Econometrics TP1

Data exercises

Patrick Waelbroeck

Telecom Paris

September 10, 2021

Getting the datasets

All files are in the archive `textfiles.zip` available on the lecture website. Each dataset is associated with 2 files.
A file with extension `.raw` contains the raw dataset. A file with extension `.des` containes the name of the variable (in column). Decompress files in your working directory. We use the standard libraries: numpy and pandas.

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

Exercise 1

Import data from `wage1.raw`.

There are several ways to do that. I recommend importing data within a panda frame.

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

It is possible to title the column using the option `names=['column title 1', 'column title 2', ...]`

Histograms

Exercice 2

Plot the histogram of the wage.

You can use the library `matplotlib`.

```
import matplotlib.pyplot as plt
wage=df[0]
plt.hist(wage,'auto')
```

Another way to plot the histogram is to use the built-in function in Pandas

```
wage.hist(bins=20)
```

Exercice 3

Compute the mean, standard deviation and maxium and minimum of the variable wage using the commands `mean`, `std`, `max`, `min` of Numpy.

Covariance and correlation

Exercice 4

Compute the covariance and the correlation between wage and eduction using the commands `cov` and `corrcoef` of Numpy

```
educ=df[1]
np.cov(wage,educ)
np.corrcoef(wage,educ)
```

Scatterplots

Exercice 5

Show a scatter plot between `wage` and `educ` using the command `scatter` form `plt`.

Note: this is a raw correlation that does not take into account other factors.

Computing averages on sub-samples

Exercice 6

Compute the average wage of men and women. Compute the average difference. Is there a gender bias?

The most universal way to select observations is to create a marker using a logical condition and then select the rows corresponding to the sub-sample.

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

Remark: You can also use the Pandas command: `loc` and `iloc`.

Computing averages on sub-samples (cont.)

Exercice 7

Compute the average wage of women who have a wage higher than the median wage using the command `np.median`.

Percentiles

Exercice 8

Compute the 5th percentile of `wage`.

There are 2 possilibities. First, you can use the built-in command `np.percentile`. Or, you can sort the pandas frame using the command `pd.sort_values(by['wage'], inplace=True)` and extract the correct observation. For instance, if there are 100 observations, the 5th percentile is the 5th observation of the sorted dataset.

Plotting averages

Exercice 9

Plot the mean of wage for each tenure year

Program a loop that computes at each iteration the average wage. Plot the results.

Removing observations

Exercice 10

Remove observations for which `wage>10`.

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

Adding observations

Exercise 11

Add observations for which `wage>10` at the end of the dataset of exercise 10.