A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

Group A.

Experiment - 04

- Title of the assignment - Create a linear regression model using Python/R to predict home prices using Boston housing dataset. (https://www.kaggle.com/c/boston-housing).

DATA ANALYSIS - I.

- Problem statement - The Boston housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset.
  The objective is to predict the value of prices of the house using given features.

- Objective of the assignment -
  Students should be able to data analysis using linear regression using python for any open source dataset.

- Pre-requisite -
  - Basic of python programming
  - Concept of regression.

- THEORY:

Linear Regression

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

1] Linear Regression -

- It is a machine learning algorithm based on supervised learning. It targets prediction values on the basis of independent variables.
- It is preferred to find out the relationship between forecasting and variables.
- A linear relationship between a dependent variable (X) is continuous; while independent variable (Y) relationship may be continuous or discrete. A linear relationship should be available in between predictor and target variable so known as linear regression.
- Linear regression is popular because the cost function is Mean Squared Error (MSE) which is equal to the average squared difference between an observation's actual and predicted values.
- It is known as equation of line like -

$$Y = mX + b + e.$$

Where b is intercepted m is slope of line and e is error term.

Fig 1. Geometry of Linear Regression.

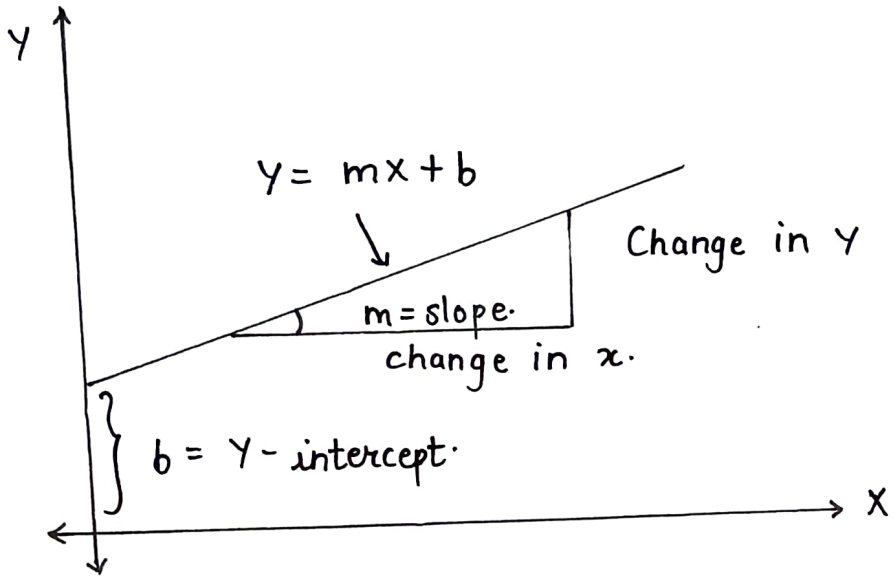
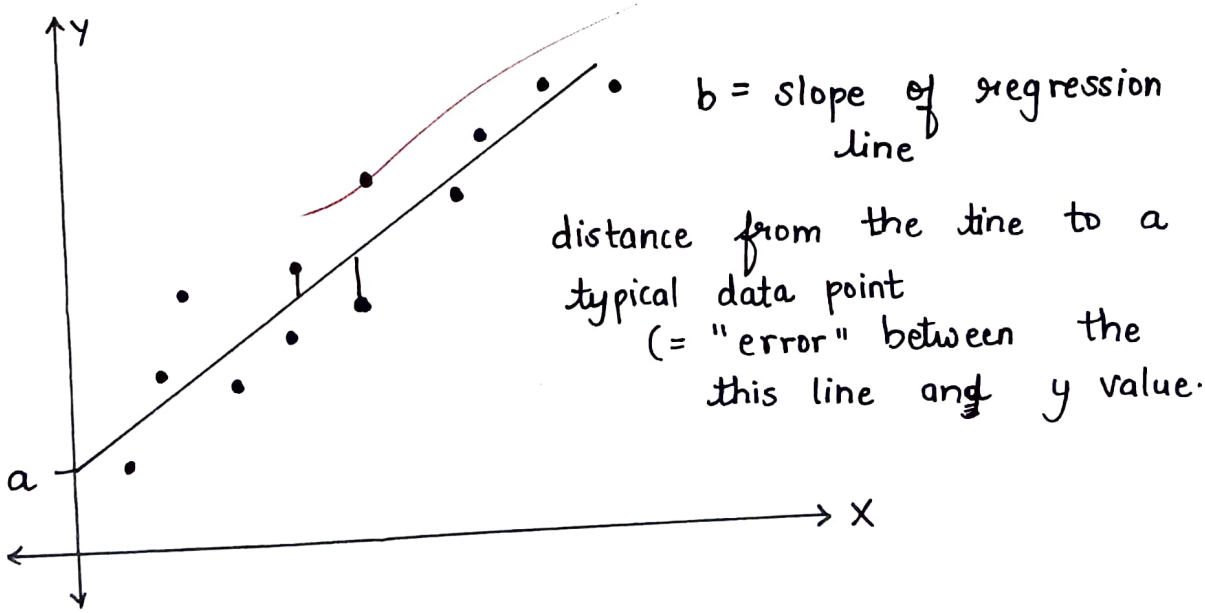


Fig. 2. Relation between weight (in Kg) and height (in cm).



A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

This equation can be used to predict the value of target variable Y based on given predictor variable (X) in Fig 1.

- Fig 2: Shown is about relation between weight in (kg) and height in (cm) a linear relation. It is used for learn relationship between quantitative variables.

- Another variable, denoted 'y' is considered as the response, outcome or dependent variable. While 'predictor' and response used to refer to these variables.

2) Multivariate Regression –

- It concerns the study of two or more predictor variables. Usually a transformation of the orignal features into polynomial features from a given degree is preferred and further linear regression is applied on it.

- A simple linear model $Y = a + bx$ is in original feature will be transformed into polynomial feature is transformed and further a linear regression applied to it and will be something like

$$Y = a + bx + cx2.$$

- If a high degree value is used in transformation the curve becomes overfitted as it captures the noise from data as well.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

3] Measuring Performance of linear regression -

- **MEAN SQUARE METHOD -**

- The MSE represents the error of the estimator or predictive model created based on the given set of observations in the sample.
- Two or more models can be compared based on their MSE.
- The lesser the MSE, the better the model is.
- When the model is trained using a given set of observations, the model with least mean sum of squares error (MSE) is selected as the best model.

- The python or R packages select the best fit model as the model with lowest MSE or lowest RMSE when training linear regression models.

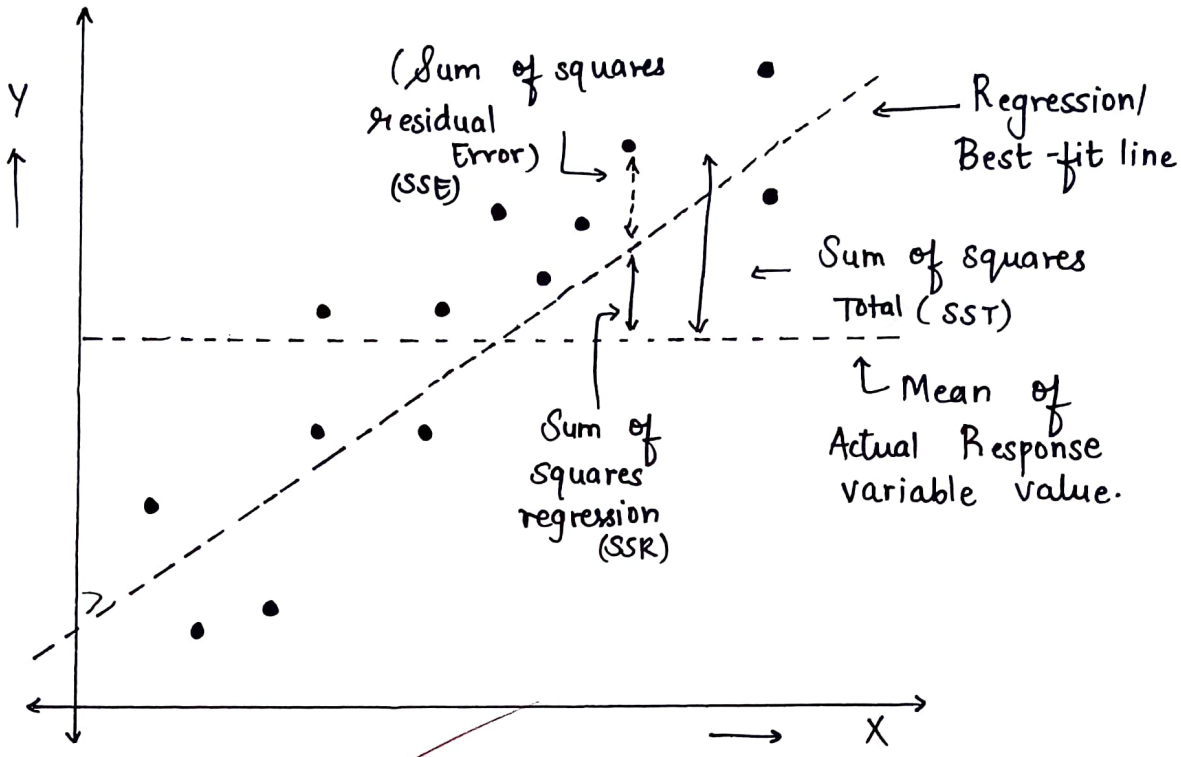
- Mathematically, MSE can be calculated as the average sum of squared difference between actual value and predicted or estimated value represented by the model. (line or plane).

$$MSE = \frac{1}{n} \sum \underbrace{(y - \bar{y})^2}$$

Square of difference between actual and predicted value.

RMSE - Least square Regression Method.

R- Squared.



$$SST = \sum_{i=1} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- An MSE of zero (0) represents the fact that the predictor is a perfect predictor.

- RMSE

- Root Mean Square Error method that basically calculates the least-squares error and takes a root of the summed values.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (\bar{y}_i - y_i)^2}$$

A value of r-squared closer to 1 would mean that the regression model covers most part of the variance of the values of the response variable and can be termed as a good model.

* Interpretation of Regression line -

Interpretation 1
- For an increase in value of x by 0.644 units there is an increase in value of y in one unit.

Interpretation 2
- Even if $x = 0$ value of independent variable, it is expected that value of y is 26.768.

Example of linear regression -

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

- Consider following data for 5 students -

Each x_i (i = 1 to 5) represents the score of i^{th} student in standard x and corresponding y_i represents the score of i^{th} student in standard XII

i) Linear regression equation best predicts standard XII score.

ii) Interpretation for the equation of linear regression

iii) If a students score is 80 in std x then what is expected in std XII. ?

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
95	85	17	8	289	136
85	95	7	18	49	126
80	70	2	-7	4	-14
70	65	-8	-12	64	96
60.	70.	-18	-7	324	126.

$\bar{x} = 78$ $\bar{y} = 77$ $\Sigma(x-\bar{x})^2 = 730.$ $\Sigma = 470.$

i) Linear regression equation -

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{\sum_{i=1} (x_i - x) - \bar{y})}{\sum_{i=1} (x_i - \bar{x})^2}$$

$$\beta_1 = 470/730 = 0.644$$

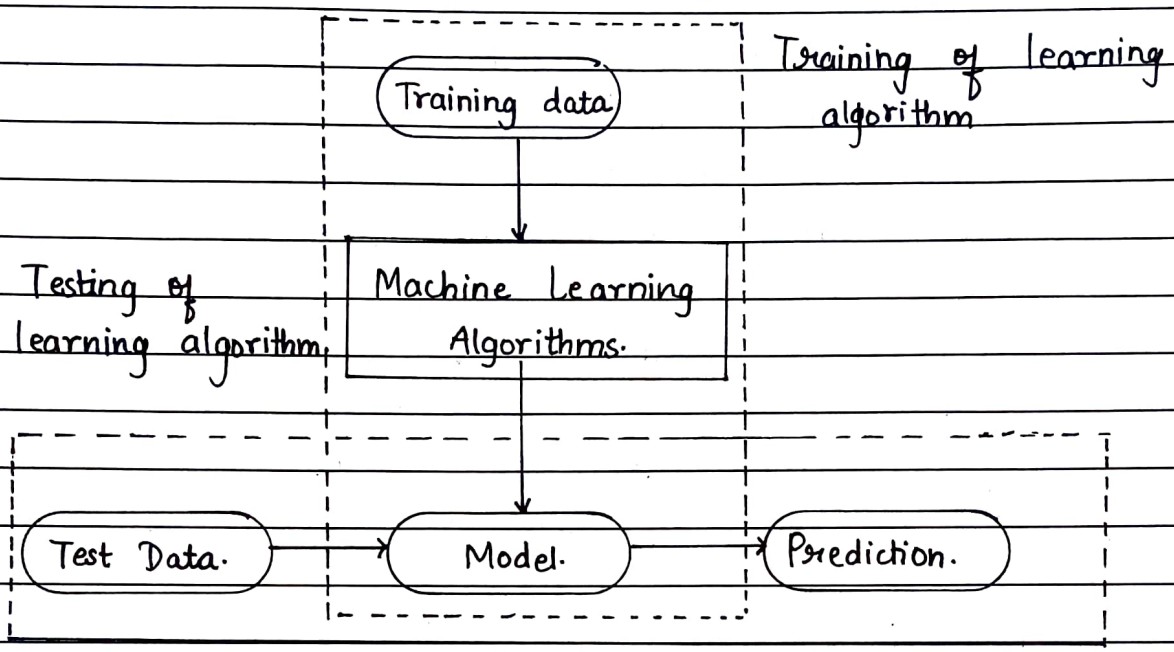
$$\beta_0 = \bar{y} - \beta_1(\bar{x})$$

$$\beta_0 = 77 - (0.644 * 78)$$

$$= 26.768.$$

$$y = 26.76 + 0.644x.$$

* Training data set and Testing data set.



(a) Training phase:

- Training dataset is provided as input to this phase.
- Training dataset is a dataset having attributes and class labels and used for training machine learning algorithms to prepare models.
- Machines can learn when they observe enough data. Using this one can model algorithms to find relationships, detect patterns, understand complex problems and make decisions.
- Training error is error that occurs by applying the model to same data from which model is trained.
- In a simple way actual output of training data and predicted output of model does not match training error.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

(b). Testing Phase -

- Testing dataset is provided as input to this phase.
- Test dataset is a dataset for which class label is unknown. It is tested using model.
- A test dataset used for assessment of finally chosen model.
- Training and testing dataset as completely different.
- The actual output of testing data and predicted output of model does not match the testing error t_{out} is said to have occured.

(c). Generalization -

- It is the prediction of future based on past system
- It needs to generalize beyond the training data to some future data that it might not have seen yet.
- The ultimate aim of machine learning is to minimize generalization error.
- The fit method is called on the training set to build the model.
- This fit method is applied to the model. on test set to estimate the target value and evaluate model's performance.

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

• Algorithm : (Boston dataset).

Step 1 : Import libraries and create alias for Pandas, Numpy, Matplotlib.

```
import numpy as np.
import pandas as pd.
import matplotlib.pyplot as ptt.
```

Step 2 : Import Boston housing dataset.

```
from sklearn.datasets import load_boston
boston = load_boston()
```

Step 3: Initialize the data frame.

```
data = pd.dataFrame( boston.data)
```

Step 4: Add feature names to dataframe.

```
data.columns = boston.feature_names.
```

Step 5 : Adding target variable to dataframe.

```
data ['PRICE'] = boston.target.
```

Step 6: Perform data pre-processing

```
data.isnull().sum().
```

Step 7: Split dependent variable and independent variables

```
x = data.drop (['PRICE'] , axis = 1)
y = data ['PRICE']
```

Step 8: Splitting data to training and testing data

A.I.S.S.M.S
COLLEGE OF ENGINEERING
NO - 1, Kennedy Road, Near R.T.O Pune - 411001

Step 9 : Use linear regression. (Train the machine) to create model.

```
lm = LinearRegression().
model = lm.fit(xtrain, ytrain).
```

Step 10: Predict the y-pred for all values of train-x and test-x.

```
ytrain-pred = lm.predict(xtrain)
ytest-pred = lm.predict(xtest).
```

Step 11: Evaluate the performance of model.

Step 12: Calculate mean-squared-error() for train-y, test-y

```
mse = mean_squared_error(ytest, ytestpred).
```

Step 13 : Plotting the linear regression model.

```
plt.scatter( ----- ).
```

* **Conclusion :** In this way we have done data analysis using linear regression for boston dataset and predict price of houses using the features of the Boston dataset.