M1 Informatique AMU – TP2 ISD — 2023/24 — Rapport. **Date : 08 novembre 2023.**
ZEMMOURI Yasmine, Groupe 3 TP# 3.

La régression linéaire par moindres carrés sous sklearn

Partie : Etude d'un régresseur linéaire simple

Le code est disponible dans le fichier **partie1_TP3.py** et **partie1_Question6_TP3.py.**

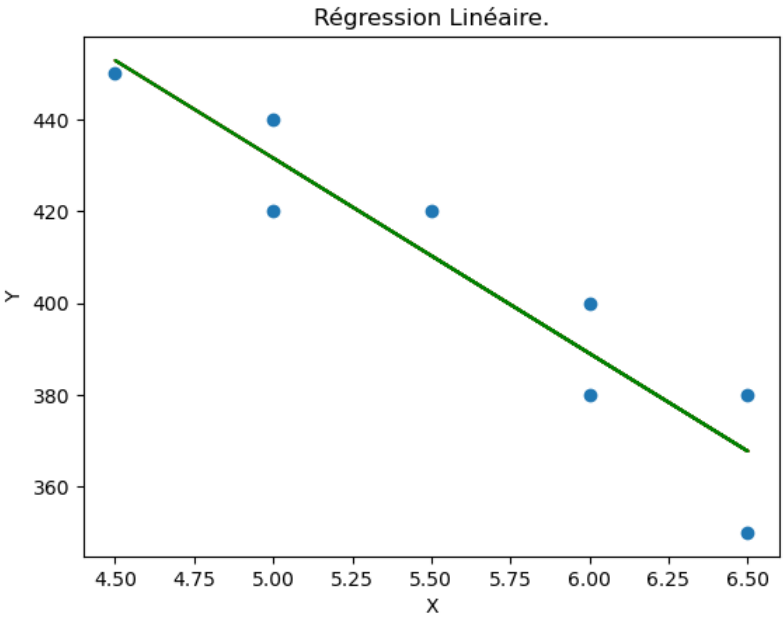
Une droite de régression linéaire s'écrit sous la forme $a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$, avec {$x_1$, $x_2 \ldots x_n$} nos données.

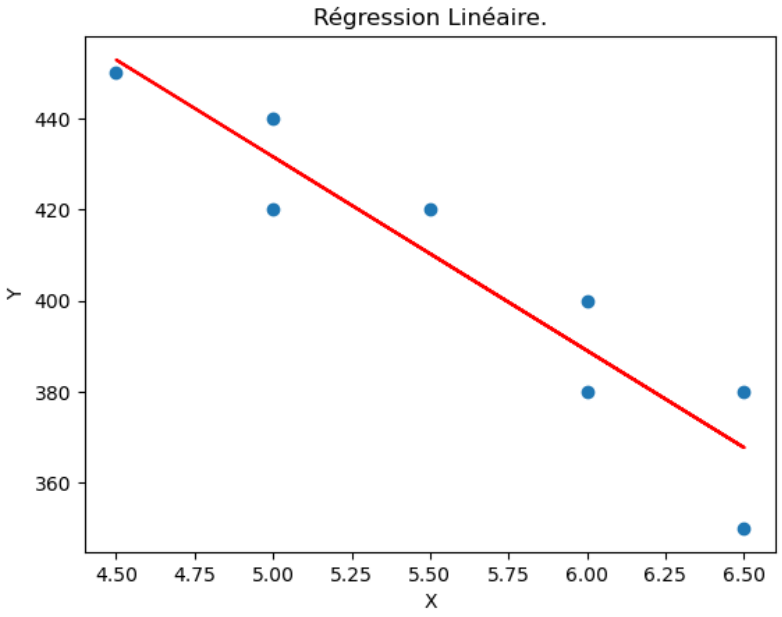
L'attribut coef_ représente les poids attribués aux caractéristiques de nos données (les coefficients a_i), et l'attribut intercept_ représente le b.

Le coefficient de détermination, le score, est un nombre qui nous indique à quel point notre modèle de régression linéaire correspond bien aux données. Plus il est proche de 1, mieux c'est. Dans notre cas, un score de 0.8782 est plutôt bon. Cela signifie que le modèle peut expliquer environ 87,82 % de la variation dans nos données.

La MSE, quant à elle, est relativement petite par rapport à l'échelle des données (121.77), cela signifie que les valeurs prédites par le modèle de régression sont très proches des valeurs réelles de l'ensemble de données. Le modèle est capable de s'ajuster de manière précise aux données.

Les graphes suivants représentent la droite apprise (en rouge), la droite de régression (en vert) et le nuage des points (initialement dans un seul plot).





On remarque les deux droites coïncident, elles sont exactement les mêmes. Elles traversent le nuage de points de façon à minimiser leur écart par rapport à chaque point.

- Lorsque le paramètre fit_intercept est mis à faux, l'attribut intercept_ n'est pas pris en considération, il est mis à 0. Ce qui revient à faire passer la droite par l'origine 0. L'équation de la droite devient alors : $a_1x_1 + a_2x_2 + \cdots + a_nx_n$.

On observe que le score obtenu est de -5.386, ce qui est très mauvais pour un modèle, de régression ou pas. Quant à la MSE, qui est de 6386.11, elle est également élevée. Le modèle a du mal à ajuster correctement les données.

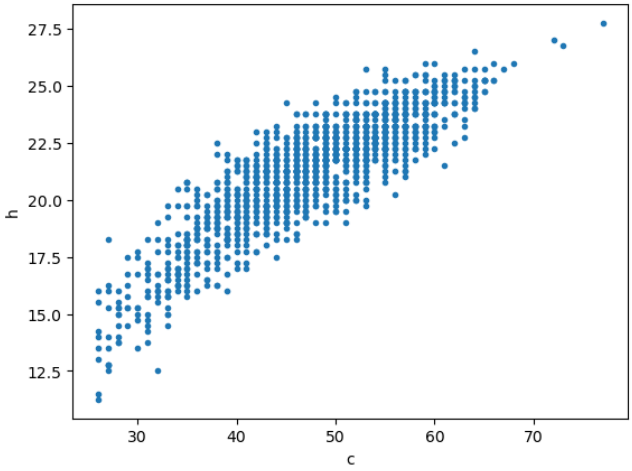
L'option `fit_intercept=False` est généralement utilisée avec des données standardisées pour garantir que la droite de régression passant par l'origine n'entraîne pas de biais dans le modèle.

Partie : Jeu de données Eucalyptus

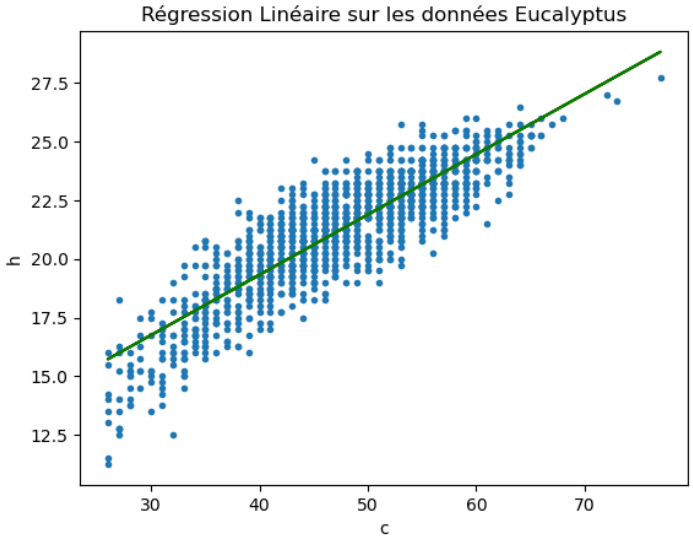
Le code produit est disponible dans le fichier **partie2_TP3.**

Visualisation du nuage de points des données Eucalyptus :

On remarque qu'il y'a une forte corrélation positive entre les deux attributs. Cela signifie qu'il existe une relation linéaire entre les deux. Ainsi, que l'un peut être calculée à partir de l'autre.

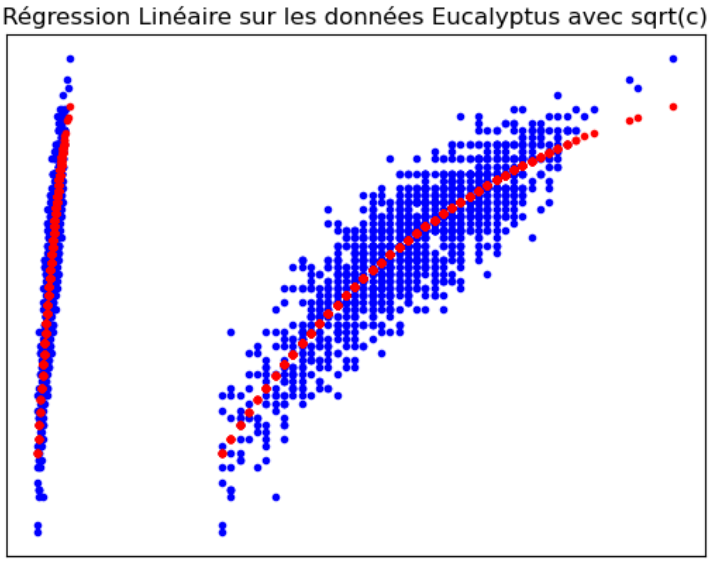


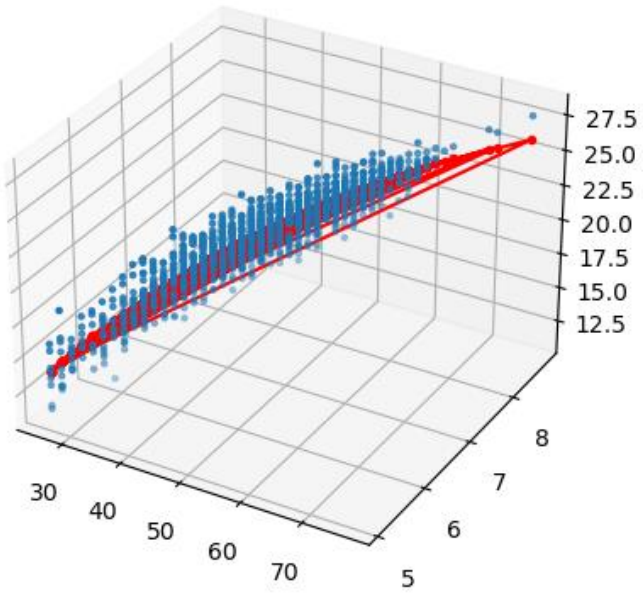
Visualisation de la droite de régression sur le nuage de points des données Eucalyptus :



Nous obtenons un score par validation croisée de 0.70247 pour ce modèle, et une MSE de 1.47 ce qui est assez bon pour un modèle de prédiction.

Visualisation de la droite de régression sur le nuage de points des données Eucalyptus complétées par sqrt(c) :





Les résultats obtenus s'améliorent: un score de 0.7328 et une MSE de 1.288.

En ajoutant le carré de c, en plus de la racine de c, les performances obtenus sont de 0,7351 pour le score et 1.276 pour la MSE.

Le meilleur modèle de régression est donc le dernier : en rajoutant des variables <u>linéairement</u> indépendantes, nous avons amélioré les performances du modèles. Il faut cependant ajouter un nombre considérables de variables afin d'éviter le surajustement (overfitting).

```
Meilleur score by cross_val = 0.7351224466101725 de la régression num : 3
```