ST.FRANCIS INSTITUTE OF TECHNOLOGY

Mount Poinsur, S.V.P. Road, Borivli (West), Mumbai - 400103

Computer Engineering Department
Experiment No. 02

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.

Real-time example

We have a dataset which contains information about relationship between ‘number of hours studied’ and ‘marks obtained’. Many students have been observed and their hours of study and grade are recorded. This will be our training data. Goal is to design a model that can predict marks if given the number of hours studied. Using the training data, a regression line is obtained which will give minimum error. This linear equation is then used for any new data. That is, if we give number of hours studied by a student as an input, our model should predict their mark with minimum error.

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.

$$\text{Error} = \sum_{i=1}^{n}(actual\_output - predicted\_output) ** 2$$

For model with one predictor,
Intercept Calculation:

$$b_0 = \bar{y} - b_1\bar{x}$$

Slope Calculation:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Exploring 'b1'

- If b1 > 0, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.
- If b1 < 0, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.

Exploring 'b0'

- If the model does not include x=0, then the prediction will become meaningless with only b0. For example, we have a dataset that relates height(x) and weight(y). Taking x=0(that is height as 0), will make equation have only b0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.
- If the model includes value 0, then 'b0' will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.

- The value of b0 guarantee that residual have mean zero. If there is no 'b0' term, then regression will be forced to pass over the origin. Both the regression co-efficient and prediction will be biased.

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.:

```
#gradient descent
def grad_descent(s_slope, s_intercept, l_rate, iter_val, x_train, y_train):

    for i in range(iter_val):
        int_slope = 0
        int_intercept = 0
        n_pt = float(len(x_train))


        for i in range(len(x_train)):
            int_intercept = - (2/n_pt) * (y_train[i] - ((s_slope * x_train[i]) + s_intercept))
            int_slope = - (2/n_pt) * x_train[i] * (y_train[i] - ((s_slope * x_train[i]) + s_intercept))

        final_slope = s_slope - (l_rate * int_slope)
        final_intercept = s_intercept - (l_rate * int_intercept)
        s_slope = final_slope
        s_intercept = final_intercept


    return  s_slope, s_intercept
```

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).

$$\text{Error} = \sum_{i=1}^{n} (\text{Predicted\_output} - \text{average\_of\_actual\_output})\text{^}2$$

2. Sum of Squared error (SSE)

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

$$\text{Error} = \sum_{i=1}^{n} (Actual\_output - predicted\_output)\text{^}2$$

3. Total sum of squares (SSTO)

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

$$\text{Error} = \sum_{i=1}^{n} (\text{Actual\_output} - \text{average\_of\_actual\_output})\text{^}2$$

4. R Square

$$R^2 = 1 - (SSE/SSTO)$$

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.

$r = (+/-) \text{ sqrt}(r^2)$

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.

ST.FRANCIS INSTITUTE OF TECHNOLOGY

Mount Poinsur, S.V.P. Road, Borivli (West), Mumbai - 400103

Computer Engineering Department

Experiment No. 02

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.

Variables will be evaluated by what they add to the prediction of the dependent variable which is different from the predictability afforded by the other predictors in the model. The F-test will be used to assess whether the set of independent variables collectively predicts the dependent variable. R-squared—the multiple correlation coefficient of determination—will be reported and used to determine how much variance in the dependent variable can be accounted for by the set of independent variables. The t test will be used to determine the significance of each predictor and beta coefficients will be used to determine the magnitude of prediction for each independent variable. For significant predictors, every one unit increase in the predictor, the dependent variable will increase or decrease by the number of unstandardized beta coefficients.

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.

ST.FRANCIS INSTITUTE OF TECHNOLOGY

Mount Poinsur, S.V.P. Road, Borivli (West), Mumbai - 400103

Computer Engineering Department
Experiment No. 02

Experiment:

Linear Regression Problem Definition:

Predict the value of NOX given: CRIM.

(Other simple linear regression and multilinear regression definition can be added after data analysis)

Data Analysis Report :

Predict the Petal Width, given: Petal Length		
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).

(// Make this report for other simple linear regression and multilinear regression)

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

- NOX value is linearly related to CRIM value.