MULTIPLE LINEAR REGRESSION

Multiple linear regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data. The steps to perform multiple linear regression are almost similar to that of simple linear regression. The difference lies in the evaluation. You can use it to find out which factor has the highest impact on the predicted output and how different variables relate to each other.

Dependent variable

multiple independent variables

$$y = b_\circ + b_1x_1 + b_2x_2 \dots\dots b_nx_n$$

ASSUMPTIONS

FOR A SUCCESSFUL REGRESSION ANALYSIS, IT'S ESSENTIAL TO VALIDATE THESE ASSUMPTIONS.

1. Linearity: The relationship between dependent and independent variables should be Linear.
2. Homoscedasticity (constant variance) of the errors should be maintained.
3. Multivariate Normality: Multiple regression assumes that the residuals are normally distributed.
4. Lack of Multicollinearity: It is assumed that there is little or no multicollinearity in the data. Multicollinearity occurs when the features (or independent variables) are not independent of each other.

NOTE

Having too many variables could potentially cause our model to become less accurate, especially if certain variables have no effect on the outcome or have a significant effect on other variables. There are various methods to select the appropriate variable like -

1. Forward Selection
2. Backward Elimination
3. Bi-directional Comparision

DUMMY VARIABLES

Gender
Female
Female
Male
Female
Male
Male
Male

Male	Female
0	1
0	1
1	0
0	1
1	0
1	0
1	0

Using categorical data in Multiple Regression Models is a powerful method to include non-numeric data types into a regression model.

Categorical data refers to data values which represent categories - data values with a fixed and unordered number of values, for instance, gender (male/female). In a regression model, these values can be represented by dummy variables - variables containing values such as 1 or 0 representing the presence or absence of the categorical value.

DUMMY VARIABLE TRAP

The Dummy Variable trap is a scenario in which two or more variables are highly correlated; in simple terms, one variable can be predicted from the others. Intuitively, there is a duplicate category: if we dropped the male category it is inherently defined in the female category (zero female value indicate male, and vice-versa).

The solution to the dummy variable trap is to drop one of the categorical variables - if there are m number of categories, use m-1 in the model, the value left out can be thought of as the reference value.

Dummy variable

Dummy variable

$$D_2 = 1 - D_1$$

$$y = b_\circ + b_1x_1 + b_2x_2 + b_3D_1$$

1 PREPROCCESS THE DATA

- Import the Libraries.
- Import the DataSet.
- Check for Missing Data.
- Encode Categorical Data
- Make Dummy Variables if necessary and avoid dummy variable trap.
- Feature Scaling will be taken care by the Library we will use for Simple Linear Regression Model.

2 FITTING OUR MODEL TO THE TRAINING SET

This step is exactly the same as for simple linear regression. To fit the dataset into the model we will use LinearRegression class from sklearn.linear_model library. Then we make an object regressor of LinearRegression Class. Now we will fit the regressor object into our dataset using fit() method of LinearRegression Class.

3 PREDICTING THE TEST RESULTS

Now we will predict the observations from our test set. We will save the output in a vector Y_pred. To predict the result we use predict() method of LinearRegression Class on the regressor we trained in the previous step.