What is Linear Regression in ML?

Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable.

The linear regression is based on certain statistical assumptions. It is crucial to check these regression assumptions before modeling the data using the linear regression approach.

What are Linear Regression assumptions?

There are 4 assumptions.

1. **Linearity:** The relationship between independent (X) and dependent (Y) must be linear.
2. **Independence:** Observations are independent of each other.
3. **Homoscedasticity:** The variance of residual is the same for any value of X.
4. **Normality** of the residuals.

Assumption # 1:

Linearity: The relationship between independent (X) and dependent (Y) must be linear.

How does it affect the model?

If the relationship is non-linear, then the predictions made by the linear regression model will not be accurate and will vary from the actual observations a lot. For the non-linear data, do not use a linear model, but rather use a non-linear model of which plenty exist.

How to find the relationship?

You can check for linear relationships easily by making a scatter plot for each independent variable with the dependent variable.

How to Fix?

To fix non-linearity, one can either do log transformation of the Independent variable, log(X) or other non-linear transformations like √X or X^2.

Assumption # 2:

No or little/low multicollinearity / Independence: Observations are independent of each other.

Multicollinearity is when independent variables in a regression model are correlated. One of the key assumptions for a regression-based model is that the independent/explanatory variables should not be correlated amongst themselves.

How do we measure Multicollinearity?

A very simple test known as the VIF test is used to assess multicollinearity in our regression model. The variance inflation factor (VIF) identifies the strength of correlation among the predictors.

How to fix Multicollinearity?
To reduce/fix multicollinearity, remove the column(s) with the highest VIF and check the results.

How it affects the model?
Although multicollinearity does not affect the regression estimates, it makes them vague, imprecise, and unreliable.

Assumption # 3:

Homoscedasticity: The variance of residual is the same for any value of X.

Homoscedastic refers to a condition in which the variance of the residual, or error term, in a regression model is constant. That is, the error term does not vary much as the value of the predictor variable changes.

Homoscedasticity means to be of “The same Variance”. In Linear Regression, one of the main assumptions is that there is a Homoscedasticity present in the errors or the residual terms (Y_Pred – Y_actual).

How to detect?

The Breusch-Pagan test is used to determine whether or not heteroscedasticity is present in a regression model.

The test uses the following null and alternative hypotheses:

Null Hypothesis (H0): Homoscedasticity is present (the residuals are distributed with equal variance)
Alternative Hypothesis (HA): Heteroscedasticity is present (the residuals are not distributed with equal variance)

If the p-value of the test is less than some significance level (i.e. α = .05) then we reject the null hypothesis and conclude that heteroscedasticity is present in the regression model.

How to deal with it?

One way to deal with heteroscedasticity is to transform the dependent variable. Perform a log transformation on the variable and check again with White’s test.

Assumption # 4:
Normality of the residuals

The residuals should follow a normal distribution. Once you obtain the residuals from your model, this is relatively easy to test using either a histogram or a QQ Plot.