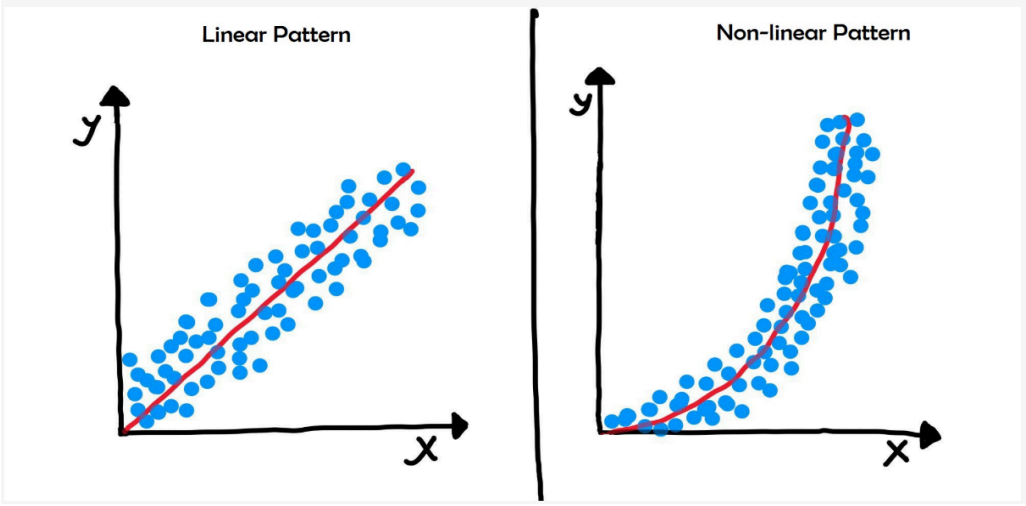
1. Explain the linear regression algorithm in detail.

Regression: It is a statistical measurement used in various disciplines that attempts to determine the strength of the relationship between one dependent variable denoted by Y and a series of other changing variables know as independent variables.

Linear Regression: It is the most basic and commonly used in predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables using a best fit straight line.

Mathematical convention:
Linear regression attempts to model the relationship between two variables by fitting a **linear** equation to observed data. A linear regression line has an equation of the form **Y = a + b*X + e**, where a is the intercept, b is the slope of line, X is the explanatory variable or independent variable and Y is the dependent variable and e is the error term.

Linear relationship between X and Y



There are two types of linear regression:

1. Simple linear regression
2. Multiple linear regression

1. Simple Linear regression:
Analysis the simplest form of regression analysis using one dependent variable and one independent variable.

Standard equation of Simple Linear Regression:

$$Y = \beta 0 + \beta 1 X,$$

Where $\beta 0$ is the intercept and $\beta 1$ is the slope of given line.

2. Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables.

The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Standard equation of Multiple Linear Regression:

$$\mathbf{Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp + \varepsilon}$$

Where β0 is the intercept and β1, β2 are the slope of given line.

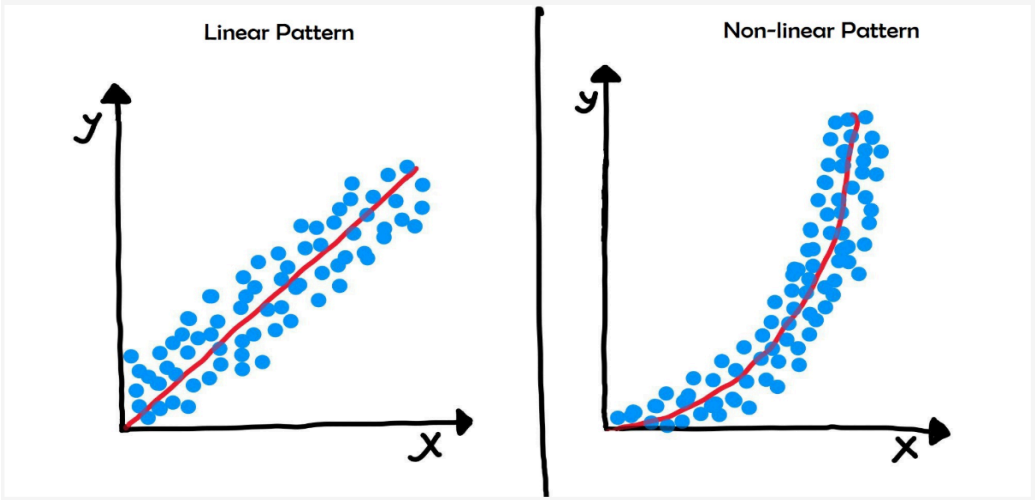
2. What are the assumptions of linear regression regarding residuals?

Certain assumptions in linear regression regarding residuals:

1. There is a linear relationship between X and Y
2. The error terms are normally distributed with zero mean (Not X and Y)
3. Error terms are independent of each other
4. Error terms has constant variance (Homoscedasticity)

1. There is a linear relationship between X and Y:

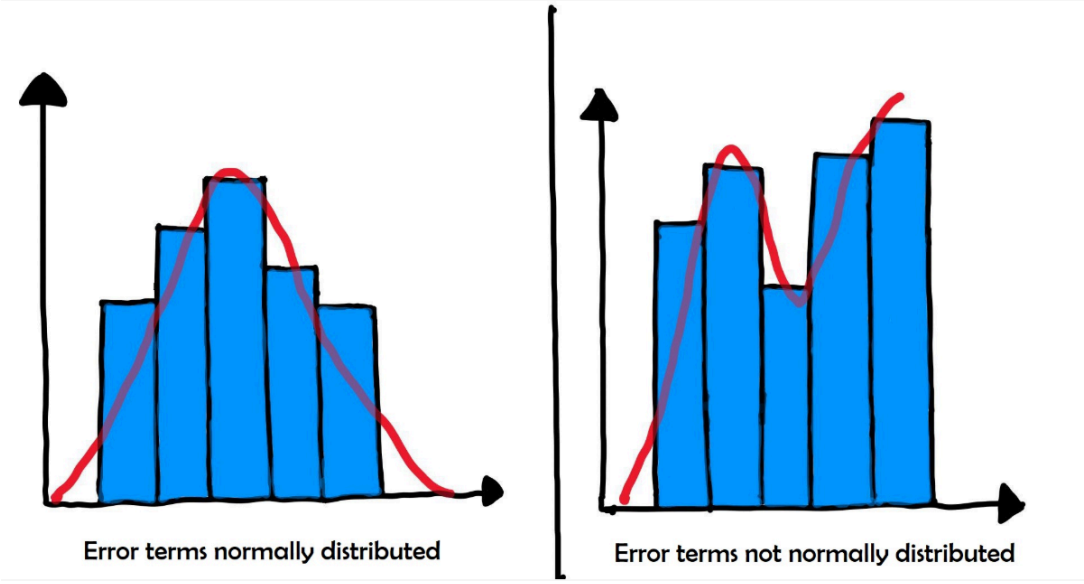
X and Y should display some sort of linear relationship; otherwise there is no use of fitting model between them.



2. The error terms are normally distributed with zero mean:

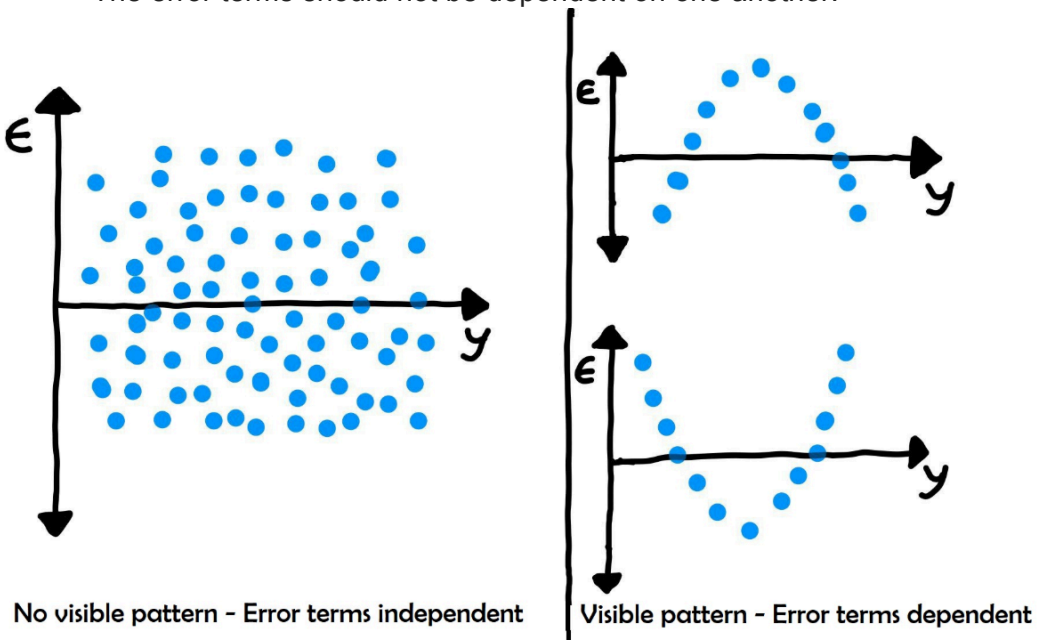
- There is no problem if the error terms are not normally distributed if you just wish to fit a line and not making further interpretations.

- When you are making some inferences on the model that you have built, you need to have notion of the distribution of the error terms.
- If the error terms not being normally distributed is that p-values obtain during the hypothesis test to determine the significance f the coefficients become unreliable.
- The assumption of normality is made, as it has been observed that the error terms follow a normal distribution with mean equal to zero.



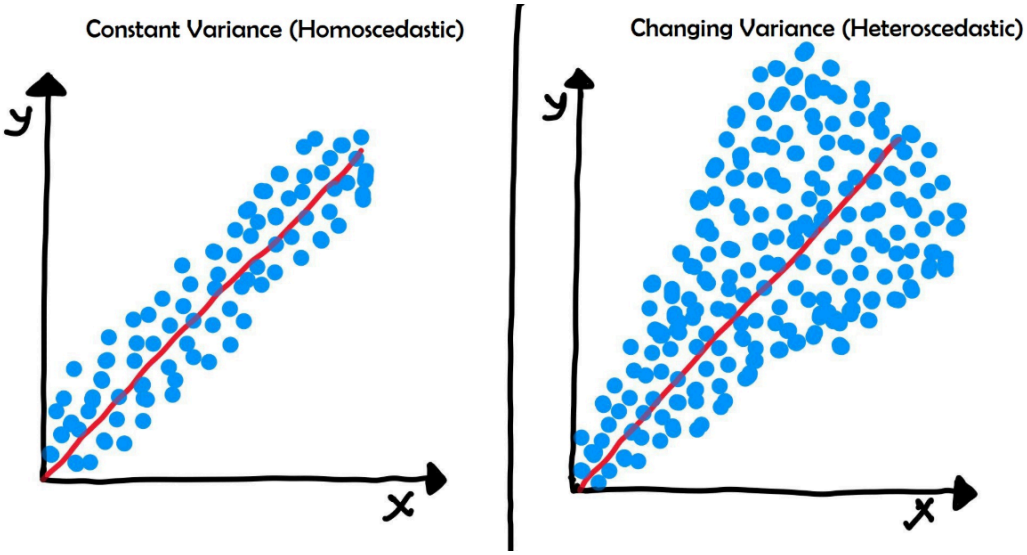
3. Error terms are independent of each other:

- The error terms should not be dependent on one another.



4. Error terms has constant variance (Homoscedasticity):

- The variance should increase or decrease as the error value changed.
- The variance should not follow any pattern as the error terms change.



3. What is the coefficient of correlation and the coefficient of determination?

Correlation coefficient is a numerical measure of some type of correlation, a statistical relationship between two variables. The variables may be two columns of a given data set of observations, or two components of a multivariate random variable with a known distribution.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range from −1 to +1, where ±1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

Correlation coefficient denoted as β.

The best way to check correlation by using "heat map".

Coefficient of determination is the proportion of variance in the dependent variable that is predictable from the independent variables.

In the context of statistic model the main purpose is either the prediction of the future outcome or the hypothesis testing on the basis of other related information.

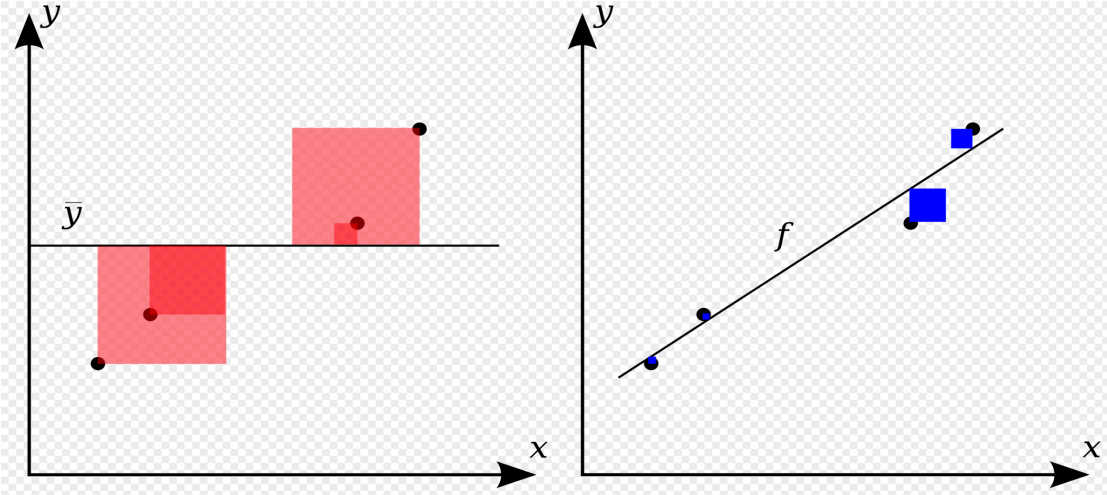
- It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcome explained by the model.
- It is denoted by r^2 or R^2 and "R- squared" and also called as "goodness-of-fit".
- The coefficient of correlation normally ranges from 0 to 1.

In general coefficient of determinant given by:

R^2 = 1 – RSS/TSS

Where RSS = Residual sum of square

TSS = Total sum of square



Observation:

- The better the linear regression (on right graph) fits the data in comparison to the simple average (on left graph), the closer the value of R^2 is 1.
- The area of blue squared represents the squared residuals with respect to linear regression.
- The areas of the red squares represent the squared residuals with respect to the average blue.

4. Explain the Anscombe's quartet in detail.

Anscombe's Quartet comprises four datasets that have nearly identical descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset contains eleven (x, y) points.

They were constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Example: Consider a dataset

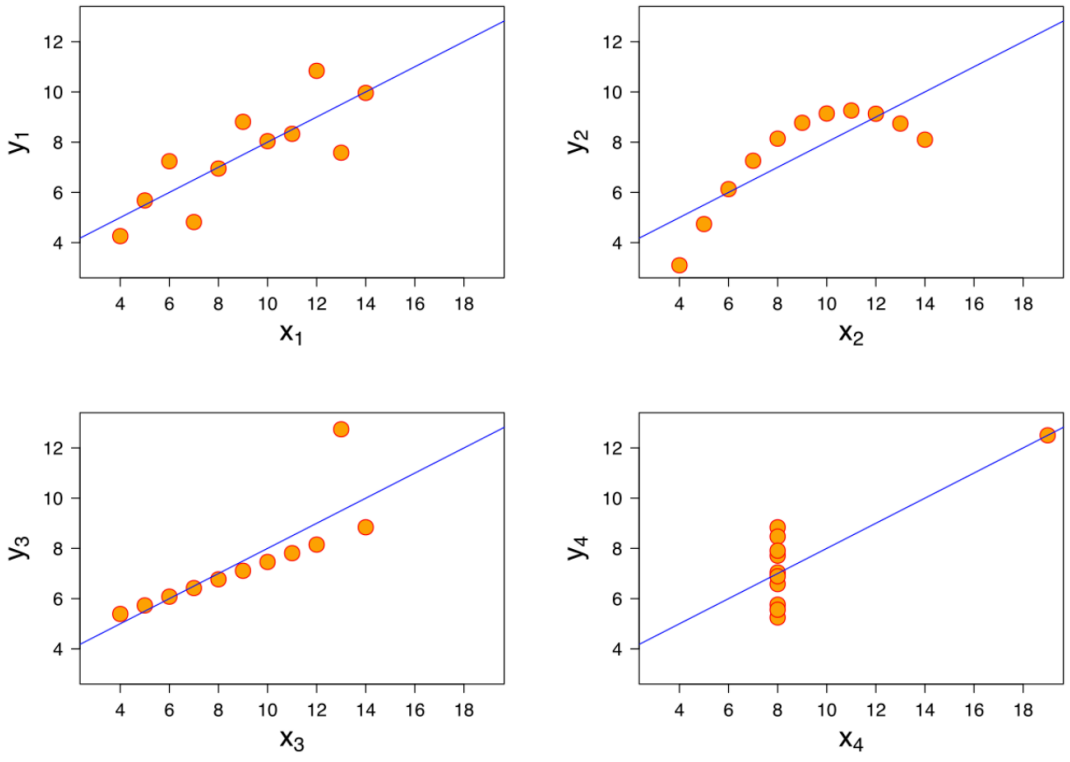
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

The summary statistics show that the mean and variance were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.85 for each dataset.
- Similarly, The variance of x is 11 and the variance of y is 4.13 for each dataset.
- The correlation coefficient (shows how strong relationship between two variables) between x and y is 0.816 for each dataset.

When we plot these four dataset on x and y plane, we can observed that they show the same regression line as well but each dataset is telling a different story.



Observation from graph:

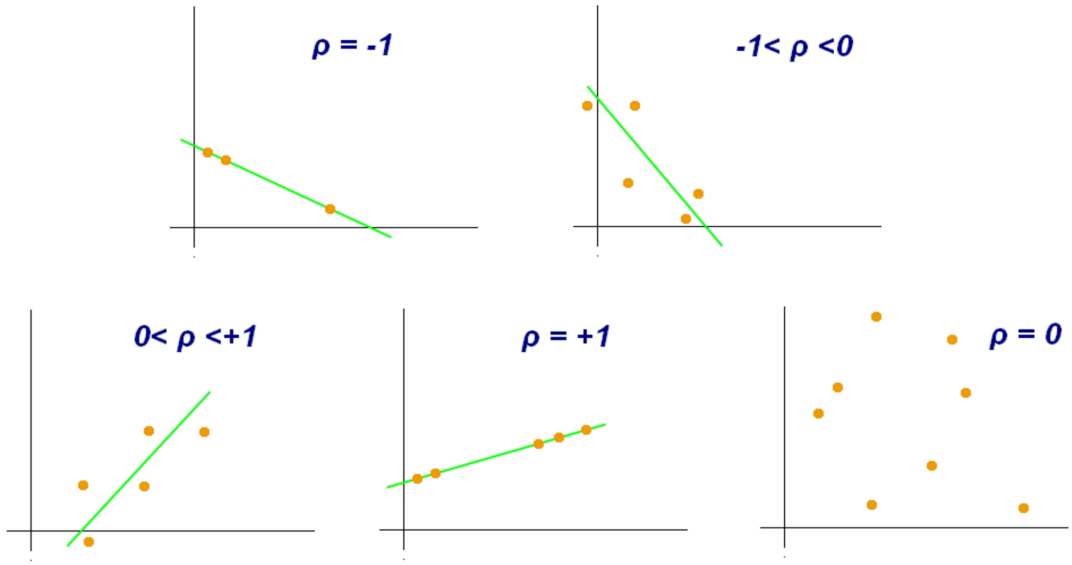
- Dataset I appear to have clean and well fitting linear model.
- Dataset II is not distributed normally.
- Dataset III the distribution is linear, but the outlier thrown off the calculated regression.
- Dataset IV shows that one outlier is enough to produce high correlation coefficient.

This quartet emphasizes the importance of data visualization in data analysis. Looking at the data revels a lot of structure and a clear picture of the dataset.

5. What is Pearson's R?

Pearson correlation coefficient also referred as Pearson's r, the Pearson Product-Moment Correlation Coefficient (PPMCC) or the bivariate correlation, is a measure of linear correlation between two variables X and Y. According to Cauchy-Schwarz inequality it has value -1 and 1 where 1 is total positive linear correlation, 0 is no linear correlation and -1 is total negative linear correlation.

Example of scatter diagrams with different value of correlation coefficients (ρ):



Pearson coefficient is the covariance of the two variables divided by the product of their standard deviations. The form definition involves product-moment that is the mean of the product of the mean-adjusted random variables.

Pearson correlation coefficient is given by:

For a population:

$$\rho(X, Y) = \text{cov}(X,Y)/\sigma X\, \sigma Y$$

Where, cov is covariance, σX is the standard deviation of X, σY is the standard deviation of Y

For a Sample:

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

where, n is the sample size

xi, yi are the individual points indexed with i

x(bar) = 1/n ∑ i=1 to n xi (the sample mean)

Scenario:

Pearson correlation is used in thousands of real life situations. For example, Scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically. The goal was to find out the evolutionary potential of rice. Pearson's correlation between two groups was analyzed. It showed a positive Pearson Product-Moment correlation between 0.783 and 0.895 for weedy rice populations. This figure quite high, which suggested a fairly strong relationship.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling: The process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as scaling.

Feature Scaling or Standardization: It is a step of Data Pre Processing, which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Normalization and Standardization: This both don't change the distribution of dataset only scale them and shift them.

- **Normalization (min-max Normalization):** Also know as min-max scaling; bring all the data in the range of 0 and 1. Selecting the target range depends on the nature of data.

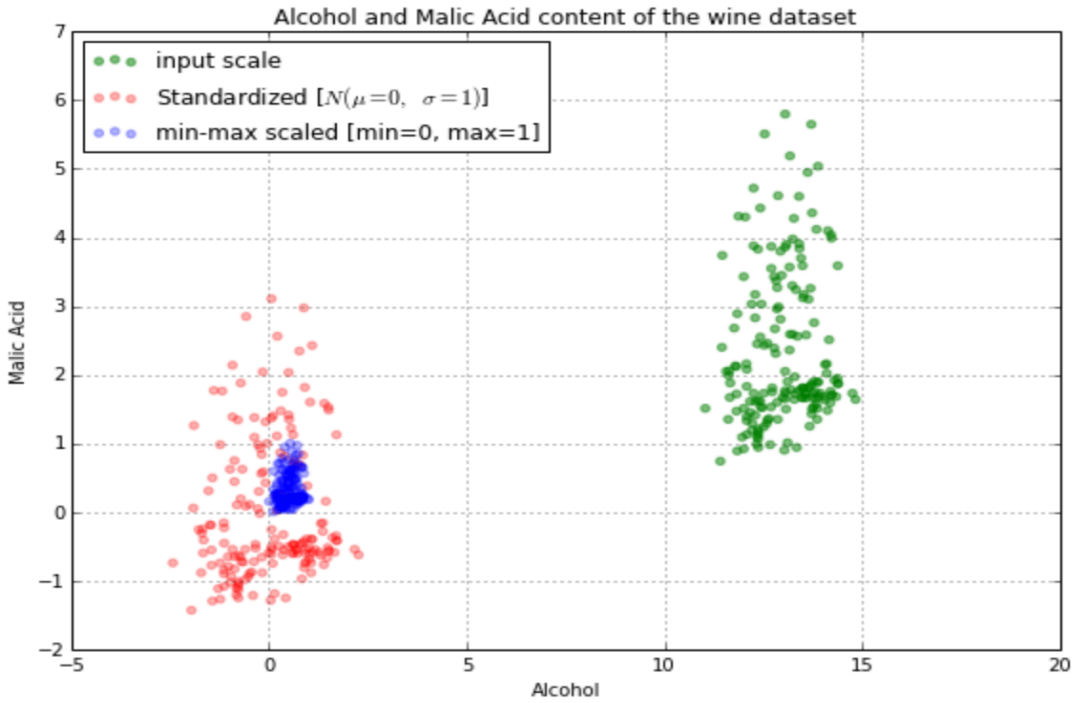
Formula is given by: **X (new) = x – x (min) / x (max) – x (min)**

- **Standardization (Z-score Normalization):** Bring all the data into a standard normal distribution with mean 0 and standard deviation 1. This method is widely used in machine learning algorithms.

Formula given by: **X (new) = x – mean (μ) / standard deviation (σ)**

NOTE: Both of these techniques have their drawbacks: If you have outliers in your data set, normalizing your data will certainly scale the "normal" data to a very small interval. And generally, most of data sets have outliers. When using standardization, your new data aren't bounded (unlike normalization).

Graph showing Normalization and Standardization:



Observations:

- The green dots showing the inputs in dataset.
- The blue color dot that is between 0 and 1 are normalized value.
- The red color dot that is scatter across certain range is standardized value.

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If all the independent variables are orthogonal to each other, then VIF= 1.0. If there is perfect correlation, then **VIF= Infinity**. A large value of VIF indicates that there is a correlation between the variables.

- It is indicated numerically as +1 and −1 for perfect Positive Correlation.
- If the values of both the variables move in the same direction with a fixed proportion is called a perfect positive correlation.

- Negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa. In statistics, a perfect negative correlation is represented by the value -1.
- A 0 indicates no correlation.

Suppose:

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that the standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this rule of thumb, in some cases we might choose to live with high VIF values if it does not affect the model results such as when we fitting a quadratic or cubic model depending on the sample size a large value of VIF may not necessarily indicate poor model.

Heuristics of VIF:

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

8. What is the Gauss-Markov theorem?

The **Guass-markov theorem** states that in a linear regression model in which the errors are uncorrelated, have equal variance and expectation value of 0, The best linear unbiased estimator (BLUE) of the coefficients given by the ordinary least square (OLS) estimator, giving the lowest variance of the estimate as compared to other unbiased linear estimator. The errors do not need to be normal nor need to be independent and identically distributed (with only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimator exists with lower variance.

Statement: Suppose we have in matrix notation,

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad (\underline{y}, \underline{\varepsilon} \in \mathbb{R}^n, \underline{\beta} \in \mathbb{R}^K \text{ and } X \in \mathbb{R}^{n \times K})$$

expanding to,

$$y_i = \sum_{j=1}^{K} \beta_j X_{ij} + \varepsilon_i \quad \forall i = 1, 2, \ldots, n$$

Where, βj is non-random but unobservable parameters, **X**ij are non-random and observable (called explanatory variables), εi is random so yi are random. The random variable εi are called disturbance or noisy or simply error.

The Guass-Markov assumptions concern the set of error random variables, εi:

- They have zero mean: E [εi] = 0
- They are homoscedastic, i.e. all have same finite variance: Var (εi) = σ^2 < ∞
- Distinct error terms are uncorrelated: Cov (εi, εj) =0

A linear estimator of βj is a linear combination

$$\widehat{\beta}_j = c_{1j} y_1 + \cdots + c_{nj} y_n$$

In which the coefficients cij are not allowed to depend on underlying coefficients βj, since those are not observable, but are allowed to depend on value Xij, since these data are observable. The estimator is said to be unbiased only if:

$$\mathrm{E}\left[\widehat{\beta}_j\right] = \beta_j$$

Regardless of the value Xij. Then the mean square error of the corresponding estimation is given by:

$$\mathrm{E}\left[\left(\sum_{j=1}^{K} \lambda_j \left(\widehat{\beta}_j - \beta_j\right)\right)^2\right]$$

It is the expectation of the square of the weighted sum (across parameters) of the difference between the estimators and corresponding parameters to be estimated.

The ordinary least squares estimator (OLS) is the function:

$$\widehat{\beta} = (X'X)^{-1} X'y$$

Of y and X where X' denoted the transpose of X, that minimizes the sum of square of residuals:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{K} \widehat{\beta}_j X_{ij} \right)^2 .$$

The theorem now states that the OLS estimator is BLUE. The main idea of the proof is that the least-squares estimator is uncorrelated with every linear unbiased estimator of 0, i.e. with every linear combination $\mathbf{a_1y_1 + \ldots + a_ny_n}$ whose coefficients do not depend upon the unobservable β but whose expected value is always zero.

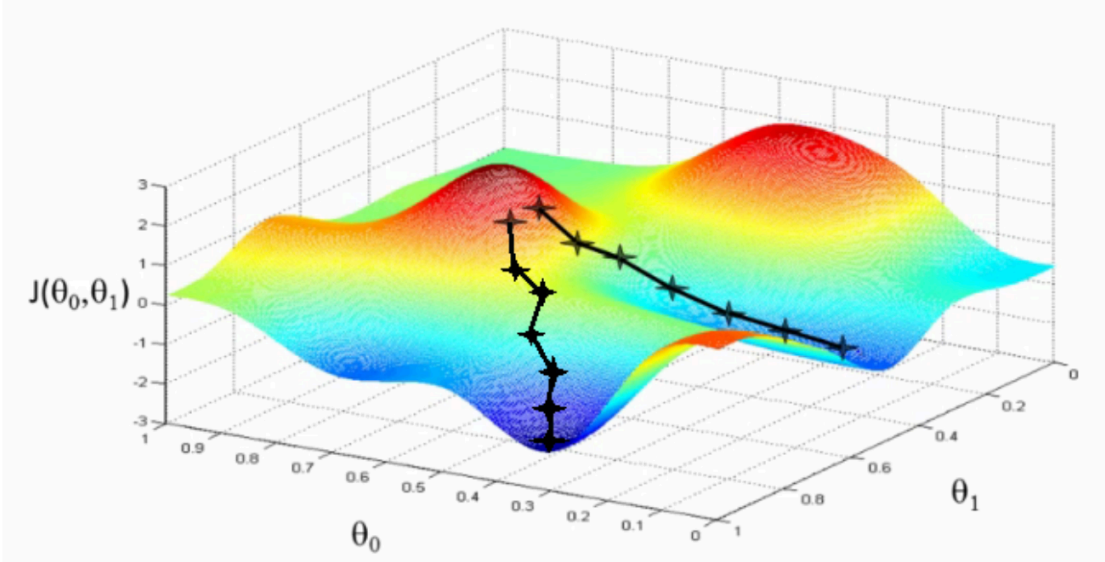
9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to find the values of the parameters (coefficients) of a function (f) that minimizes a given cost function.

It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters.

Scenario:

Suppose you are at the top of a mountain, and you have to reach a lake, which is at the lowest point of the mountain. The thing is that you are blindfolded and you have zero visibility to see where you are headed. So, how do you approach to reach the lake?
The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.



- A standard approach to solving this type of problem is to define an error function (also called a cost function) by incorporating gradient descent function we can achieve that measures how a line is good fit.
- This function will take in an (m, b) or (β_0, β_1) pair and return an error value based on how well the line fits our data. To compute this error for a given line, we will iterate through each (x, y) point in our data set and sum the square distances between each point's y value and the candidate line y value (**y = mx + b**).
- It is conventional to square this distance to ensure that it is positive and to make our error function differentiable.

The Sum of Square Error Equation:

$$\text{Error}_{(\text{m,b})} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$$

To run gradient descent on this error function, we first need to compute its gradient. The gradient will act like a direction and always point us downhill. To compute, we will need to differentiate our error function. Since our function is defined by two parameters (m and b), we need to compute a partial derivative for each.

Sum of Square Error Equation Partial Derivatives with respect to m and b:

$$\frac{\partial}{\partial \text{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \text{b}} = \frac{2}{N}\sum_{i=1}^{N} - (y_i - (mx_i + b))$$

Pick a value for the learning rate α. The learning rate determines how much big step will be on each iteration.

- If α is very small, it will take long time to converge and become computationally expensive.
- If α is large, it may fail to coverage and overshoot the minimum.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plots or (Quantile – Quantile) plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

Example, the median is a quantile where 50% of the data fall below that point and 50% lies above it.

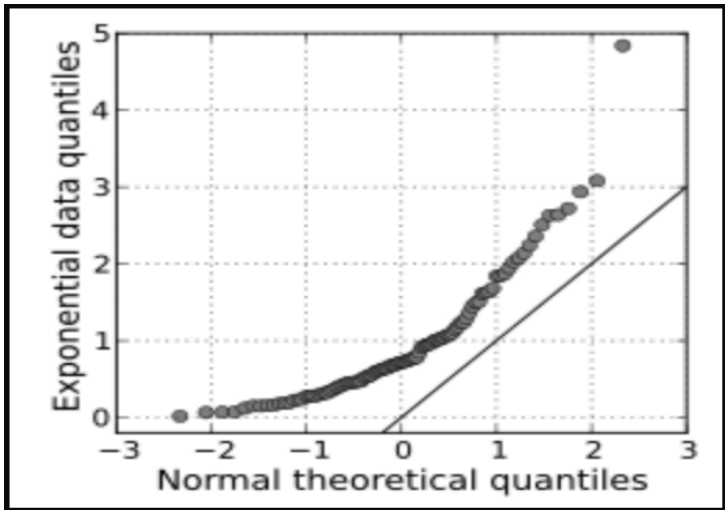
In general, the quantile-quantile plot is graphical technique for determining if two data sets come from populations with a common distribution.

The purpose of Q-Q plots is to find out, if two data sets come from the same distribution.

Importance:

- A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.
- Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions.
- A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples.
- Q–Q plots are also used to compare two theoretical distributions to each other. Since Q–Q plots compare distributions; there is no need for the values to be observed as pairs, as in a scatter plots, or even for the numbers of values in the two groups being compared to be equal.

Scenario:



The image above shows the quantiles from a theoretical normal distribution on the horizontal axis. It is compared to set a data on the y-axis. This particular type of Q-Q plot is called a normal quantile-quantile plot.