Linear Regression

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Linear Regression

- Linear regression is a simple statistical regression method used for predictive analysis that shows the relationship between the continuous independent variable (X-axis) and the continuous dependent variable (Y-axis).
- In regression, we try to calculate the best fit line which describes the relationship between the predictors and dependent variable.
- The equation of best fit line is **Y=a+b*X + e,** where a is intercept, b is slope of the line and e is error term.
- When we have one independent variable, we call it Simple Linear Regression (SLR). If the number of independent variables is more than one, we call it Multiple Linear Regression (MLR).
- SLR Example: You are a social researcher interested in the relationship between income and happiness.
- MLR Example: The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Assumptions

Simple Linear Regression:

- **Linearity:** The relationship between independent variables and the mean of the dependent variable is linear.
- **Homoscedasticity:** The variance of residuals should be equal.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of an independent variable, the dependent variable is normally distributed.

Multi Linear Regression:

- **Linearity:** The relationship between independent variables and the mean of the dependent variable is linear.
- **Multicollinearity:** There should not be high correlation between two or more independent variables. Multicollinearity can be checked using correlation matrix, Tolerance and Variance Inflation Factor (VIF).

$$VIF_i = \frac{1}{1 - R_i^2} = \frac{1}{Tolerance}$$

- **Homoscedasticity:** The variance of residuals should be equal.
- **Multivariate Normality:** Residuals should be normally distributed.
- **Categorical Data:** Any categorical data present should be converted into dummy variables.
- **Minimum records:** There should be at least 20 records of independent variables.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaj@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Linear Regression Model Representation

- The representation is a linear equation that combines a set of input values (x) and predicted output (y).
- As such, both the input values (x) and the output value are numeric.
- For example, in a simple regression problem (a single x and a single y), the form of the model would be

$$Y = \beta 0 + \beta 1x$$

- For example, in a multi linear regression problem, the form of the model would be

$$Y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 \ldots\ldots + \beta nxn$$

- In higher dimensions when we have more than one input (x), the line is called a plane or a hyperplane
- The coefficients β0,β1… βn are estimated using the Ordinary Least Square (OLS). The goal behind this is to minimize the squared difference between actual and predict values.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Performance of Regression

Mean Absolute Error (MAE)

By using MAE, we calculate the average absolute difference between the actual values and the predicted values.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

Mean Square Error (MSE)

Mean Square Error (MSE) is defined as Mean or Average of the square of the difference between actual and estimated values.

$$MSE = \frac{1}{N}\sum_{i}^{n}(Y_i - y_i)^2$$

Root Mean Square Error (RMSE)

RMSE calculates the square root average of the sum of the squared difference between the actual and the predicted values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

R-squared values

R-square depicts the percentage of the variation in the dependent variable explained by the independent variable in the model.

$$\text{R-squared} = 1 - \frac{RSS}{TSS}$$

RSS = Residual sum of squares: It is the measure of the difference between the expected and the actual output. It is also defined as follows:

$$RSS = \sum (y_i - \hat{y}_i)^2$$

TSS = Total sum of squares: It is the sum of errors of the data points from the mean of the response variable.

$$TSS = \sum (y_i - \bar{y})^2$$

Higher the R-square value better the model. The value of R^2 increases if we add more variables to the model irrespective of the variable contributing to the model or not. This is the disadvantage of using R^2.

Adjusted R-squared values

Adjusted R^2 value will improve only if the added variable is making a significant contribution to the model. Adjusted R^2 value adds penalty in the model.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

where R^2 is R-square value, n = total number of observations, and k = total number of variables used in the model.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Advantages and Disadvantages

Advantages

- Linear regression performs exceptionally well for linearly separable data
- Easy to implement and train the model
- It can handle overfitting using dimensionality reduction techniques and cross validation and regularization

Disadvantages

- Sometimes Lot of Feature Engineering Is required
- If the independent features are correlated it may affect performance
- It is often quite prone to noise and overfitting

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

1. When you will use linear regression?

Linear Regression analysis is used when you want to predict a continuous dependent variable from a number of independent variables.

2. How do you know explain the linear regression to layman terms?

Linear regression is a way to explain the relationship between a dependent variable and one or more explanatory variables using a straight line.

3. What is heteroscedasticity?

Heteroscedasticity is exactly the opposite of homoscedasticity, which means that the error terms are not equally distributed. To correct this phenomenon, usually, a log function is used.

4. What is the difference between R square and adjusted R square?

R square and adjusted R square values are used for model validation in case of linear regression. R square indicates the variation of all the independent variables on the dependent variable. i.e. it considers all the independent variable to explain the variation. In the case of Adjusted R squared, it considers only significant variables(P values less than 0.05) to indicate the percentage of variation in the model.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

5. What are the possible ways of improving the accuracy of a linear regression model?

There could be multiple ways of improving the accuracy of a linear regression, most commonly used ways are as follows:

Outlier Treatment: Regression is sensitive to outliers, hence it becomes very important to treat the outliers with appropriate values. Replacing the values with mean, median, mode or percentile depending on the distribution can prove to be useful.

6. What if data is not normally distributed?

Using transformation like Power Law transformation, Log Normal, Box-Cox, Exponential transformation etc.,

7. Whether feature scaling is required in linear regression?

Yes, feature scaling is required if you are using gradient descent for creating linear regression.

8. How the best fit line is selected?

The least Sum of Squares of Errors is used as the cost function for Linear Regression. For all possible lines, calculate the sum of squares of errors. The line which has the least sum of squares of errors is the best fit line.

Proprietary content. ©Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited
This file is meant for personal use by venkatabaji@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.