Assignment - 3

TITLE
Statistical Modeling

PROBLEM STATEMENT:
Load the dataset.

i) Test the association of mother's (lwt) age and birth weight using the correlation test and linear regression.

ii) Test the association of mother's weight (lwt) and birth weight using correlation test and linear regression.

iii) Produce two scaler plot of i) age of birth weight ii) mother's weight by birth weight. Elaborate the conclusion.

OBJECTIVES:

- To understand role of computation as a tool of discovery in data analysis.
- Compute and interpret correlation coefficient.
- Compute and interpret coeff. in linear regression analysis.

OUTCOMES:
Design and analyse real world engineering problems by applying various modeling techniques.

PREREQUISITE:
~~Concep~~ Concept of data distribution.

Date: / /20

THEORY:

A correlation or simple linear regression analysis can determine if two numeric variables are significantly linear related. For 2 related variables, it measures the association between the 2 variables. In contrast, linear regression is used for prediction of values of one variable to another.

1) Correlation

The correlation coeff. between 2 variables answers the question - If one variable changes, does other also change? . The correlation between two variables is a measures of linear relationship between them. The correlation between two random variable X and Y is measure of degree of linear association between 2 variables.

Two variables are highly correlated if they move well together. It is indicated by correlation coeff.

The population correlation coeff. is denoted by ρ. It can take on any value -1, through 0 to 1.

The possible values of ρ and their interpretation are given below:

- When ρ is equal to zero, there is no correlation.
- When $\rho = 1$ there is perfect, positive linear relationship between two variables. Whenever one variable, X or Y, increases, other one also increases and whenever it decreases, the other must also decrease.

- When $\rho = -1$, there is perfect negative relationship between X and Y. When X and Y increases, the other decreases and vice-versa.
- When value of ρ is between 0 to 1 in absolute value, it reflects relative strength of linear relationship between two variables. For eg. a correlation of 0.9 implies relatively strong relation while correlation of -0.7 is relatively weaker between X and Y.

In correlation analysis we will assume that both X and Y are distributed random variables with means μ_x & μ_y, S.D. as σ_x and σ_y respectively. We define covariance X and Y as follows:

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

The population correlation coeff. can take any value from -1 to +1.

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

Like all population parameter, value of ρ is not known to us. We need to estimate it from our random sample of (X,Y) observation pair. It turns out that sample estimater of $Cov(X,Y)$ is $SS_{xy}/(n-1)$ an estimater of σ_x is $\sqrt{SS_{xx}/(n-1)}$ and estimater

of σ_y is $\sqrt{SS_{yy}/(n-1)}$. Substituting these estimators, we get sample correlation coeff., denoted by r. The estimate of ρ also referred as Pearson product-moment correlation coeff.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

2) Linear Regression

The eq. of straight line is $Y = A + BX$, where A is intercept and B is slope of line. In simple regression, we model the relationship between 2 variables X and Y as a straight line. So, our model must contain two parameters, an intercept parameter and a slope parameter. The usual notation is β_0, and notation for slope is β_1. If we include error term ε, the population regression model is

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

The model parameters are:
β_0 is y intercept of straight lines
β_1 is slope of line Y.
The simple linear regression model applies only if relation between 2 variable X and Y is a straight-line relationship. If it is curved then it is curvilinear relationship.

CONCLUSION:

Hence correlation and line regression for the given dataset birthwt Risk Factors Associated with Low Infant Birth Wieght calculated and produced the scatter plot.

Combine your Documents and In | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH

Apps | HackerRank | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

+ Code + Text

Connecting

Load the dataset: birthwt Risk Factors Associated with Low Infant Birth Weight at
https://raw.github.com/neurospin/pystatsml/master/datasets/birthwt.csv

1. Test the association of mother's (bwt) age and birth weight using the correlation test and linear regeression.
2. Test the association of mother's weight (lwt) and birth weight using the correlation testand linear regeression.
3. Produce two scatter plot of: (i) age by birth weight; (ii) mother's weight by birth weight. Elaborate the Conclusion ?

```
[ ] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    from sklearn.linear_model import LinearRegression
```

```
[ ] df = pd.read_csv("/content/drive/MyDrive/data/birthwt.csv")
```

```
[ ] def calc_covariance(dataset1,dataset2):
        '''
     Def : Covariance measures the relationship trend
     between two sets of data.
     Formula : 1) Σ((X - X_mean)*(Y - Y_mean)) / n
        '''

        mean1 = np.mean(dataset1)
        mean2 = np.mean(dataset2)
```

TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain

ENG 13:34 13-01-2022

Combine your Documents and I | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH

Apps | HackerRenk | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

..

sample_data

+ Code + Text

Initializing Editing

```
[ ]     mean1 = np.mean(dataset1)
        mean2 = np.mean(dataset2)
        return np.sum(np.multiply(dataset1-mean1,dataset2-mean2))/len(dataset1)

    def correlation(dataset1,dataset2):
        '''
     Def : Covariance measures the relationship trend between two sets of data.
     Formula : 1) cov(x,y)/(std(x)*std(y))
     '''
        cov =calc_covariance(dataset1,dataset2)
        sd1 = np.std(dataset1)
        sd2 = np.std(dataset2)

        return cov/(sd1*sd2)
```

Test the association of mother's (bwt) age and birth weight using the correlation test and linear regeression.

1)Using correlation coefficients test :

```
[ ] # Age of mother
    age = df["age"]
    age = age.to_numpy()

    # Birth weight in grams
    birthwt = df["bwt"]
    birthwt = birthwt.to_numpy()
```

TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain ENG 13:34 13-01-2022

Combine your Documents and I | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH

Apps | HackerRenk | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment | Share

Files

..

sample_data

+ Code + Text

Initializing | Editing

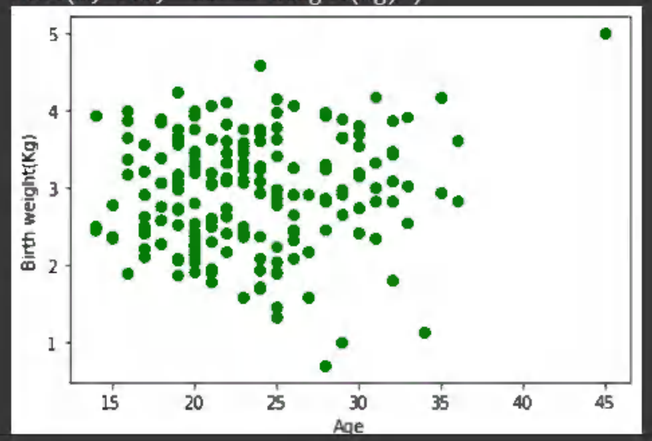
```
# Birth weight in grams
birthwt = df["bwt"]
birthwt = birthwt.to_numpy()
```

```
correlation(age, birthwt)
```

```
0.0903178136685326
```

```
# Converting birth weight from gram to kg for better scaling
plt.scatter(age,birthwt/1000,c ="green")
plt.xlabel("Age")
plt.ylabel("Birth weight(Kg)")
```

```
Text(0, 0.5, 'Birth weight(Kg)')
```



TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain

ENG

13:34

13-01-2022

Combine your Documents and I | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH

Apps | HackerRenk | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

..

sample_data

Mounting Google Drive...

+ Code + Text

Initializing Editing

Conclusion:

The corellation value is 0.09 which is very low, this means the correlation is non-existent between the maternal age and birth weight.

Using simple linear regression :

```
[ ] lr = LinearRegression()
    age = age.reshape(-1,1)
    lr.fit(age,birthwt)
```

```
LinearRegression()
```

```
y = lr.predict(age)
print("Coefficients :",lr.coef_[0])
print("intercept :",lr.intercept_)
```

```
Coefficients : 12.429712027714634
intercept : 2655.744469705171
```

```
[ ] plt.plot(age,y,color= "red")
    plt.scatter(age,birthwt,c= "green")
    plt.xlabel("Age")
    plt.ylabel("Birth weight(g)")
    plt.show()
```

TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain ENG 13:34 13-01-2022

Combine your Documents and I | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH

Apps | HackerRenk | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

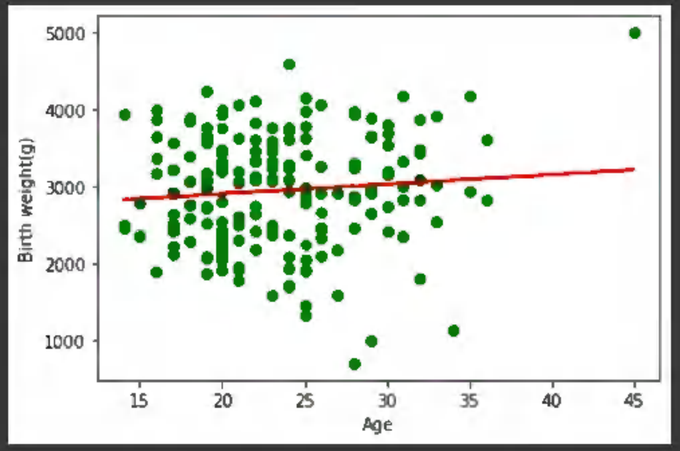
..

sample_data

+ Code + Text

Initializing | Editing

[ ]



Test the association of mother's weight (lwt) and birth weight using the correlation testand linear regeression.

Using correlation coefficients test :

[ ]
```
# Mother's weight during last menstrual period.(in pounds)
motherswt = df["lwt"]
motherswt =motherswt.to_numpy()

# converting in grams to pounds
birthwt = birthwt/454
```

Mounting Google Drive...

TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain

ENG

13:34

13-01-2022

Combine your Documents and I | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH

Apps | HackerRank | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment | Share

Files

..

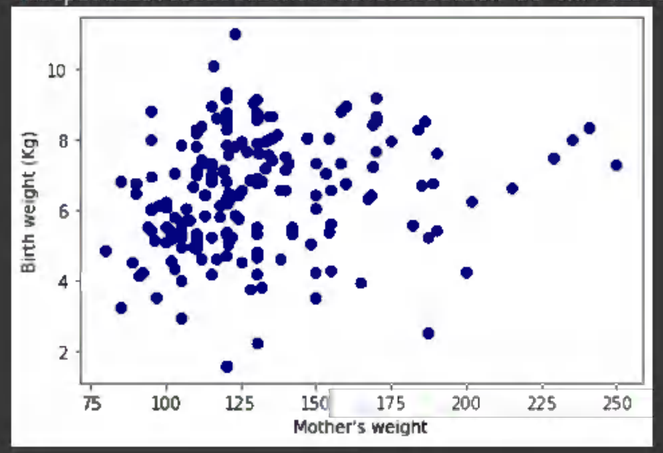
sample_data

+ Code + Text

Connected | Editing

```
[ ] plt.xlabel("Mother's weight")
    plt.ylabel("Birth weight (Kg)")
    plt.scatter(motherswt,birthwt,c = "darkblue")
```

```
<matplotlib.collections.PathCollection at 0x7fa1373d4d50>
```



Conclusion:

The correlation value is 0.18573328444909923 which is positive correlation, but the value is small which means the correlation is positive and small between the maternal weight and birth weight.

Using simple linear regression :

TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain

ENG

13:34

13-01-2022

Combine your Documents and Ir | AIML Assignment3.ipynb - Colab

colab.research.google.com/drive/1keTf6BNWpz0oUT9om2m8TwSQn1biyJvH#scrollTo=2ba70bqJeq_Q

Apps | HackerRank | W3Schools | Elearn | Codeforces | Django | ERDPlus | LeetCode | Tinkercad | Coursera | My Captain | Google Tech Dev G... | Reading list

AIML Assignment3.ipynb

File Edit View Insert Runtime Tools Help Last [redacted] on [redacted]uary 10

Comment Share

RAM Disk

Files

..

sample_data

Disk 65.98 GB available

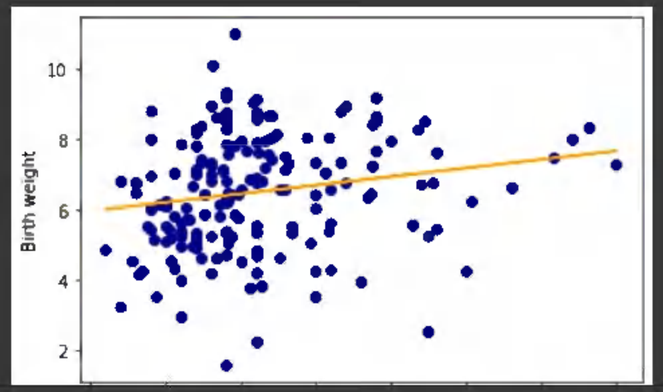
+ Code + Text

```
LinearRegression()
```

```
[ ] z = lr.predict(motherswt)
    print("Coefficients :",lr.coef_[0])
    print("intercept :",lr.intercept_)
```

```
Coefficients : 0.009755743626323136
intercept : 5.219435061396471
```

```
plt.plot(motherswt,z,c="orange")
plt.scatter(motherswt,birthwt,c ="darkblue")
plt.xlabel("Mother's weight")
plt.ylabel("Birth weight")
plt.show()
```



TCOB41 AIML Assi....pdf

Show all

Type here to search

22°C Light rain

ENG

13:34

13-01-2022