solution-test-excercise-1

August 5, 2023

[1]:
```
import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt
```

[2]:
```
data = pd.read_excel('DataSet.xls')
```

1 Simple Regression

[3]:
```
X = data.Age
Y = data.Expenditures
```

1.1 Question 1

1.1.1 Coefficients

[4]:
```
b = (X*Y - X*Y.mean()).sum()/(X*X - X*X.mean()).sum()
a = Y.mean() - b*X.mean()
a, b
```

[4]:
```
(114.24110795493151, -0.3335960966062749)
```

1.1.2 Standard Error

[5]:
```
error            = Y - (a + b*X)
Sum_Square_Error = ( error**2).sum()
n = data.shape[0]
stdev = math.sqrt(1/(n-2)*Sum_Square_Error)
# logging.info(f'standard error is {round(standard_error, 3)}')
print(f'standard error is {round(stdev, 3)}')
```

```
standard error is 5.073
```

[6]:
```
C = 1/(X-X.mean()).sum()
```

[7]:
```
s_b_2 = stdev ** 2 / ((X-X.mean())**2).sum()
print(f's_b square is {s_b_2}')
```

```
s_b square is 0.00909528102577286
```

1.1.3 t-value of b

[8]:
```
#t-test on H_0: beta is 0 based on t_b = b/s_b
beta = 0 #b - (C*error).sum()

t_b = (b-beta)/math.sqrt(s_b_2)
print(f't-value of b is {t_b}')
```

```
t-value of b is -3.497944376283516
```

1.1.4 Answer 1

[9]:
```
print('Answer of question 1:')
print( f'Value of intercept a is {round(a, 4)}')
print( f'Value of coefficient b is {round(b, 4)}')
print( f'Standard Error is {round(stdev, 4)}')
print( f't-value of b is {t_b}')
```

```
Answer of question 1:
Value of intercept a is 114.2411
Value of coefficient b is -0.3336
Standard Error is 5.0733
t-value of b is -3.497944376283516
```

1.1.5 Summarize solution 1 into function for following questions

[10]:
```
def calc_q1(df_data, group):
    X = df_data.Age
    Y = df_data.Expenditures

    b = (X*Y - X*Y.mean()).sum()/(X*X - X*X.mean()).sum()
    a = Y.mean() - b*X.mean()

    error             = Y - (a + b*X)
    Sum_Square_Error = ( error**2).sum()
    n = df_data.shape[0]
    stdev = math.sqrt(1/(n-2)*Sum_Square_Error)

    C = 1/(X-X.mean()).sum()

    s_b_2 = stdev ** 2 / ((X-X.mean())**2).sum()

    #t-test on H_0: beta is 0 based on t_b = b/s_b
    beta = 0 #b - (C*error).sum()

    t_b = (b-beta)/math.sqrt(s_b_2)
```

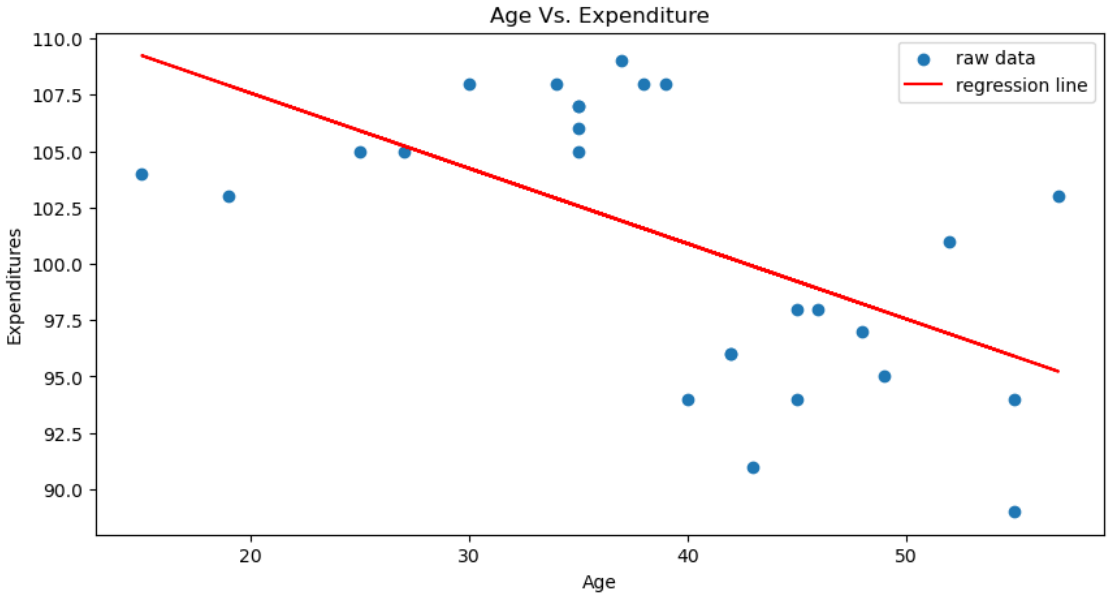
```
    print(f't-value of b is {t_b}')

    print( f'Present result for Age {group}')
    print( f'Value of intercept a is {round(a, 4)}')
    print( f'Value of coefficient b is {round(b, 4)}')
    print( f'Value of beta is {round(beta, 4)}')
    print( f'Standard Error is {round(stdev, 4)}')
    print( f't-value of b is {t_b}')

    return a, b, stdev, t_b
```

1.2 Question 2

[11]:
```
plt.figure(figsize=(10, 5))
plt.scatter(X, Y, label='raw data')
plt.plot(X, a+b*X, color='r', linestyle='solid', label='regression line')
plt.xlabel('Age')
plt.ylabel('Expenditures')
plt.title('Age Vs. Expenditure')
plt.legend()
plt.show()
```



1.2.1 Answer 2

1. Based on regresion line, expense decreases as age increases
2. Based on raw data points, these data can be separated into two clusters and each group can be modeled separately.

1.3 Question 3

[12]:
```
# Split group based on Age
df_g1 = data[data.Age >= 40]
df_g2 = data[data.Age < 40]
```

1.3.1 Answer 3

[13]:
```
a1, b1, stdev1, t_b1 = calc_q1(df_g1, group='>= 40')
```

```
t-value of b is 0.7420587155705188
Present result for Age >= 40
Value of intercept a is 88.8719
Value of coefficient b is 0.1465
Value of beta is 0
Standard Error is 3.8329
t-value of b is 0.7420587155705188
```

[14]:
```
a2, b2, stdev2, t_b2 = calc_q1(df_g2, group='< 40')
```

```
t-value of b is 4.4604533501562305
Present result for Age < 40
Value of intercept a is 100.2323
Value of coefficient b is 0.198
Value of beta is 0
Standard Error is 1.1531
t-value of b is 4.4604533501562305
```

1.4 Question 4 and Answer 4

1. in a) we can see the Age and Expenditure are negative related. However, the behaviors of two Age clusters are different when modeling them separetly.
2. However, in c), for Age <= 40, as t-value > 2, we can see a possitive relation between Age and Expenditure. For Age > 40, at t-value are within (-2, 2), the Age and Expenditure are not significantly related.
3. As assumption 6 is violated, we should consider Age above 40 and Age below 40 as separate observations when forming the strategies.