**“Rough notes – for reference only”**

**Understanding Linear Regression**

In the most simple words, 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(y) and independent variable(x).

Linear Regression is of two types: Simple and Multiple.

Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable

Whereas, In Multiple Linear Regression there are more than one independent variables for the model to find the relationship.

Equation of Simple Linear Regression, where bois the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

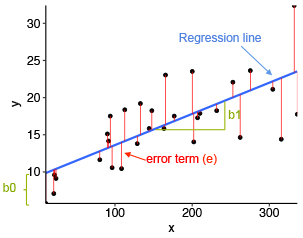
linear regression 1

Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

linear regression 2

A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.  
Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Let’s understand this with the help of a diagram.



linear regression 1

Image Source: Statistical tools for high-throughput data analysis

In the above diagram,

* x is our independent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
* Black dots are the data points i.e the actual values.
* bo is the intercept which is 10 and b1 is the slope of the x variable.
* The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
* The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

**Mathematical Approach:**

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

i.e

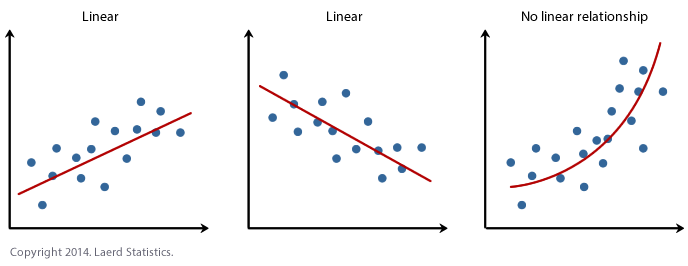
linear regression

Rsq, AdjRsq, MSE,RMSE,MAE – 5 evaluation metrics

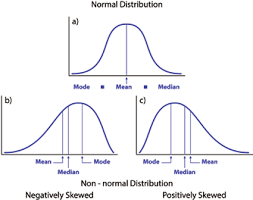
**Assumptions of Linear Regression –**

The basic assumptions of Linear Regression are as follows:

1. Linearity: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

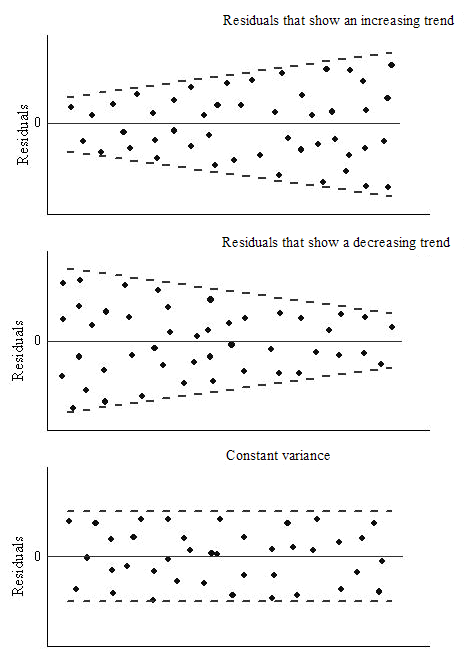


2. Normality: The X and Y variables should be normally distributed.



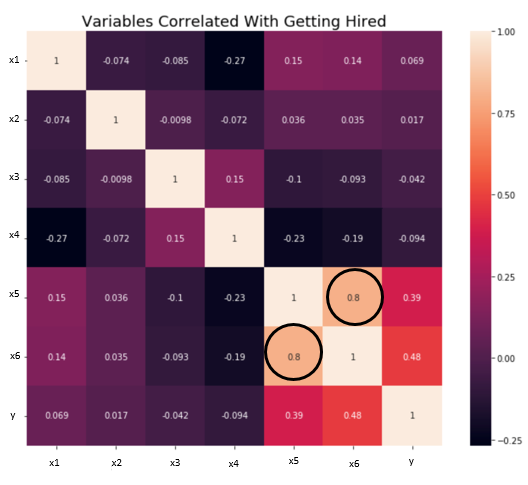
3. Homoscedasticity: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.

Error Term : y act – y pred



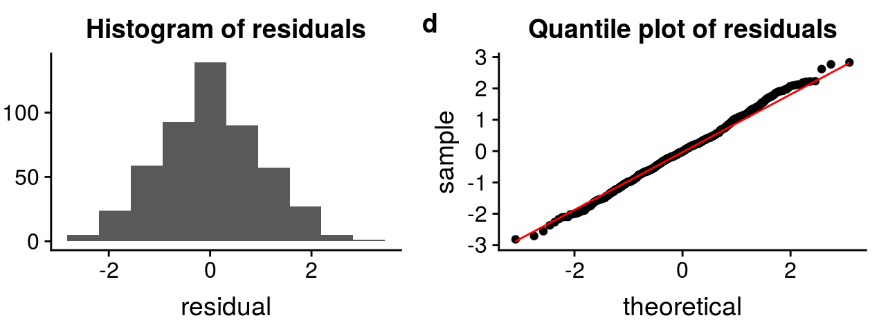
4. **Independence/No Multicollinearity**: The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.

In the below image, a high correlation is present between x5 and x6 variables.

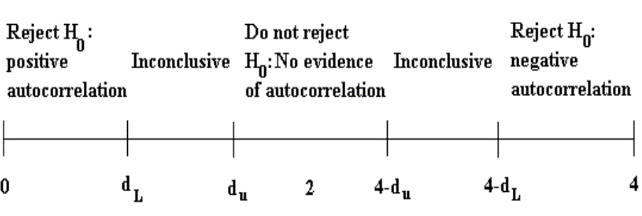


VIF / RFE ->

5. The error terms should be normally distributed. Histograms can be used to check the distribution of error terms.

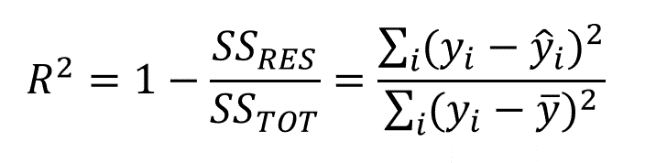


6. **No Autocorrelation**: The error terms(yact – ypred) should be independent of each other. Autocorrelation can be tested using the **Durbin Watson test**. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.



**Evaluation Metrics for Regression Analysis**

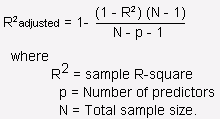
1. R squared or Coefficient of Determination: The most commonly used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.



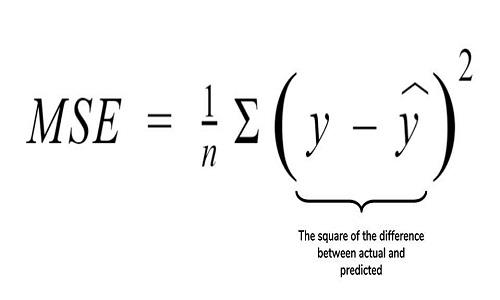
1 – (RSS/TSS)

where SSRES is the Residual Sum of squares and SSTOT is the Total Sum of squares

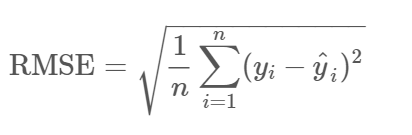
2. Adjusted R squared: It is the improvement to R squared. The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model and shows the real improvement of the model.  
Adjusted R2 is always lower than R2.



3. Mean Squared Error (MSE): Another Common metric for evaluation is Mean squared error which is the mean of the squared difference of actual vs predicted values.



4. Root Mean Squared Error (RMSE): It is the root of MSE i.e Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn’t.



5. MAE (Mean Absolute Error)