Rémi Colas 1D1
Nathan Filaux Groupe d_12

PARTIE 3 BDD

BUT Informatique
SAE 2.04 - Livrable 3 partie
2022-2023

0 Les données Parcoursup1 - Problématique

(a) Présentation des données
Le fichier **parcoursup.csv** contient plusieurs séries statistiques sur l'ensemble de toutes les formations répertoriées dans Parcoursup :

- La population est l'ensemble des formations, représentées par leur code cod_aff et leur nom.
- La 1e série correspond au code de la formation
- La 2e série correspond à l'année de la session
- La 3e série correspond à la sélectivité
- La 4e série correspond à la capacité
- La 5e série correspond à l'effectif total des candidats
- La 6e série correspond à l'effectif total des candidates
- La 7e série correspond à l'effectif total de proposition d'admissions
- La 8e série correspond à l'effectif d'admis
- La 9e série correspond à l'effectif d'admises
- La 10e série correspond à l'effectif des admis boursiers étant neo bac
- La 11e série correspond à l'effectif admis dans le même établissement
- La 12e série correspond à l'effectif admises dans le même établissement
- La 13e série correspond à l'effectif admis dans la même académie
- La 14e série correspond à l'effectif admises dans le même académie

(b) Problématique
En utilisant ces données, on va essayer de répondre à la problématique suivante :
Parmi les données de notre fichier, certaines peuvent-elles permettre d'expliquer l'effectif d'admis dans les différentes formations ?

(c) Utilisation de la régression linéaire multiple : comment ?
En choisissant la 1e série statistique comme variable endogène et les autres séries comme variables explicatives, la régression linéaire multiple nous permettrait d'obtenir une estimation de l'effectif d'étudiants admis dans chaque formation en fonction des autres informations sur ces formations.

(d) Utilisation de la Régression linéaire multiple : pour quoi ?

Les paramètres de la régression linéaire multiple nous informeront des descripteurs qui influencent le plus le nombre de personnes admises. En observant si cette estimation est proche de la réalité, on aura une réponse à la problématique.

1 Import des données, mise en forme

(a) Importer les données en Python
On importe notre vue sous forme de **Data Frame** (“pd.read_csv(<fichier>)”) tout en enlevant les **valeurs nulles** (“.dropna()”) avec la commande suivante :

```
parcoursupDF = pd.read_csv("parcoursup.csv").dropna()
```

(b) Mise en forme
On transforme notre DataFrame en **Array** :

```
parcoursupArray = parcoursupDF.to_numpy()
```

(c) Centrer-réduire
On enlève les **3 premières** colonnes (à l’indice 0, 1, 2) de notre tableau, qui ne sont pas des données statistiques exploitables :

```
parcoursupArray = parcoursupArray[:,[i for i in range(len(parcoursupArray[0])) if i > 2]]
```

Puis on utilise la fonction “**Centreduire**” sur notre tableau :

```
def Centreduire(array):
    array = np.array(array,dtype = np.float64)
    res = [[0 for j in range(len(array[0]))] for i in range(len(array))]

    moy = np.average(array, axis = 0)
    eca = np.std(array, axis=0)

    for i in range(len(array)):
        for j in range(len(array[0])):
            res[i][j] = (array[i][j]- moy[j])/eca[j]
    return np.array(res)
```

```
parcoursupCR = Centreduire(parcoursupArray)
```

2 Choix des variables explicatives

(a) Démarche
Dans cette partie, on réduit le nombre de variables explicatives pour ne garder que les plus pertinentes. On commence par calculer la matrice de covariance :

```
MatriceCov = np.cov(parcoursupCR, rowvar=False)
```

(b) Matrice de covariance

On obtient la **matrice** suivante :

	0	1	2	3	4	5	6	7	8	9	10
0	1.00016	0.142352	0.144352	0.473618	0.592977	0.546498	0.222697	0.270119	-0.00253728	0.0148299	0.115781
1	0.142352	1.00016	0.89379	0.644005	0.585819	0.484731	0.425063	0.652338	-0.00897735	0.0640801	0.450767
2	0.144352	0.89379	1.00016	0.569377	0.515911	0.641683	0.43262	0.550399	-0.0464433	0.151167	0.364837
3	0.473618	0.644005	0.569377	1.00016	0.803485	0.643349	0.325785	0.75402	-0.0110473	0.0489808	0.549922
4	0.592977	0.585819	0.515911	0.803485	1.00016	0.792748	0.51094	0.877367	0.183748	0.191347	0.702471
5	0.546498	0.484731	0.641683	0.643349	0.792748	1.00016	0.473221	0.66264	0.0671584	0.334096	0.495382
6	0.222697	0.425063	0.43262	0.325785	0.51094	0.473221	1.00016	0.487966	0.236182	0.28753	0.516397
7	0.270119	0.652338	0.550399	0.75402	0.877367	0.66264	0.487966	1.00016	0.243047	0.224996	0.846592
8	-0.00253728	-0.00897735	-0.0464433	-0.0110473	0.183748	0.0671584	0.236182	0.243047	1.00016	0.692884	0.387033
9	0.0148299	0.0640801	0.151167	0.0489808	0.191347	0.334096	0.28753	0.224996	0.692884	1.00016	0.333553
10	0.115781	0.450767	0.364837	0.549922	0.702471	0.495382	0.516397	0.846592	0.387033	0.333553	1.00016

(c) Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible l'effectif de candidats admis dans les formations, qui se trouve dans la **colonne 4** de parcoursupCR. La colonne 4 de MatriceCov donne les coefficients de corrélation de l'effectif d'admis avec chacune des autres variables/colonnes de **parcoursupCR**. On va choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec le nombre total de candidat admis.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 4 de MatriceCov sont : **0.8034 | 0.7927 | 0.8773**. Ils correspondent aux variables numéros **3, 5, 7 .** Les colonnes 3, 5 et 7 de parcoursupCR correspondent aux :

- **effectif_total_proposition_admission**
- **effectif_total_admises**
- **effectif_total_admis_neo_bac**

On choisit donc ces 3 variables comme variables explicatives.

3 Régression linéaire multiple pour parcoursup.csv

(a) Regression lineaire multiple

On fait maintenant la régression linéaire multiple avec la série du nombre de candidats admis comme variable endogène, et les 3 variables explicatives trouvées ci-dessus.

```
Y = parcoursupCR[:,4]
X =  parcoursupCR[:,[3,5,7]]

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

Y contient notre variable endogène
X contient les 3 variables explicatives

(b) Paramètres, interprétation

Calculons maintenant les coefficient de corrélation linéaire (paramètres)

Code :

```
a=linear_regression.coef_
```

Résultat :

```
[0.22249161 0.31983428 0.49758898]
```

Interprétation :

- On peut voir que le nombre de candidats qui demandent à être admis influent légèrement sur le nombre total d'admis (≈0.222).
- On voit également que le nombre de candidates admises influent sur le nombre total d'admis (≈0.32)
-Et finalement que le nombre de candidats admis après avoir obtenu le bac influent fortement sur le nombre total d'admis (≈0.49).

(c) Coefficient de corrélation multiple, interprétation

On calcule le coefficient de corrélation multiple :

```
coefM = linear_regression.score(X, Y)
```

Résultat :

```
0.868745912728556 5
```

Interprétation:

On voit que la qualité de notre régression linéaire est très forte cela peut s'expliquer par le fait qu'on a utilisé les variables les plus corrélées de la matrice.

4 Conclusions

(a) Réponse à la problématique

Alors parmi les données de notre fichier, certaines peuvent-elles permettre d'expliquer l'effectif d'admis dans les différentes formations ?
En effet, plusieurs séries permettent d'expliquer l'effectif d'admis que ce soit la capacité ou d'autres effectifs comme le nombre de candidates, ce qui paraît logique. Ils ont donc une forte corrélation avec le nombre d'effectifs admis.

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

D’après les résultats de la régression linéaire on voit que les 3 séries que l’on a retenues influent positivement (car les coefficients sont positifs) sur l’effectif d’admis, ce qui veut dire que plus il y a d’admises, de néo bac admis et de nombre de demandes plus l’effectif sera grand.

(c) Interprétations personnelles

Pour conclure il fait sens que plus il y a d’effectif admis de manière générale (néo bac et filles), plus l’effectif total sera élevé car il comprend ces séries statistiques.de plus le nombre de demande laisse penser que la formation est populaire et donc qu’elle peut accueillir plus de personnes. On voit d’ailleurs, lors de la corrélation que la capacité de la formation (1ère colonne) à un fort taux de corrélation.