Régression linéaire au sens des moindres carrés

D'après http://exo7.emath.fr/

En mathématiques, les méthodes de régression servent à déterminer une relation entre les données d'une ou plusieurs variables indépendantes, dites explicatives, et celles d'une variable dépendante, dite expliquée. En particulier, la régression linéaire étudie l'existence d'une relation linéaire entre les variables explicatives et la variable expliquée.

1. Régression linéaire simple

La régression linéaire est dite **simple** lorsque l'on ne considère qu'une seule variable explicative.

Contexte. On considère :

1. un nombre $N \geqslant 1$ de données **d'entraînement** constitué de :
   (a) données d'entrée $\{x_1, \dots, x_N\}$, où $x_i \in \mathbb{R}$,
   (b) données de sortie $\{y_1, \dots, y_N\}$, où $y_i \in \mathbb{R}$,
2. une relation (un modèle) linéaire entre l'entrée x_i et la sortie $y_i = g(x_i)$ définie par la fonction :
$$g : x_i \mapsto ax_i + b,$$
où a et b sont respectivement la pente du modèle et son ordonnée à l'origine que l'on va chercher à déterminer,
3. et une fonction d'erreur E dont le but est de mesurer l'erreur commise par notre modèle sur les données d'entraînement :
$$E(a,b) = \sum_{i=1}^{N} (y_i - (ax_i + b))^2.$$

Résoudre la régression linéaire simple revient à trouver la valeur du couple $(a,b) \in \mathbb{R}^2$ qui minimise la fonction d'erreur E.

Mini-exercices 1.

1. On considère les données d'entraînement sous la forme de couples (x_i, y_i) suivants :
$$\mathscr{D} = \{(1, 11.5), (2, 10), (3, 8.4), (4, 5.1), (5, 4.7), (6, 3.6)\}.$$
   (a) Donner la valeur de N.
   (b) Représenter dans le plan les données d'entraînement.
   (c) On considère le modèle linéaire $g : x \mapsto -x + 10$.
      i. Calculer l'erreur commise par ce modèle.
      ii. Représenter dans le plan la droite de régression, et interpréter graphiquement l'erreur commise.

2. Résolution de la régression linéaire simple

Proposition (Équation normale). *Résoudre la régression linéaire revient à résoudre le système d'inconnues (a,b) suivant, appelé **équation normale** :*
$$\begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix}.$$

Proposition (Solution). *Lorsque les x_i ne sont pas tous égaux, la régression linéaire a pour solution :*
$$\begin{cases} a &= \dfrac{Cov(x,y)}{Var(x)} \\ b &= \overline{y} - a\overline{x} \end{cases}.$$

Rappel. On a les formules :
$$\mathrm{Cov}(x,y) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}),$$
$$\mathrm{Var}(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Mini-exercices 2.

1. On reprend les données de l'exercice précédent. Donner la solution au problème de régression linéaire, calculer l'erreur commise dans ce cas et représenter dans le plan la droite de régression.

3. Écriture matricielle de la régression linéaire simple

Notations. Dans cette partie, on note $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$ et $\theta = \begin{pmatrix} b \\ a \end{pmatrix}$ notre inconnue.

Dans ce cas, **résoudre** la régression linéaire simple revient à trouver la valeur de θ qui minimise $\|y - X\theta\|^2$.

Proposition (Équation normale). *Résoudre la régression linéaire revient à résoudre l'**équation normale** :*
$${}^tXX\theta = {}^tXy.$$

Proposition (Solution). *Lorsque les x_i ne sont pas tous égaux, la régression linéaire a pour solution :*
$$\theta = ({}^tXX)^{-1}\,{}^tXy.$$

4. Régression linéaire multiple

La régression linéaire est dite **multiple** lorsque l'on considère un nombre $p \geqslant 2$ de variables explicatives.

Contexte. On considère :

1. un nombre $N \geqslant 1$ de données **d'entraînement** constitué de :
   (a) données d'entrée $\left\{(x_1^1, x_2^1, \dots, x_p^1), \dots, (x_1^N, x_2^N, \dots, x_p^N)\right\}$, où $x_j^i \in \mathbb{R}$,
   (b) données de sortie $\{y_1, \dots, y_N\}$, où $y_i \in \mathbb{R}$,
2. une relation (un modèle) linéaire entre l'entrée $x^i = (x_1^i, x_2^i, \dots, x_p^i)$ et la sortie $y_i = g(x^i)$ définie par la fonction :
$$g : x^i \mapsto \theta_0 + \theta_1 x_1^i + \dots \theta_p x_p^i,$$
où $\theta = (\theta_0, \theta_1, \dots, \theta_p)$ est le paramètre du modèle que l'on va chercher à déterminer,
3. et une fonction d'erreur E dont le but est de mesurer l'erreur commise par notre modèle sur les données d'entraînement :
$$E(\theta_0, \theta_1, \dots, \theta_p) = \sum_{i=1}^{N} (y_i - (\theta_0 + \theta_1 x_1^i + \dots \theta_p x_p^i))^2.$$

Notations. On note $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$, $X = \begin{bmatrix} 1 & x_1^1 & \cdots & x_p^1 \\ 1 & x_1^2 & \cdots & x_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & \cdots & x_p^N \end{bmatrix}$ et $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$ notre inconnue.

Dans ce cas, **résoudre** la régression linéaire simple revient à trouver la valeur de θ qui minimise $\|y - X\theta\|^2$.

Proposition (Équation normale). *Résoudre la régression linéaire revient à résoudre l'**équation normale** :*
$${}^tXX\theta = {}^tXy.$$

Proposition (Solution). *Lorsque la matrice ${}^tXX$ est inversible, la régression linéaire a pour solution :*
$$\theta = ({}^tXX)^{-1}\,{}^tXy.$$

Mini-projet : étude de la régression linéaire au sens des moindres carrés

Ce travail est à réaliser en binôme ou en trinôme et à rendre sur Moodle pour le 31/03.

1. Étude théorique de la régression linéaire simple

1. Calculer les dérivées partielles d'ordre 1 de E, fonction définie dans la section 1 de la page précédente.
2. Déterminer la hessienne de E. On admettra qu'elle est définie positive : que peut-on en déduire à propos des points critiques de E ?
3. Montrer que les points critiques de E sont solution du système d'inconnues (a, b)

$$\begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix} \begin{pmatrix} b \\ a \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix}.$$

4. On note dans cette question $\overline{x}$ et $\overline{y}$ les moyennes respectives de x et y.
   (a) Montrer que $\mathrm{Var}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2$.
   (b) Montrer que $\mathrm{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\,\overline{y}$.
   (c) En déduire la solution de la régression linéaire simple est donnée par :

$$\begin{cases} a &= \dfrac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \\ b &= \overline{y} - a\overline{x} \end{cases}$$

5. (Facultative) Dans cette question, on note $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}$ et $\theta = \begin{pmatrix} b \\ a \end{pmatrix}$ notre inconnue.
   (a) Montrer que $\|y - X\theta\|^2 = E(a, b)$.
   (b) Montrer que résoudre la régression linéaire simple revient à résoudre l'équation normale :

$$^tXX\theta \;=\; ^tXy.$$

   (c) Montrer que lorsque la matrice ^tXX est inversible, l'équation normale a pour solution :

$$\theta = (^tXX)^{-1}\ ^tXy,$$

   et que cette solution est équivalente à celle trouvée à la question 4 (c).

2. En pratique, un exemple de régression linéaire multiple

On considère les données du taux de criminalité de 50 villes étasusiennes [1], regroupées dans le fichier `crime.csv` à télécharger sur Moodle.

city	crime rate	violent crime rate	funding	hs	not-hs	college	college4
1	478	184	40	74	11	31	20
2	494	213	32	72	11	43	18
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
50	940	1244	66	67	26	18	16

Détail des variables :
- `crime rate` : taux global de criminalité déclaré pour 1 million d'habitants
- `violent crime rate` : taux de criminalité violente déclaré pour 100 000 habitants
- `funding` : financement annuel de la police en $/habitant
- `hs` : pourcentage de personnes âgées de 25 ans et plus ayant suivi 4 années d'études secondaires
- `not-hs` : pourcentage de jeunes de 16 à 19 ans n'ayant pas terminé leurs études secondaires et n'ayant pas obtenu de diplôme d'études secondaires.
- `college` : pourcentage de jeunes de 18 à 24 ans dans l'enseignement supérieur
- `college4` : pourcentage de personnes âgées de 25 ans et plus ayant suivi au moins 4 années d'études supérieures

Le but de cet exercice est de construire un modèle de régression linéaire au sens des moindres carrés pour expliquer la variable `crime rate` à partir des variables explicatives `funding`, `hs`, `not-hs`, `college`, et `college4`.

1. Le modèle de régression linéaire est-il simple ou multiple ? Justifier.
2. Donner l'expression du modèle linéaire et de la fonction d'erreur associés, en prenant soin de bien expliciter les notations utilisées.
3. On utilise Python dans cette question pour résoudre la régression linéaire.
   (a) Compléter le script `projet_squelette.py` afin de calculer la solution au problème de régression linéaire.
   (b) Quelle valeur de la variable `crime rate` renvoie ce modèle si l'on donne en entrée les valeurs des variables explicatives de la ville 1 ? Commenter.
   (c) Pour quelle ville la valeur de la variable `crime rate` que renvoie ce modèle est-elle la plus proche de la valeur réelle ?
   (d) Quelle variable explicative semble avoir le plus d'effet sur la variable expliquée ? Comment pourrait-on l'interpréter ?

1. https://hastie.su.domains/StatLearnSparsity/data.html

Basile LE THIEC

RT31

SAE - Poursuite d'étude - Mini - Projet

I) Étude théorique de la régression linéaire simple :

1) $E(a,b) = \sum_{i=1}^{N} (y_i - (ax_i + b))^2$, $N \geq 1$

$$\frac{\partial E}{\partial a} = \frac{\partial}{\partial a} \sum_{i=1}^{N} (y_i - (ax_i + b))^2$$

$$\frac{\partial}{\partial a} (y_i - (ax_i + b))^2 = -2(y_i - (ax_i + b))(x_i)$$

$$= -2 \sum_{i=1}^{N} x_i (y_i - ax_i - b)$$

$$\frac{\partial E}{\partial b} = \frac{\partial}{\partial b} \sum_{i=1}^{N} (y_i - (ax_i + b))^2$$

$$\frac{\partial}{\partial b} (y_i - (ax_i + b))^2 = -2(y_i - (ax_i + b))(1)$$

$$= -2 \sum_{i=1}^{N} (y_i - ax_i - b)$$

2) $$H = \begin{bmatrix} \frac{\partial^2 E}{\partial a^2} & \frac{\partial^2 E}{\partial a \partial b} \\ \frac{\partial^2 E}{\partial b \partial a} & \frac{\partial^2 E}{\partial b^2} \end{bmatrix}$$

$$\frac{\partial^2 E}{\partial a^2} = -2 \sum_{i=1}^{N} x_i \frac{\partial}{\partial a} (y_i - ax_i - b)$$

$$= -2 \sum_{i=1}^{N} x_i (-x_i)$$

$$= 2 \sum_{i=1}^{N} x_i^2$$

$$\frac{\partial^2 E}{\partial b^2} = -2 \sum_{i=1}^{N} \frac{\partial}{\partial b} (y_i - a x_i - b)$$

$$= -2 \sum_{i=1}^{N} (-1)$$

$$= 2N$$

$$\frac{\partial^2 E}{\partial a \partial b} = \frac{\partial^2 E}{\partial b \partial a} = \frac{\partial}{\partial b} \underbrace{\left(-2 \sum_{i=1}^{N} x_i (y_i - a x_i - b) \right)}_{\frac{\partial E}{\partial a}}$$

$$= -2 \sum_{i=1}^{N} x_i \frac{\partial}{\partial b} (y_i - a x_i - b)$$

$$= -2 \sum_{i=1}^{N} x_i (-1)$$

$$= 2 \sum_{i=1}^{N} x_i$$

$$H = \begin{bmatrix} 2 \sum_{i=1}^{N} x_i^2 & 2 \sum_{i=1}^{N} x_i \\ 2 \sum_{i=1}^{N} x_i & 2N \end{bmatrix}$$

On admet que H est définie positive (valeur propre > 0), nous pouvons donc en déduire que tout point critique de la fonction d'erreur (E) correspond à un minimum local.

3) Les points critiques de E sont :

- $\frac{\partial E}{\partial a} = -2 \sum_{i=1}^{N} x_i (y_i - a x_i - b) = 0$

- $\frac{\partial E}{\partial b} = -2 \sum_{i=1}^{N} (y_i - a x_i - b) = 0$

On a donc le système suivant :

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

Qui peut être exprimé sous forme matricielle :

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \boxed{\begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}}$$

4) On a $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

ⓐ $\mathrm{Var}(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$

$$(x_i - \bar{x})^2 = x_i^2 - 2 x_i \bar{x} + \bar{x}^2$$

$$\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - 2\bar{x} \sum_{i=1}^{N} x_i + N\bar{x}^2 \quad , \text{avec} \sum_{i=1}^{N} x_i = N\bar{x}$$

$$= \sum_{i=1}^{N} x_i^2 - 2N\bar{x}^2 + N\bar{x}^2$$

$$= \sum_{i=1}^{N} x_i^2 - N\bar{x}^2$$

$$\Rightarrow Var(x) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

(b) $$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

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

$$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - \bar{y}\sum_{i=1}^{N} x_i - \bar{x}\sum_{i=1}^{N} y_i + N\bar{x}\bar{y}$$

$$-2N\bar{x}\bar{y}$$

avec $\sum_{i=1}^{N} x_i = N\bar{x}$ et $\sum_{i=1}^{N} y_i = N\bar{y}$

$$\Rightarrow \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}$$

$$\Rightarrow Cov(x,y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

(c) $$\begin{cases} a = \dfrac{Cov(x,y)}{Var(x)} \\ b = \bar{y} - a\bar{x} \end{cases}$$

sachant que $E(a,b)$ sont obtenus de cette manière (minimisés E) :

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{bmatrix}$$

Cmq :

① $N b + a \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$

② $b \sum_{i=1}^{N} x_i + a \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$

isoler b de ① : $b = \dfrac{\sum_{i=1}^{N} y_i - a \sum_{i=1}^{N} x_i}{N}$

substitution de b dans ② :

$$\left(\frac{\sum_{i=1}^{N} y_i - a \sum_{i=1}^{N} x_i}{N} \right) \sum_{i=1}^{N} x_i + a \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$$

isolons a : $$a = \frac{\sum_{i=1}^{N} (x_i y_i) - \frac{\sum_{i=1}^{N} x_i}{N} \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}$$

En simplifiant avec les resultats précédent et $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$,

$$\boxed{a = \frac{Cov(x, y)}{Var(x)}}$$

$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$:

De même pour b, $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$:

$$\boxed{b = \bar{y} - a\bar{x}}$$

II) En pratique, régression linéaire multiple

1) Il s'agit d'une régression linéaire multiple. En effet, la variable expliquée (crime rate) est modélisée par plusieurs variables (funding, hs, not-hs, college, college4).

2) $$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \theta_3 x_{i3} + \theta_4 x_{i4} + \theta_5 x_{i5} + \varepsilon_i$$

y_i : crime rate pour $i^{ème}$ ville

x_{i1} : funding

x_{i2} : hs

x_{i3} : not-hs

x_{i4} : college

x_{i5} : college4

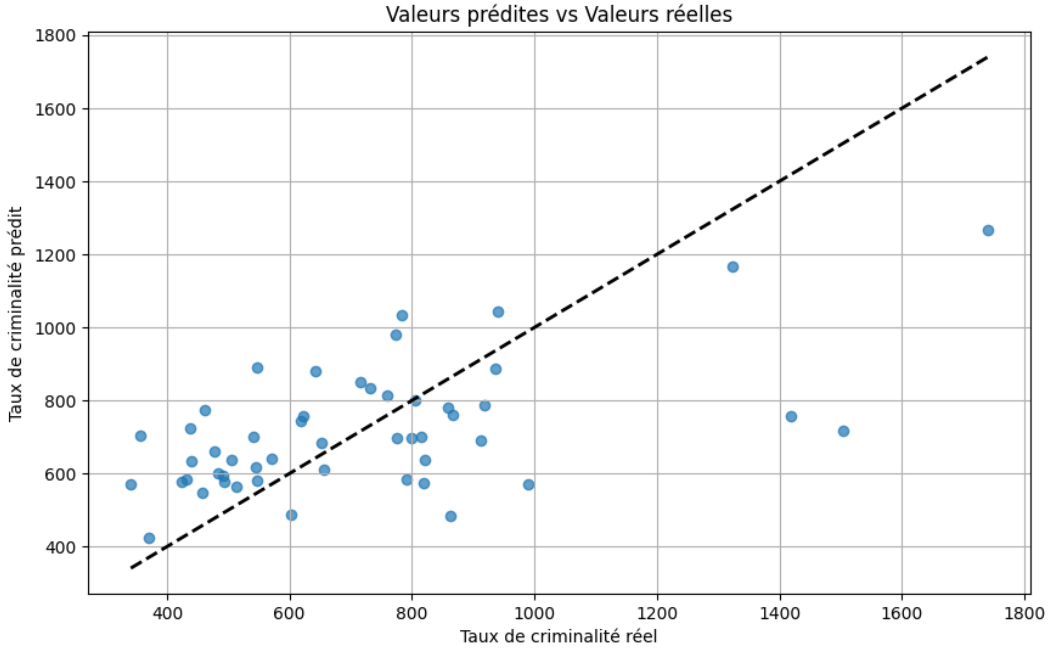
θ_0 : constante

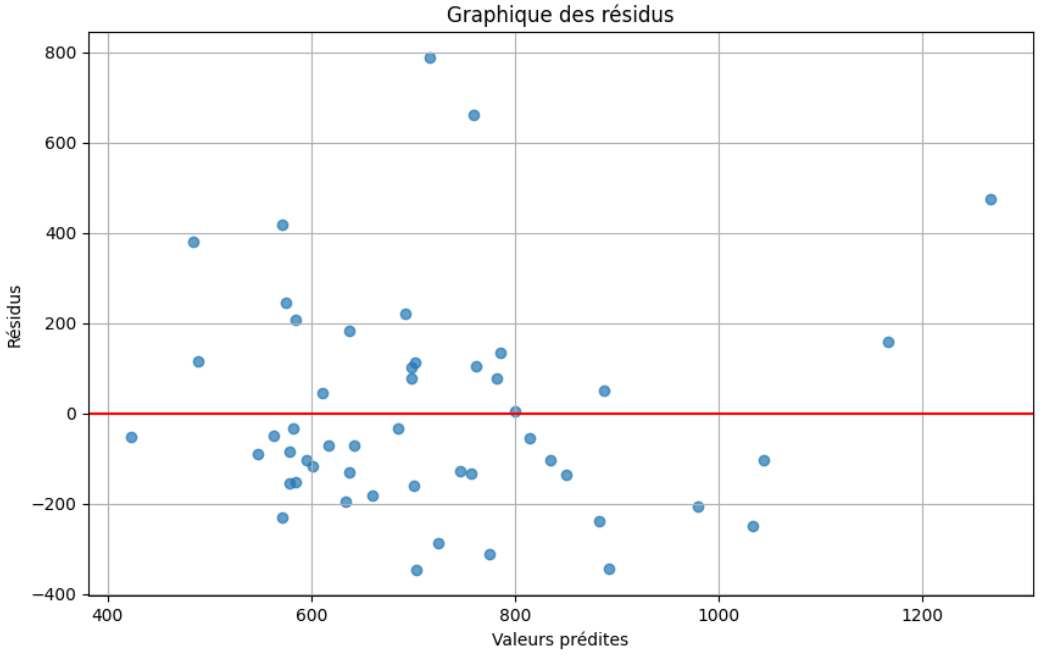
$\theta_1, \theta_2, \theta_2, \theta_4, \theta_5$: coefficients des variables explicatives

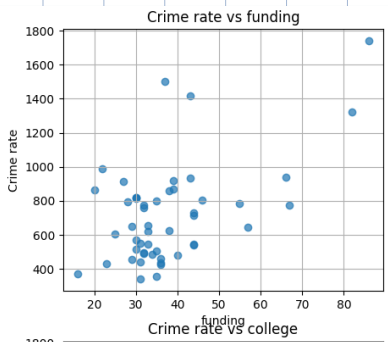
ε_i : erreur aléatoire

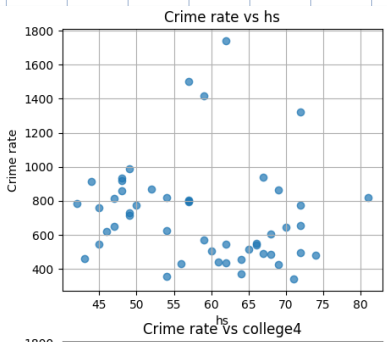
$$E(\theta) = \sum_{i=1}^{N} (y_i - \theta_0 - \theta_1 x_{i1} - \theta_2 x_{i2} - \theta_3 x_{i3} - \theta_4 x_{i4} - \theta_5 x_{i5})^2$$

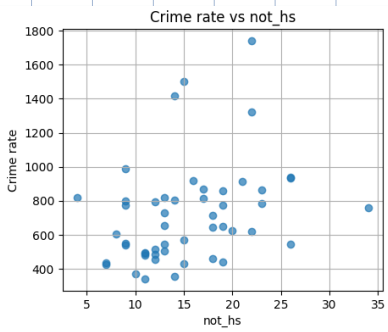
3) @

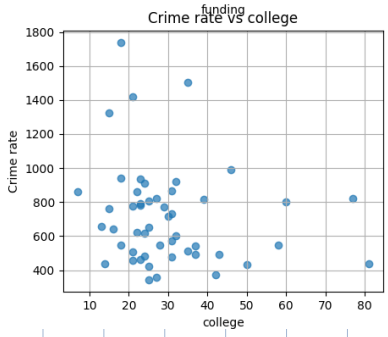

Valeurs prédites vs Valeurs réelles
Taux de criminalité prédit
Taux de criminalité réel
1800
1600
1400
1200
1000
800
600
400
400
600
800
1000
1200
1400
1600
1800

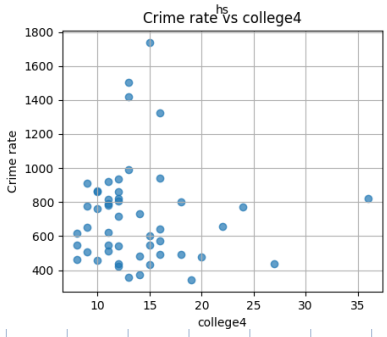

Graphique des résidus
Résidus
Valeurs prédites
800
600
400
200
0
−200
−400
400
600
800
1000
1200











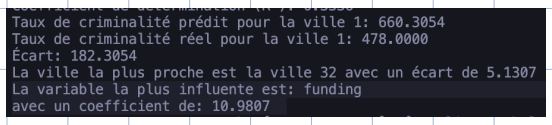

```
Coefficients de régression (theta):
Constante: 489.6486
funding: 10.9807
hs: -6.0885
not-hs: 5.4803
college: 0.3770
college4: 5.5005
Erreur quadratique moyenne (MSE): 56425.1892
Coefficient de détermination (R²): 0.3336
```

(b)
- Taux de criminalité prédit pour la ville 1: 660,3054
- Taux de criminalité réel pour la ville 1 : 478,0000
- Ecart : 182,3054

Nous observons que la prédiction est nettement supérieure à la valeur réelle. Cela indique que le modèle surestime le taux de criminalité pour la ville 1.

(c) La ville la plus proche du résultat du modèle est la ville 32 avec un écart de 5,1307.

(d) La variable la plus influente est : funding avec un coefficient de 10,9807. Cela signifie qu'une hausse d'une unité de la variable funding entraine une augmentation de 10,98 du taux de criminalité prédit. Funding exerce une forte influence dans le modèle.



https://github.com/basilelt/mini-projet