Correlation Workshop

September 13, 2020

CSV files

The **CSV** file format is very common for ‘medium data’ i.e. structured data consisting of approximately dozens of columns and approximately thousands of rows. The CSV file format is very simple

```
data11, data12, data13, ..., data1M
data21, data22, data23, ..., data2M
...
dataN1, dataN2, dataN3, ..., dataNM
```

CSV stands for ‘column separated values’. You many also encounter TSV, which stands for ‘tab separated values’. The python library `csv` is useful for reading in CSV files.

```
import csv
with open("corr.csv", 'r') as infile:
  csvreader = csv.reader(infile, delimiter=',')
  for row in csvreader:
    ...
```

Note the delimiter argument here is a comma, since we are working with rows ‘delimited’, that is separated, by a comma.

Exercise 1: The file `corr.csv` contains three columns of data: population, population density and total hours worked per week. Compute the correlation co-efficient of the every column against every other column (use scipy). One of these correlations is high, why is that?

Exercise 2: The two data sets below were recorded each year between 2000 and 2009.

```
X = [29.8,30.1,30.5,
30.6,31.3,31.7,32.6,
33.1,32.7,32.8 ]
Y = [327,456,509,
497,596,573,661,
741,809,717]
```

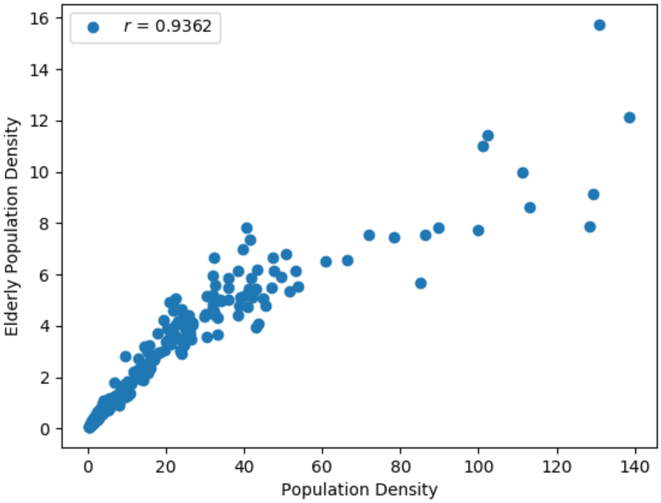
Plot X against Y and compute the correlation of X and Y.

X is the US per capita consumption of cheese per year and Y is the number of people who died by becoming tangled in their bedsheets in the same year.

Exercise 3: Have a look at this website `http://tylervigen.com/old-version.html` and use it to compute some correlations. What do you think are the reasons behind these spurious correlations?

Exercise 4: Use Google Trends to find search terms which are (positively or negatively) correlated and speculate about the cause of the high correlation.

Exercise 5: The file `old.csv` is aggregated population data. The first column contains population density i.e. number of people per hectare and the second is number of elderly residents per hectare. Calculate the correlation coefficient between the two series. Make a scatter plot of the data. Comment on the plot and the value of the correlation coefficient.



Non-linear data

Often data which is related, but not linearly related, can be **transformed** to be linearly related. A very common situation is when Y is a power of X.

$$y = Ax^b \tag{1}$$

Exercise 6: Plot $\log y$ against $\log x$. Compute the correlation coefficient.

Exercise 7: If the relationship in equation 1 holds, explain (by doing some simple algebra) why this transformation (taking logs) puts the data into a linear relationship.

Exercise 8 (Hard): Compute the correlation coefficient (again by doing some simple algebra) for data x_i and y_i where $y_i = mx_i + c$ for some constants m and c.