BCSE00133 & 731

Multi Variable Linear Regression

Machine Learning Lab

DR GAURAV KUMAR

ASST. PROF, CEA, GLA UNIVERSITY

Implementing Multi Linear Regression Algorithm

Multivariable Linear Regression

Regressions

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Implementing Multi Linear Regression Algorithm

R&D Spend	Administration	Marketing Spend	State	Profit

Startups profit prediction using Multiple Linear Regression

Implementing Multi Linear Regression Algorithm

Startups profit prediction using Multiple Linear Regression

- We will use dataset that includes sample data of **1000 startup companies** operating cost and their profit. (download it form Kaggle)
- We will use the concept of **Multiple linear Regression** to predict the profit of startups companies.

Implementing Multi Linear Regression Algorithm

GLA UNIVERSITY MATHURA
Accredited with A Grade by NAAC
12-B Status from UGC

Startups profit prediction using Multiple Linear Regression

R&D Spend	Administration	Marketing Spend	State	Profit

Dataset holds data from 1000 startups of New York, California, and Florida states. The features in this dataset are R&D spending, Administration Spending, Marketing Spending, location, and Profit.

1. R&D spending: The amount which startups are spending on Research and development.

2. Administration spending: The amount which startups are spending on the Admin panel

3. Marketing spending: The amount which startups are spending on marketing strategies

4. State: To which state that particular startup belongs.

5.**Profit:** How much profit that particular startup is making.

Implementing Multi Linear Regression Algorithm

Multivariable Linear Regression

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as pt
import seaborn as sn
```

```
companies = pd.read_csv("1000_companies.csv")
companies.head()
```

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

Implementing Multi Linear Regression Algorithm

Exploring Dataset

Numerical/Statistical analysis of the dataset

companies.describe()

	R&D Spend	Administration	Marketing Spend	Profit
count	50.000000	50.000000	50.000000	50.000000
mean	73721.615600	121344.639600	211025.097800	112012.639200
std	45902.256482	28017.802755	122290.310726	40306.180338
min	0.000000	51283.140000	0.000000	14681.400000
25%	39936.370000	103730.875000	129300.132500	90138.902500
50%	73051.080000	122699.795000	212716.240000	107978.190000
75%	101602.800000	144842.180000	299469.085000	139765.977500
max	165349.200000	182645.560000	471784.100000	192261.830000

- **Note-** Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.

Or

- Percentiles indicate **the percentage of scores that fall below a particular value**. It tells you **where a score stands relative to other** scores.

- For example, a person with an IQ of 120 is at the 91 percentile, which indicates that their IQ is higher than 91 percent of other scores.

Implementing Multi Linear Regression Algorithm

Exploring Dataset

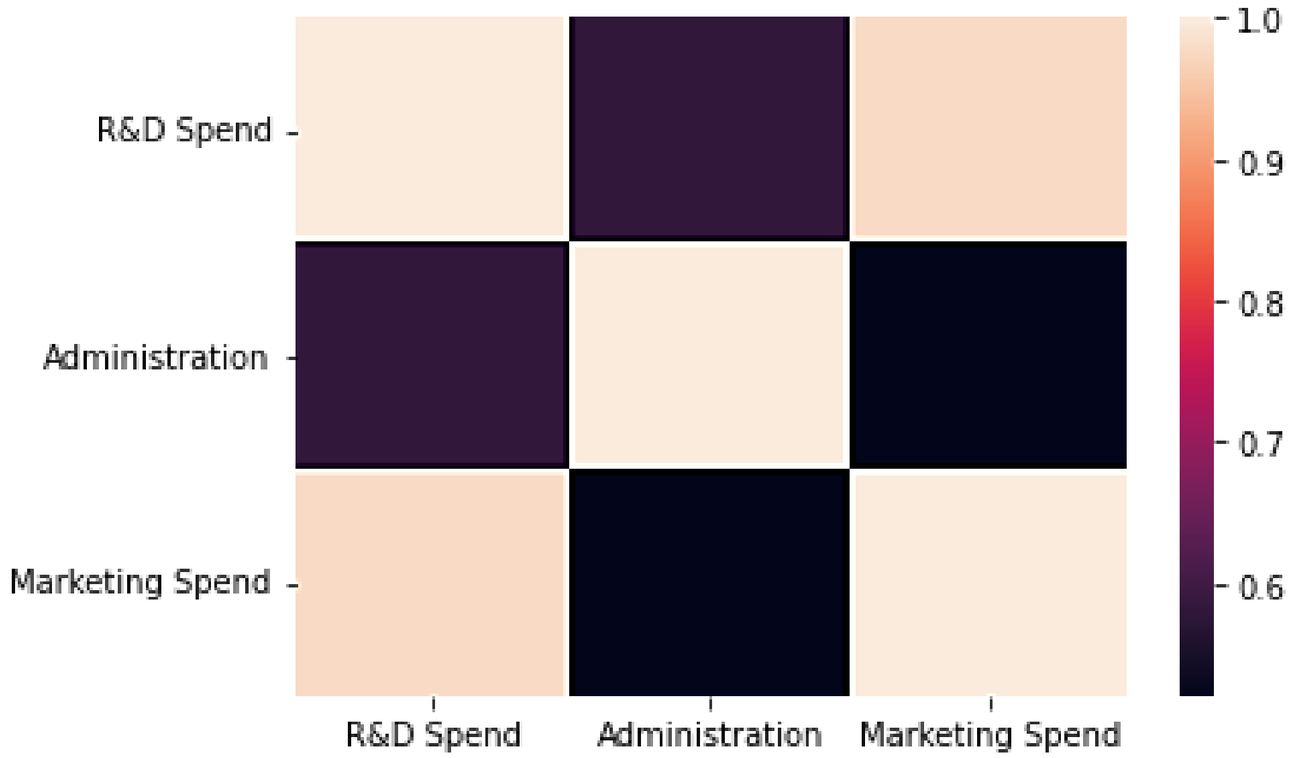
Numerical/Statistical analysis of the dataset

companies[['R&D Spend', 'Administration', 'Marketing Spend']].corr()

	R&D Spend	Administration	Marketing Spend
R&D Spend	1.000000	0.582434	0.978407
Administration	0.582434	1.000000	0.520465
Marketing Spend	0.978407	0.520465	1.000000

Implementing Multi Linear Regression Algorithm

sns.heatmap(companies[['R&D Spend','Administration','Marketing Spend']].corr())



	R&D Spend	Administration	Marketing Spend
R&D Spend	1.000000	0.582434	0.978407
Administration	0.582434	1.000000	0.520465
Marketing Spend	0.978407	0.520465	1.000000

Implementing Multi Linear Regression Algorithm

Categorizing Independent and Dependent Variable

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

```
X=companies.iloc[:,:-1].values

y=companies.iloc[:,4].values
```

Implementing Multi Linear Regression Algorithm

GLA UNIVERSITY MATHURA
Accredited with A Grade by NAAC
12-B Status from UGC

removing column which is highly correlated

```
[7] companies.drop(['Marketing Spend'],axis=1,inplace=True)
```

```
companies
```

	R&D Spend	Administration	State	Profit
0	165349.20	136897.800	New York	192261.83000
1	162597.70	151377.590	California	191792.06000
2	153441.51	101145.550	Florida	191050.39000
3	144372.41	118671.850	New York	182901.99000
4	142107.34	91391.770	Florida	166187.94000
...	...	...	...	...
995	54135.00	118451.999	California	95279.96251
996	134970.00	130390.080	California	164336.60550

	R&D Spend	Administration	Marketing Spend
R&D Spend	1.000000	0.582434	0.978407
Administration	0.582434	1.000000	0.520465
Marketing Spend	0.978407	0.520465	1.000000

Implementing Multi Linear Regression Algorithm

Encoding Data

```
dummies=pd.get_dummies(companies.State)
```

```
companies=pd.concat([companies,dummies],axis=1)
```

```
companies
```

	R&D Spend	Administration	State	Profit	California	Florida	New York
0	165349.20	136897.800	New York	192261.83000	0	0	1
1	162597.70	151377.590	California	191792.06000	1	0	0
2	153441.51	101145.550	Florida	191050.39000	0	1	0
3	144372.41	118671.850	New York	182901.99000	0	0	1
4	142107.34	91391.770	Florida	166187.94000	0	1	0
...	...	...	...	...	...	...	...
995	54135.00	118451.999	California	95279.96251	1	0	0
996	134970.00	130390.080	California	164336.60550	1	0	0
997	100275.47	241926.310	California	413956.48000	1	0	0
998	128456.23	321652.140	California	333962.19000	1	0	0

Implementing Multi Linear Regression Algorithm

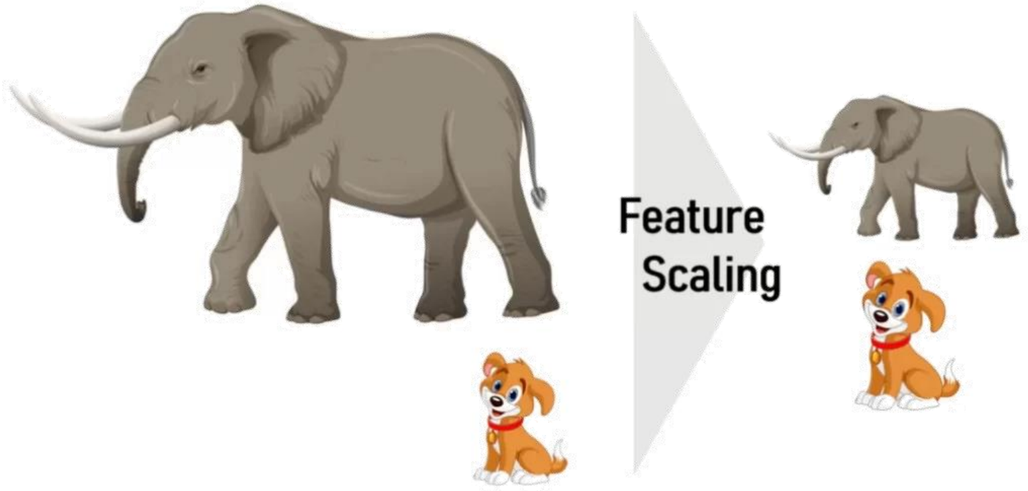

```
companies.drop(['State'],axis=1,inplace=True)
companies
```

	R&D Spend	Administration	Profit	California	Florida	New York
0	165349.20	136897.800	192261.83000	0	0	1
1	162597.70	151377.590	191792.06000	1	0	0
2	153441.51	101145.550	191050.39000	0	1	0
3	144372.41	118671.850	182901.99000	0	0	1
4	142107.34	91391.770	166187.94000	0	1	0
...	...	...	...	...	...	...
995	54135.00	118451.999	95279.96251	1	0	0
996	134970.00	130390.080	164336.60550	1	0	0
997	100275.47	241926.310	413956.48000	1	0	0
998	128456.23	321652.140	333962.19000	1	0	0
999	161181.72	270939.860	476485.43000	0	0	1

1000 rows × 6 columns

Implementing Multi Linear Regression Algorithm

Feature Scaling



sklearn.preprocessing.MinMaxScaler

- Transform features by scaling each feature to a given range.
- This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.

The transformation is given by:

```
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
X_scaled = X_std * (max - min) + min
```

where min, max = feature_range.

Implementing Multi Linear Regression Algorithm

Feature Scaling

```
from sklearn.preprocessing import MinMaxScaler

scale=MinMaxScaler()

companies[['R&DSpend','Administration']]=scale.fit_transform(companies[['R&D Spend','Administration']])
```

```
companies
```

	R&D Spend	Administration	Profit	California	Florida	New York
0	1.000000	0.316659	192261.83000	0	0	1
1	0.983359	0.370214	191792.06000	1	0	0
2	0.927985	0.184424	191050.39000	0	1	0
3	0.873136	0.249247	182901.99000	0	0	1
4	0.859438	0.148348	166187.94000	0	1	0
...	...	...	...	...	...	...
995	0.327398	0.248434	95279.96251	1	0	0
996	0.816272	0.292589	164336.60550	1	0	0
997	0.606447	0.705122	413956.48000	1	0	0
998	0.776878	1.000000	333962.19000	1	0	0

Implementing Multi Linear Regression Algorithm

Specifying X (Independent) and y (dependent Variable)

```
y=companies.iloc[:,2].values
y
```

```
array([192261.83   , 191792.06   , 191050.39   , 182901.99   ,
       166187.94   , 156991.12   , 156122.51   , 155752.6    ,
       152211.77   , 149759.96   , 146121.95   , 144259.4    ,
       141585.52   , 134307.35   , 132602.65   , 129917.04   ,
       126992.93   , 125370.37   , 124266.9    , 122776.86   ,
       118474.03   , 111313.02   , 110352.25   , 108733.99   ,
       108552.04   , 107404.34   , 105733.54   , 105008.31   ,
       103282.38   , 101004.64   ,  99937.59   ,  97483.56   ,
        97427.84   ,  96778.92   ,  96712.8    ,  96479.51   ,
        90708.19   ,  89949.14   ,  81229.06   ,  81005.76   ,
        78239.91   ,  77798.83   ,  71498.49   ,  69758.98   ,
        65200.33   ,  64926.08   ,  49490.75   ,  42559.73   ,
```

```
companies
```

	R&D Spend	Administration	Profit	California	Florida	New York
0	1.000000	0.316659	192261.83000	0	0	1
1	0.983359	0.370214	191792.06000	1	0	0
2	0.927985	0.184424	191050.39000	0	1	0
3	0.873136	0.249247	182901.99000	0	0	1
4	0.859438	0.148348	166187.94000	0	1	0
...	...	...	...	...	...	...
995	0.327398	0.248434	95279.96251	1	0	0
996	0.816272	0.292589	164336.60550	1	0	0
997	0.606447	0.705122	413956.48000	1	0	0
998	0.776878	1.000000	333962.19000	1	0	0

Implementing Multi Linear Regression Algorithm

Specifying X (Independent) and y (dependent Variable)

```
companies.drop(['Profit'],axis=1,inplace=True)
companies
```

	R&D Spend	Administration	California	Florida	New York
0	1.000000	0.316659	0	0	1
1	0.983359	0.370214	1	0	0
2	0.927985	0.184424	0	1	0
3	0.873136	0.249247	0	0	1
4	0.859438	0.148348	0	1	0
...	...	...	...	...	...
995	0.327398	0.248434	1	0	0
996	0.816272	0.292589	1	0	0
997	0.606447	0.705122	1	0	0

```
companies
```

	R&D Spend	Administration	Profit	California	Florida	New York
0	1.000000	0.316659	192261.83000	0	0	1
1	0.983359	0.370214	191792.06000	1	0	0
2	0.927985	0.184424	191050.39000	0	1	0
3	0.873136	0.249247	182901.99000	0	0	1
4	0.859438	0.148348	166187.94000	0	1	0
...	...	...	...	...	...	...
995	0.327398	0.248434	95279.96251	1	0	0
996	0.816272	0.292589	164336.60550	1	0	0
997	0.606447	0.705122	413956.48000	1	0	0
998	0.776878	1.000000	333962.19000	1	0	0

Implementing Multi Linear Regression Algorithm

Specifying X (Independent) and y (dependent Variable)

```
X=companies.iloc[:,:]
```

	R&D Spend	Administration	California	Florida	New York
0	1.000000	0.316659	0	0	1
1	0.983359	0.370214	1	0	0
2	0.927985	0.184424	0	1	0
3	0.873136	0.249247	0	0	1
4	0.859438	0.148348	0	1	0
...	...	...	...	...	...
995	0.327398	0.248434	1	0	0
996	0.816272	0.292589	1	0	0
997	0.606447	0.705122	1	0	0

```
X=companies.iloc[:,:].values
X
```

```
array([[1.        , 0.31665857, 0.        , 0.        , 1.        ],
       [0.98335946, 0.37021423, 1.        , 0.        , 0.        ],
       [0.92798459, 0.18442355, 0.        , 1.        , 0.        ],
       ...,
       [0.60644666, 0.70512215, 1.        , 0.        , 0.        ],
       [0.77687845, 1.        , 1.        , 0.        , 0.        ],
       [0.97479589, 0.81243308, 0.        , 0.        , 1.        ]])
```

Implementing Multi Linear Regression Algorithm

Splitting dataset into train and test

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.25,random_state=0)
```

Fitting model to training set

```
from sklearn.linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(X_train,y_train)
```

Predicting the test dataset

```
y_pred = lin_reg.predict(X_test)
```

Implementing Multi Linear Regression Algorithm

Testing Accuracy

```
import sklearn.metrics as sm
print("Mean squared error =", sm.mean_squared_error(y_test, y_pred))
print("Explain variance score =", round(sm.explained_variance_score(y_test, y_pred), 2))
print("R2 score =", round(sm.r2_score(y_test, y_pred), 2))
```

```
Mean squared error = 160388568.10952258
Explain variance score = 0.92
R2 score = 0.92
```

Assignment for Analysis

1. Change the Training and Testing Size (4 variations) and Analyze the Accuracy

2. Reduce the Data Set Size and See the Results and analyze with original data set size

3. What will happen if we do not use Feature Scaling, Analyze the result with feature scaling methods

4. Use all features without dropping any column and analyze the result with the case if we drop the least significant column.

THANKYOU

gaurav.kumar@gla.ac.in

8586968801