Data Mining Applications – Frederik Darwin

M10601836

Homework 1 – Linear Regression

Problem:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Revenue(m)(X)	7	2	6	4	14		16	12	14	20	15	7
Profit(m)(Y)	0.15	0.10	0.13	0.15	0.25	0.27	0.24	0.20	0.27	0.44	0.34	0.17

The following table shows the monthly revenues and the corresponding profits for a franchise company in 2017. Please write a computer program to find the linear regression model and predict the profit for January, 2018 if its revenue is 10 million dollars. There is a missing value in the data. Try to solve this problem yourself. Any kind of computer language is allowed for this homework.

Language Used: Python 3.6 in Jupyter Notebook

1. The missing value
   Can be easily replaced by calculating the total average of all revenue from that year, then the resulting average is used to fill the empty value.

```
In [1]:
1 X = [7, 2, 6, 4, 14, 16, 12, 14, 20, 15, 7]

In [2]:
1 mis_val = sum(X)/len(X)

In [3]:
1 X = [7, 2, 6, 4, 14, mis_val, 16, 12, 14, 20, 15, 7]
2 Y = [0.15, 0.10, 0.13, 0.15, 0.25, 0.27, 0.24, 0.20, 0.27, 0.44, 0.34, 0.17]
```

2. Regression Model

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

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

Will try to use these formulas in the code to find a and b for the $y = a + bx$ equation.

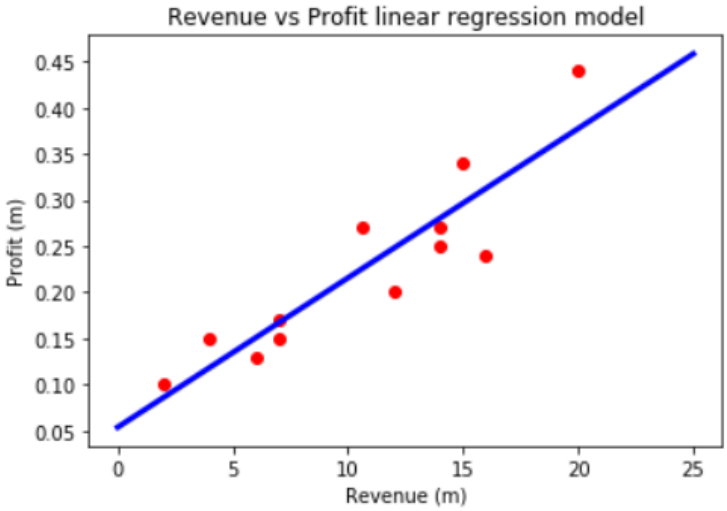
With these formulas, the *a* and *b* is found to have value of *a = 0.0536* and *b = 0.0162*

```
In [9]:  1  a

Out[9]: 0.05361426064655464

In [10]:  1  b

Out[10]: 0.01619153674832962
```



3. Prediction of January 2018
   With Revenue of 10 million, which mean x = 10, then the predicted value of the profit is around 0.21552963

```
In [7]:  1  y = a + (b*10)

In [8]:  1  y

Out[8]: 0.21552962812985085
```

4. Source Code

```
# coding: utf-8

# LINEAR REGRESSION

# In[1]:

X = [7, 2, 6, 4, 14, 16, 12, 14, 20, 15, 7]

# In[2]:
```

```
mis_val = sum(X)/len(X)

# In[3]:

X = [7, 2, 6, 4, 14, mis_val, 16, 12, 14, 20, 15, 7]
Y = [0.15, 0.10, 0.13, 0.15, 0.25, 0.27, 0.24, 0.20, 0.27, 0.44, 0.34, 0.17]

# In[4]:

sigma_x_2 = 0
sigma_y_2 = 0
sigma_xy = 0

# In[5]:

for i in range(len(X)):
    sigma_x_2 = sigma_x_2 + (X[i]**2)
    sigma_y_2 = sigma_y_2 + (Y[i]**2)
    sigma_xy = sigma_xy + (X[i]*Y[i])

# In[6]:

b = ((len(X)*(sigma_xy)) - (sum(X)*sum(Y))) / ((len(X)*(sigma_x_2)) - ((sum(X))**2))
mean_Y = sum(Y)/len(Y)
mean_X = sum(X)/len(Y)
a = mean_Y - (b*mean_X)

# In[7]:

y = a + (b*10)

# In[8]:

y

# In[9]:

a

# In[10]:

b
```