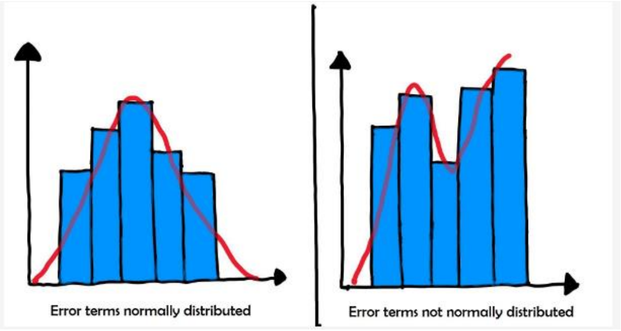
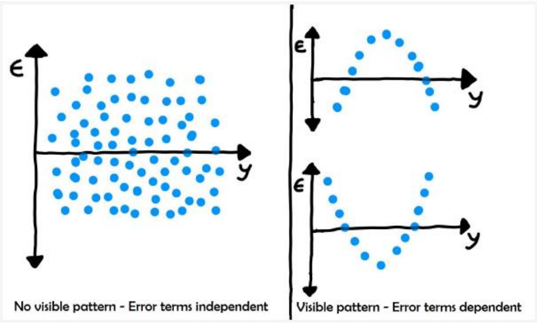
1. What are the assumptions of linear regression regarding residuals?

Answer: Assumptions of linear regression regarding residuals as below:

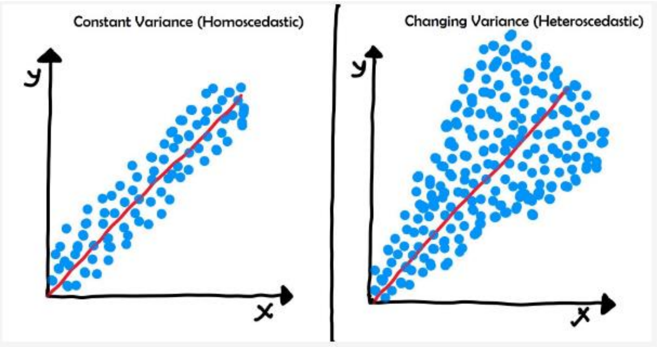
1. Error terms are normally distributed with mean zero(not with variables).



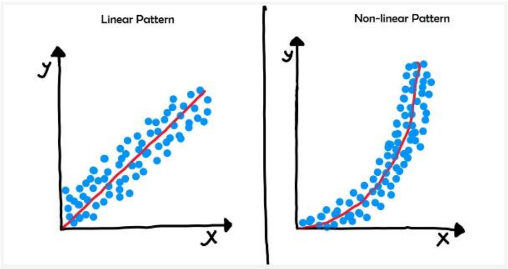
2. Error terms are independent of each other. The error terms should not be dependent on one another (like in a time-series data).



3. Error terms have *constant variance* (homoscedasticity).



4. The independent variables and residuals are not correlated.
5. There is a linear relationship between X and Y



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

Answer:

Coefficient correlation:

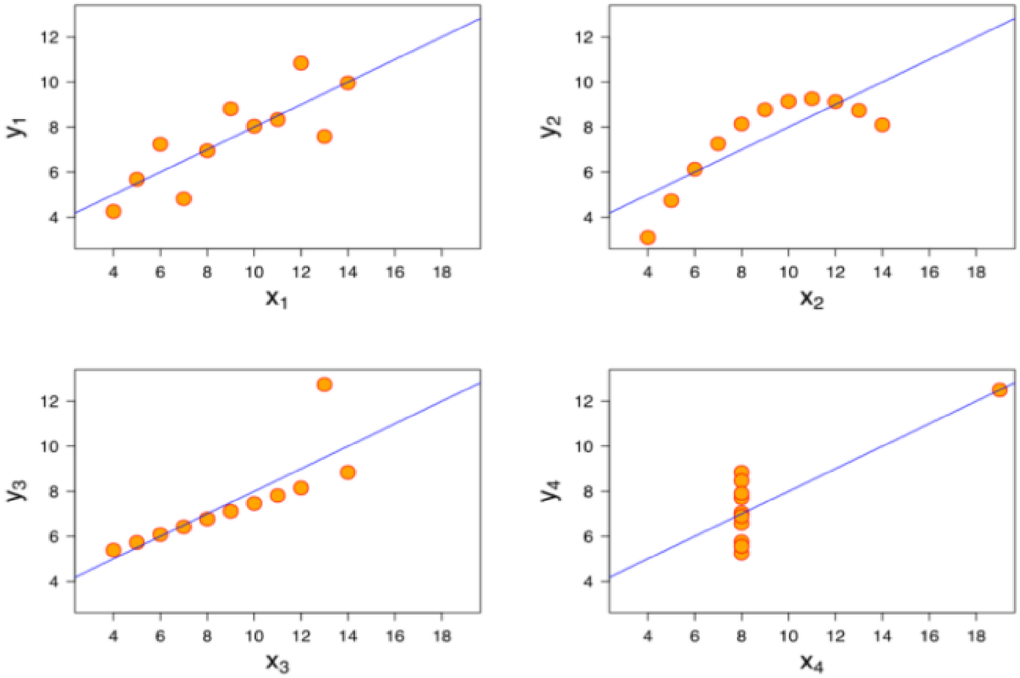
- It’s denoted by quantity r in regression. It determines the direction (positive and negative) and strength of the linear relationship between two variables(x and y).
- The value of coefficient of determination is always in between -1 to +1. If it’s positive (+), then x and y are in positive linear relationship and if it’s negative, then x and y are in negative linear relationship.
- If r value is close to +1, then the relationship is strongly positive linear correlation means if x increases then y also increases.
- If r value is close to -1, then the relationship is strongly negative linear correlation means if x decreases then y also decreases.
- If r value is 0, there is no correlation linear relationship between x and y. Both are independent variables.

Coefficient of determination:

- It’s denoted by r-squared. It determines the strength of the linear association between variables (x and y). It results proportion of variance of one variable(y) , which can be predictable from other variable(x).
- R-squared gives the percentage of data points, that are closest to the best fit line.
- If the regression line passes through all the data point in scatter plot, it is able to explain all the variance otherwise it is not able to explain all the variance.

3. Explain the Anscombe’s quartet in detail.

Answer:



- Anscombe's quartet suggests us that we should visualize the data prior to start of analysis and during analysis of data, outliers should be removed and it was developed by statistician Francis Anscombe.
- If we consider four datasets, each containing eleven(x,y) pairs with many similar statistical properties, the graphical representation of it (Anscombe's quartet) shows huge difference from one another.
- In First graph, the variables have best fitting linear relationship.
- In Second graph, the variables should not be analyzed with a linear regression because it's not normally distributed.
- In third graph, the data points distribution is linear but there is an outlier also.
- In Fourth Graph, It seems, X variable is constant except one outlier.
- Anscombe's quartet tells us that, how the visualization is important in Data Analysis.

4. What is Pearson's R?

   Answer:

   - In statistics, the Pearson correlation coefficient is also called as Pearson's r. Pearson's correlation coefficient is considered as best method of measuring the association between variables of interest because it is based on the method of covariance.
   - It measures the statistical relationship or association as well as the direction of the relationship between two continuous variables.
   - If r value is close to +1, then the relationship is strongly positive linear correlation means if x increases then y also increases.
   - If r value is close to -1, then the relationship is strongly negative linear correlation means if x decreases then y also decreases.
   - If r value is 0, there is no correlation linear relationship between x and y. Both are independent variables.

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

   Answer:

   - Feature scaling is applied to all independent variables or features of the data in step of data pre-processing. It is used to normalize the data in specific range. It is also known as data normalization.
   - By doing scaling, All the predictor variables are in comparable stage, so that all the co-efficient of variables can be comparable. This is called as interportabilty of co-efficients.
   - If the rescaling of the variable is done in between 0 and 1, then the optimization which happens behind the scenes becomes much faster. The gradient decent algorithm tries to minimize the cost function.

     <u>Normalized Scaling(Min Max normalization):</u>
   - Here the data is scaled to a fixed range, usually in between 0 to 1.
   - Formula : $$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- It handles the outliers in the variables.
- In python, “from sklearn.preprocessing import MinMaxScaler” is used to perform the normalization of the variables.

Standardized scaling (**Z-score normalization**):

- Here the features are rescaled to have properties of a standard normal distribution with μ=0 and σ=1. The distribution will have values between -1 to 1.
  Where μ=population mean and σ=standard deviation
- Formula: $z = \frac{x - \mu}{\sigma}$
- It can be used for algorithms, which has 0 centric data.

- In real time, most of the cases, Normalization is preferred because the values are in between 0 and 1 and it handles outliers.

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

Answer:

- VIF calculates, how well one independent variable is explained by all the other independent variable combined. If VIF is high, it is redundant in the presence of the other variables.
- VIF is infinite because, when the variables have perfect correlation (unity), the corresponding VIF becomes infinite.
- If the variables are highly correlated , the standard error will equal to infinity.