Régression linéaire simple

Y : variable quantitative "à expliquer" endogène (dépendante)

X : variable quantitative "explicative" (exogène (effet fi

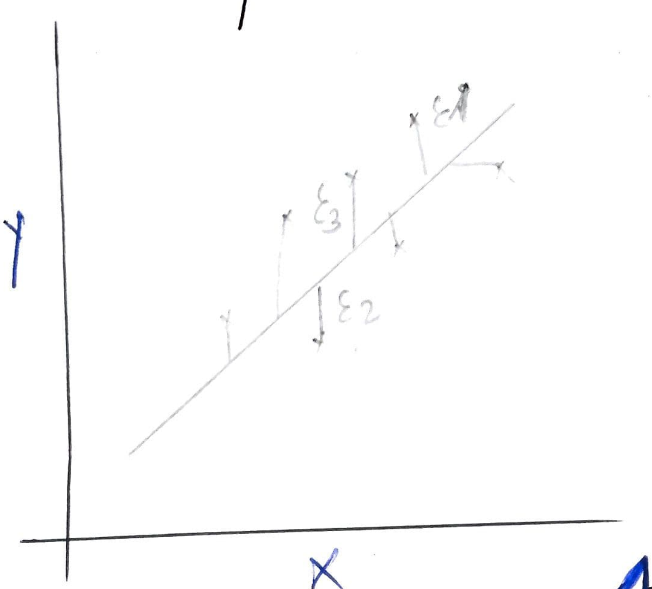
objectif

expliquer/exprimer une variable quantitative Y en fonction d'une variable quantitative X

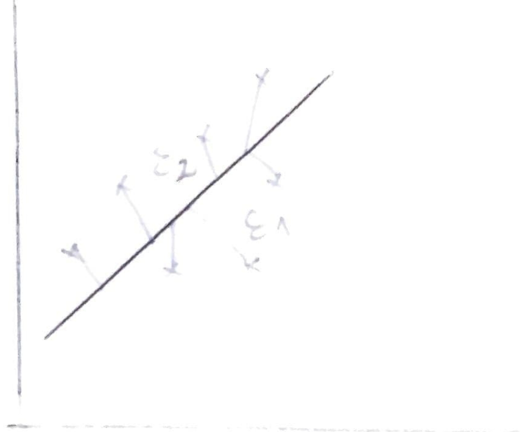
On suppose que en moyenne $\mathbb{E}[Y]$ est un fct° de X.

$\mathbb{E}[Y] = aX + b$

$$Y = aX + b + \varepsilon$$ (Y comme une fonction affine de X).



! "la droite affine ne passe pas nécessairement de tous les points"



Y_i

X_i

$$Y_i = b + aX_i + \varepsilon_i$$

modele statistique

ε_i : erreur qui represente les (Inférence) autres facteurs qui peuvent influencer la variable Y autre que X.

$\boxed{\text{minimiser les } \varepsilon_i}$

Hypothèses du modèle :

1) La distribution de ε est indépendante de X.

[Pas de dépendance entre ε et X
$cov(\varepsilon, X) = 0$ H°]

2) Homoscédasticité : Erreur centrée et de variance constante.

H° . $E[\varepsilon_i] = 0$, $V[\varepsilon_i] = \sigma^2$

s'annule au moyenne. même variance.

3) Pas de rupture du modele :
a et b sont deux constantes

4) ε_i indépendante.
et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ $\forall i$.

⟹ Estimation des paramètres du modèle : a ? b ?

On cherche a et b. ?

Estimation par maximum de vraisemblance.

Estimation par les moindres carrés ordinaires

Plus facile ⟹ Moindre carrés

minimiser la somme des carrés.

$$\min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} (y_i - a x_i - b)^2.$$

on cherche a et b tels que $\sum_{i=1}^{n} \varepsilon_i^2$ soit minimale.

On cherche a et b tq

+ la droite la "plus proche" du nuage des points

+ la somme des carrés des erreurs ε_i soit minimale.

$$\min_{(a,b)} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{(a,b)} \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

On pose :
$$F(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$$\begin{cases} \dfrac{\partial F}{\partial a} = \displaystyle\sum_{i=1}^{n} -2(y_i - ax_i - b)\, x_i = -2 \sum_{i=1}^{n} x_i(y_i - ax_i - b) \\ \dfrac{\partial F}{\partial b} = \displaystyle\sum_{i=1}^{n} -2(y_i - ax_i - b) = -2 \sum_{i=1}^{n} (y_i - ax_i - b) \end{cases}$$

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0 \\ \dfrac{\partial F}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} -2 \displaystyle\sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0 \\ -2 \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \displaystyle\sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0 \\ \displaystyle\sum_{i=1}^{n} (y_i - ax_i - b) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a \sum_{i=1}^{n} x_i - nb = 0 \end{cases}$$

$$\begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \frac{1}{n} a \sum_{i=1}^{n} x_i^2 - b \overbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)}^{\bar{x}} = 0 \\ \underbrace{\frac{1}{n} \sum_{i=1}^{n} y_i}_{\bar{y}} - a \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)}_{\bar{x}} - b = 0 \end{cases}$$

$$\begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i y_i - a \frac{1}{n} \sum_{i=1}^{n} x_i^2 - b \bar{x} = 0 & (1) \\ \bar{y} - a \bar{x} - b = 0 & (2) \end{cases}$$

$$(2) \Rightarrow \boxed{b = \bar{y} - a\bar{x}}$$

$$(1) \quad \frac{1}{n} \sum_{i=1}^{n} x_i y_i - a \frac{1}{n} \sum_{i=1}^{n} x_i^2 - (\bar{y} - a\bar{x}) \bar{x} = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i - a \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{y}\bar{x} + a\bar{x}^2 = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} = a \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right)$$

$$a = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{Cov(X,Y)}{\sigma_X^2}$$

$$\begin{cases} Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ \quad = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \\ \mathbb{V}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \end{cases}$$

a : pente de la droite

$$\boxed{a = \frac{Cov(X,Y)}{\sigma_X^2} \quad ; \quad b = \bar{y} - a\bar{x}}$$

a et b sont les estimations pour les réalisations.

$$\hat{y}_i = a x_i + b.$$

<u>on définit</u> :

<u>Résidus</u> (résidus estimés)

$$e_i = y_i - \hat{y}_i$$

" La méthode des moindres carrés conduit à des résidus avec une sommes des carrés minimale "

Qualité d'ajustement :

y_i : observation réelle.

$\hat{y}_i$: estimation par la méthode des moindres carrés.

$$e_i = y_i - \hat{y}_i$$

Une bonne qualité d'ajustement :

$$\sum_{i=1}^{n} e_i^2 = 0 \Rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0$$

Somme des carrés résiduelle

$$SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Prédiction parfaite $\Rightarrow$ $SCR = 0$

Mais à partir de quelle valeur des SCR peut-on dire que la régression est mauvaise ?

On définit alors

la somme des carrés totale

$$\underline{SCT} = SCR + SCE$$

somme des carrés résiduelle

somme des carrés expliquée.

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

la différence entre les y_i et la valeur de référence $\bar{y}$ qui représente le centre de gravité du nuage de points

variabilité totale de y.

$$\underline{SCE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

la différence entre l'estimation et la valeur de référene

variabilité expliquée par le modèle, la variation de y expliquée par x.

$$\underbrace{SCR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\text{variabilité non expliquée par le modèle}}$$

différence entre les valeurs observé et les estimations.

Imaginant

- Situation extrême 1 : Prédiction parfaite.

$SCR = 0 \Rightarrow SCT = SCE$.

"Les variations de y sont complètement expliquées par celles de X."

$\Longrightarrow$ modèle parfait 

- Situation extrême 2 :

$SCE = 0 \rightarrow SCT = SCR$.

$\rightarrow$ X n'apporte aucune information sur Y.

"Coefficient de détermination"

$$R^2 = \frac{SCE}{SCT} = \frac{SCT - SCR}{SCT}$$
$$= 1 - \frac{SCR}{SCT}.$$

"Variabilité totale expliquée par le modèle"

Plus que $R^2 \longrightarrow 1$, plus que le modèle tend vers la perfection.

- $R^2 \longrightarrow 1 \Rightarrow \frac{SCE}{SCT} \longrightarrow 1.$
$\Rightarrow SCE \longrightarrow SCT.$

- $R^2 \longrightarrow 1 \Rightarrow 1 - \frac{SCR}{SCT} \longrightarrow 1 \rightarrow \frac{SCR}{SCT} \rightarrow 0$
$\Rightarrow SCR \longrightarrow 0.$

- <u>R^2 proche de 1</u> : modèle parfait $\Rightarrow$ la connaissance de X nous permet de préciser Y.
- <u>R^2 proche de 0</u> : X n'apporte pas d'information sur Y.

Coefficient de Corrélation linéaire

$$R \text{ tq } \quad a = R \frac{\sigma_Y}{\sigma_X}.$$

$$R = \sqrt{R^2}.$$

$$a = \frac{\mathrm{cov}(X,Y)}{\sigma_X^2} = R \frac{\sigma_Y}{\sigma_X}.$$

$$\Rightarrow \quad R = \frac{\mathrm{cov}(X,Y)}{\sigma_X^2} \times \frac{\sigma_X}{\sigma_Y}.$$

$$\boxed{R = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}}$$