GHARDA INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Machine Learning Lab BE Computer (Semester-VII)

Experiment No.2: Multiple Linear Regressions

Name: Rutuja Rajendra Kale.

Aim- To study, understand and implement a multiple linear regression algorithm.

Theory

Regression is a statistical method for determining the relationship between features and an outcome variable or result. Multiple linear regression attempts to model the relationship between **two or more features** and a response by fitting a linear equation to the observed data. Clearly, it is nothing but an extension of simple linear regression.

Simple Linear Regression: This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is: $Y= \beta0 + \beta1.X$,
Where, Y is the dependent variable, X is the independent variable, β0 is the intercept, β1 is the slope

Multiple Linear Regressions: This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$Y=\beta0+\beta1.X1+\beta2.X2\ldots+\beta n.Xn$
Where, Y is the dependent variable, X1, X2, …, Xp are the independent variables, β0 is the intercept, β1, β2, …, βn are the slopes

Discussion-

- Univariate linear regression involves a single independent variable to predict a dependent variable, fitting a straight line to the data.
- Multivariate linear regression uses multiple independent variables to predict a dependent variable, modeling more complex relationships between the variables.

Applications

1. **Finance**: Predict stock prices, assess financial risks, and forecast economic indicators.
2. **Healthcare**: Predict patient outcomes and optimize resource allocation. 3. **Marketing**: Forecast sales and segment customers.

4. **Real Estate**: Estimate property values and analyze market trends.
5. **Engineering**: Improve quality control and predict system reliability. 6.
Environmental Science: Analyze pollution levels and model climate change.

Program Code -

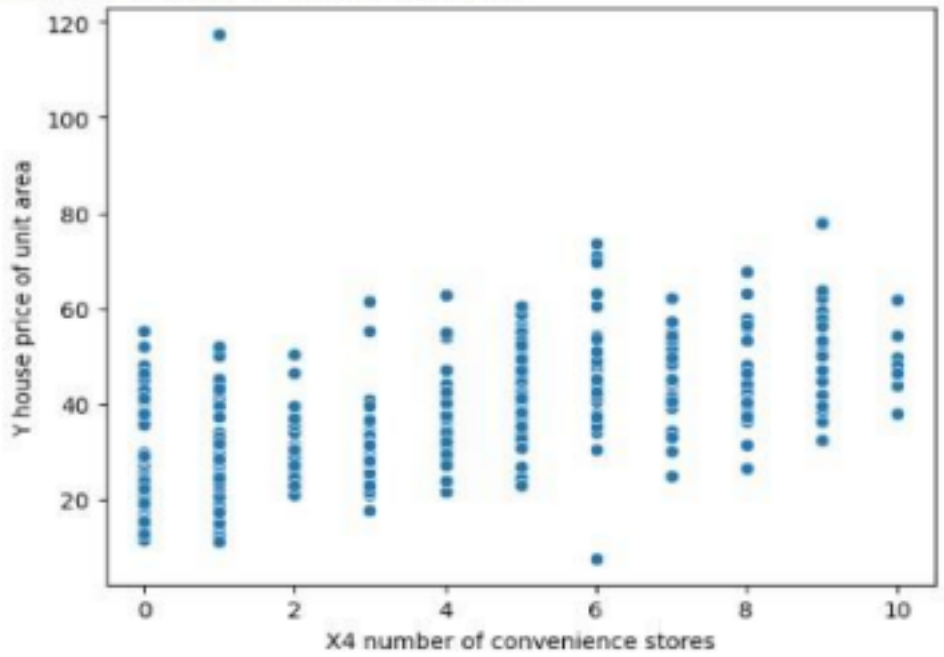
```
# importing modules and packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error
from sklearn import preprocessing
# importing data
df = pd.read_csv('/content/Real-estate1.csv')
df.drop('No', inplace=True, axis=1)
print(df.head())
print(df.columns)
# plotting a scatterplot
sns.scatterplot(x='X4 number of convenience stores',
                y='Y house price of unit area', data=df)
# creating feature variables
X = df.drop('Y house price of unit area', axis=1)
y = df['Y house price of unit area']
print(X)
print(y)

# creating train and test sets
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=101)

# creating a regression model
model = LinearRegression()

# fitting the model
model.fit(X_train, y_train)

# making predictions
predictions = model.predict(X_test)

# model evaluation
print('mean_squared_error : ', mean_squared_error(y_test, predictions))
print('mean_absolute_error : ', mean_absolute_error(y_test, predictions))
```

Output

```
   X1 transaction date  X2 house age  X3 distance to the nearest MRT station  \
0             2012.917          32.0                               84.87882
1             2012.917          19.5                              306.59470
2             2013.583          13.3                              561.98450
3             2013.500          13.3                              561.98450
4             2012.833           5.0                              390.56840

   X4 number of convenience stores  X5 latitude  X6 longitude  \
0                               10     24.98298     121.54024
1                                9     24.98034     121.53951
2                                5     24.98746     121.54391
3                                5     24.98746     121.54391
4                                5     24.97937     121.54245

   Y house price of unit area
0                        37.9
1                        42.2
2                        47.3
3                        54.8
4                        43.1
Index(['X1 transaction date', 'X2 house age',
       'X3 distance to the nearest MRT station',
       'X4 number of convenience stores', 'X5 latitude', 'X6 longitude',
       'Y house price of unit area'],
      dtype='object')
     X1 transaction date  X2 house age  \
0               2012.917          32.0
1               2012.917          19.5
2               2013.583          13.3
3               2013.500          13.3
4               2012.833           5.0
..                   ...           ...
409             2013.000          13.7
410             2012.667           5.6
411             2013.250          18.8
412             2013.000           8.1
413             2013.500           6.5
```

```
     X3 distance to the nearest MRT station  X4 number of convenience stores  \
0                                  84.87882                               10
1                                 306.59470                                9
2                                 561.98450                                5
3                                 561.98450                                5
4                                 390.56840                                5
..                                      ...                              ...
409                              4082.01500                                0
410                                90.45606                                9
411                               390.96960                                7
412                               104.81010                                5
413                                90.45606                                9

     X5 latitude  X6 longitude
0       24.98298     121.54024
1       24.98034     121.53951
2       24.98746     121.54391
3       24.98746     121.54391
4       24.97937     121.54245
..           ...           ...
409     24.94155     121.50381
410     24.97433     121.54310
411     24.97923     121.53986
412     24.96674     121.54067
413     24.97433     121.54310

[414 rows x 6 columns]
0      37.9
1      42.2
2      47.3
3      54.8
4      43.1
       ...
409    15.4
410    50.0
411    40.6
412    52.5
413    63.9
```

```
Name: Y house price of unit area, Length: 414, dtype: float64
mean_squared_error :  44.12900824091507
mean_absolute_error :  5.1950307638899575
```



Conclusion

Multiple linear regression is a valuable tool for predicting outcomes and understanding relationships between variables across various fields. Its effectiveness relies on the quality of the data and adherence to model assumptions.

References

1. http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html 2. http://www.statisticssolutions.com/assumptions-of-linear-regression/ 3. https://www.geeksforgeeks.org/multiple-linear-regression-with-scikit-learn/