# Linear Regression:

Approach to model the relationship between scalar dependent ‘y’ and one or more exploratory variables ‘X’.

1. It’s a prediction of real continuous number
2. **Linearity:**

In linear regression we assume that there is a linear relationship between dependent variable and independent variables. This is called Null hypothesis. To check we go with:

P value < 0.05. T-value is more, than the variable is significant

If the P values is above 0.05 we will not include that variable in the model.

1. **Multi Collinearity:** This can be figured out using VIF value, which should be less than 5 (actually between -5 and 5). This will help to identify redundant variables and highly correlated variables can be ignored.
2. **Homoscedacity**: If Error is uniform across the data than data is Homoscedastic, if not it’s **Heteroscedastic:** Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model’s performance.
3. **Normality: How error is distributed? Error terms must be normally distributed (residual values).**
4. **We have AIC/BIC:** values to compare two or more models. The bigger the AIC /BIC value the better the model is
5. Two compare two different models for the same dataset with different variables :

**AIC (Akaic Information Criteria) and BIC (Bayesian Information Criteria):** Gives info about R square value and error term. It is used to compare 2 or more models. Bigger the value better the model is.

1. **R**- Squared value gives how much variance model has learnt or explained and statistical measure of how close the data are to be the fitted regression
2. **Adjusted R Squared**: Modified version of the R-Squared that has been adjusted for the predictors in the model

If we increase the R2 value there should be increase in Adjusted R2, else it is not significant increase. To increase the R2 value we generally add more rows or remove some data and Change model parameters like transforming the variables (Applying Log or square or cube to build polynomial equation)

**How do you check the model performance?**

1. Cross Validation: In cross validation we split the data set into train and test data. (70-30)

* We train the model on *Train dataset* (Which has X,y)and then we take the model and apply it on *Test dataset(X).*

*Note: X can be any number of variables (both categorical and numerical)*

* We will also find MAPE – the error percentage should be less or almost 0.

