**Steps in Regression Model Building**

* Collect/Extract Data
* Pre-process the Data
* Divide the Data into Training and Validation Data Sets
* Define the Functional Form of Relationship
* Estimate the Regression Parameters
* Perform Regression Model Diagnostics
* Model Deployment

**The Assumptions in Regression Models**

* The regression model is linear in regression parameters.
* The expected value of the residuals is zero
* The residuals follow a normal distribution. For estimation of regression parameters, the assumption of normal distribution for errors is not necessary. However, it is essential for testing hypotheses such as whether there is a statistically significant association relationship between the outcome variable and the features.
* The variance of the residuals is constant for all values of X . When the variance of the residuals is constant for different values of X , it is called homoscedasticity. A non-constant variance of residuals is called heteroscedasticity.

**THE MODEL DIAGNOSITICS**

**Hypothesis Test for Regression Co-Efficient**

The regression co-efficient (b1 ) captures the existence of a linear relationship between the response variable and the explanatory variable. If b1 = 0, we can conclude that there is no statistically significant linear relationship between the two variables.

The null and alternative hypotheses for the SLR model can be stated as follows:

H0 : There is no relationship between X and Y

HA: There is a relationship between X and Y

b1 = 0 would imply that there is no linear relationship between the response variable Y and the explanatory variable X. Thus, the null and alternative hypotheses can be restated as follows:

H0 : b1 = 0

HA: b1 ≠ 0

If the p-value is less than 0.05 (or an appropriate significance value), we reject the null hypothesis and conclude that there is significant evidence suggesting a linear relationship between X and Y.

(remember, the p-value gets smaller as the test statistic calculated from the data gets further away from the center which is zero as predicted by the null hypothesis)

**What is Homoskedasticity?**

Refers to a condition in which the variance of the residual, or[error term](https://www.investopedia.com/terms/e/errorterm.asp), in a regression model is constant. That is, the error term does not vary much as the value of the predictor variable changes. Another way of saying this is that the [variance](https://www.investopedia.com/terms/v/variance.asp) of the data points is roughly the same for all data points.

This suggests a level of consistency and makes it easier to model and work with the data through regression; however, the lack of homoskedasticity may suggest that the regression model may need to include additional predictor variables to explain the performance of the dependent variable.

**What is Heterocedasticity**

Heteroskedasticity happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant.

With heteroskedasticity, the tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out (errors increase as the X or Y variable increases in magnitude)

**What is co-efficient of determination (R-squared)?**

The primary objective of regression is to explain the variation in Y using the knowledge of X. The coefficient of determination (or R-square or R2 ) measures the percentage of variation in Y explained by the model (b0 + b1 X).

R2 is the proportion of variation in response variable Y explained by the regression model. Coefficient of determination (R2 ) has the following properties:

* The value of R2 lies between 0 and 1.
* Higher value of R2 implies better fit, but one should be aware of spurious regression
* Mathematically, the square of correlation coefficient is equal to coefficient of determination.
* We do not put any minimum threshold for R-squared ; higher value of R-squared implies better fit

**Calculation of R-Squared**

R-Squared = SSR/SST

SSR: is the sum of squares due to regression (explained sum of squares)

SST: is the total sum of squares

Outlier Analysis:

The following distance measures are useful in identifying the influential observations:

* Z-Score
* Cook’s Distance
* Leverage Values

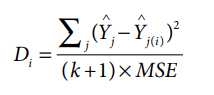
**Z-Score**

Z-score is the standardized distance of an observation from its mean value. For the predicted value of the dependent variable Y, the Z-score is given by

Ypred – Ymean/Std-Y

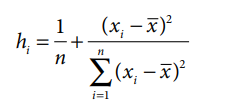
**Cook’s Distance**

Cook’s distance measures how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters.



**Leverage Value**

Leverage value of an observation measures the influence of that observation on the overall fit of the regression function.



Leverage value of more than 2/n or 3/n is treated as highly influential observation.

**F-Statistic**

Using the Analysis of Variance (ANOVA), we can test whether the overall model is statistically significant.

The null and alternative hypothesis for F-test are given by

H0 : There is no statistically significant relationship between Y and any of the explanatory variables (i.e., all regression coefficients are zero).

H1 : Not all regression coefficients are zero.

Alternatively:

H0 : All regression coefficients are equal to zero.

HA: Not all regression coefficients are equal to zero.

The F-statistic is given by

F = [SSR/k] / [SSE/n-k-1] = MSR/MSE

Where k is no. of parameters, n is no. of observations.

**T-Distribution**

The*t*-distribution, also known as Student’s *t*-distribution, is a way of describing data that follow a bell curve when plotted on a graph, with the greatest number of observations close to the [mean](https://www.scribbr.com/statistics/mean/) and fewer observations in the tails.



It is a type of [normal distribution](https://www.scribbr.com/statistics/normal-distribution/) used for smaller sample sizes, where the [variance](https://www.scribbr.com/statistics/variance/) in the data is unknown.

The *t*-distribution is used when data are *approximately*normally distributed, which means the data follow a bell shape but the population variance is unknown. The variance in a *t*-distribution is estimated based on the degrees of freedom of the data set (total number of observations minus 1).

It is a more conservative form of the [**standard normal distribution**](https://www.scribbr.com/statistics/standard-normal-distribution/), also known as the *z*-distribution. This means that it gives a lower probability to the center and a higher probability to the tails than the standard normal distribution.