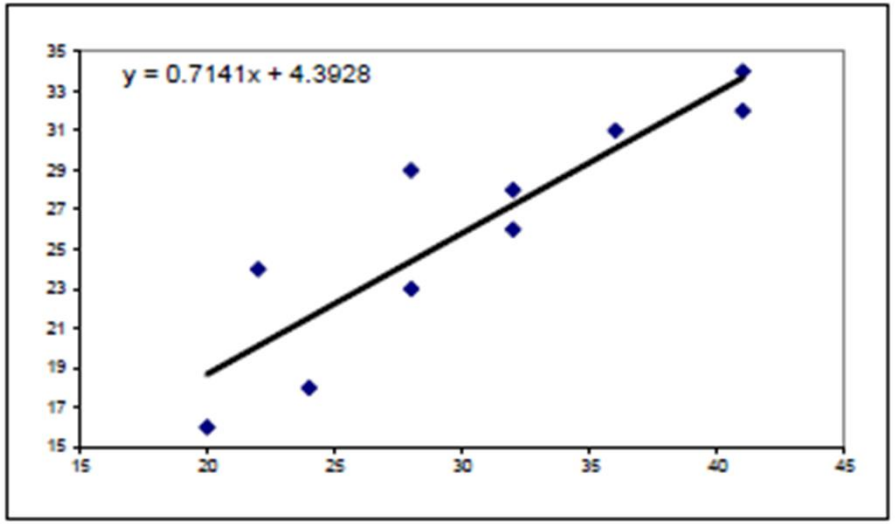
Estimateurs des moindres carrés

Voir détail des calculs...

Exemple des rendements agricoles

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB)^2
1	16	20	-10.1	-10.4	105.04	108.160
2	18	24	-8.1	-6.4	51.84	40.960
3	23	28	-3.1	-2.4	7.44	5.760
4	24	22	-2.1	-8.4	17.64	70.560
5	28	32	1.9	1.6	3.04	2.560
6	29	28	2.9	-2.4	-6.96	5.760
7	26	32	-0.1	1.6	-0.16	2.560
8	31	36	4.9	5.6	27.44	31.360
9	32	41	5.9	10.6	62.54	112.360
10	34	41	7.9	10.6	83.74	112.360
Moyenne	**26.1**	**30.4**		**Somme**	**351.6**	**492.4**

$$\begin{cases} \hat{a} = \dfrac{351.6}{492.4} = 0.714 \\ \hat{b} = 26.1 - 0.714 \times 30.4 = 4.39 \end{cases}$$



Propriétés

Droite de régression au sens des moindres carrés

La droite de régression au sens des moindres carrés a pour expression :

$$\widehat{y}_i = b_1 x_i + b_0$$

C'est une estimation du modèle de régression par la méthode des moindres carrés.

Les erreurs observées sur l'échantillon sont appelés résidus.

$$e_i = (y_i - \widehat{y}_i) = y_i - b_1 x_i - b_0$$

Equation d'analyse de la variance

$$y_i - \bar{y} = (\widehat{y_i} - \bar{y}) + (y_i - \widehat{y_i})$$

$$(y_i - \bar{y})^2 = (\widehat{y_i} - \bar{y})^2 + (y_i - \widehat{y_i})^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Somme des carrés totale SCT}} = \underbrace{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2}_{\text{Somme des carrés expliquée SCE}} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}_{\text{Somme des carrés résiduelle SCR}}$$

Le coefficient de détermination R^2 est défini par

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2} = \frac{\text{variabilité expliquée (SCE)}}{\text{variabilité totale (SCT)}} = 1 - \frac{SCR}{SCT}$$

Remarque. On a la formule "classique" de l'analyse de la variance nous donnant la décomposition suivante :

$$\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$$

$$\text{variabilité totale} = \text{variabilité résiduelle} + \text{variabilité expliquée}$$

Commentaire. Le coefficient R^2 donne la proportion de variabilité de y qui est expliquée par le modèle. Plus le R^2 est proche de 1, meilleure est l'adéquation du modèle aux données.

Evaluation globale de la régression

Tableau d'analyse de variance - Test de significativité globale

Le test F permet d'évaluer la significativité globale de la régression.

$$\begin{cases} \mathcal{H}_0 : \text{La variabilité expliquée est identique à la variabilité résiduelle} \\ \mathcal{H}_1 : \text{La variabilité expliquée est supérieure à la variabilité résiduelle} \end{cases}$$

Sous $\mathcal{H}_0$

$$F = \frac{CME}{CMR} \sim \mathcal{F}_{1,n-2 \text{ ddl}}$$

Interprétation :

$$\begin{cases} \mathcal{H}_0 : \text{"Le modèle est non explicatif"} \\ \mathcal{H}_1 : \text{"Le modèle est explicatif"} \end{cases}$$

Test de significativité globale du modèle

H0 : Le modèle n'amène rien dans l'explication de Y

H1 : Le modèle est pertinent (globalement significatif)

Tableau d'analyse de variance

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyen
Régression (expliqués)	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$\frac{SCE}{1}$
Résidus	$SCR = \sum_i (\hat{y}_i - y_i)^2 = \sum_i \hat{e}_i^2$	n - 2	$\frac{SCR}{n-2}$
Total	$SCT = \sum_i (y_i - \bar{y})^2$	n - 1	

Statistique de test

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} \equiv F(1, n-2)$$

Remarque : Ecriture de F à partir du R^2

$$F = \frac{R^2}{\frac{(1-R^2)}{(n-2)}}$$

Région critique au risque α

$$F > F_{1-\alpha}(1, n-2)$$

Rendements agricoles – Tests de significativité globale

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB)^2	(Y-YB)^2	Y^	Résidus	Résidus^2
1	16	20	-10.1	-10.4	105.04	108.160	102.010	18.674	-2.674	7.149
2	18	24	-8.1	-6.4	51.84	40.960	65.610	21.530	-3.530	12.461
3	23	28	-3.1	-2.4	7.44	5.760	9.610	24.386	-1.386	1.922
4	24	22	-2.1	-8.4	17.64	70.560	4.410	20.102	3.898	15.195
5	28	32	1.9	1.6	3.04	2.560	3.610	27.242	0.758	0.574
6	29	28	2.9	-2.4	-6.96	5.760	8.410	24.386	4.614	21.286
7	26	32	-0.1	1.6	-0.16	2.560	0.010	27.242	-1.242	1.544
8	31	36	4.9	5.6	27.44	31.360	24.010	30.099	0.901	0.812
9	32	41	5.9	10.6	62.54	112.360	34.810	33.669	-1.669	2.785
10	34	41	7.9	10.6	83.74	112.360	62.410	33.669	0.331	0.110
Moyenne	26.1	30.4		Somme	351.6	492.4	314.9		Somme	63.83874898
							SCT			SCR

ESTIMATION

a	0.714053615
b	4.392770106

Tableau d'analyse de variance

Source de variation	SC	DDL	CM
Expliqués (Régression)	251.061251	1	251.061251
Résidus	63.83874898	8	7.979843623
Total	314.9	9	

F calculé	31.46192618

rejet de H0

DDL1	1
DDL2	8
F théorique (à 5%)	5.317655063

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} = \frac{251.06}{7.9798} = 31.4619$$

$$F_{1-\alpha}(1,8) = F_{0.95}(1,8) = 5.37655$$

Puisque $F > F_{1-\alpha}$

Rejet de H0 c.-à-d. on conclut que le modèle est globalement significatif

Remarque :

$$\sqrt{F} = \sqrt{31.4619} = 5.609 = t_a$$

Evaluation des coefficients - β_1

Test de significativité de β_1

Idée : tester la nullité de β_1.

$$\begin{cases} \mathcal{H}_0 : \beta_1 = 0 & \text{"X n'a aucun pouvoir explicatif sur Y"} \\ \mathcal{H}_1 : \beta_1 \neq 0 & \text{"X a un pouvoir explicatif sur Y"} \end{cases}$$

Nous savons que $\frac{B_1 - \beta_1}{\widehat{\sigma}_{B_1}} \sim \mathcal{T}_{n-2}$, par conséquent sous $\mathcal{H}_0$

$$\frac{B_1}{\widehat{\sigma}_{B_1}} \sim \mathcal{T}_{n-2}$$

Intervalle de confiance de β_1

$$\text{IC}_{\beta_1}^{1-\alpha} = \left[b_1 \pm t_{(1-\alpha/2;n-2)} \frac{s_{n-2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right]$$

Rendements agricoles – Tests de significativité des coefficients

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB)^2	(Y-YB)^2	Y^	Résidus	Résidus^2
1	16	20	-10.1	-10.4	105.04	108.160	102.010	18.674	-2.674	7.149
2	18	24	-8.1	-6.4	51.84	40.960	65.610	21.530	-3.530	12.461
3	23	28	-3.1	-2.4	7.44	5.760	9.610	24.386	-1.386	1.922
4	24	22	-2.1	-8.4	17.64	70.560	4.410	20.102	3.898	15.195
5	28	32	1.9	1.6	3.04	2.560	3.610	27.242	0.758	0.574
6	29	28	2.9	-2.4	-6.96	5.760	8.410	24.386	4.614	21.286
7	26	32	-0.1	1.6	-0.16	2.560	0.010	27.242	-1.242	1.544
8	31	36	4.9	5.6	27.44	31.360	24.010	30.099	0.901	0.812
9	32	41	5.9	10.6	62.54	112.360	34.810	33.669	-1.669	2.785
10	34	41	7.9	10.6	83.74	112.360	62.410	33.669	0.331	0.110
Moyenne	26.1	30.4		Somme	351.6	492.4	314.9		Somme	63.83874898
							SCT			SCR

sigma²(epsilon) 7.979843623

ESTIMATION

a	0.714053615
b	4.392770106

sigma²(a^)	0.016206019	sigma(a^)	0.127302862
sigma²(b^)	15.77493863	sigma(b^)	3.971767696

ddl	8
t théorique (bilatéral à 5%)	2.306004133

t(a^)	5.609093169	rejet H0
t(b^)	1.10599875	acceptation H0

$$t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} = \frac{0.714}{0.127} = 5.609$$

$$t_{1-\alpha/2}(8) = t_{1-0.05/2}(8) = t_{0.975}(8) = 2.306$$

Puisque $|t_{\hat{a}}| > t_{1-\alpha/2}$

Rejet de H0 : a = 0

4. Conclusion

La régression linéaire Simple est une méthode facile a mettre en œuvre, les résultats qu'elle donne sont satisfaisants lorsque les données traites se prêtent bien a l'analyse, mais il est indispensable de vérifier le bien fonde des hypothèse effectuées, en particulier la linéarité de la liaison