

***SAÉ 2.04 Exploitation
d'une base de données***



Compte rendu

***Yanis PONTTHOU
LUCAS DESPERROIS***

0) Les données <Nom de fichier> - Problématique

(a) Présentation des données

Le fichier "Regression.csv" contient plusieurs séries statistiques sur l'ensemble des formations répertoriées dans Regression:

- La population est l'ensemble des formations, représentées par leur code "cod_aff" et leur nom.
- La première série correspond la capacité de chaque formation
- La deuxième série correspond à l'effectif total de candidats (Tous sexe confondu) qui ont postulé à la formation
- La troisième série correspondent à l'effectif total de candidates uniquement qui ont postulé à la formation
- La quatrième série correspondent à l'effectif total de proposition d'admission que la formation a faite.
- La cinquième série correspond à l'effectif de total des admis qui sont boursier.
- La sixième série correspond à l'effectif d'admis venant de la même académie de la formation
- La septième série correspond pourcentage d'admis selon le nombre de candidat pour la formation.

(b) Problématique

En utilisant ces données, nous allons essayer de répondre à la problématique suivante :

Est-ce que l'effectif total des candidats, l'effectif total des candidates, l'effectif total des propositions d'admission, le pourcentage d'admis selon le nombre de candidats et l'effectif total des admis boursiers peuvent être utilisés pour prédire la capacité d'un établissement ?

(c) Utilisation de la régression linéaire multiple - Comment ?

Nous avons choisi la variable statistique 1, la capacité d'une formation comme variable endogène, et on va essayer d'expliquer si l'on peut prédire la capacité en fonction des variables explicatives que nous avons sélectionnée dans notre vue.

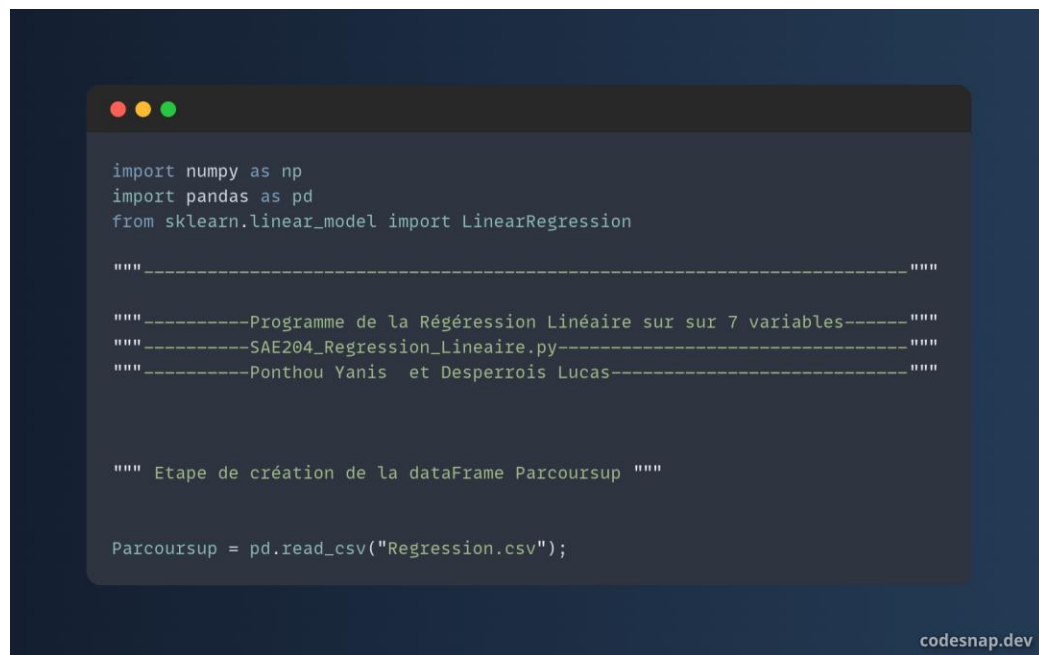
(d) Utilisation de la régression linéaire multiple - Pourquoi ?

Les paramètres de la régression linéaire multiple nous informeront des coefficients qui influencent le plus la capacité des formations. En observant si cette estimation est fondée, ce qui nous permettra d'avoir une réponse à notre problématique

1) Import des données, mise en forme

(a) Importer les données en Python

Nous importons notre vue sous forme de DataFrame avec la commande suivante :



```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

"""-----"""

"""-----Programme de la Régression Linéaire sur 7 variables-----"""
"""-----SAE204_Regression_Lineaire.py-----"""
"""-----Ponthou Yanis et Desperrois Lucas-----"""

""" Etape de création de la dataframe Parcoursup """

Parcoursup = pd.read_csv("Regression.csv");
```

codesnap.dev

(b) Mise en forme

Nous avons quand même vérifié s'il n'y avait pas de case vide. On transforme notre DataFrame en Array :



```
"""-----"""

""" Transformation des cases vides et transformation en Array """

Parcoursup = Parcoursup.dropna();
ParcoursupArray = Parcoursup.to_numpy();

"""-----"""
```

codesnap.dev

Index	capacite	tif_total_candi	tif_total_candi	al_proposition	_admis_boursi	admis_meme_	e_admis_selor
0	24	495	32	125	7	15	4.0404
1	40	566	153	163	9	10	6.89046
2	63	1782	1149	81	15	2	4.15264
3	24	407	44	265	0	15	12.0393
4	15	17	15	13	2	2	17.6471
5	48	3140	1574	541	12	12	1.52866
6	15	167	127	56	3	1	8.98204
7	40	405	342	71	2	7	9.62963
8	190	9397	2785	5048	16	34	2.00064
9	16	204	74	35	1	5	3.43137
10	30	892	433	397	2	11	2.24215
11	8	35	5	23	1	1	14.2857
12	24	277	154	90	2	10	7.94224
13	24	206	25	76	2	12	10.1942

(C) Centrer-réduire

```

""" ----- """

""" Définition de la fonction Centre_Reduire et creation de la matrice centré réduire """

def centre_reduire(T):
    T=np.array(T,dtype=np.float64);
    moyennes = np.mean(T, axis=0);
    ecarts_types = np.std(T, axis=0);
    Res = (T - moyennes) / ecarts_types;
    return Res;

ParcoursupArrayCR = centre_reduire(ParcoursupArray);
""" ----- """

```

	0	1	2	3	4	5	6
0	-0.381636	-0.26213	-0.476403	-0.320305	-0.0945031	-0.198934	-0.615622
1	-0.157062	-0.210213	-0.342553	-0.232922	0.0664903	-0.354156	-0.184404
2	0.165764	0.678948	0.759218	-0.421485	0.54947	-0.602513	-0.598641
3	-0.381636	-0.326477	-0.463129	0.00163268	-0.65798	-0.198934	0.594625
4	-0.50796	-0.611652	-0.495208	-0.577855	-0.496987	-0.602513	1.44309
5	-0.0447746	1.67194	1.22935	0.636309	0.30798	-0.292067	-0.995653
6	-0.50796	-0.501969	-0.371314	-0.478974	-0.41649	-0.633557	0.132055
7	-0.157062	-0.32794	-0.133482	-0.44448	-0.496987	-0.44729	0.230037
8	1.94833	6.24717	2.56895	11.0004	0.629967	0.390913	-0.924242
9	-0.493924	-0.474914	-0.429943	-0.527264	-0.577483	-0.509379	-0.707769
10	-0.297421	0.0281636	-0.0328186	0.305174	-0.496987	-0.323112	-0.8877
11	-0.606211	-0.59849	-0.50627	-0.554859	-0.577483	-0.633557	0.934509
12	-0.381636	-0.421535	-0.341447	-0.400789	-0.496987	-0.354156	-0.0252684
13	-0.381636	-0.473452	-0.484146	-0.432983	-0.496987	-0.292067	0.315453

2) Choix des variables explicatives

(a) Démarche

Nous avons sélectionné nos variables en voulant les inclure dans la matrice de covariance donc nous n'avons pas réduit le nombre de variables explicatives

```

""" Obtention de la Matrice de CO-Variance """

MatriceCov = np.cov(ParcoursupArrayCR,rowvar=False);

```

codesnap.dev

(b) Matrice de covariance

On obtient la matrice suivante :

	0	1	2	3	4	5	6
0	1.002	0.617861	0.573172	0.692471	0.867072	0.797366	0.143526
1	0.617861	1.002	0.920515	0.791252	0.567228	0.425624	-0.283471
2	0.573172	0.920515	1.002	0.614472	0.528823	0.380518	-0.235962
3	0.692471	0.791252	0.614472	1.002	0.553278	0.524945	-0.0966662
4	0.867072	0.567228	0.528823	0.553278	1.002	0.773826	0.0756376
5	0.797366	0.425624	0.380518	0.524945	0.773826	1.002	0.109895
6	0.143526	-0.283471	-0.235962	-0.0966662	0.0756376	0.109895	1.002

(c) Variables explicatives les plus pertinentes

Notre objectif est de trouver les variables qui expliquent le mieux possible la capacité d'un établissement, qui se trouve dans la colonne 0 de `ParcoursupArrayCR`. La colonne 0 de `MatriceCov` donne les coefficients de corrélation de la capacité d'une formation avec chacune des autres variables. On va choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec la capacité d'une formation.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 0 de `MatriceCov` sont : 0.86, 0.79, 0.69. Ils correspondent aux variables 4, 5, 3. Les colonnes 4, 5 et 3 correspondent aux:

- 4 : série correspond à l'effectif de total des admis qui sont boursier.
- 5 : série correspond à l'effectif d'admis venant de la même académie de la formation
- 3 : série correspondent à l'effectif total de proposition d'admission que la formation a faite.

Ont choisi donc ces 3 variables explicatives

3) Régression linéaire multiple pour "Regression.csv"

(a) Régression linéaire multiple

On fait maintenant la régression linéaire multiple avec la série concernant la capacité comme variables endogène, et les 3 variables explicatives trouvées ci-dessus.

(b) Paramètres, interprétation

```
"""-----"""

""" Regression Linéaire des 3 variables """

Y = ParcoursupArrayCR[:,0]      #variable endogene
X = ParcoursupArrayCR[:,[4,5,3]] #3 Variables Explicatives

linear_regression = LinearRegression();
linear_regression.fit(X, Y)

a=linear_regression.coef_

"""-----"""
```

codesnap.dev

	0
0	0.522019
1	0.250272
2	0.271725

Premièrement les variables sont toutes les 3 positifs donc elles influencent positivement la capacité (Variable Y)

De plus nous constatons que les variables ont été centré réduite avant la régression puisque les coefficients a_i sont compris entre -1 et 1 .

Nous pouvons conclure que c'est a_0 qui influence assez fortement la capacité d'une formation car c'est la valeur la plus proche de 1. C'est la série correspond à l'effectif de total des admis qui sont boursier. On remarque cependant qu'il y a une influence entre moyenne et faible des variables explicatives des cases 1 et 2 ce qui correspond aux séries correspondant à l'effectif d'admis venant de la même académie de la formation et la série correspondant à l'effectif total de proposition d'admission que la formation a faite sur la capacité

(c) Coefficient de corrélation multiple, interprétation

```
"""-----"""

""" Regression Linéaire des 3 variables et Coefficient de Correlation_multiple """

Y = ParcoursupArrayCR[:,0]      #variable endogene
X = ParcoursupArrayCR[:,[4,5,3]] #3 Variables Explicatives

linear_regression = LinearRegression();
linear_regression.fit(X, Y)

a=linear_regression.coef_

Coefficient_correlation_multiple = linear_regression.score(X,Y);

"""-----"""
```

codesnap.dev

Coefficient_correlation_multiple	float64	1	0.8386673950901073
----------------------------------	---------	---	--------------------

Nous obtenons une valeur de 0,83, qui est proche de 1, indiquant ainsi une forte corrélation linéaire entre la variable endogène et les variables explicatives. Étant donné que nous avons une corrélation positive, cela signifie que lorsque les variables explicatives augmentent, la variable endogène a tendance à augmenter également. Ainsi, selon nos résultats, une augmentation du nombre de propositions d'admission, du nombre total d'admis boursiers ou du nombre d'admis provenant de la même académie est associée à une augmentation de la capacité d'une formation.

Il ne faut pas oublier que c'est une interprétation sur le coefficient de corrélation mais que ce n'est peut être pas une vérité générale

4) Conclusions

(a) Réponse à la problématique

En analysant les résultats de la régression linéaire multiple, nous allons pouvoir répondre à la problématique qui était : Est-ce que l'effectif total des candidats, l'effectif total des candidates, l'effectif total des propositions d'admission, le pourcentage d'admis

selon le nombre de candidats et l'effectif total des admis boursiers peuvent être utilisés pour prédire la capacité d'un établissement ?

La réponse est oui, c'est variable peuvent être utilisées pour prédire la capacité d'un établissement, du moins dans le contexte des données fournies dans le fichier de notre vue.

(b) Argumentation à partir des résultats de la régression linéaire

L'analyse de la régression linéaire multiples nous a permis d'obtenir les coefficients des variables explicatives. Les variables les plus influentes pour prédire la capacité d'un établissement sont :

1. L'effectif de total des admis qui sont boursier.
2. L'effectif total de proposition d'admission que la formation a faite.
3. L'effectif d'admis venant de la même académie de la formation

Si on prend en compte juste les résultats nous pouvons comprendre que le nombre d'admis boursiers et le nombre d'admis de la même académie et le nombre de propositions d'admissions effectuées par l'établissement ont un impact sur sa capacité. Cela peut être dû au fait que les admis boursiers et les admis de la même académie ont un rôle dans l'attractivité et la renommée de la formation. Mais concernant le nombre de proposition d'admissions ça dépend aussi de la politique d'admissions parcoursup de la formation mais si la formation fait x proposition d'admissions peut être qu'il x nombre de place.

(c) Interprétations personnelles

Au-delà des résultats de la régression linéaire, il faut aussi comprendre qu'il y a d'autres facteurs et limitations à prendre en compte pour interpréter les résultats. Il y a certainement d'autres variables qui auraient pu influencer positivement la capacité d'une formation.

En résumé, selon nos analyses, nous pouvons prédire la capacité d'une formation en fonction du nombre total d'admis boursiers et du nombre de propositions d'admission et du nombre d'admis venant de la même académie. Cependant, il est important de noter que la capacité d'une formation est définie par un processus logistique, ce qui signifie que l'inverse n'est pas forcément vrai. D'autres facteurs, tels que la réputation de l'établissement, la qualité des programmes, ou d'autres caractéristiques spécifiques aux formations, peuvent également jouer un rôle important dans la détermination de leur capacité.