Linear Regression

Epoch IIT Hyderabad

Arin Aggarwal
MA22BTECH11006

Contents

1 Introduction

Linear Regression is a Machine Learning Algorithm used to predict value of a variable (called the dependent variable) using one or more variables(called the independent variables). The aim is to find the best fit line for the distribution of variables. This line is decided upon using the concept of least squares which is calculated relatively as r-squared (R^2) with (R^2) of 1 being best and 0 being worst. The algorithm gives the line in form of θ_s , the coefficients of regression line

2 Assumptions of Regression

1) **Linearity** : Linear regression needs the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatter plots.
2) **Outliers** : Make sure that there are no significant outliers in the data as they may skew the results.
3) **Multicollinearity** : The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity. Multicollinearity can be checked by 2 main criterions:
   (i **Correlation Matrix** : When computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be close to 0. +1 indicates perfect positive correlation and -1 perfect negative correlation.
   (ii **Variance Inflation Factor(VIF)** : A VIF value of >10 indicates presence of multicollinearity and thus the given variable should be removed.
4) **Homoscedasticity** : The variance of the errors is constant across all levels of the independent variables. This can be checked using Goldfeld-Quandt Test or Breusch-Pagan Test.
5) **Autocorrelation** : Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of $y(x + 1)$ is not independent from the value of $y(x)$. It can be checked using Durbin-Watson Statics that scores from 0.0 to 4.0 with score of 2.0 indicating no autocorrelation.
6) **Normality** : Data needs to be multi-variate normal. This can be checked using Histograms(by checking if the skew is close to 0) or using QQ-Plots. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.

3 Mathematical Formulations

3.1 Line Formula

The Dependent Variable is denoted by $h_\theta(x)$ given by

$$h_\theta(x) = \theta_0 + \theta_1 x$$

where x is the independent variable,θ_0 denotes the y-intercept and θ_1 denotes the slope (for single independent variable regression).

3.2 R^2 Formula

R^2 requires calculation of Residual Sum of Squares (SS_{res}) and Total Sum of Squares (SS_{tot}).

$$SS_{res} = \sum_i (\theta_0 + \theta_1 x - y_i)^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

where

$$Mean(\bar{y}) = \frac{1}{n} \sum_1^n (y_i)$$

then,

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

4 Optimal Conditions

Check p-values of the regression to see if some of the coefficients are insignificant. Features with p-value > 0.05 are insignificant. Models should be made by removing these features and further checked using Bayesian Information Criterion(BIC). The lesser the BIC value, the better the model.