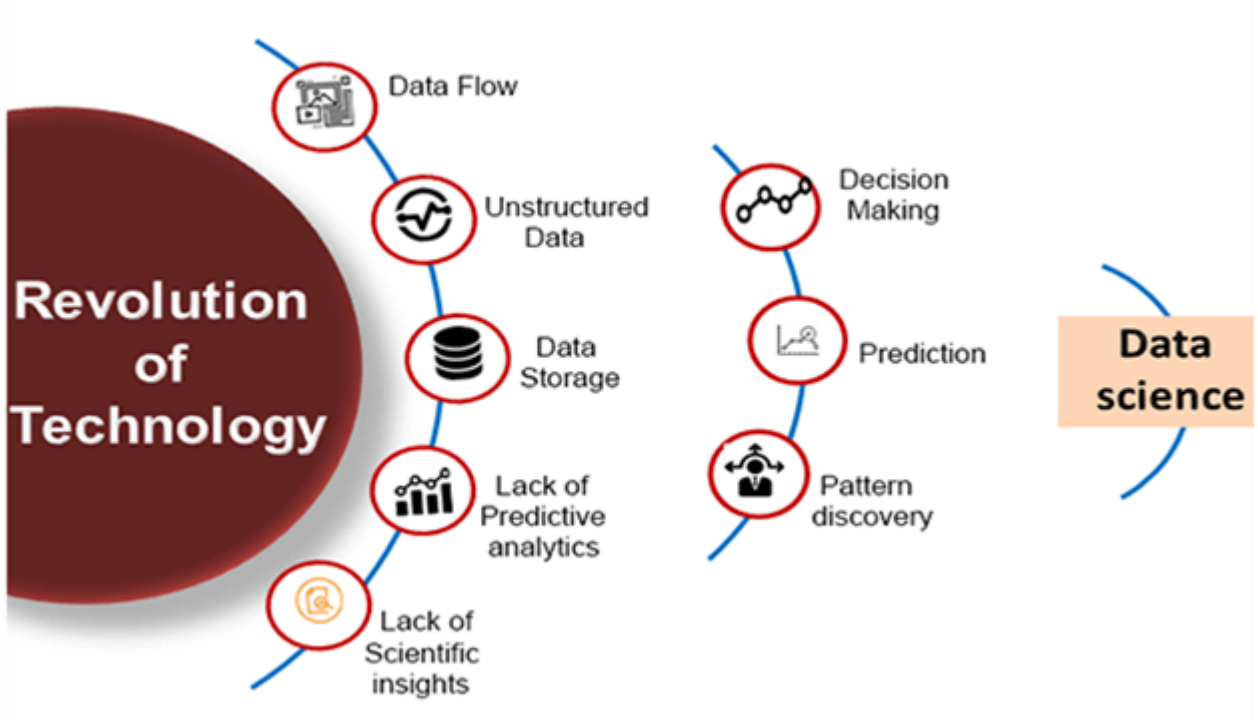
Introduction to Datascience

Introduction

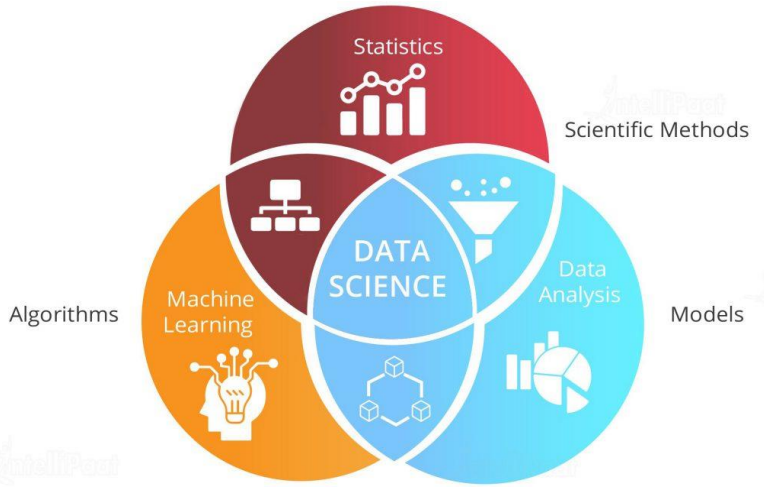
What is Datascience?

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms

Why do we need Datascience?

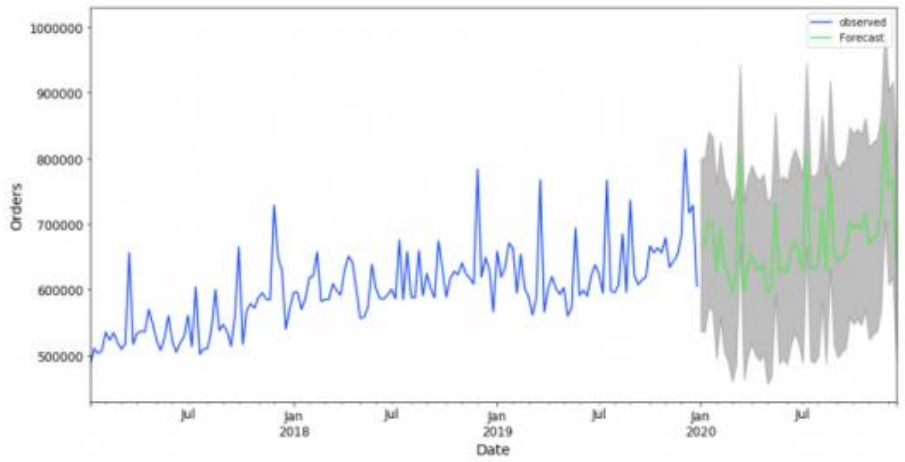


Datascience Overview

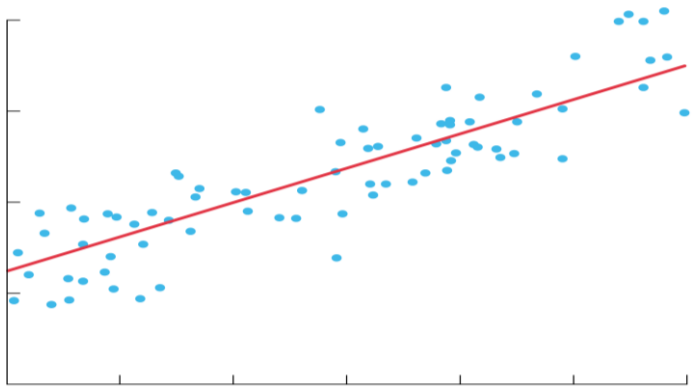


Possible Usecases

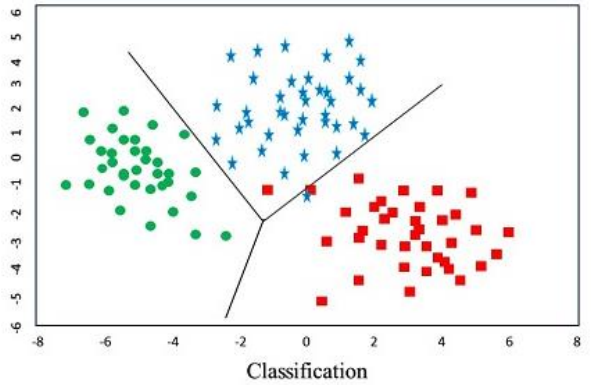
Demand Forecasting



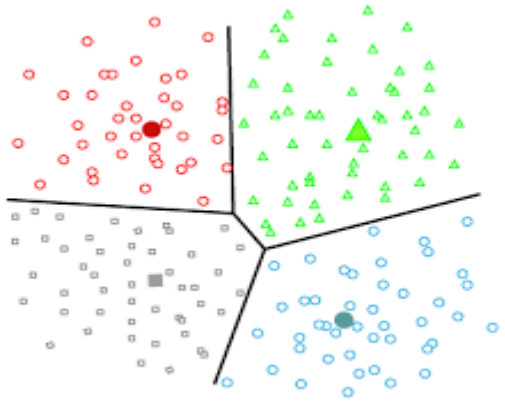
Claim Prediction



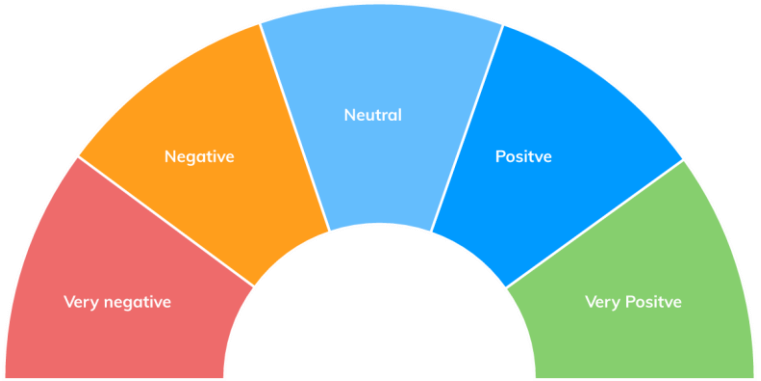
Churn Prediction



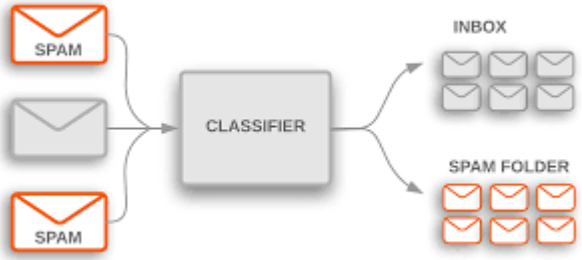
Customer Segmentation



Sentiment Analysis



Email Classification



Linear Regression

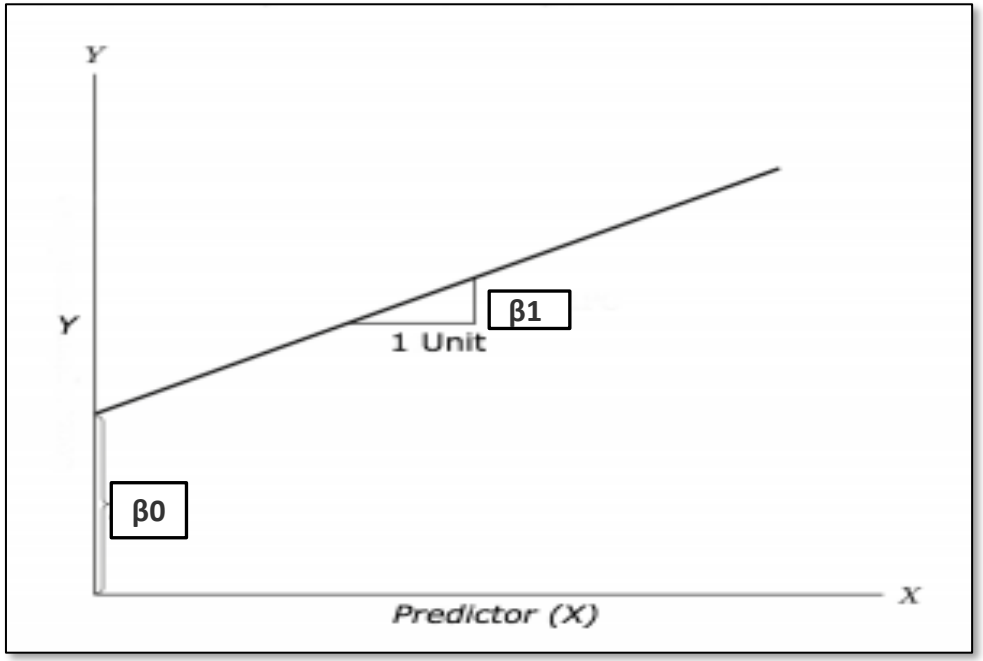
➢ What is Regression?

Regression attempts to predict one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables, usually denoted by X).

Linear Regression is a way of predicting a response Y on the basis of a single predictor variable X. It is assumed that there is approximately a linear relationship between X and Y. Mathematically, we can represent this relationship as:

$$Y = \beta_0 + \beta_1 X$$

where, **β0** is the intercept and **β1** is the slope.
Collectively, they are called regression coefficients and **ϵ** is the error term, the part of **Y** the regression model is unable to explain.

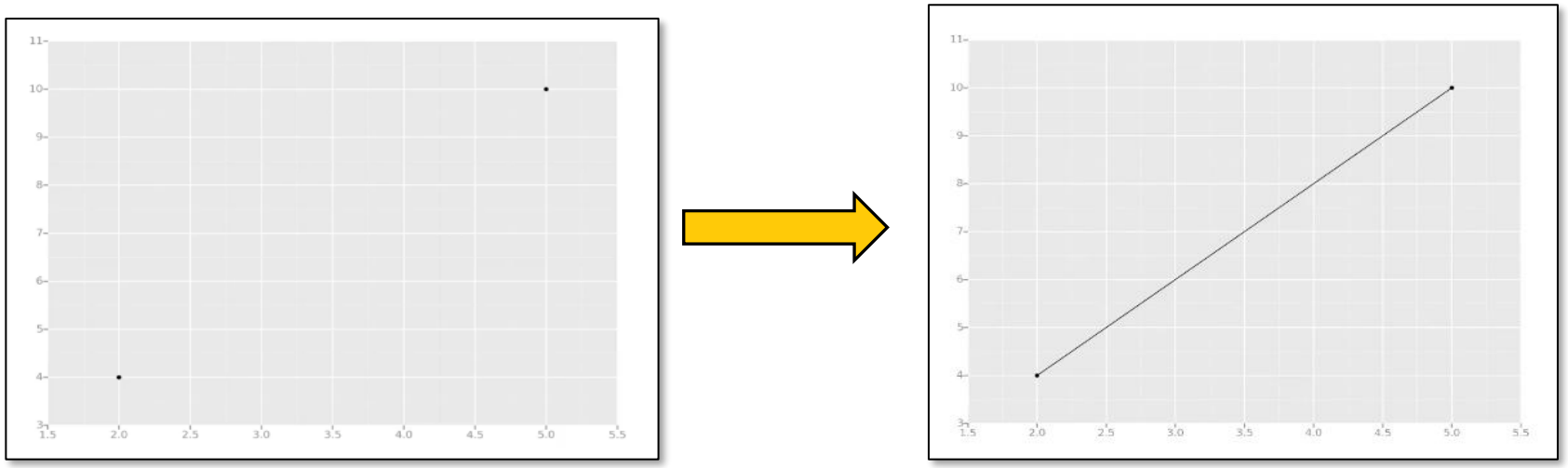


To find the parameters, we need to minimize the **least squares** or the **sum of squared errors**. Of course, the linear model is not perfect and it will not predict all the data accurately, meaning that there is a difference between the actual value and the prediction. The error is easily calculated with:

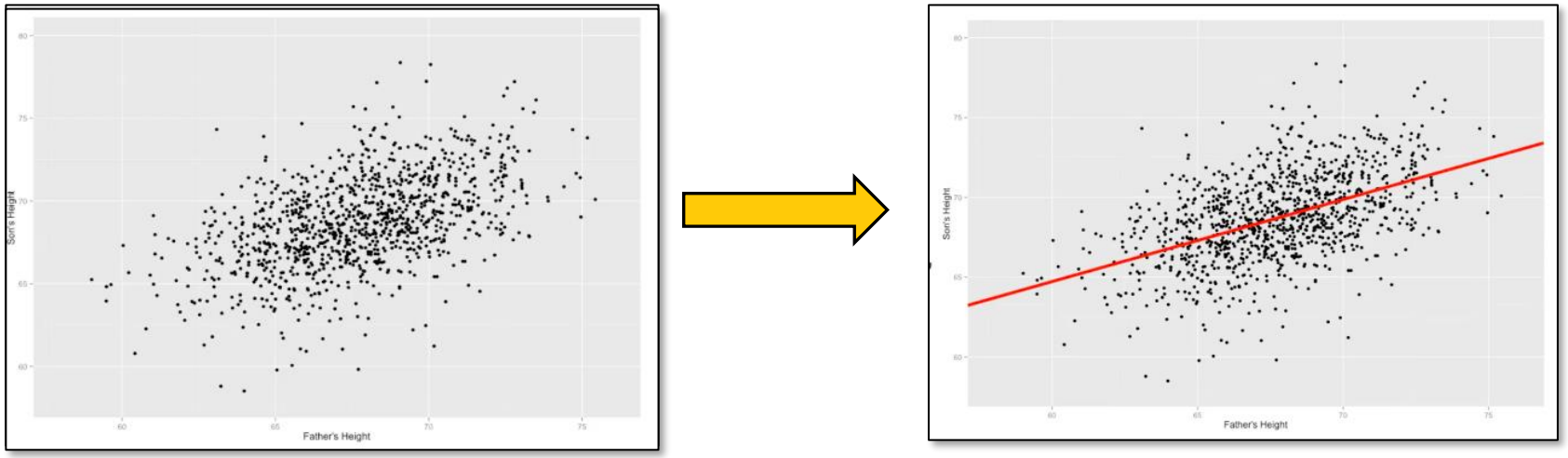
$$e_i = y_i - \hat{y}_i$$

Linear Regression

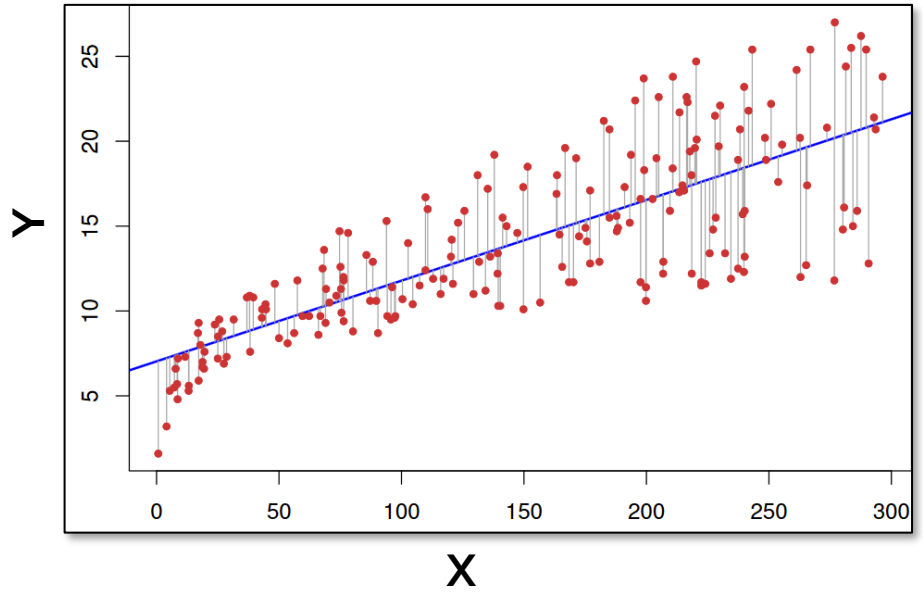
- Calculate the regression with only two data points. Here we have 2 data points represented by two black points. All we are trying to do when we calculate our regression line is draw a line that is as close to every point as possible.



- By applying linear regression we can take multiple X's and predict the corresponding Y values.



Linear Regression



In the graph above, the red dots are the true data and the blue line is linear model. The grey lines illustrate the errors between the predicted and the true values. The blue line is thus the one that minimizes the sum of the squared length of the grey lines.

➢ Coefficient estimation

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where *x bar* and *y bar* represent the mean

➢ Accuracy of Model

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

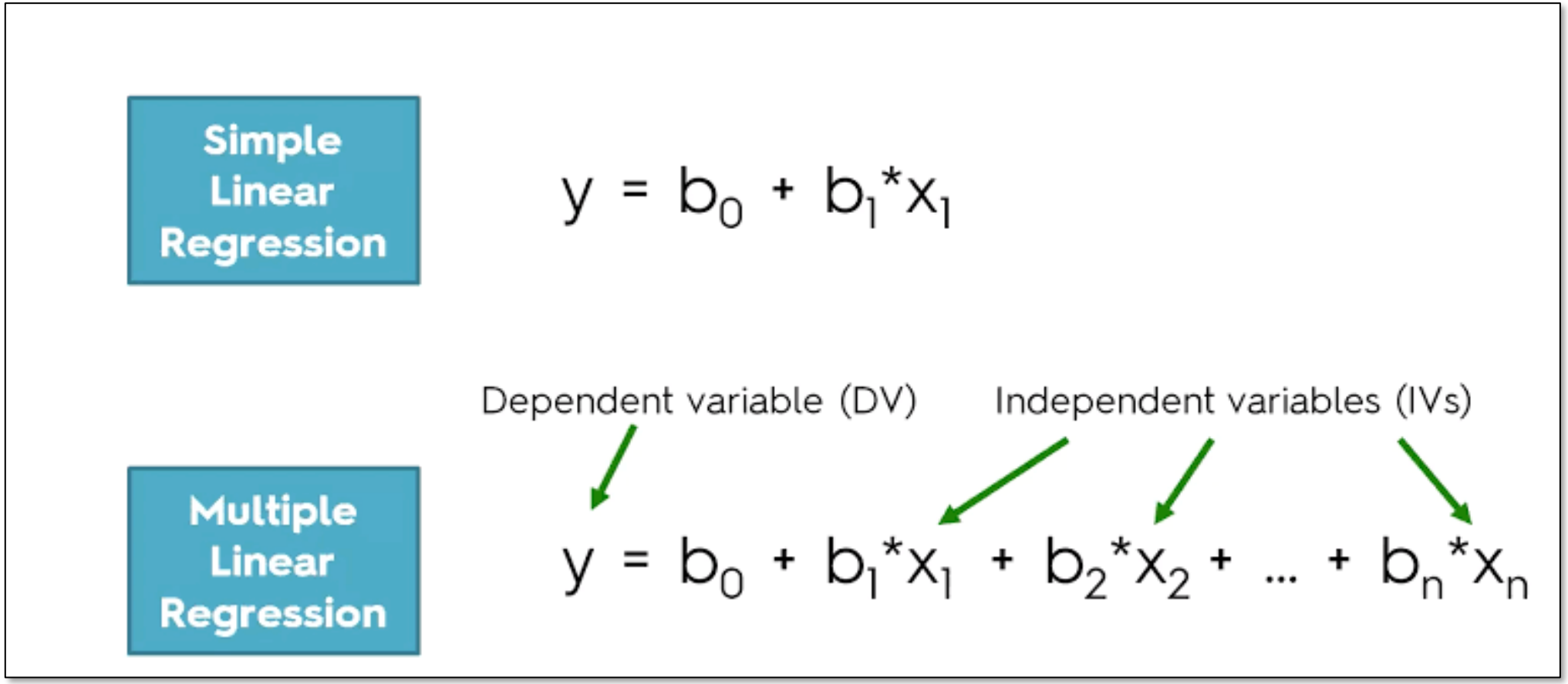
The lower the residual errors, the better the model fits the data (in this case, the closer the data is to a linear relationship).

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \quad TSS = \sum (y_i - \hat{y}_i)^2$$

As for the R^2 metric, it measures the **proportion of variability in the target that can be explained using a feature X**.

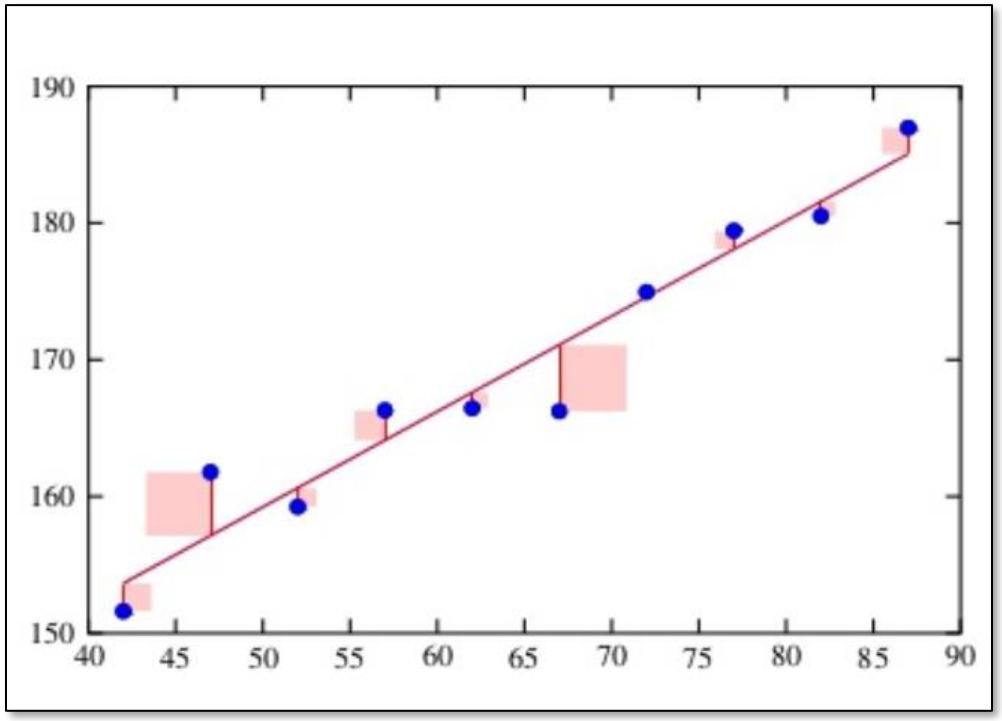
Multiple Linear Regression

- Multiple linear regression (MLR) is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.



Ordinary Least Square

- Ordinary least squares (OLS) or linear least squares is a method for estimating the unknown parameters in a linear regression model.
- The goal of OLS is to minimize the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data.
- Visually this is seen as the sum of the vertical distances between each data point in the set and the corresponding point on the regression line.
- The smaller the differences (square size), the better the model fits the data



Outliers

- **What is Outlier?**

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs.

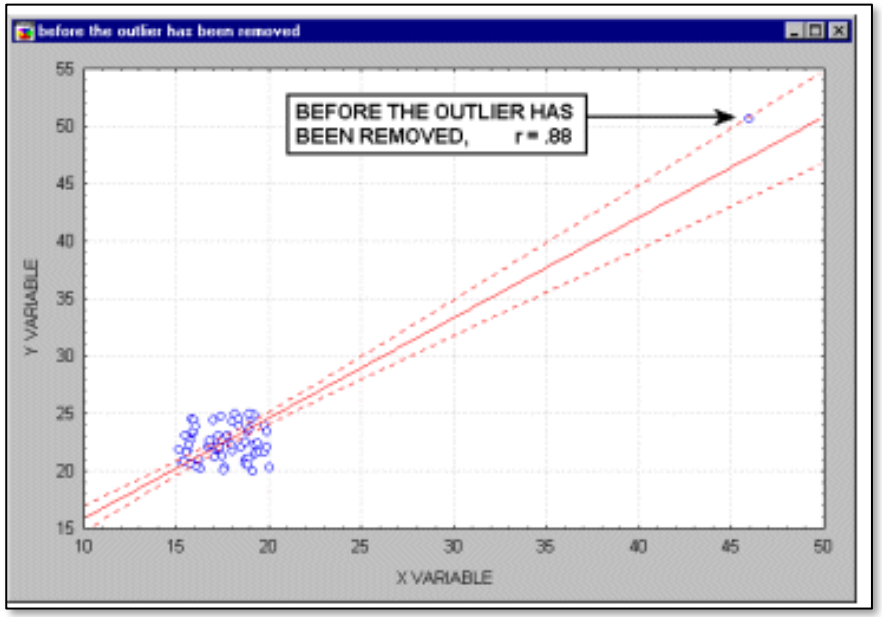
- **Detection**

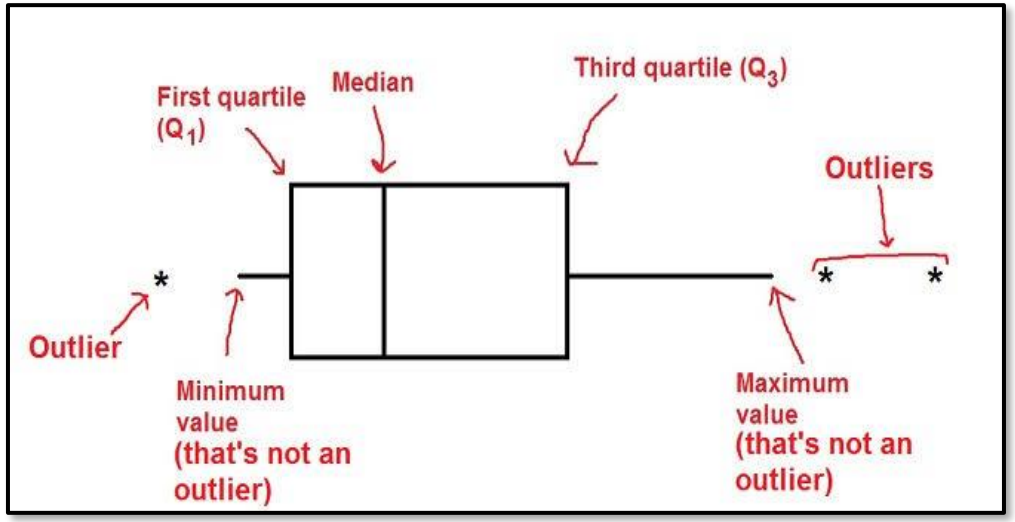
With a boxplot we can use a formula to determine if we have outliers in our distribution. We use the values that we found in our 5 number summary

Interquartile Range (IQR) is found by taking Q3 – Q1

Q1 – 1.5 IQR : any values less than this are considered outliers

Q3 + 1.5 IQR: any values greater than this are considered outliers



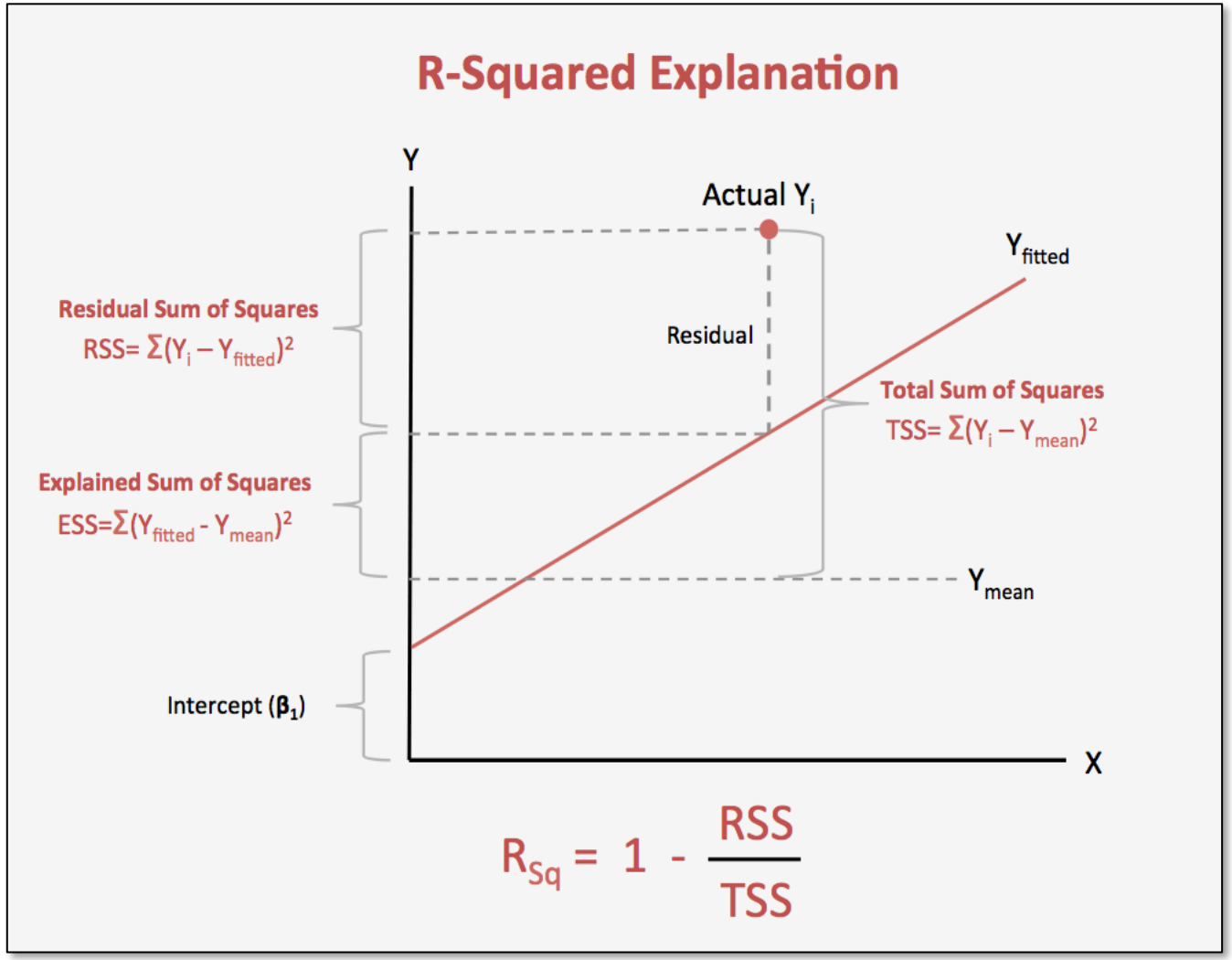


R Square

- **R-squared (R^2)** is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

The Formula for R-Squared Is

$$R^2 = 1 - \frac{\text{Explained Variation}}{\text{Total Variation}}$$



Adjusted R-Square

- The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.
- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

- In the simplified Best Subsets Regression output below, you can see where the adjusted R-squared peaks, and then declines. Meanwhile, the R-squared continues to increase.

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

Assumptions of Regression

➢ Multicollinearity

In regression, "multicollinearity" refers to predictors that are correlated with other predictors. Multicollinearity occurs when your model includes multiple factors that are correlated not just to your response variable, but also to each other. In other words, it results when you have factors that are a bit redundant.

➢ Consequences of Multicollinearity

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant. Without multicollinearity (and thus, with lower standard errors), those coefficients might be significant.

➢ Methods of detecting Multicollinearity

- An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables. If the correlation coefficient, r, is exactly +1 or -1, this is called perfect multicollinearity. If r is close to or exactly -1 or +1, one of the variables should be removed from the model if at all possible.
- A measure that is commonly available in software to help diagnose multicollinearity is the variance inflation factor (VIF).

VIF	Status of predictors
VIF = 1	Not correlated
1 < VIF < 5	Moderately correlated
VIF > 5 to 10	Highly correlated

$$VIF_i = \frac{1}{1 - R_i^2}$$

Assumptions of Regression

➢ Remedial measures of Multicollinearity

Remove highly correlated predictors from the model. If you have two or more factors with a high VIF, remove one from the model. Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared.

Principal Components Analysis, regression methods that cut the number of predictors to a smaller set of uncorrelated components.

Ridge Regression Penalize the magnitude of coefficients of features. Minimize the error between the actual and predicted observations.

Assumptions of Regression

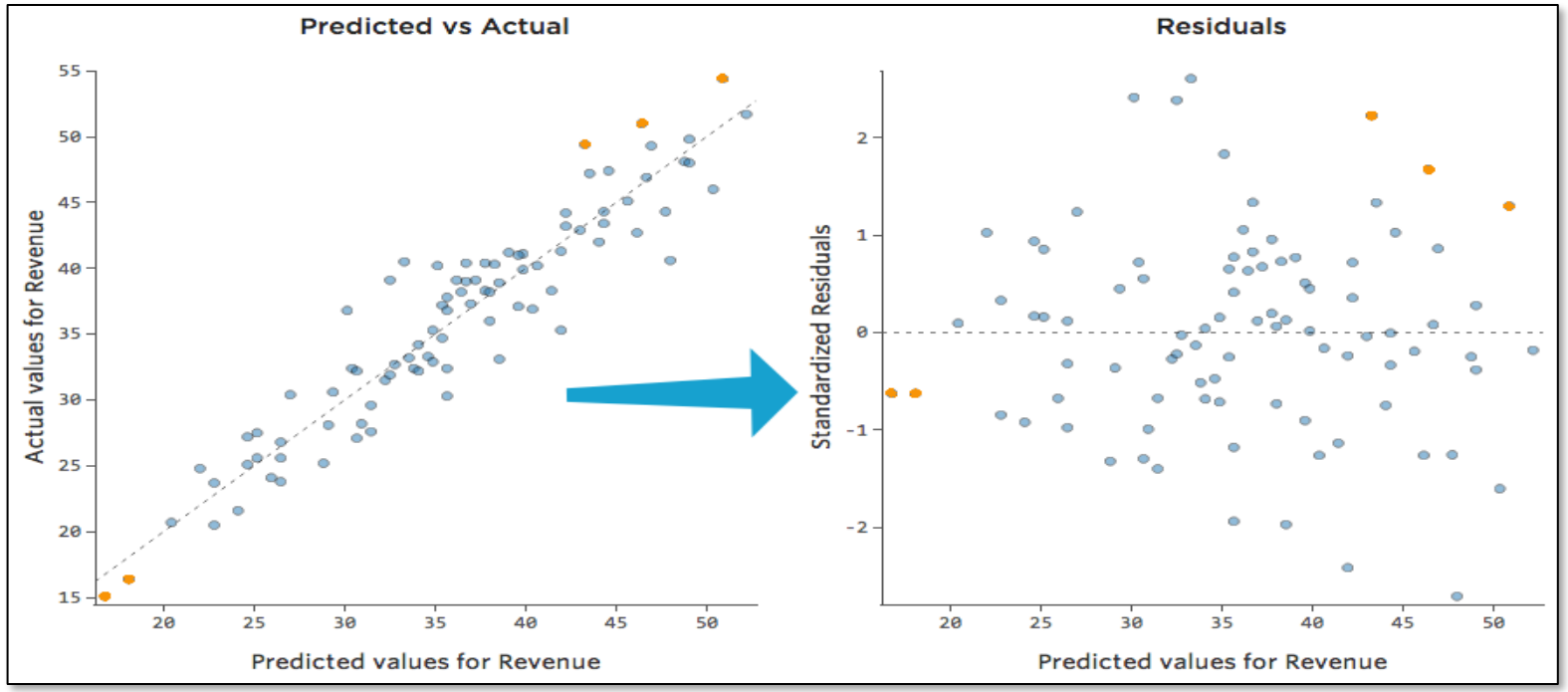
➢ Residual Analysis

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

Residual = Observed value - Predicted value

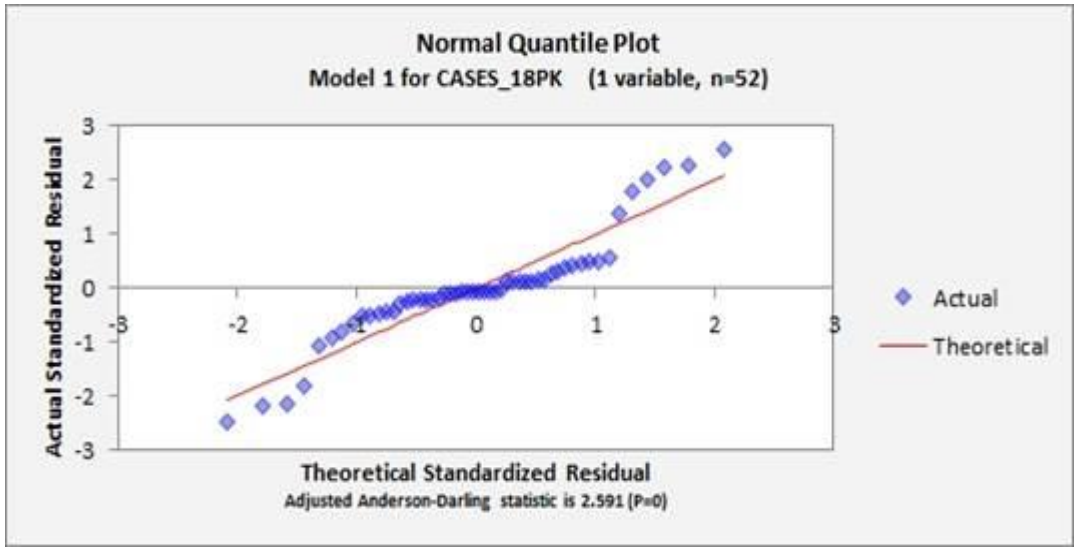
$e = y - \hat{y}$

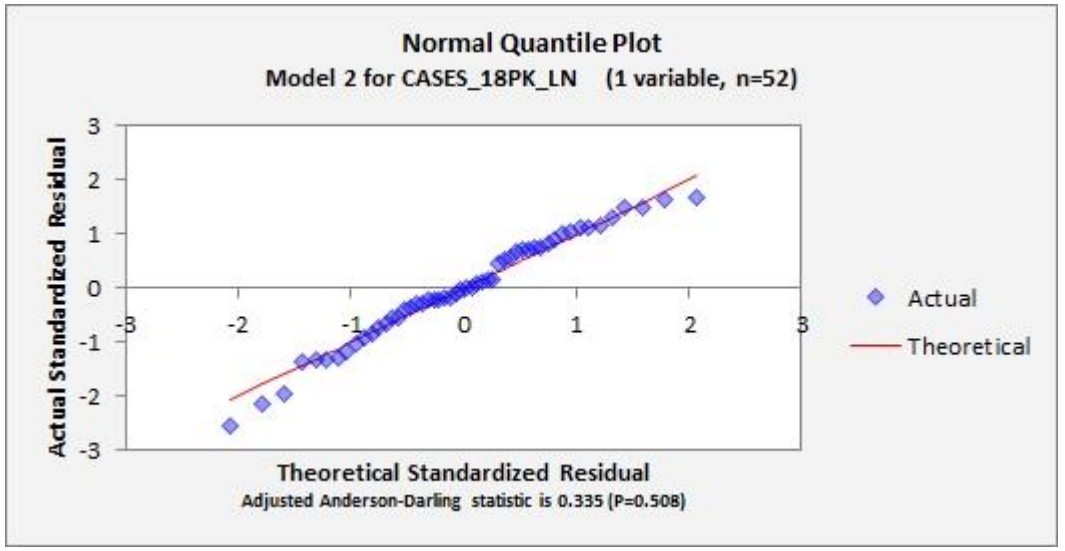
- **Residual plots that is plotting residuals against predicted value.**
- **Check for normality of errors using statistical plots and test like Q-Q plot and Shapiro Wilk test**



Assumptions of Regression

- **Remedial Measures**



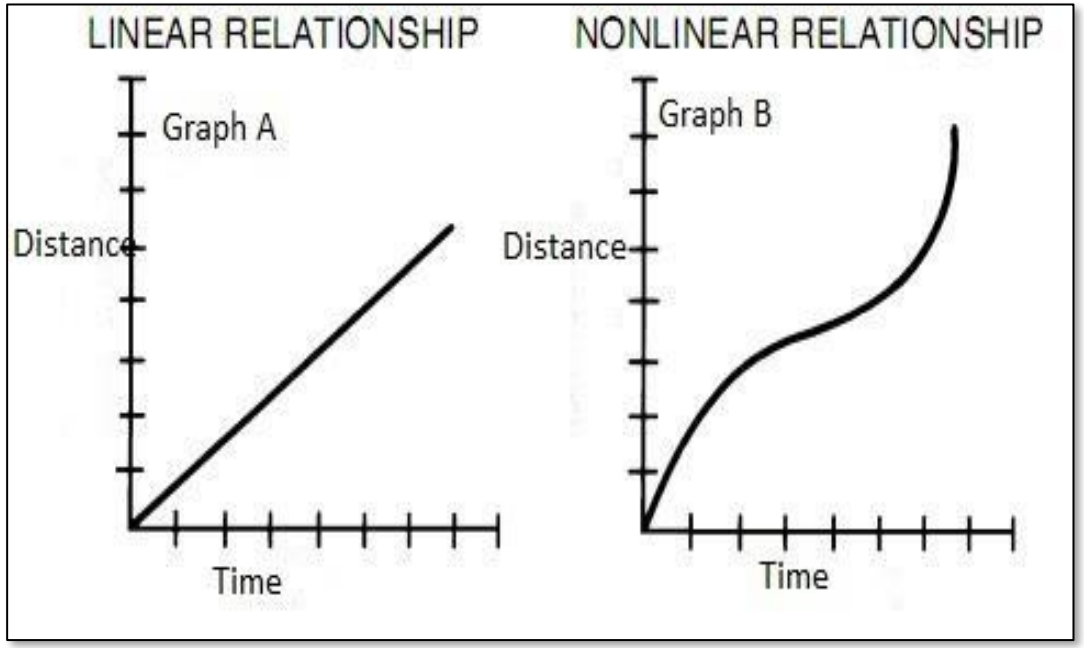


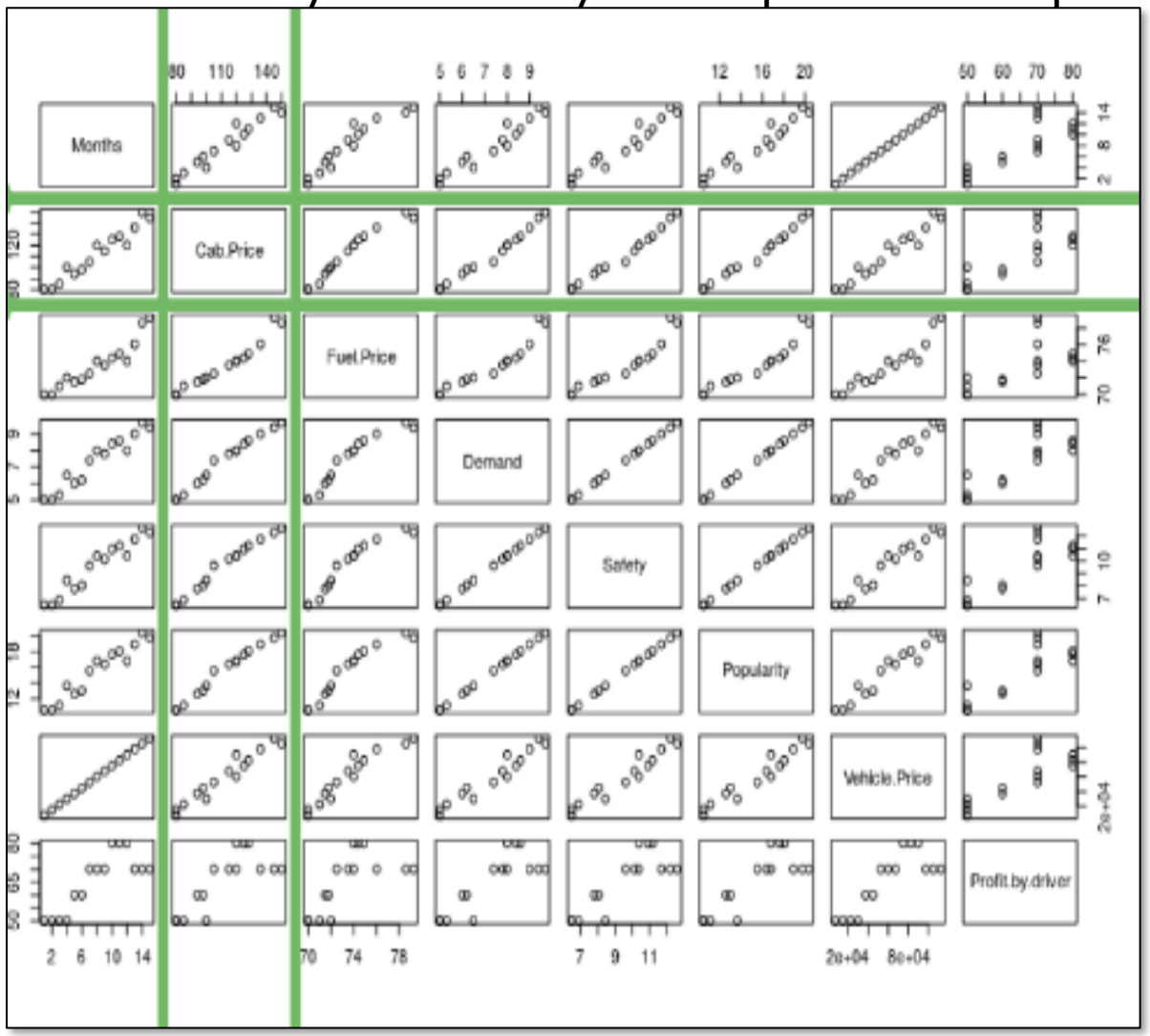
- Violations of normality often arise either because (a) the *distributions of the dependent and/or independent variables* are themselves significantly non-normal, and/or (b) the *linearity assumption* is violated. In such cases, a nonlinear transformation of variables might cure both problems. In the case of the two normal quantile plots above, the second model was obtained applying a natural log transformation to the variables in the first one.

Assumptions of Regression

➢ Linear Functional Form

Relationship between your independent and dependent variables should always be linear i.e. you can depict a relationship between two variables with help of a straight line.

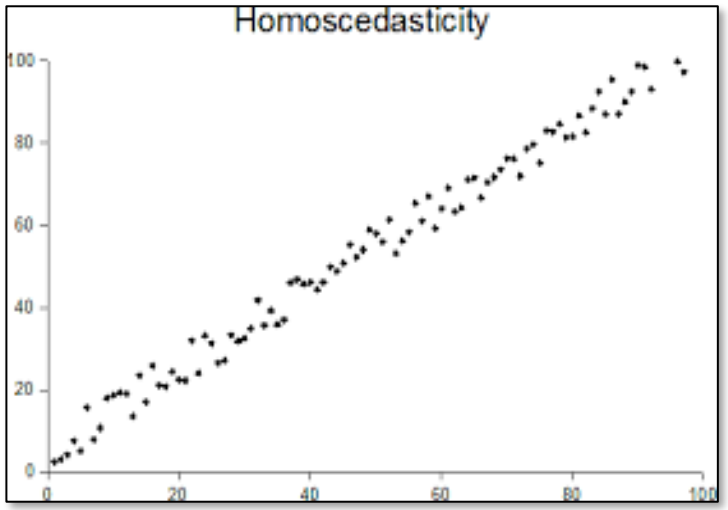


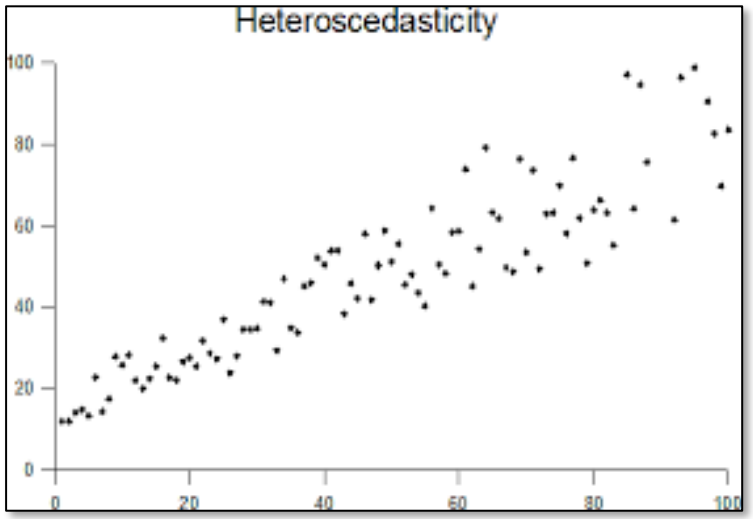


Assumptions of Regression

➢ Homoscedasticity

- There should be **homoscedasticity** or equal variance in our regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X). The sample plot below shows a violation of this assumption. For the lower values on the X-axis, the points are all very near the regression line. For the higher values on the X-axis, there is much more variability around the regression line.
- To check homoscedasticity, we make a plot of residual values on the y-axis and the predicted values on the x-axis. If we see a bell curve, then we can say that there is no homoscedasticity. It means that the variability of a variable is unequal across the range of values of a second variable that predicts it.





- **Breush-Pagan test** can be used to detect presence of heteroscedasticity in the data.

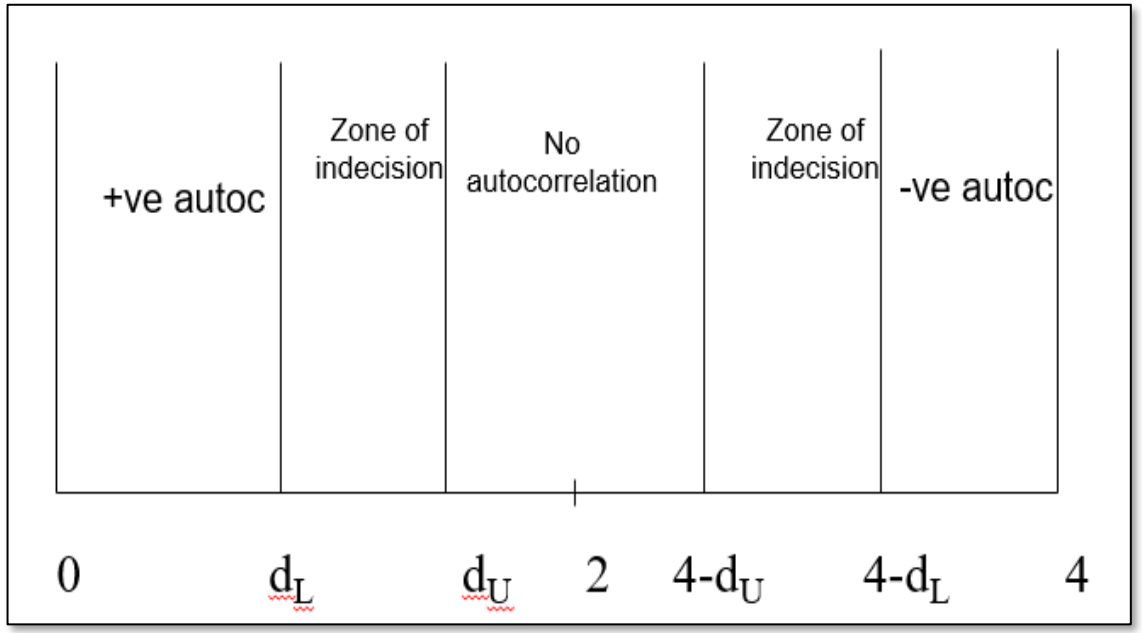
Assumptions of Regression

➢ Autocorrelation

- Autocorrelation represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- Autocorrelation measures the relationship between a variable's current value and its past values.
- An autocorrelation of +1 represents a perfect positive correlation, while an autocorrelation of negative 1 represents a perfect negative correlation.
- Technical analysts can use autocorrelation to see how much of an impact past prices for a security have on its future price.

Durbin Watson Test

Step 1: Estimate the model by OLS and obtain the residuals
Step 2: Calculate the DW statistic
Step 3: Construct the table with the calculated DW statistic and the d_U, d_L, 4-d_U and 4-d_L critical values.
Step 4: Conclude



Interpreting Regression Model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.760
Method:	Least Squares	F-statistic:	69.47
Date:	Sun, 29 Dec 2019	Prob (F-statistic):	6.99e-49
Time:	22:04:58	Log-Likelihood:	0.20817
No. Observations:	174	AIC:	17.58
Df Residuals:	165	BIC:	46.02
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	7.6218	0.278	27.368	0.000	7.072	8.172
boreratio	0.6407	0.087	7.360	0.000	0.469	0.813
drivewheel_rwd	0.2416	0.054	4.456	0.000	0.135	0.349
enginetype_dohcv	-0.5925	0.309	-1.918	0.057	-1.203	0.018
enginetype_ohcv	-0.1697	0.111	-1.523	0.130	-0.390	0.050
cylindernumber_eight	0.5594	0.177	3.168	0.002	0.211	0.908
cylindernumber_four	-0.6337	0.057	-11.064	0.000	-0.747	-0.521
cylindernumber_three	-0.9392	0.256	-3.675	0.000	-1.444	-0.435
cylindernumber_twelve	0.5295	0.269	1.971	0.050	-0.001	1.060

Omnibus:	10.049	Durbin-Watson:	2.089
Prob(Omnibus):	0.007	Jarque-Bera (JB):	21.509
Skew:	0.110	Prob(JB):	2.13e-05
Kurtosis:	4.708	Cond. No.	64.4

Interpreting Regression Model

- **R-Squared – 0.771**

i.e. almost 77% of the total variance is explained by the independent variables.

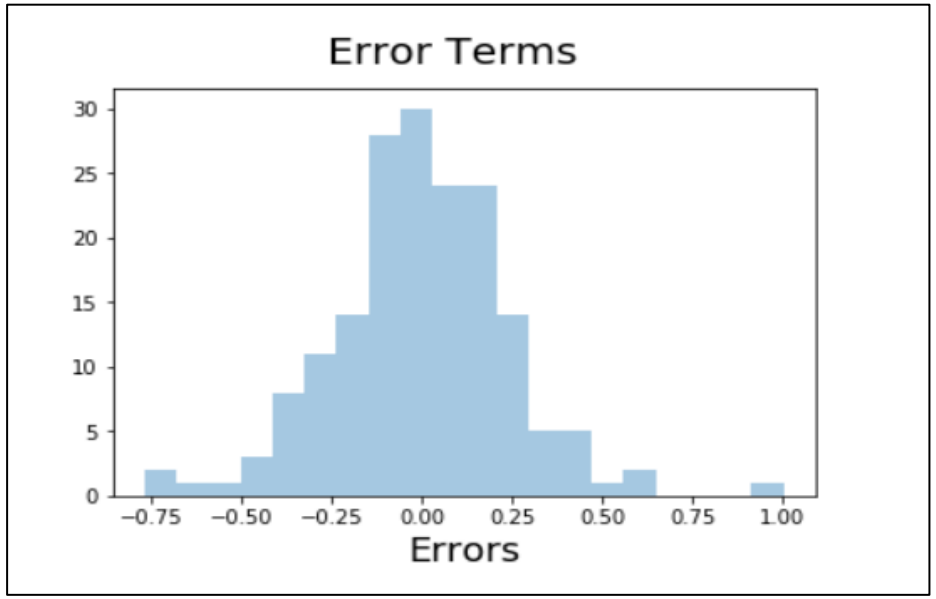
- **Adjusted R-Squared - 0.760**

i.e. almost 76% of the total variance is explained by independent variables which actually affect dependent variable.

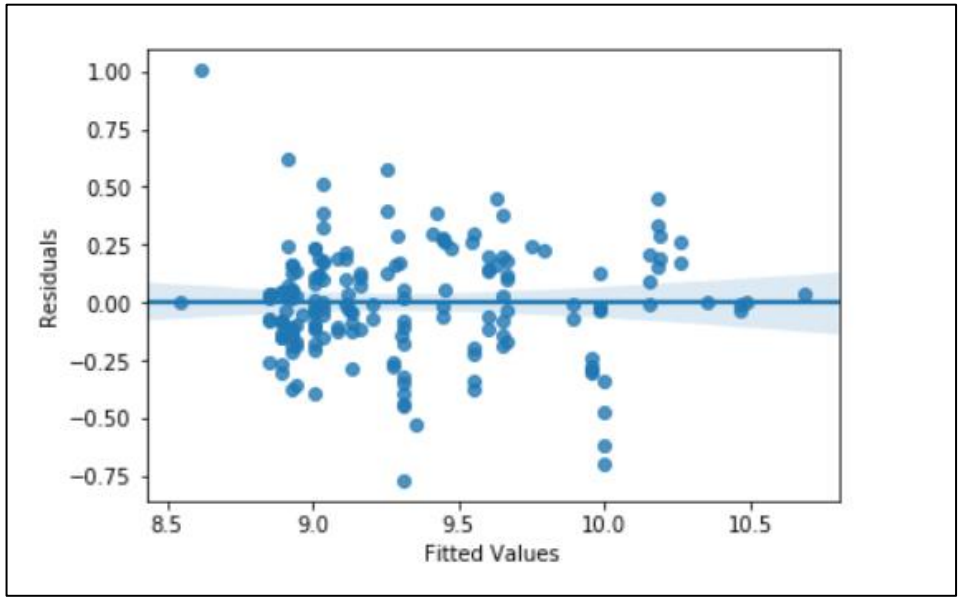
- **Durbin Watson – 2.089**

i.e. Durbin Watson test statistic is approximately 2 which means absence of Autocorrelation.

Distribution of Residuals terms



Fitted values v/s Residuals



Interpreting Regression Model

- Regression coefficients Interpretation

The case when dependent variable is log transformed regression coefficients can be interpreted as follows:

Take exponent of the beta coefficient and subtract 1 from it.

To interpet the amount of change in the original metric of the outcome, we first exponentiate the coefficient of variable to obtain exp(Beta coefficient) = a (some constant). To calculate the percent change, we can subtract one from this number and multiply by 100(let say m percent). Thus, for a one unit increase in the variable, the dependent variable increases by m percent.

In our case for variable boreratio is having regression coefficient 0.6407 i.e. e^0.6407 = 1.897

(1.897-1)*100 = 89.7%

Therefore it can be interpreted as for 1 unit increase in boreratio the price of car changes by 89.7%.

For dummy variable cylindernumber_four regression coefficient is -0.6337 i.e e^(-0.6337) = 0.5306

(0.5306 – 1) * 100 = -46.94

Therefore it can be interpreted as percentage change in price would reduce by 46.94% when number of cylinders is four.

Logistic Regression Model

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Logistic Regression is used when the dependent variable (target) is categorical.
For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

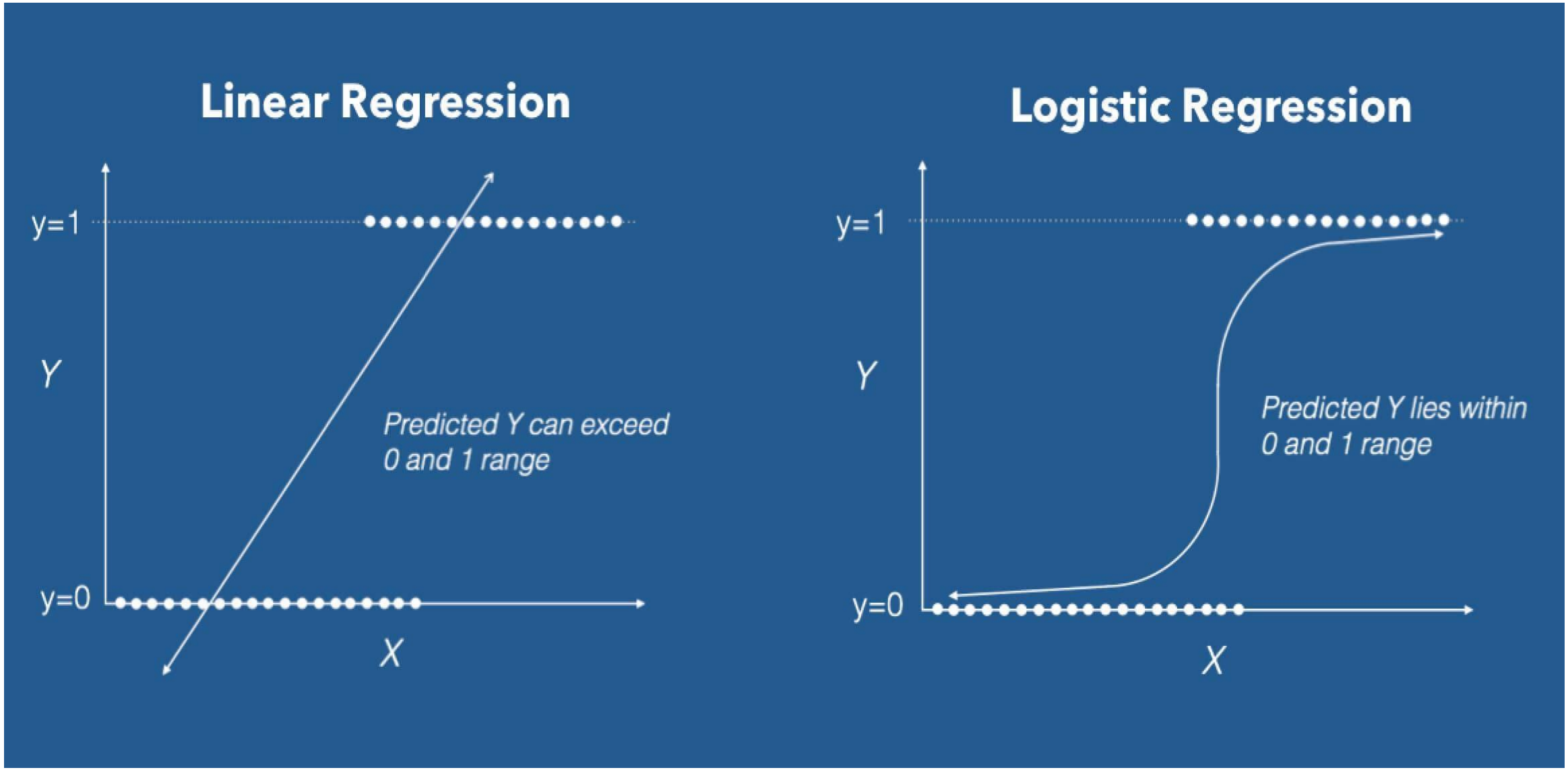
Types of Logistic Regression:

1. **Binary Logistic Regression**: The categorical response has only two 2 possible outcomes. E.g.: Spam or Not

2. **Multinomial Logistic Regression:** Three or more categories without ordering. E.g.: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

3. **Ordinal Logistic Regression:** Three or more categories with ordering. E.g.: Movie rating from 1 to 5

Logistic Regression Model



Logistic Regression Model

The logistic regression classifier can be derived by analogy to the ***linear regression hypothesis*** *which is*:

$$h_\theta(\mathbf{x}) = \theta^\top \mathbf{x}$$

However, the logistic regression hypothesis *generalizes* from the linear regression hypothesis in that it uses the ***logistic function***:

$$h_\theta(\mathbf{x}) = g(\theta^\top \mathbf{x})$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

The result is the logistic regression hypothesis:

$$h_\theta(\mathbf{x}) = \frac{1}{1 + e^{-\theta^\top \mathbf{x}}}$$

Logistic Regression Model

$$g(y) = \beta o + \beta(Age) \qquad \text{---- (a)}$$

Since probability must always be positive, we'll put the linear equation in exponential form. For any value of slope and dependent variable, exponent of this equation will never be negative.

$$p = \exp(\beta o + \beta(Age)) = e^{(\beta o + \beta(Age))} \qquad \text{------- (b)}$$

To make the probability less than 1, we must divide p by a number greater than p. This can simply be done by:

$$p = \exp(\beta o + \beta(Age)) / \exp(\beta o + \beta(Age)) + 1 = e^{(\beta o + \beta(Age))} / e^{(\beta o + \beta(Age))} + 1 \qquad \text{----- (c)}$$

Using (a), (b) and (c), we can redefine the probability as:

$$p = e^y / 1 + e^y \qquad \text{--- (d)}$$

where p is the probability of success. This (d) is the Logit Function

If p is the probability of success, 1-p will be the probability of failure which can be written as:

$$q = 1 - p = 1 - (e^y / 1 + e^y) \qquad \text{--- (e)}$$

Logistic Regression Model

On dividing, (d) / (e), we get,

$$\frac{p}{1-p} = e^{y}$$

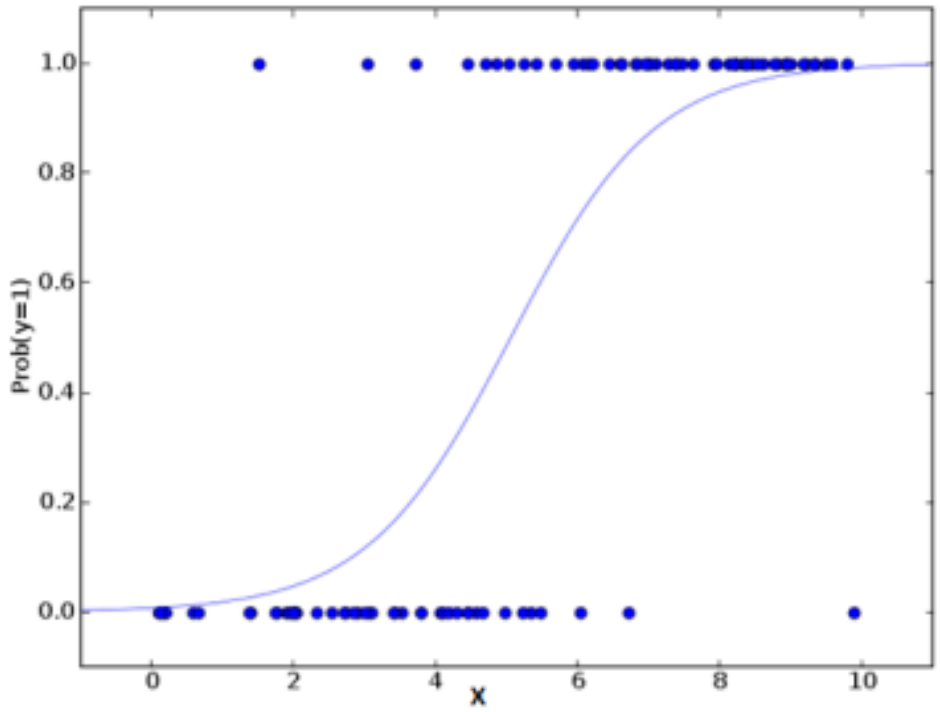
After taking log on both side, we get,

$$\log\left(\frac{p}{1-p}\right) = y$$

log(p/1-p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way.
After substituting value of y, we'll get:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta(\text{Age})$$

This is the equation used in Logistic Regression. Here (p/1-p) is the odd ratio. Whenever the log of odd ratio is found to be positive, the probability of success is always more than 50%. A typical logistic model plot is shown below. You can see probability never goes below 0 and above 1.

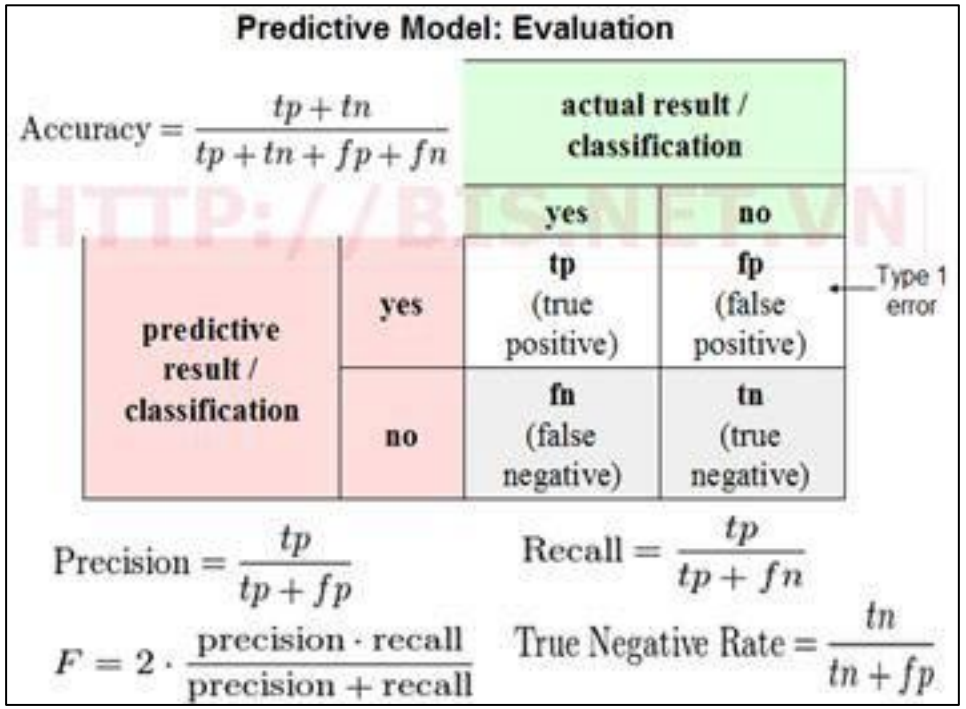


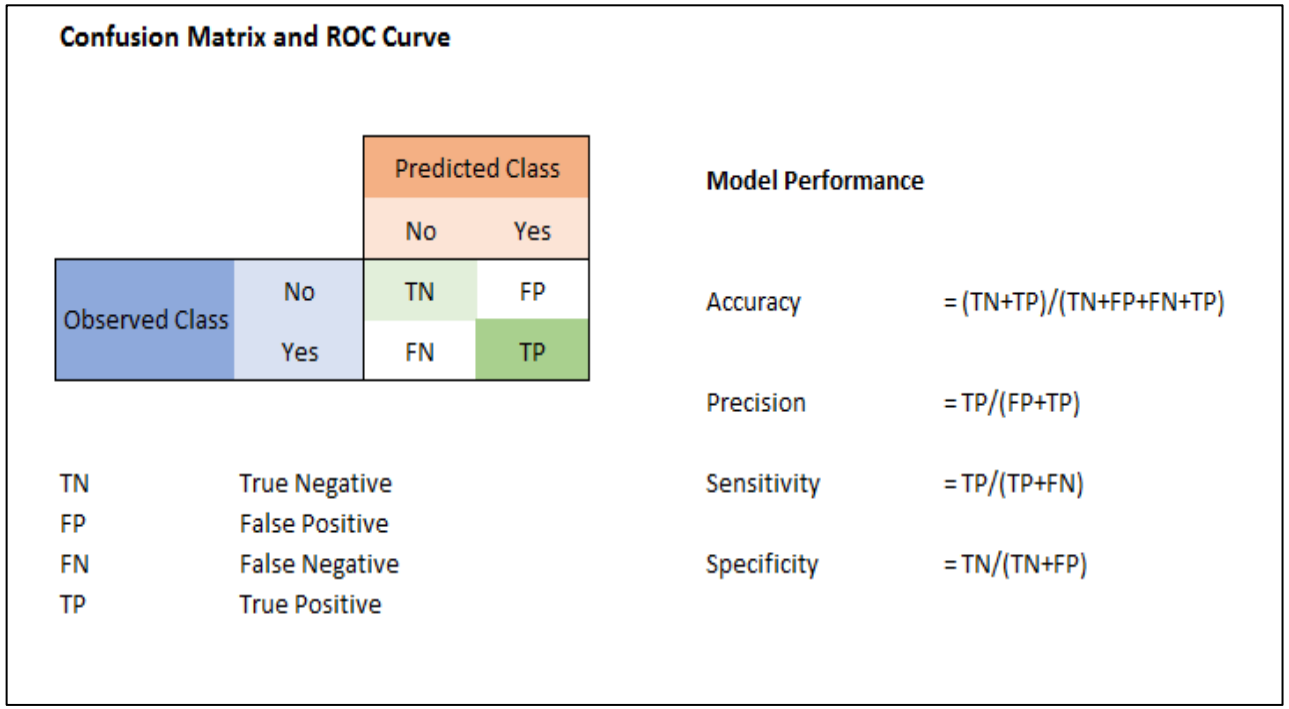
Logistic Regression Model

- Performance of Logistic Regression model

Confusion Matrix:

It is nothing but a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting.





Logistic Regression Model

➢ Performance of Logistic Regression model

Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity). For plotting ROC, it is advisable to assume p > 0.5 since we are more concerned about success rate. ROC summarizes the predictive power for all possible values of p > 0.5. The area under curve (AUC), referred to as index of accuracy(A) or concordance index, is a perfect performance metric for ROC curve. Higher the area under curve, better the prediction power of the model. Below is a sample ROC curve. The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph.