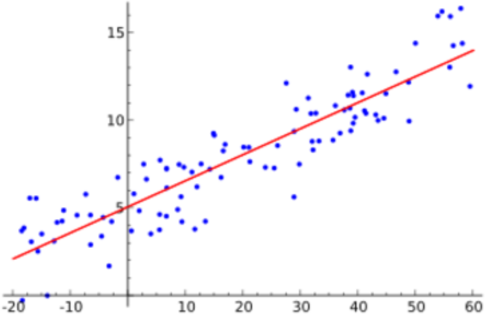
Analyse de données

Rakotoarimalala Tsinjo Tony

Cours 3: Régréssion linéaire multiple

- La **régression** recouvre plusieurs méthodes d'analyse statistique permettant d'approcher une variable à partir d'autres qui lui sont corrélées.
- un modèle de **régression linéaire** est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.
- la **régression linéaire multiple** est une méthode de régression mathématique étendant la régression linéaire simple pour décrire les variations d'une variable endogène (expliquée) associée aux variations de plusieurs variables exogènes (explicatives).



Droite de régression

- en abscisse **la variable explicative** (on n'a qu'une seule dans cet exemple), et en ordonnée la **variable expliqué**
- En bleu on a des nuages de points dans le plan
- En rouge le modèle de régression linéaire

Formalisation du problème

- Étant donné un échantillon $(y_i, X_{i1}, \ldots, X_{ip})$ pour $i \in \{1, n\}$, on cherche à expliquer, avec le plus de précision possible, les valeurs prises par y_i, à partir d'une série de variables explicatives $X_{i1}, ..., X_{ip}$.
- Le modèle théorique, formulé en termes de variables aléatoires, prend la forme

$$y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + \ldots + a_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

- Les coefficients $a_0, a_1, \ldots, a_p$ sont les paramètres à estimer.
- La forme complète est donc

$$\begin{cases} y_1 = a_0 + a_1 x_{1,1} + \ldots + a_p x_{1,p} + \varepsilon_1 \\ y_2 = a_0 + a_1 x_{2,1} + \ldots + a_p x_{2,p} + \varepsilon_2 \\ \ldots \\ y_n = a_0 + a_1 x_{n,1} + \ldots + a_p x_{n,p} + \varepsilon_n \end{cases}$$

Estimateur de moindre carré

- L'estimateur utilisé est donc un estimateur linéaire de la forme

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{i,1} + \cdots + \hat{a}_p x_{i,p}, \quad i = 1 \cdots n$$

- Les résidus estimés $\hat{\epsilon}_i \equiv y_i - \hat{y}_i$ sont la différence entre la valeur de y observée et estimée
- L'objectif est de choisir les $\hat{a}_i$ qui minimise la somme des carrées des résidus

$$(\hat{a}_0, ., \hat{a}_p) = \operatorname{argmin} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \underset{\hat{a}_0,.,\hat{a}_p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{a}_0 - \hat{a}_1 x_{i,1} - \cdots - \hat{a}_p x_{i,p})^2$$

- Les $\hat{\epsilon}_i$ peuvent être interpréter par la distance de la valeur réelle et la valeur donnée par le modèle

Estimateur de moindre carré

- Minimiser $S = \sum_{i=1}^{n} \hat{\epsilon}_i^2$ revient à chercher des solutions de $\frac{\partial(\sum \hat{\epsilon}_i^2)}{\partial \hat{a}_j} = 0$ pour j allant de 0 à p
- On a pour tout $j = 0, \cdots, p$:

$$\frac{\partial(\sum \hat{\epsilon}_i^2)}{\partial \hat{a}_j} = 0 \Leftrightarrow \sum_{i=1}^{n} x_{i,j}(y_i - \hat{a}_0 - \hat{a}_1 x_{i,1} - \cdots - \hat{a}_p x_{i,p}) = 0$$

Sous forme matricielle

$$\underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,p} & x_{1,p} & \cdots & x_{n,p} \end{pmatrix}}_{X^T} \left(\underbrace{\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix}}_{Y} - \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_p \end{pmatrix}}_{A} \right) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

ou encore

$$X^T \left(Y - X\hat{A} \right) = 0$$

Estimateur de moindre carré

Donc il nous suffit de résoudre l'équation suivante sur A

$$X^T\left(Y - X\hat{A}\right) = 0$$

C'est-à-dire

$$X^TY = X^TXA \Leftrightarrow A = \left(X^TX\right)^{-1}X^TY$$

Cette dernière suppose que X^TX est **inversible** c'est-à-dire X de rang $p+1$ (**pas de colinéarité entre les colonnes** (les variables) de X). Dans la pratique on supprime tout simplement les colonnes colinéaires.

Coefficient de détermination

On définit alors les notions suivantes:

- Somme de carrées résiduelle

$$SCR = \sum_{n} \left((y_i - \hat{y}_i)^2 \right)$$

- Somme de carrées expliquée

$$SCE = \sum_{n} \left((\hat{y}_i)^2 - \bar{y} \right)$$

- Somme de carrées totale

$$SCT = SCR + SCE$$

- La coefficient de détermination

$$R^2 = \frac{SCE}{SCT}$$

- On a $0 \leq R^2 \leq 1$. Si R^2 est proche de 0 alors le pouvoir prédictif du modèle est faible et s'il est proche de 1 son pouvoir prédictif est fort