Cours-TP2

Régression Linéaire Multiple

sur l'exemple des sangliers

Statistique pour la SAé 2.04

Tiphaine Jézéquel

2023-2024

Plan

1. Régression Linéaire Multiple : définitions *(rappels)*

2. Analyse des paramètres obtenus

3. Coefficient de corrélation multiple

1. Régression Linéaire Multiple : définition *(rappels)*

- avec 2 variables X et Y, équation de droite $y = ax + b$
- avec 3 variables X_1, X_2 et Y, équation de plan $y = a_1x_1 + a_2x_2 + b$
- avec 4 variables X_1, X_2, X_3 et Y, équation $y = a_1x_1 + a_2x_2 + a_3x_3 + b$
- ...

Définition

On suppose qu'on a $n + 1$ variables statistiques : $X_1, X_2, \ldots, X_n$ et Y définies sur une même population.

Faire la **régression linéaire multiple** de Y en fonction de $X_1, X_2, \ldots, X_n$ consiste à trouver les coefficients $a_1, a_2, \ldots, a_n$ et b tels que l'équation

$$y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$$

soit la meilleure approximation possible de la relation entre Y et les variables $X_1, X_2, \ldots, X_n$.

Définitions

Les variables statistiques $X_1, X_2, \ldots, X_n$ sont appelées les **variables explicatives**.
La variable statistique Y est appelée la **variable endogène**.
Les coefficients $a_1, a_2, \ldots, a_n$ et b de l'équation trouvée sont appelés les **paramètres**.

- **Exemple des âges des enfants**

Enfant	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Taille	0.84	0.98	1.14	1.32	1.44	0.8	0.95	1.2	1.25	1.4	0.9	0.99	0.98	1.12
Crocro	0	0.3	0.8	0.9	1	0.1	0.4	0.7	1	1	0.1	0.35	0.5	0.6
Moucher	0	0.2	0.9	0.8	1	0.1	0.3	0.8	0.9	1	0.1	0.4	0.4	0.5
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7

- la variable endogène est .....
- les variables explicatives sont .....
  .....
- En faisant la régression linéaire multiple, on obtiendra les coefficients $a_1 = 7.5, a_2 = 3.6, a_3 = 1.6$ et $b = -4$.

$$\text{Age} \approx 7.5 \times \text{Taille} + 3.6 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

2. Analyse des paramètres obtenus

A partir de 4 variables statistiques, on obtient une équation $y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$, qui est l'équation de ce qu'on appelle un *hyperplan*... qu'on ne peut pas représenter graphiquement !
Il faut donc savoir interpréter directement les **paramètres**.

Interprétation des paramètres

On suppose qu'on a fait la régression linéaire multiple d'une série Y en fonction de $X_1, X_2, \ldots, X_n$, et qu'on a obtenu les paramètres $a_1, a_2, \ldots, a_n$ et b.

- si $a_i > 0$ est positif, cela signifie que la variable X_i influence positivement la variable Y.
- si $a_i < 0$, la variable X_i influence négativement Y.

De plus, si les variables $X_1, X_2, \ldots, X_n$ et Y ont été centrées-réduites avant d'effectuer la régression:

- $b = 0$
- plus $|a_i|$ est grand, plus la variable X_i a une influence forte sur Y.

5 [5]

- **Exemple des âges des enfants**

  • En faisant la régression linéaire multiple *sans centrer-réduire*, on obtient les coefficients $a_1 = 7.5, a_2 = 3.6, a_3 = 1.6$ et $b = -4$.

  $$\text{Age} \approx 7.5 \times \text{Taille} + 3.6 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

  Interprétation : les 3 variables influencent positivement l'age. C'est-à-dire que l'âge augmente quand la taille, la prononciation et la capacité de mouchage augmentent.

  • En faisant la régression linéaire multiple après avoir centré-réduit, on obtient $a_1 = 0.47, a_2 = 0.37, a_3 = 0.17$ et $b = 0$.

  Interprétation : c'est la taille qui influence le plus l'âge, et la capacité de mouchage qui influence le moins.

3. Coefficient de corrélation multiple

Pour les corrélations à 2 variables, on avait appris à calculer le coefficient de corrélation, qui indiquait si les données étaient proches de la droite.

Pour estimer si la régression linéaire multiple est de bonne qualité, on va calculer un coefficient de corrélation... multiple.

Coefficient de corrélation multiple

Soit $X_1, X_2 \ldots, X_n$ et y des variables statistiques définies sur une même population P de N individus.
On suppose qu'on a effectué la régression linéaire multiple de Y en fonction de $X_1, X_2, \ldots, X_n$, et qu'on a obtenu les paramètres $a_1, a_2, \ldots, a_n$ et b.

• on introduit la variable statistique Y_{pred} qui, pour chaque individu i de la population P, donne la prédiction donnée par la régression linéaire multiple :

$$Y_{pred}(i) = a_1 X_1(i) + a_2 X_2(i) + \ldots + a_n X_n(i) + b$$

• le coefficient de corrélation multiple de cette régression est alors

$$\big(Cor((X_1, \ldots, X_n), Y)\big)^2 = 1 - \frac{\sum_{i \in P} (Y_{pred}(i) - Y(i))^2}{N\, Var(Y)}$$

• **Exemple des âges des enfants**

On avait obtenu l'équation :

$$\text{Age} \approx 7.5 \times \text{Taille} + 3.6 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

Donc pour chaque enfant l'age prédit est

$$Age_{pred} = 7.5 \times \text{Taille} + 3.6 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

Enfant	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7
Âge prédit	1.5	3.9	7.9	9.7	11.7	2.2	4.7	8.5	10.1	11.4	3	5	5.5	7

On rappelle que la variance de l'âge est de 11.2. Donc

$\big(Cor((Taille, Crocro, Moucher), Age)\big)^2 = \ldots\ldots$

Statistique pour la SAé 2.04 - TP2
Régression Linéaire Multiple - sur les Sangliers

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 langage camouflage) 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 suivante :

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 ?

En choisissant la 2e séries statistique comme **variable endogène** et les 3 autres séries comme **variables explicatives**, la **régression linéaire multiple** nous permettra d'obtenir une estimation du montant des dégats en fonction des 3 autres séries statistiques.

Les **paramètres** de la régression linéaire multiple nous informeront des descripteurs qui influencent le plus le montant des dégâts.
Le **coefficient de corrélation multiple** nous permettra d'apporter une réponse à la problématique.

Régression linéaire multiple sur données Sangliers

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
```

On va avoir besoin de ce qui a été fait dans le TP précédent, il peut être judicieux de rédiger ce TP à la suite.

1. A partir du tableau `SangliersOAr` créé dans le TP1, créer un np.array `SangliersOAr_CR` contenant toutes les données centrées-réduites.
2. Créer un np.array `Y` à une colonne, contenant les données sur les montants des dégâts causés par les sangliers (la **variable endogène**).
3. A partir du tableau `SangliersOAr_CR`, créer un np.array `X` à 3 colonnes, contenant les 3 autres séries statistiques (les **variables explicatives**).
4. Pour faire la régression linéaire multiple, exécuter les commandes suivantes :
   ```
   linear_regression = LinearRegression()
   linear_regression.fit(X, Y)
   a=linear_regression.coef_
   ```
5. Afficher ce que contient `a`. A l'aide de l'encadré **Interprétation des paramètres** du cours ci-dessus, proposez une interprétation des paramètres contenus dans `a`.
6. 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

7. 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.

8. Calculer le coefficient de corrélation multiple `CorSangliers`, puis comparer votre résultat `CorSangliers` à la commande suivante :
   ```
   CorS=linear_regression.score(X,Y)
   ```

9. Pensez-vous que la corrélation entre les dégâts causés par les sangliers et les 3 autres variables est forte ?

... 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
```

10. Créer une matrice `X1` qui contient la matrice `X` créée ci-dessus à laquelle on ajoute une colonne de 1 à gauche.

11. 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 :

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