MACHINE LEARNING

ASSIGNMENT – 3

Submitted by

Saritha H

Roll no:25

Linear Regression

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, "how much additional sales income do I get for each additional $1000 spent on marketing?"

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, "what will the price of gold be in 6 months?"

Linear Regression Python Implementation

Introduction

The data set contains information about money spent on advertisement and their generated sales. Money was spent on TV, radio and newspaper ads.The objective is to use linear regression to understand how advertisement spending impacts sales.

Import libraries

The advantage of working with Python is that we have access to many libraries that allow us to rapidly read data, plot the data, and perform a linear regression.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import statsmodels.api as sm
```

Read the data

Assuming that you downloaded the data set, place it in a data directory within your spyder file. Then, read the data like so:

```
data = pd.read_csv("C:\\Users\\Saritha H\\.spyder-py3\\Book1.csv")
```

To see what the data looks like, we do the following:

```
print(data)
```

And you should see this:

```
    Unnamed: 0     TV  radio  newspaper  sales
0             1  230.1   37.8       69.2   22.1
1             2   44.5   39.3       45.1   10.4
```

```
2             3   17.2   45.9       69.3    9.3
3             4  151.5   41.3       58.5   18.5
4             5  180.8   10.8       58.4   12.9
..          ...    ...    ...        ...    ...
195         196   38.2    3.7       13.8    7.6
196         197   94.2    4.9        8.1    9.7
197         198  177.0    9.3        6.4   12.8
198         199  283.6   42.0       66.2   25.5
199         200  232.1    8.6        8.7   13.4

[200 rows x 5 columns]
print(data.head())
print(data.columns)
Unnamed: 0     TV  radio  newspaper  sales
0           1  230.1   37.8       69.2   22.1
1           2   44.5   39.3       45.1   10.4
2           3   17.2   45.9       69.3    9.3
3           4  151.5   41.3       58.5   18.5
4           5  180.8   10.8       58.4   12.9
Index(['Unnamed: 0', 'TV', 'radio', 'newspaper', 'sales'], dtype='object')
```

As you can see, the column Unnamed: 0 is redundant. Hence, we remove it.

```
print(data.drop(['Unnamed: 0'], axis=1))
        TV  radio  newspaper  sales
0    230.1   37.8       69.2   22.1
1     44.5   39.3       45.1   10.4
2     17.2   45.9       69.3    9.3
3    151.5   41.3       58.5   18.5
4    180.8   10.8       58.4   12.9
..     ...    ...        ...    ...
195   38.2    3.7       13.8    7.6
```

```
196   94.2    4.9        8.1    9.7
197  177.0    9.3        6.4   12.8
198  283.6   42.0       66.2   25.5
199  232.1    8.6        8.7   13.4

[200 rows x 4 columns]
```

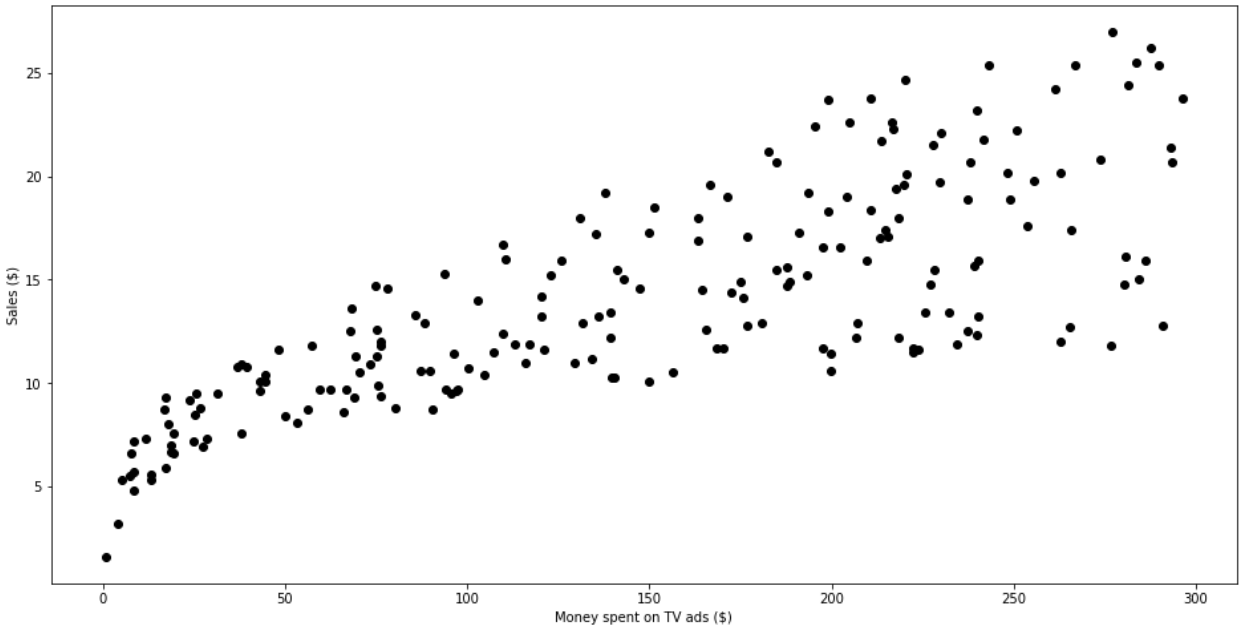
Modeling

For simple linear regression, let's consider only the effect of TV ads on sales. Before jumping right into the modeling, let's take a look at what the data looks like.

We use matplotlib , a popular Python plotting library to make a scatter plot.

```
plt.figure(figsize=(16, 8))
plt.scatter(data['TV'],data['sales'],c='black')
plt.xlabel("Money spent on TV ads ($)")
plt.ylabel("Sales ($)")
plt.show()
```

Run this cell of code and you should see this graph:



As you can see, there is a clear relationship between the amount spent on TV ads and sales.

Let's see how we can generate a linear approximation of this data.

```
X = data['TV'].values.reshape(-1,1)

y = data['sales'].values.reshape(-1,1)

reg = LinearRegression()

reg.fit(X, y)

print(reg.coef_[0][0])

print(reg.intercept_[0])

print("The linear model is: Y = {:.5} + {:.5}X".format(reg.intercept_[0],
reg.coef_[0][0]))
```

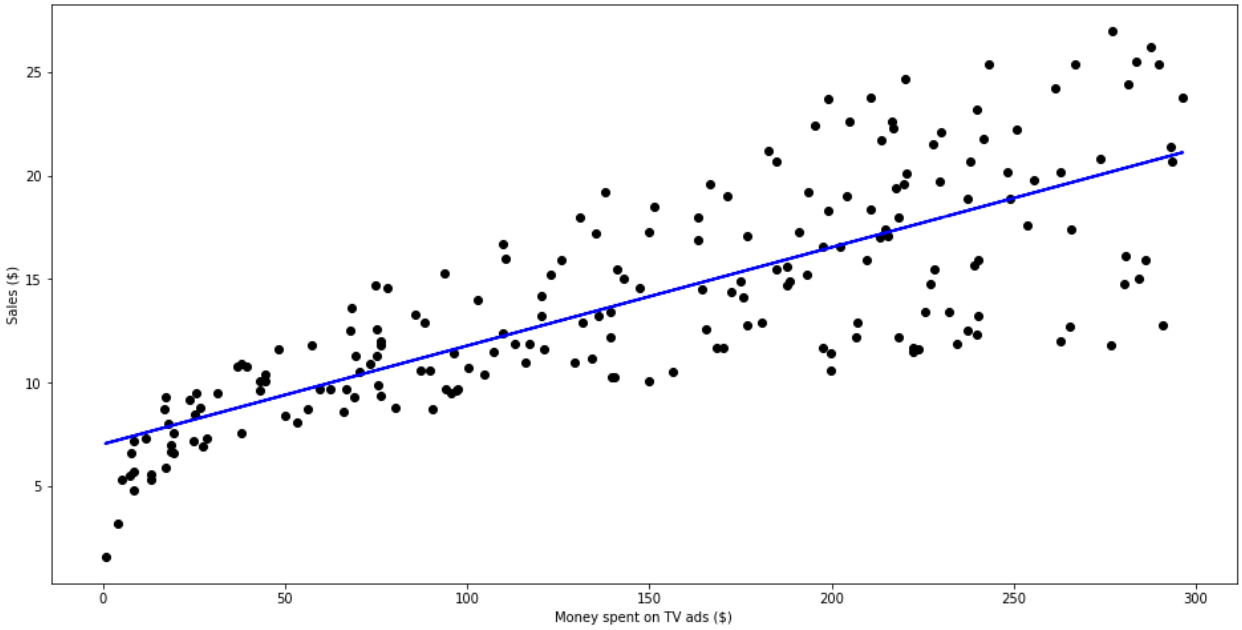
It is that simple to fit a straight line to the data set and see the parameters of the equation. In this case, we have

```
reg.coef_[0][0]))

0.047536640433019764

7.032593549127693

The linear model is: Y = 7.0326 + 0.047537X
```

$$\text{Sales} = 7.0326 + 0.047537(\text{TV})$$

Let's visualize how the line fits the data.

```
predictions = reg.predict(X)

plt.figure(figsize=(16, 8))

plt.scatter(data['TV'], data['sales'],c='black')

plt.plot(data['TV'],predictions,c='blue',linewidth=2)

plt.xlabel("Money spent on TV ads ($)")

plt.ylabel("Sales ($)")

plt.show()
```



From the graph above, it seems that a simple linear regression can explain the general impact of amount spent on TV ads and sales.

Assessing the relevancy of the model

We need to look at the R^2 value and the p-value from each coefficient.

Here's how we do it:

```
X = data['TV']

y = data['sales']

X2 = sm.add_constant(X)

est = sm.OLS(y, X2)

est2 = est.fit()

print(est2.summary())
```

```
                            OLS Regression Results

==============================================================================
=

Dep. Variable:                      sales   R-squared:
0.612

Model:                              OLS   Adj. R-squared:
0.610
```

```
Method:                 Least Squares   F-statistic:
312.1

Date:                Fri, 27 Mar 2020   Prob (F-statistic):           1.47e-
42

Time:                        13:14:02   Log-Likelihood:                -
519.05

No. Observations:                 200   AIC:
1042.

Df Residuals:                     198   BIC:
1049.

Df Model:                           1

Covariance Type:            nonrobust

==============================================================================
=

                 coef    std err          t      P>|t|      [0.025
0.975]

------------------------------------------------------------------------------
-

const          7.0326      0.458     15.360      0.000       6.130
7.935

TV             0.0475      0.003     17.668      0.000       0.042
0.053
```

```
Omnibus:                        0.531   Durbin-Watson:
1.935

Prob(Omnibus):                  0.767   Jarque-Bera (JB):
0.669

Skew:                          -0.089   Prob(JB):
0.716

Kurtosis:                       2.779   Cond. No.
338.
```

Looking at both coefficients, we have a p-value that is very low . This means that there is a strong correlation between these coefficients and the target .

Then, looking at the R^2 value, we have 0.612. Therefore, about 60% of the variability of sales is explained by the amount spent on TV ads. This is okay,

but definitely not the best we can to accurately predict the sales. Surely, spending on newspaper and radio ads must have a certain impact on sales.