SES722 économétrie TP Python 1

Cours de Patrick Waelbroeck
Telecom Paris

Modules de base

On utilise les modules numpy et pandas

```
import numpy as np
import pandas as pd
```

Exercice 1 : Importer les données

Les données sont dans l'archive `textfiles`
Chaque jeu de données contient le fichier de données `.raw` et et un fichier descriptif `.des` .

Pour importer les données:

```
df = pd.read_csv('wage1.raw', delim_whitespace=True,
                 header=None)
```

On peut passer comme option le nom de chaque colonne

```
names=['nom de la colonne 1', 'nom de la colonne 2',
...]
```

Exercice 2 : Histogramme du salaire

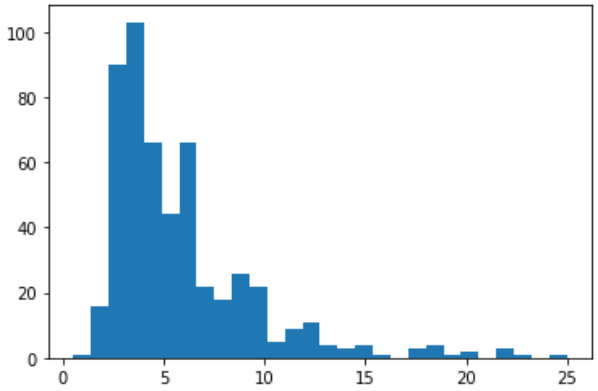
Le salaire est dans la première colonne

`wage=df[0]`

Deux possibilités en utisant le module `matplotlib`

```
import matplotlib.pyplot as plt
```

1. `plt.hist(wage,'auto')`
2. `wage.hist(bins=20)`



Exercice 3 : Statistiques descriptives du salaire

Pour calculer la moyenne, l'écart-type, le maximum, … du salaire, on utiliser les commandes `mean` , `std` , `max` , … de `numpy`

```
np.mean(wage)
np.std(wage)
np.max(wage)
```

Exercice 4 : Corrélation entre salaire et éducation

L'éducation est dans la colonne 2

```
educ=df[1]
```

Les commandes `cov` et `corrcoef` donne respectivement la covariance et la corrélation entre deux variables

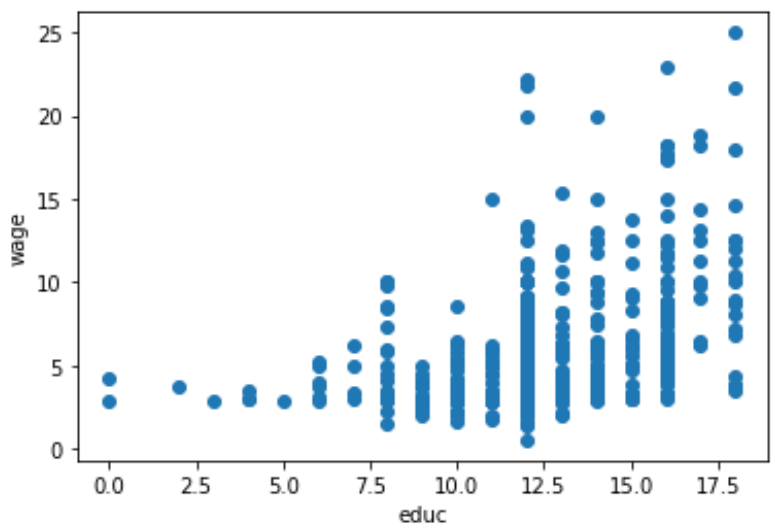
```
np.cov(wage,educ)
np.corrcoef(wage,educ)
```

Remarque : `corr()` est également un attribut de `df`

Exercice 5 : Faire un graphique en nuage de points entre salaire et educ

Utiliser la commande `scatter` de `plt`

```
plt.scatter(educ,wage)
plt.xlabel("educ")
plt.ylabel("wage")
plt.show()
```



Exercice 6 : Calculer le salaire moyen des femmes

La variable indicatrice des femmes est dans la colonne 6. On sélectionne ensuite les lignes correspondant aux femmes.

```
femme=df[5]
np.sum(femme)
s=femme==1
np.mean(wage[s])
```

Remarque : on peut également utiliser les fonctions `loc` et `iloc` de `pandas`

Exercice 7 : Calculer la différence moyenne de salaire entre les hommes et les femmes

Sélectionner les lignes correspondant aux hommes et appliquer la méthode précédente.

Exercice 8 : Supprimer les observations avec wage > 10

On convertit le data frame en array pour faire les opérations sur les lignes et les colonnes

```
s=wage<=10
df1=np.array(df)
df2=df1[s,:]
df2.shape
```

Remarque : il existe des modules qui permettent d'ajouter, de supprimer des ligne; on peut également utiliser les fonctions de `pandas` .

Exercice 9 : supprimer les 15 premières et les 15 dernières lignes

La fonction shape retourne le nombre de lignes et de colonnes

```
n=np.shape(df1)[0]
```

ou

```
n,k=np.shape(df1)
```

Ensuite, sélectionner les lignes à conserver

```
df3=df1[15:n-15,:]
df3.shape
```