Statistiques pour la SAé 2.04

TP2 -Régression Linéaire Multiple

Dans ce TP, on va **apprendre à effectuer une régression linéaire multiple** en utilisant la fonction `LinearRegression` et en l'utilisant sur les données Sangliers importées dans le TP précédent.

Les données Sangliers - Problématique

Cette partie est une présentation complète des données et de la problématique, qui suit les consignes qui vous seront données (+ tard) pour le rendu de SAé.

Le fichier Sangliers.csv contient plusieurs séries statistiques sur un même ensemble d'années :

- La **population** est l'ensemble des années de 2000 à 2015.
- La **1e série** correspond au nombre de sangliers prélevés (=tués, en langue de bois) par des chasseurs chaque année en France.
- La **2e** correspond au montant total payé par les assurances pour indemniser les dégâts causés par des sangliers en France, en euros.
- La **3e série** correspond au nombre de M de tonnes de viande de porc consommée dans les pays de l'OCDE.
- La **4e série** correspond au nombre de permis de chasses validés en France pour chaque année.

En utilisant ces données, on va essayer de répondre à la problématique :

Problematique

La quantité de dégâts causés par les sangliers est-elle une conséquence de leur colère face au nombre de morts causées dans leurs rangs par les chasseurs et/ou de la consommation de viande de porc ?

1. Pour répondre à cette problématique avec la régression linéaire multiple, quelle sera la variable endogène, et quelles seront les variables explicatives ?

Régression linéaire multiple

On va **apprendre à effectuer une régression linéaire multiple** en utilisant la fonction `LinearRegression` du package `sklearn`. On commence donc par importer le package :

```
from sklearn.linear_model import LinearRegression
```

2. On utilise à nouveau le fichier `Sangliers.csv` téléchargeable sur Moodle. L'importer dans Python sous forme d'un DataFrame `Sa_DF`, puis le transformer en un NdArray `Sa_Ar`.
3. Créer ensuite le NdArray `Sa0_Ar` correspondant au tableau `Sa_Ar` sans la colonne des années, qui n'est pas une de nos variables mais la population.
4. Créer le tableau `Sa0_Ar_N` contenant la normalisation (centrage, réduction) de `Sa0_Ar`.
5. Créer un np.array `Sa_Y` à une colonne, contenant les données (normalisées) sur les montants des dégâts causés par les sangliers (la **variable endogène**).
6. A partir du tableau `Sa0_Ar_N`, créer un NdArray `Sa_X` à 3 colonnes, contenant les 3 autres séries statistiques (les **variables explicatives**).
7. Pour faire la régression linéaire multiple, exécuter les commandes suivantes :
```
linear_regression = LinearRegression()
linear_regression.fit(Sa_X, Sa_Y)
a=linear_regression.coef_
```
8. Afficher ce que contient `a`. A l'aide de l'encadré **Interprétation des paramètres** du cours, proposer une interprétation des paramètres contenus dans `a`.
9. Comparer votre interprétation à celle faite par Gwendal, Jean-Baptiste, et Théo en 2019 :

On remarque qu'il y a une forte évolution pour la valeur « a_1 » (~ 1,0). On en déduit que plus il y a de sangliers prélevés par les chasseurs plus les dégâts causés par des sangliers augmentent car ils se vengent en détruisant les cultures.

On remarque cependant qu'il y a une évolution plutôt moyenne des valeurs « a_2 » (~ -0,49) et « a_3 » (~ 0,44). On peut en déduire pour la valeur « a_2 » que plus il y a viande de porc consommée plus les sangliers sont calmes car les cochons et les sangliers ne font pas partie de la même famille.

Pour la valeur « a_3 » on en déduit que moins il y a de chasseurs plus les dégâts des sangliers augmentent car les sangliers ont moins peur des chasseurs.

Coefficient de corrélation multiple pour Sangliers

10. A partir des valeurs de `a` et des variables explicatives, créer le tableau à une colonne `Ypred`, de prédiction des montants des dégats causés par les sangliers.
11. Calculer le coefficient de corrélation multiple `CorSangliers`, puis comparer votre résultat `CorSangliers` à la commande suivante :
    ```
    CorS=linear_regression.score(Sa_X,Sa_Y)
    ```
12. La corrélation entre les dégâts causés par les sangliers et les 3 autres variables est-elle forte ?
13. Proposer une réponse à la problématique. Avez-vous une idée de facteur confondant ?

... Pour les plus rapides ...

Calcul des paramètres par opération matricielle

Le calcul des paramètres contenus dans le tableau `a` ci-dessus n'est pas si compliqué avec un calcul matriciel. Dans cette partie, nous allons le faire en quelques étapes.
On importe d'abord le package d'algèbre linéaire qui permet de faire l'inversion et la transposition de matrices :

```
import numpy.linalg as la
```

14. Créer une matrice `X1` qui contient la matrice `Sa_X` créée ci-dessus à laquelle on ajoute une colonne de 1 à gauche.
15. Le tableau `a` ci-dessus est ensuite obtenu par le calcul suivant, où ${}^tX1$ est la transposée de la matrice `X1`, `Y` est la matrice créée ci-dessus, et "." est le produit matriciel :

$$a = ({}^tX1.X1)^{-1} \,.\, {}^tX1.y$$