SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE CODE: DJ19TEL7014 **DATE:** 4/10/23

COURSE NAME: Machine Learning **CLASS:** Final Year B.Tech

EXPERIMENT NO. 3

CO Measured:

CO1 Solve real-world problems using suitable machine learning techniques.

TITLE: Model-building using Multiple Linear Regression

AIM / OBJECTIVE:

To perform Multiple Linear Regression and find the correlation matrix and discuss on error.

DESCRIPTION OF EXPERIMENT:

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for i=n observations:

yi = dependent variable

xi = explanatory variables

β0 = y-intercept (constant term)

βp = slope coefficients for each explanatory variable

ϵ = the model’s error term (also known as the residuals)

Advantages of Multiple Regression over a Simple OLS Regression

A dependent variable is rarely explained by only one variable. In such cases, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. The model, however, assumes that there are no major correlations between the independent variables.

PROCEDURE:

exp3

October 4, 2023

[1]:
```
#dataset : https://www.kaggle.com/datasets/mdrazakhan/linear-regression-dataset
```

[16]:
```
import pandas as pd
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
import matplotlib.pyplot as plt
```

[3]:
```
df=pd.read_csv('dataset\cars.csv')
```

[4]:
```
df.head()
```

[4]:

	Car	Model	Volume	Weight	CO2
0	Toyoty	Aygo	1000	790	99
1	Mitsubishi	Space Star	1200	1160	95
2	Skoda	Citigo	1000	929	95
3	Fiat	500	900	865	90
4	Mini	Cooper	1500	1140	105

[7]:
```
x=df[['Weight', 'Volume']]
y=df['CO2']
```

[19]:
```
combined_df = pd.concat([x, y], axis=1)

# Calculate the correlation matrix
correlation_matrix = combined_df.corr()
print(correlation_matrix)
```

```
          Weight    Volume       CO2
Weight  1.000000  0.753537  0.552150
Volume  0.753537  1.000000  0.592082
CO2     0.552150  0.592082  1.000000
```

[ ]:
```
'''
Weight and CO2:
    The correlation coefficient between 'Weight' and 'CO2' is approximately 0.
↪5522.
```

```
    This suggests a moderate positive correlation between 'Weight' and 'CO2'.
    As the weight of an object increases, the CO2 emissions tend to increase to␣
 ↪some extent.

Volume and CO2:
    The correlation coefficient between 'Volume' and 'CO2' is approximately 0.
 ↪5921.
    This also indicates a moderate positive correlation between 'Volume' and␣
 ↪'CO2'.
    As the volume of an object increases, the CO2 emissions tend to increase to␣
 ↪some extent.
'''
```

[11]:
```
#train test split
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size=0.2,␣
 ↪random_state=42)
```

[12]:
```
regrLine=linear_model.LinearRegression()
regrLine.fit(X_train, y_train)
```

[12]: LinearRegression()

[13]:
```
# Coefficient
print("Coefficient : ",end=' ')
print(regrLine.coef_)
```

```
Coefficient :  [0.00804928 0.00428741]
```

[17]:
```
# Make predictions on the test data
y_pred = regrLine.predict(X_test)

# Calculate mse to evaluate the model's performance
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
```

```
Mean Squared Error: 58.07928583657771
```

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

OBSERVATIONS:

1. Differentiate between Simple Linear Regression and Multiple Linear regression.

<table>
<tr><th></th><th>Simple Linear Regression</th><th>Multiple Linear regression</th></tr>
<tr><td>Number of Independent Variables</td><td>In simple linear regression, there is only one independent variable (predictor variable) that is used to predict the dependent variable. The relationship between the two variables is modeled as a straight line (hence "simple" linear regression).</td><td>In multiple linear regression, there are two or more independent variables that are used to predict the dependent variable. The relationship is modeled as a linear combination of these variables, allowing for a more complex and multidimensional analysis.</td></tr>
<tr><td>Equation</td><td>The equation for simple linear regression is typically represented as:
$Y = \beta0 + \beta1*X + \varepsilon$
Where:
Y is the dependent variable.
X is the independent variable.
$\beta0$ is the intercept (constant).
$\beta1$ is the coefficient of the independent variable.
ε represents the error term.</td><td>The equation for multiple linear regression extends to include multiple independent variables:
$Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \beta n*Xn + \varepsilon$
Where:
Y is the dependent variable.
X1, X2, ..., Xn are the independent variables.
$\beta0$ is the intercept (constant).
$\beta1, \beta2, ..., \beta n$ are the coefficients of the independent variables.
ε represents the error term.</td></tr>
<tr><td>Purpose</td><td>Simple linear regression is used when you want to understand the relationship between two variables and predict one based on the other. It is suitable for scenarios where there is a single predictor variable</td><td>Multiple linear regression is used when you want to predict a dependent variable based on the influence of multiple independent variables. It allows you to analyze the collective impact of several predictors on the outcome.</td></tr>
<tr><td>Example</td><td>Predicting a student's exam score (dependent variable) based on the number of hours they studied (independent variable)</td><td>Predicting a house's sale price (dependent variable) based on multiple factors such as square footage, number of bedrooms, and neighborhood quality (multiple independent variables)</td></tr>
</table>

SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)

CONCLUSION:

In conclusion, the multiple regression analysis conducted in this experiment has unveiled crucial insights into the interplay between our chosen independent variables and the dependent variable. The model effectively captures the combined impact of these predictors, revealing their significance and the direction of their influence. The statistical tests conducted confirm the model's overall significance and the relevance of specific predictors, offering valuable information for decision-making and forecasting in various domains.

This experiment underscores the utility of multiple regression analysis as a powerful tool for data-driven decision-making and enhanced understanding of complex relationships.