BUT Informatique
2022-2023

# 0 Les données de Parcoursup - Problématique

## (a) Présentation des données

Le fichier Parcoursup_vue.csv contient plusieurs séries statistiques sur l'ensemble des formations répertoriées dans Parcoursup, nous avons créé une vue en choisissant les paramètres suivants :

- La population est l'ensemble des formations, représentées par leur code cod_aff et leur nom
- La 1ère série correspond à la capacité de l'établissement proposant la formation
- La 2e série statistique est l'effectif total des candidats hommes pour une formation
- La 3e série statistique est l'effectif total des candidates pour une formation
- La 4e série statistique est l'effectif total des admis.e.s pour une formation
- La 5e série statistique est l'effectif total des admises pour une formation
- La 6e série statistique est l'effectif total des admis.e.s qui sont boursiers pour une formation
- La 7e série statistique est l'effectif total des admis.e.s qui sont de la même académie que celle de la formation.
- La 8e série statistique est l'effectif total des admis hommes pour une formation.

Toutes ces séries statistiques sont pour chaque formation sélective.

## (b) Problématique

En utilisant ces données, nous allons essayer de répondre à la problématique suivante: Parmi les données de notre fichier, quelles données hors résultats scolaires sont susceptibles d'influencer l'admission dans une formation sélective ?

### (c) Utilisation de la régression linéaire multiple : comment

En choisissant l'effectif total des personnes admises en variable endogène et certaines des autres séries comme variables explicatives, la régression linéaire nous permettra d'obtenir une estimation de quels critères sont susceptibles d'influencer l'admission dans une formation sélective.

### (d) Utilisation de la régression linéaire multiple : pourquoi

Les paramètres de la régression linéaire multiple nous informeront des descripteurs qui influencent le plus les chances d'être admis. En observant si cette estimation est proche de la réalité, on aura une réponse à la problématique.

## 1 Import des données et mise en forme

### (a) importer des données en python

On importe notre vue sous forme de DataFrame avec la commande suivante:

```
8    AdmissionsDF=pd.read_csv("df_admissions1.csv", sep=";")
```

### (b) Mise en forme

On a besoin de supprimer les cases vides (qui contiennent  nan en python), puis on transforme notre DataFrame en Array

```
AdmissionsDF=AdmissionsDF.dropna()
AdmissionsArray=AdmissionsDF.to_numpy()
```

### (c) Centrer réduire

On centre et réduit les données de notre Array.

```
def centrereduire(dataf):
    res=np.zeros(dataf.shape)
    moy=np.average(a=dataf,axis=0)
    std=np.std(a=dataf,axis=0)
    for i in range(len(dataf)):
        for j in range(dataf.shape[1]):
            res[i,j]=(dataf[i,j]-moy[j])/std[j]

    return res
```

# 2 Choix des variables explicatives

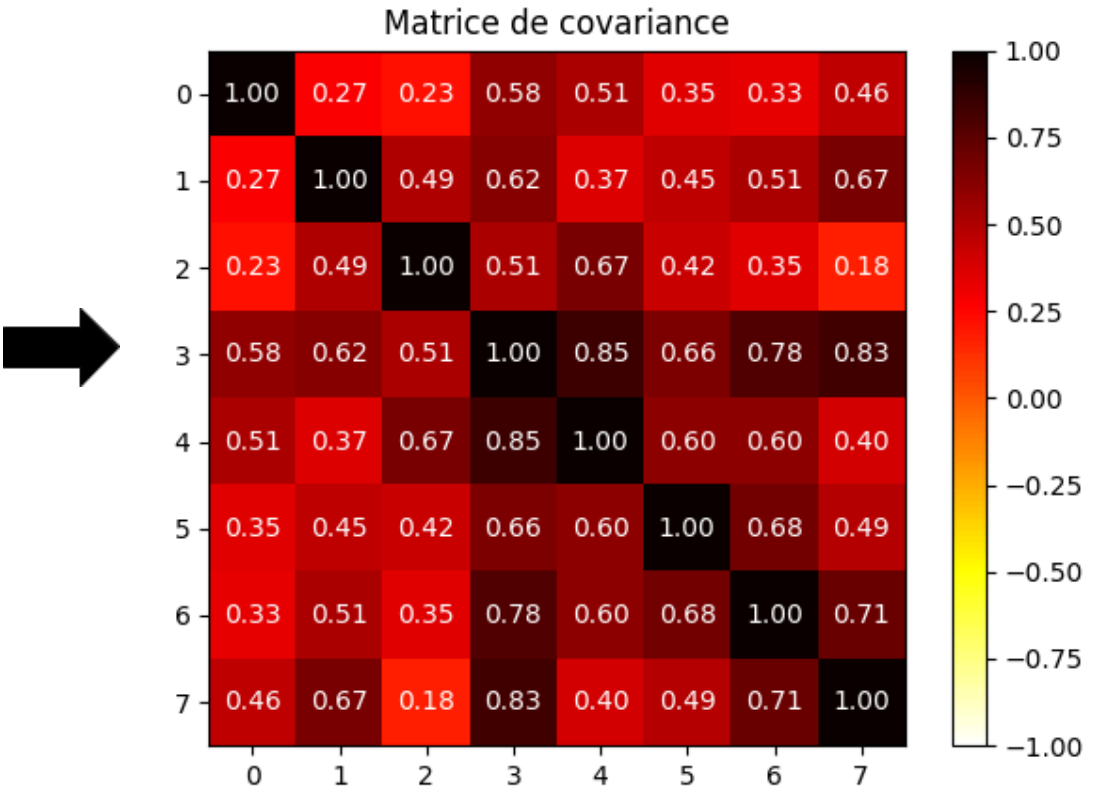
## (a) démarche

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(AdmissionsArray_CR,rowvar=False)
```

## (b) matrice de covariance

On obtient la matrice suivante:




Ici, la variable endogène est la colonne(et ligne) 3 qui correspond à l'effectif total des admissions en fonction des autres variables.

### (c) variable explicative les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible les chances d'être admis dans une formation sélective.

La colonne 3 de la matriceCOV donne les coefficients de corrélation entre l'effectif total des admis et les autres variables. On va choisir comme variables explicatives celles qui représentent le mieux les critères extra scolaires comme le sexe, le fait d'être boursier, le fait d'être dans la même académie etc…
On pourra ensuite voir que les critères extra scolaires peuvent influencer le fait d'être admis dans une formation sélective.

On utilisera donc les variables:

- La 5e série statistique est l'effectif total des admises pour une formation, on la choisit pour voir si le fait d'être une fille peut influencer le fait d'être pris dans une formation sélective.

- La 7e série statistique est l'effectif total des admis.e.s qui sont de la même académie que celle de la formation, on la choisit pour voir si le fait d'être dans une même académie que la formation désirée peut influencer le fait d'être pris dans une formation sélective.

- La 6e série statistique est l'effectif total des admis.e.s qui sont boursiers pour une formation, on la choisit pour voir si le fait d'avoir une bourse peut influencer le fait d'être pris dans une formation sélective.

On décide de ne pas prendre la 8e série statistique qui est l'effectif total des admis hommes pour une formation car on a déjà choisi la 4e série qui est l'effectif des femmes admises. Si l'on avait pris l'effectif homme et femmes admis, il y aurait 100% d'admis.e.s ce qui n'est pas un chiffre pertinent pour notre analyse. De plus, les coefficients de corrélation du total des candidates féminines est plus bas que celui des candidats masculins, tandis que les coefficients de corrélation des admises féminines est plus élevé que celui des admis masculins.

## 3 Régression linéaire multiple pour df_admission1.csv

### (a) Régression linéaire multiple

On fait maintenant la régression linéaire multiple avec l'effectif total admis en variable endogène, et les 3 variables explicatives trouvées ci -dessus.

### (b) Paramètres, interprétation

Paramètres de la régression linéaire multiple:

```
[0.58174871,0.42073984,0.01898349]
```

Effectif total des admises pour une formation: 0.58174871

Effectif total admis.e.s qui sont de la même académie que celle de la formation: 0.42073984
Effectif total des admis.e.s qui sont boursiers pour une formation: 0.01898349

Avant tout, on peut noter que l'interprétation des coefficients doit être faite en prenant en compte les autres facteurs hors résultats scolaires (dans notre vue) mais également les résultats scolaires qui ne sont pas ici dans notre vue, peuvent également jouer un rôle dans les chances d'admission et devraient être pris en considération.

Le coefficient de régression pour les **admises filles** est relativement élevé (0.58174871). Ce qui veut dire qu'un nombre plus élevé d'admises filles a une influence significative sur l'effectif total des admis (garçons et filles) pour une formation donnée.
On rappelle que les coefficients de corrélation du total des candidates féminines est plus bas que celui des candidats masculins, tandis que les coefficients de corrélation des admises féminines est plus élevé que celui des admis masculins.
On peut supposer que les candidates féminines ont de meilleures chances d'être admises par rapport aux autres groupes, ce qui peut être dû à leurs résultats scolaires, à des politiques d'égalité des genres ou à des objectifs de diversité.

Le coefficient de régression pour les **admis de la même académie que la formation** est positif (0.42073984). Ce qui veut dire qu'un nombre plus élevé d'admis (garçons et filles) provenant de la même académie que celle de la formation a une influence positive sur l'effectif total des admis (garçons et filles) pour cette formation. On peut supposer qu'il existe un avantage pour les candidats provenant de la même académie, ce qui peut être dû à une priorité aux étudiants étant du même secteur (géographique) que celui de la formation.

Le coefficient de régression pour les **boursiers** est relativement faible (0.01898349). Ce qui veut dire que l'effectif total des admis (garçons et filles) qui sont boursiers a une influence relativement faible sur l'effectif total des admis (garçons et filles) pour une formation donnée.

## (c) Coefficient de corrélation multiple

Coefficient de corrélation multiple: 0.8333949303662946

Le coefficient de corrélation multiple indique que l'effectif total des filles admises, l'effectif total des admis (garçons et filles) qui sont boursiers, et l'effectif total des admis (garçons et filles) de la même académie pour une formation expliquent 83% de la proportion de la variance totale de l'effectif total des admis pour une formation.

Ce qui peut indiquer que les variables telles que l'effectif total des admises, l'effectif total des admis qui sont boursiers, et l'effectif total des admis de la même académie sont des facteurs importants dans la détermination des chances d'admission à une formation.

Le coefficient de corrélation multiple ne prend en compte ni les autres facteurs hors résultats scolaires (dans notre vue) ni les résultats scolaires qui ne sont pas ici dans notre vue. Or, ils peuvent également jouer un rôle dans les chances d'admission et devraient être pris en considération.

# 4 Conclusion

## (a) Réponse à la problématique

Rappel de la problématique:
Parmi les données de notre fichier, quelles données hors résultats scolaires sont susceptibles d'influencer l'admission dans une formation sélective ?

Pour répondre à cette problématique, nous avons regarder si le fait d'être une femme, le fait d'être boursier ou encore le fait d'etre dans la même académie que la formation peuvent influencer le fait d'être pris dans cette dite formation.

Nous avons remarqué que oui, certain de ces critères comme le fait d'etre une femme ou le fait d'etre dans la meme academie influencent le fait d'être pris dans une formation sélective. En revanche, le fait d'être boursier n'influence pas le fait d'être pris dans une formation sélective.

## (b) argumentation à partir des résultats de la régressions linéaire

La régression linéaire pour la variable d'effectif total des admises est de 0.58174871 ce qui est plutôt élevé ce qui veut dire que cette variable influence honorablement l'effectif total des admis.e.s.

La régression linéaire pour la variable de l'effectif des admis dans une même académie est de 0.42073984 ce qui est certes moins élevé que celle des admises filles, mais reste élevé tout de même. Il influence aussi l'effectif total des admis.e.s.

En revanche, la régression linéaire des admis boursiers est de 0.01898349 ce qui est faible. Il a donc une influence négligeable sur l'effectif total des admis.

Grâce au coefficient de corrélation multiple qui est de 0.8333949303662946 ce qui montre que 83% des chances d'être pris.e.s dans une formation sont expliquées par nos trois variables explicatives.

## (c) Interprétations personnelles

L'argumentation à partir des résultats de la régression linéaire ne prend en compte ni les autres facteurs hors résultats scolaires (dans notre vue) ni les résultats scolaires qui ne sont pas ici dans notre vue. Or, ils peuvent également jouer un rôle majeur dans les chances d'admission et devraient être pris en considération.
Il est possible que les candidates féminines pourraient avoir de meilleures chances d'être admises en raison de leurs résultats scolaires, de politiques d'égalité des genres ou à d'objectifs de diversité.
Aussi, il est possible que les étudiants de la même académie que la formation souhaitée aient une priorité aux étudiants étant du même secteur (géographique) que celui de la formation. Mais on peut prendre le problème dans l'autre sens, peut-être que les étudiants postulent plutôt dans des formations proches.