**Aim:**

To visualize the data and find the regression application in the dataset

To implement linear regression in statistical method

To implement linear regression in scikit learn

**Theory:**

**What is Linear Regression?**

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.

**Regression Methods**

**Simple Linear Regression**

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship.

Y(pred) = b0 + b1\*x

The values b0 and b1 must be chosen so that they minimize the error. If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

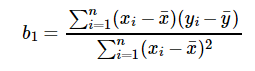
https://cdn-images-1.medium.com/max/800/1*Utp8sgyLk7H39qOQY9pf1A.png

For model with one predictor,

Intercept Calculation:

https://cdn-images-1.medium.com/max/800/1*1evY0PuCUENCpDP_QRplig.png

Slope Calculation:



**Optimizing using gradient descent:**

Complexity of the normal equation makes it difficult to use, this is where gradient descent method comes into picture. Partial derivative of the cost function with respect to the parameter can give optimal co-efficient value.:

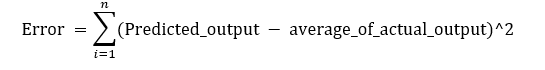
***Metrics for model evaluation***

***R-Squared value***

This value ranges from 0 to 1. Value ‘1’ indicates predictor perfectly accounts for all the variation in Y. Value ‘0’ indicates that predictor ‘x’ accounts for no variation in ‘y’.

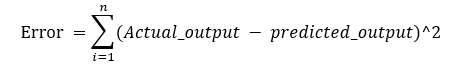
1. Regression sum of squares (SSR)

This gives information about how far estimated regression line is from the horizontal ‘no relationship’ line (average of actual output).



2. Sum of Squared error (SSE)

How much the target value varies around the regression line (predicted value).



3. Total sum of squares (SSTO)

This tells how much the data point move around the mean.



4. R Square



5. ***Correlation co-efficient (r)***

This is related to value of ‘r-squared’ which can be observed from the notation itself. It ranges from -1 to 1.

r = (+/-) sqrt(r²)

If the value of b1 is negative, then ‘r’ is negative whereas if the value of ‘b1’ is positive then, ‘r’ is positive. It is unitless.

**Multiple Linear Regression**

To examine the research question, a multiple linear regression will be conducted to assess if the independent variables predict the dependent variable (criterion). A multiple linear regression assesses the relationship among a set of dichotomous, or ordinal, or interval/ratio predictor variables on an interval/ratio criterion variable. In this instance, the independent variables include independent variable 1, independent variable 2, and independent variable 3 and the dependent variable is dependent variable. The following regression equation (main effects model) will be used: y = b1\*x1 + b2\*x2 +b3\*x3+…+ c; where Y = estimated dependent variable, c = constant (which includes the error term), b = regression coefficients and x = each independent variables.

The assumptions of multiple regression—linearity, homoscedasticity and multicollinearity—will be assessed. Linearity assumes a straight line relationship between the predictor variables and the criterion variable, and homoscedasticity assumes that scores are normally distributed about the regression line. Linearity and homoscedasticity will be assessed by examination of a scatter plot. The absence of multicollinearity assumes that predictor variables are not too related and will be assessed using Variance Inflation Factors (VIF). VIF values over 10 will suggest the presence of multicollinearity.

**Experiment:**

**Linear Regression Problem Definition:**

Predict the value of NOX given: CRIM.

Data Analysis Report :

|  |  |  |
| --- | --- | --- |
| Predict the NOX, given: CRIM | | |
| Mean\_Squared\_Error: | 0.012 | Less value indicates the model is good |
| R2\_Score: | 0.98 | High value indicates the variables are highly correlated |
| Karl Pearson Coefficient: | 0.99999999999978 | High value indicates the variables are highly correlated and positive indicates that the slope of the line is positive (NOX value increases with value of CRIM). |

**Conclusion:** The data analysis was performed on IRIS dataset and the following linear relationship was identified:

* NOX value is linearly related to CRIM value.