**“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.

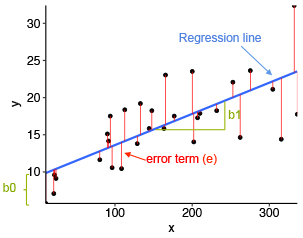
linear regression 1

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

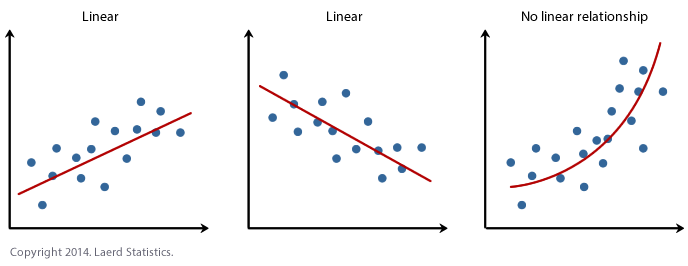
linear regression

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

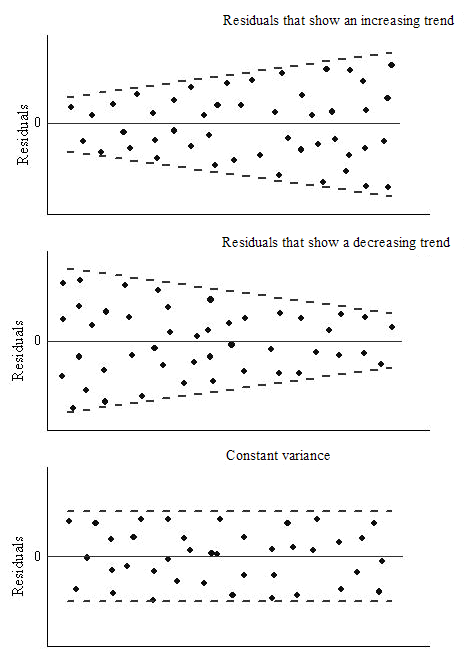
**Assumptions of Linear Regression –**

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.



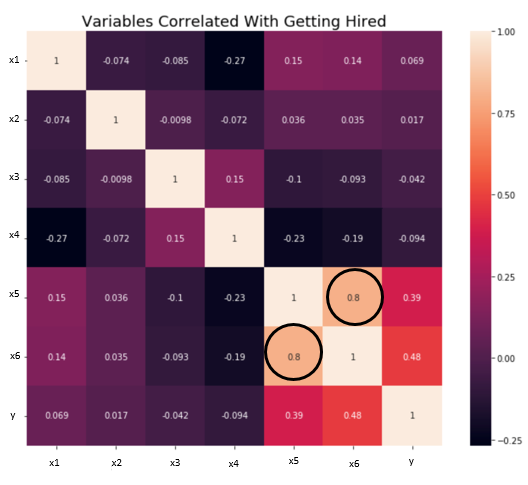
2. Homoscedasticity: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.



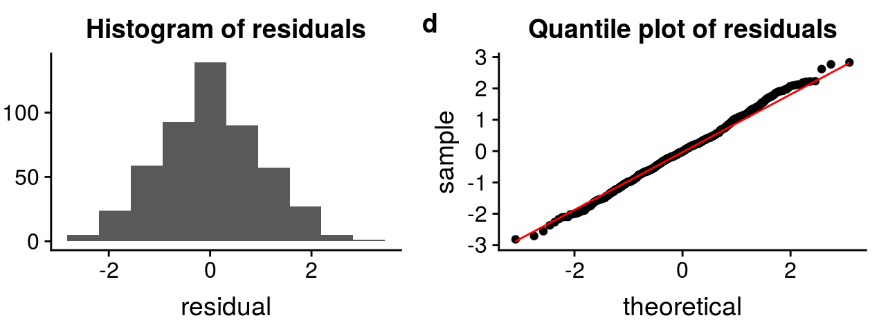
Error Term : y act – y pred

3. **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.

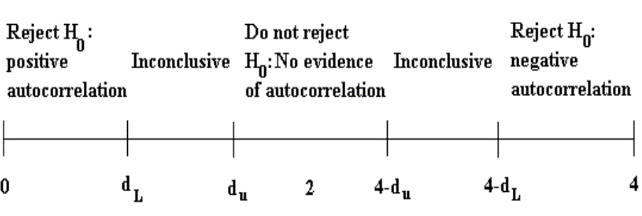
In the below image, a high correlation is present between x5 and x6 variables.



4. The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.



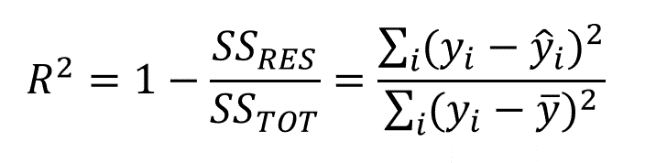
5. **No Autocorrelation**: The error terms(yact – ypred) should be independent of each other. Autocorrelation can be tested using the **Durbin Watson test**. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there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



—-----------—----------—----------—----------—----------—----------—----------—----------—----------—------

**Evaluation Metrics for Regression Analysis**

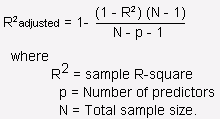
1. R squared or Coefficient of Determination: The most commonly used metric for model evaluation in regression analysis is R squared. It can be defined as a Ratio of variation to the Total Variation. The value of R squared lies between 0 to 1, the value closer to 1 the better the model.



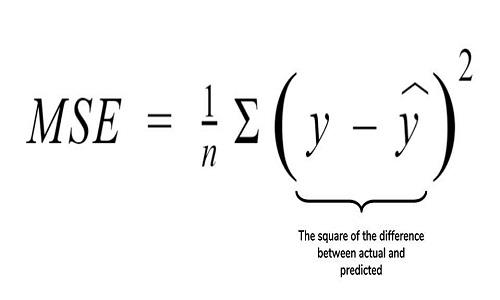
1 – (RSS/TSS)

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

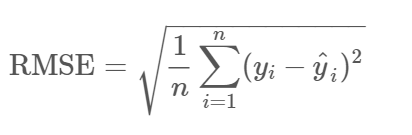
2. Adjusted R squared: It is the improvement to R squared. The problem/drawback with R2 is that as the features increase, the value of R2 also increases which gives the illusion of a good model. So the Adjusted R2 solves the drawback of R2. It only considers the features which are important for the model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

Mean absolute error (MAE) is a common measure of how far the predicted values are from the actual values in a dataset. It is the average absolute difference between the predicted and actual values.

The formula for calculating the MAE is:

MAE = (1/n) \* Σ|i=1 to n| (|actual\_i - predicted\_i|)

where n is the total number of observations, actual\_i is the actual value of the i-th observation, and predicted\_i is the predicted value of the i-th observation.

In simple terms, MAE measures the average distance between the predicted values and the actual values in a dataset, and it is often used in regression analysis and machine learning to evaluate the performance of a model. The lower the value of MAE, the better the model's performance.

**—-------------------—-------------------—-------------------—-------------------—-------------------—--------------**

**RMSE v/s MSE :**

In the context of RMSE (Root Mean Squared Error) and MSE (Mean Squared Error), "penalize" means that RMSE gives more weight to larger errors compared to smaller errors. This is because the RMSE takes the square root of the average of the squared differences between the predicted and actual values, which amplifies the impact of larger errors on the overall score.

In other words, RMSE considers the magnitude of the errors and gives them more weight in the calculation of the score, while MSE treats all errors equally, regardless of their size. Penalizing large errors means that they have a greater impact on the overall evaluation of the model's performance, and the model will be penalized more heavily if it makes larger errors.

**Durbin Watson :**

Autocorrelation is a statistical term that refers to the correlation of a signal with a delayed copy of itself over time. In other words, it is a measure of how a variable is correlated with its past values. Autocorrelation is often used in time-series analysis to identify patterns in the data and to make predictions about future values.

The Durbin-Watson test is a statistical test used to detect the presence of autocorrelation in a time series. It is based on the idea that if there is no autocorrelation, the residuals from a regression model will be random and uncorrelated with each other. The Durbin-Watson test statistic measures the degree of correlation between adjacent residuals in a regression model. The test statistic ranges from 0 to 4, with values close to 2 indicating no autocorrelation, values less than 2 indicating positive autocorrelation, and values greater than 2 indicating negative autocorrelation. The Durbin-Watson test is commonly used in econometrics and other fields where time-series analysis is important.

**Constant Variance (Homoscedasticity):**

Homoscedasticity is a statistical term that describes a situation where the variance of the errors (or residuals) of a linear regression model is constant across all levels of the predictor variable(s). In simpler terms, it means that the spread of the residuals is the same for all values of the independent variable.

In a linear regression model, the residuals represent the difference between the predicted values and the actual observed values. If the variance of these residuals is constant for all values of the independent variable(s), then the model is said to be homoscedastic. On the other hand, if the variance of the residuals is not constant, but instead changes with the level of the independent variable(s), then the model is said to be heteroscedastic.

Homoscedasticity is an important assumption of linear regression models. Violations of this assumption can lead to biased and inefficient estimates of the model parameters, as well as inaccurate predictions.

–

Homoscedasticity refers to the assumption that the variance of the errors (or residuals) in a linear regression model is constant across all levels of the predictor variable(s). In other words, the spread of the residuals should be roughly equal across the range of values of the independent variable(s).

If the assumption of homoscedasticity is violated, and the residuals show a pattern of increasing or decreasing variance across the range of the independent variable(s), it can lead to problems with the validity and reliability of the linear regression model. Specifically, heteroscedasticity can lead to biased coefficient estimates, inflated standard errors, and lower statistical power.

When heteroscedasticity is present, the least squares estimation method used to fit the linear regression model may not produce the best estimates of the coefficients, leading to biased predictions. In addition, if the standard errors of the coefficients are incorrectly estimated, confidence intervals and hypothesis tests may be unreliable.

Therefore, it is important to test for homoscedasticity before relying on the results of a linear regression model. If heteroscedasticity is present, there are several methods to address it, such as using weighted least squares regression, transforming the dependent or independent variables, or using a different regression model altogether, such as a robust regression model.