0 Les données Parcoursup1 - Problématique

(a) Présentation des données

Le fichier Parcoursup1.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 pourcentage de filles parmi les personnes admises dans la formation,
- La 2e correspond au pourcentage de filles parmi les personnes candidates à la formation,
- La 3e et la 4e série correspond au pourcentage de bacs technos ayant été respectivement classés (3e série) et admis (4e série) dans la formation
- La 5e série correspond au nombre d'étudiant.e.s pouvant être accueillis dans la formation,
- La 6e série au nombre d'étudiant.e.s ayant candidaté pour cette formation,
- La 7e série au nombre d'étudiant.e.s admis.e.s dans la formation,
- Les séries suivantes correspondent aux pourcentages d'admis pour chaque mention au bac.

(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 la proportion d'étudiantEs admises dans les différentes formations ?

(c) Utilisation de la Régression linéaire multiple : comment ?

En choisissant la 1e séries statistique comme **variable endogène** et certaines des autres séries comme **variables explicatives**, la **régression linéaire multiple** nous permettrait d'obtenir une estimation de la proportion d'étudiantEs dans chaque formation en fonction d'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 montant des dégâts. 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 DataFrame avec la commande suivante :

```
Parcoursup1DF=pd.read_csv("Parcoursup1.csv")
```

(b) Mise en forme

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

```
Parcoursup1DF = Parcoursup1DF.dropna()

Parcoursup1Array=Parcoursup1DF.to_numpy()
```

(c) Centrer-réduire

On enlève les 2 premières colonnes de notre tableau, qui contiennent les noms et codes des formations, et ne sont donc pas des données statistiques :

```
def Centreduire(T):
    T=np.array(T,dtype=np.float64)
    TMoy=np.mean(T,axis=0)
    TEcart=np.std(T,axis=0)
    (n,p)=T.shape
    res=np.zeros((n,p))
    for j in range(p):
        res[:,j]=(T[:,j]-TMoy[j])/TEcart[j]
    return res

Parcoursup1ArrayCR=Centreduire(Parcoursup1Array[:,2:])
```

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(Parcoursup1ArrayCR,rowvar=False)
```

(b) Matrice de covariance

On obtient la matrice suivante :

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	0.917	0.0457	0.0225	-0.095	-0.077	0.0888	0.175	0.113	0.0584	-0.206	-0.111	0.0155	0.07	0.0465
1	0.917	1	0.0525	0.0335	-0.106	-0.0862	0.117	0.173	0.14	0.0759	-0.191	-0.13	-0.0211	0.0342	0.0277
2	0.0457	0.0525	1	0.276	-0.0446	-0.0415	0.0182	0.0256	0.0174	0.201	-0.0535	-0.0293	0.00737	0.0104	0.0144
3	0.0225	0.0335	0.276	1	-0.0414	-0.0474	0.0151	-0.0137	0.0113	0.27	-0.0448	-0.0197	-0.0085	-0.00622	-0.00241
4	-0.095	-0.106	-0.0446	-0.0414	1	0.739	-0.0366	-0.0791	-0.0551	-0.0813	0.119	0.0771	-0.0464	-0.0896	-0.0543
5	-0.077	-0.0862	-0.0415	-0.0474	0.739	1	-0.0352	-0.0516	-0.0401	-0.06	0.109	0.085	-0.0353	-0.0716	-0.0447
6	0.0888	0.117	0.0182	0.0151	-0.0366	-0.0352	1	0.42	0.797	0.119	-0.057	-0.0308	0.0066	0.0148	0.0203
7	0.175	0.173	0.0256	-0.0137	-0.0791	-0.0516	0.42	1	0.581	0.0902	-0.239	-0.0329	0.182	0.247	0.172
8	0.113	0.14	0.0174	0.0113	-0.0551	-0.0401	0.797	0.581	1	0.157	-0.0847	-0.00227	0.0526	0.0567	0.0466
9	0.0584	0.0759	0.201	0.27	-0.0813	-0.06	0.119	0.0902	0.157	1	-0.0443	-0.0199	0.0157	0.00879	-0.00529
10	-0.206	-0.191	-0.0535	-0.0448	0.119	0.109	-0.057	-0.239	-0.0847	-0.0443	1	0.00205	-0.509	-0.467	-0.231
11	-0.111	-0.13	-0.0293	-0.0197	0.0771	0.085	-0.0308	-0.0329	-0.00227	-0.0199	0.00205	1	-0.0379	-0.332	-0.234
12	0.0155	-0.0211	0.00737	-0.0085	-0.0464	-0.0353	0.0066	0.182	0.0526	0.0157	-0.509	-0.0379	1	0.359	0.0238
13	0.07	0.0342	0.0104	-0.00622	-0.0896	-0.0716	0.0148	0.247	0.0567	0.00879	-0.467	-0.332	0.359	1	0.507
14	0.0465	0.0277	0.0144	-0.00241	-0.0543	-0.0447	0.0203	0.172	0.0466	-0.00529	-0.231	-0.234	0.0238	0.507	1

(c) Variables explicatives les plus pertinentes

Notre objectif est de trouver des variables qui expliquent le mieux possible le pourcentage de filles admises dans les formations, qui se trouve dans la colonne 0 de Parcoursup1Array. La colonne 0 de MatriceCov donne les coefficients de corrélation du pourcentage de filles admises avec chacune des autres variables/colonnes de Parcoursup1Array. On va choisir comme variables explicatives celles qui ont le coefficient de corrélation le plus grand (en valeur absolue) avec le pourcentage de filles admises.

Les coefficients de corrélation les plus grands en valeur absolue dans la colonne 0 de MatriceCov sont : 0.917, 0.175 et -0.206. Ils correspondent aux variables numéro 1, 7 et 10. Les colonnes 1, 7, et 10 de Parcoursup1Array correspondent aux :

- pourcentages de filles candidates
- effectif total de candidat.e.s
- pourcentage d'admis sans mention au bac

On choisit donc ces 3 variables comme variables explicatives.

3 Régression linéaire multiple pour Parcoursup1.csv

(a) Régression linéaire multiple

On fait maintenant la régression linéaire multiple avec la série des pourcentages de filles admises comme **variable endogène**, et les 3 variables **variables explicatives** trouvées ci-dessus.

...

(b) Paramètres, interprétation

...

(c) Coefficient de corrélation multiple, interprétation

...

4 Conclusions

(a) Réponse à la problématique

...

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

...

(c) Interprétations personnelles

...