Data Science Homework: Linear Regression

Won Kim

2022

Exercise 1

- Using hand calculation, derive and interpret a covariance matrix for the following dataset.

Person	Age	Income	Yrs worked	Vacation
1	30	200	10	4
2	40	300	20	4
3	50	800	20	1
4	60	600	20	2
5	40	300	20	5

Using Numpy (1/3): Creating a Population Covariance Matrix

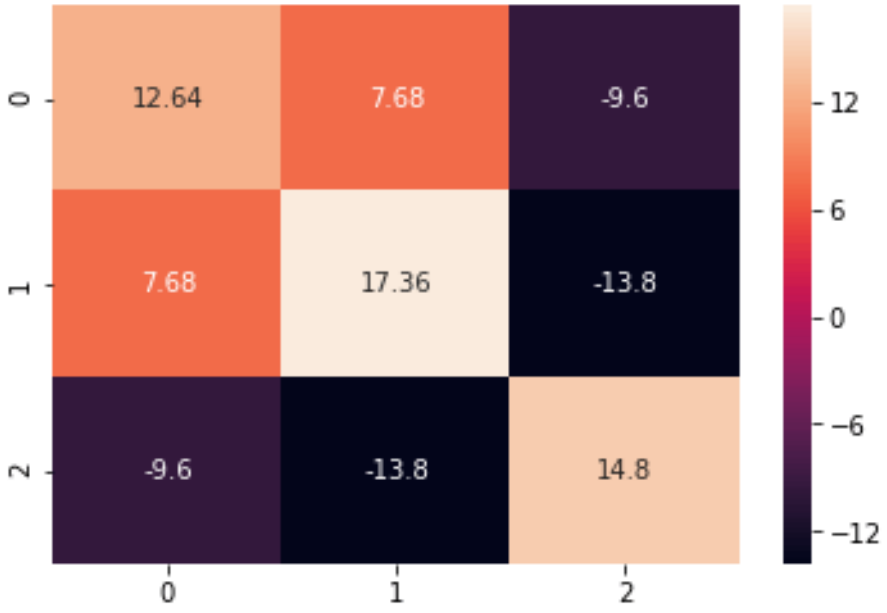
- https://datatofish.com/covariance-matrix-python/

```
import numpy as np
# input data
A = [45,37,42,35,39]
B = [38,31,26,28,33]
C = [10,15,17,21,12]

data = np.array([A,B,C])
# population covariance matrix (N)
covMatrix = np.cov(data,bias=True)
print (covMatrix)
```

```
[[ 12.64   7.68  -9.6 ]
 [  7.68  17.36 -13.8 ]
 [ -9.6  -13.8   14.8 ]]
```

Using Numpy and Seaborn (2/3): Visualizing a Covariance Matrix

```
import seaborn as sn
import matplotlib.pyplot as plt

sn.heatmap(covMatrix, annot=True, fmt='g')
plt.show()
```



Using Numpy (3/3): Creating a Sample Covariance Matrix

```
# sample covariance matrix (N-1)
covMatrix = np.cov(data,bias=False)
print (covMatrix)
```

```
[[ 15.8    9.6   -12.   ]
 [  9.6   21.7   -17.25]
 [-12.   -17.25   18.5 ]]
```

Using Pandas (1/2): Creating a Sample Covariance Matrix

```
import pandas as pd

data = {'A': [45,37,42,35,39],
        'B': [38,31,26,28,33],
        'C': [10,15,17,21,12]
        }

df = pd.DataFrame(data,columns=['A','B','C'])
# sample covariance matrix
covMatrix = pd.DataFrame.cov(df)
print (covMatrix)
```

Using Pandas and Seaborn (2/2): Visualizing a Covariance Matrix

```
import seaborn as sn
import matplotlib.pyplot as plt

sn.heatmap(covMatrix, annot=True, fmt='g')
plt.show()
```

Exercise 2

- As shown previously, using NumPy and Pandas (and Seaborn), create a covariance matrix and visualize it. For this exercise, use the dataset used for Exercise 1.
  - A population covariance matrix
  - A sample covariance matrix

Linear Regression Formula

$$\hat{y} = a + bX$$

a: intercept
b: slope

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

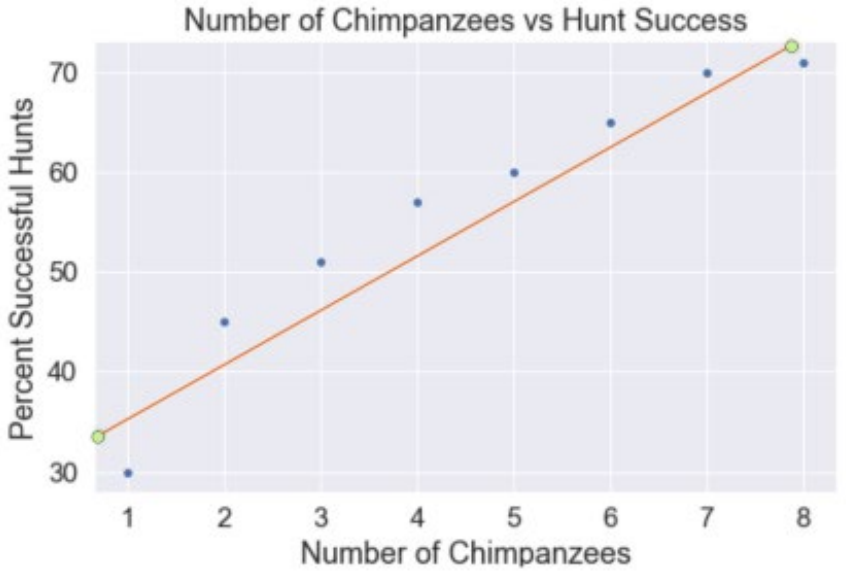
$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Walkthrough Example

- https://towardsdatascience.com/linear-regression-by-hand-ee7fe5a751bf
- Dataset: #of chimpanzees and hunting success

	Number of Chimpanzees	Percent Successful Hunts
0	1	30
1	2	45
2	3	51
3	4	57
4	5	60
5	6	65
6	7	70
7	8	71



First, Calculate All the Terms

Number of Chimpanzees (x)	Percent Successful Hunts (y)	xy	x^2	y^2
1	30	30	1	900
2	45	90	4	2025
3	51	153	9	2601
4	57	228	16	3249
5	60	300	25	3600
6	65	390	36	4225
7	70	490	49	4900
8	71	568	64	5041
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum y^2$
36	449	2249	204	26541

Next, Plug the Values into the Formulas

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b = \frac{\Sigma y - m(\Sigma x)}{n}$$

$$m = \frac{8(2249) - (36)(449)}{8(204) - (36)^2} \qquad b = \frac{449 - 5.4405(36)}{8}$$

$$m = 5.4405 \qquad b = 31.6429$$

$$y = mx + b$$

$$y = 5.4405x + 31.6429$$

Homework

- The following dataset is the amount a person spends on recreation and the person's income.
- 1. Using the following dataset, hand calculate the least squares regression line. Then predict the income of two new persons who spend 3500 and 5300.
- 2. Using scikit-learn and seaborn library, find the regression line and also draw the line and a scatter plot of the dataset.

spends	income
2400	41200
2650	50100
2350	52000
4950	66000
3100	44500
2500	37700
5106	73500
3100	37500
2900	56700
1750	35600

End of Homework