Assignment 3

May 8, 2016

1 Week 3 - The Pearson correlation coefficient (r)

1.1 Research question: Is there association between alcohol consumption and emloyment rate?

Here we are interested in relationship between alcohol consumption per capita (in liters) and employment rate (expressed as %). We want to identify if there is an relationship and , if it exists,whether it is negative (higher alcohol consumption means lower employment rate) or positive (higher alcohol consumption means lower employment rate). As both variables are quantitative, we will analyze the potential relationship by calculating the Pearson coefficient.

1.2 Data

We use the data provided in the course material - “gapminder” data. The data were downloaded from the following link: Gapminder data files If you want to reproduce same output as here, please save this jupyter notebook and the above .csv file in the same folder.

1.3 Code

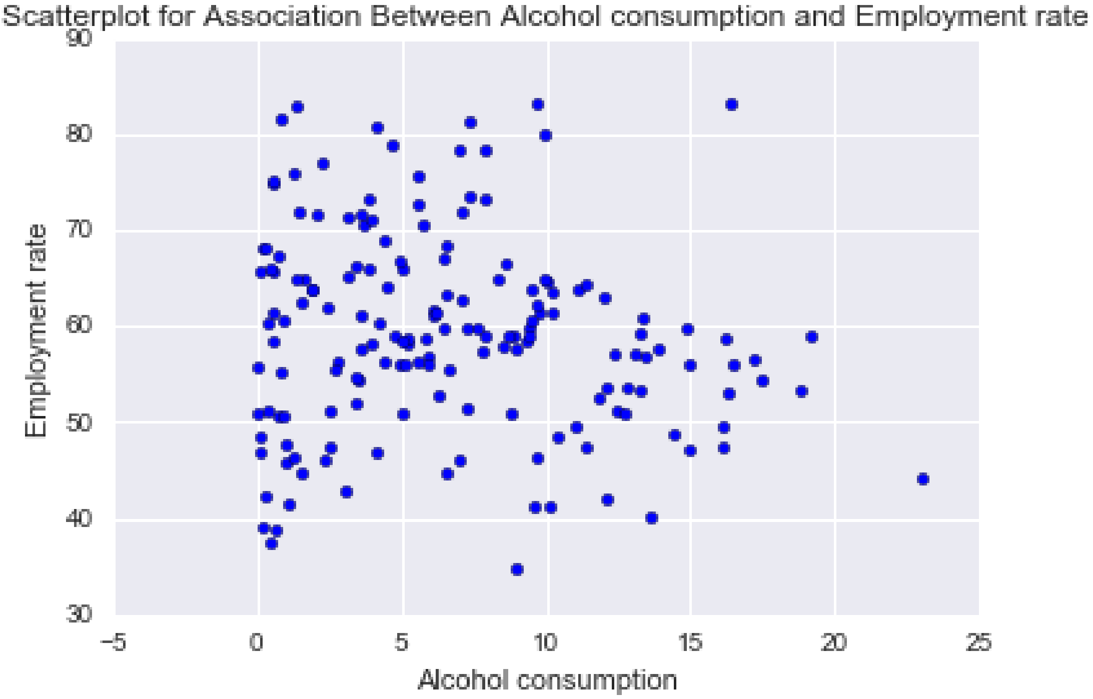
In the following lines I will include my python code (generated via jupyter notebook). The code will be splitted into blocks / cells , where each block is one step of a code. The step description will be added as comment within the code. Here is the code:

```
In [22]: %matplotlib inline
         # importing libraries
         import pandas
         import numpy
         import seaborn
         import scipy
         import matplotlib.pyplot as plt
```

```
In [23]: # Reading data set - using the data set from the course
         data = pandas.read_csv(’gapminder.csv’, low_memory=False)
```

```
In [25]: #setting relevant variables to numeric
         data[’alcconsumption’] = pandas.to_numeric(data[’alcconsumption’],errors=’raise’)
         data[’employrate’] = pandas.to_numeric(data[’employrate’],errors=’raise’)
         # setting NAs values
         data[’alcconsumption’]=data[’alcconsumption’].replace(’ ’, numpy.nan)
         data[’employrate’] = data[’employrate’].replace(’ ’, numpy.nan)
         #cleaning data -removing of NAs
         data_clean=data.dropna()
```

```
In [26]: # Scatter plot for the variables of interest
         plt.figure(1)
         plt.scatter(data_clean['alcconsumption'], data_clean['employrate'])
         plt.xlabel('Alcohol consumption')
         plt.ylabel('Employment rate')
         plt.title('Scatterplot for Association Between Alcohol consumption and Employment rate')
         plt.show()
```



The scatterplot above seems to show no (LINEAR) relationship between these two variables. Let us verify it with the Pearson coefficient. Here is the code calculating the coefficient:

```
In [28]: # Pearson coefficient for the two variables and its p-value
         print ('Association between alcconsumption and employrate - \
                The Pearson coefficient (r) and its p-value')
         print (scipy.stats.pearsonr(data_clean['alcconsumption'], data_clean['employrate']))
```

```
Association between alcconsumption and employrate - The Pearson coefficient (r) and its p-value
(-0.13383831023376283, 0.085596367946780993)
```

1.4 Interpretation & Summary

The Pearson coefficient (r) is ca. -0.13, so very close to zero, what would mean that, **if there is a linear relationship** then it is a weak relation - or no association. However, the p-value is 0.085.. > 0.05 , meaning that the correlation coefficient is **not significant**. One of the reason for high p-value could be low sample size.

1.5 Conclusion

There is no statistically significant linear relationship between alcohol consumption and employment rate, as the p-value is higher than 0.05.

1.6 Note for Coeficient of Determination or Rˆ2

Since the Pearson coefficient is not significant, it makes noi sense to calculate the coefficient of determination as we can not interpret it correctly.