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