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

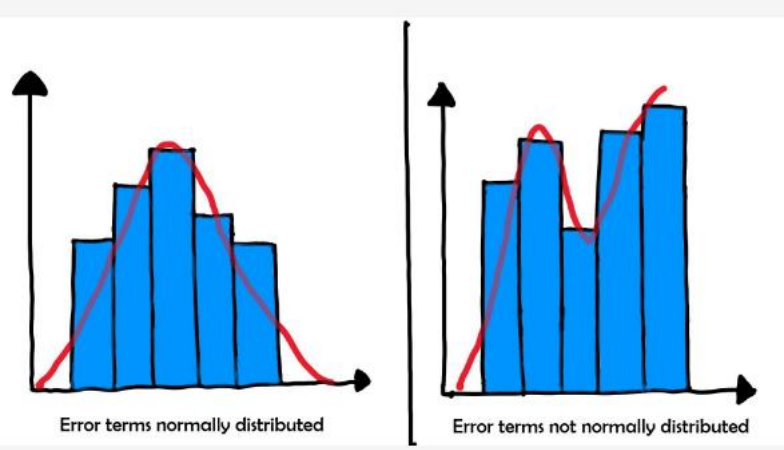
Ans. **Linear Regression**: It explains whether one or more predictor variable is able to explain the dependent variable. It should have linear relationship between dependent and independent variables. The output variable to be predicted is a continuous variable. Eg., cost of a house

There can be 2 types of Linear Regression model:

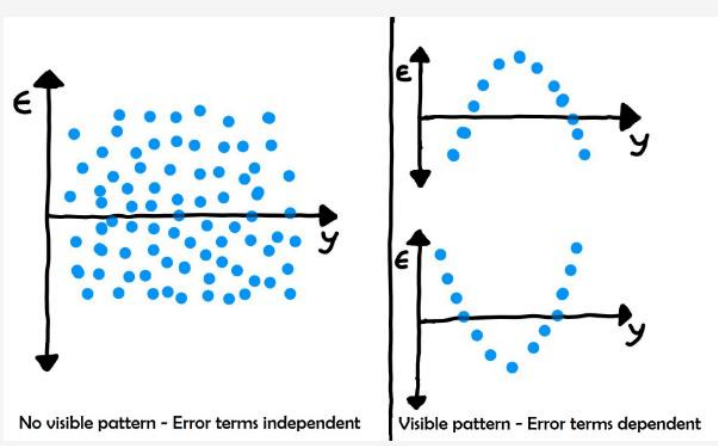
- **Simple Linear Regression**
- **Multiple Linear Regression**

Assumptions of Linear Regression residuals(errors) are:

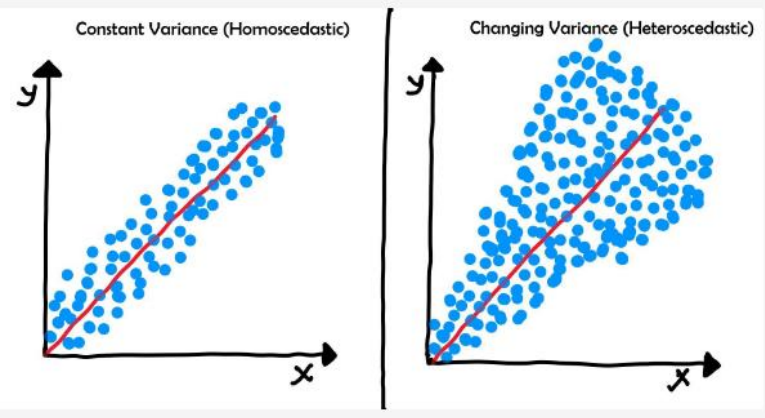
a. **Residuals must be normally distributed with mean 0**:
   - It is fine if the error term is not normally distributed only if we have to fit a line and not make any interpretations
   - But to derive the inference from the model, error terms must be normally distributed. And because of that p-value obtained during the hypothesis testing to define the significance of coefficients becomes unreliable.
   - If mean is not zero then it means it carry some information of dependent variable, therefore it has to be zero.



b. Residuals must be independent to each other:
   - If residuals are dependent then that means they are carrying some information about the model and hence it won't be able to accurately predict the model.



c. **<u>Homoscedasticity (constant variance) :</u>**
   - The variance should not increase or decrease as the error value changes
   - The variance should not follow any pattern as the error value changes.



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

Ans: **<u>Coefficient of correlation</u>**:

- It measures the direction and strength of linear relationship between dependent and independent variables which is denoted by **r.**
- It is sometimes referred to as **Pearson product moment correlation coefficient** after its developed Karl Pearson.
- The mathematical **formula** of r is given as:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of pairs of data.

- The value of r lies between -1 and +1. The positive sign (+) indicates positive correlation and the negative sign (-) indicates negative correlation.
- **Positive correlation**: it indicates the relationship between x and y where y increases as the value of x increases.
- **Negative correlation**: it indicates the relationship between x and y where y decreases as the value of x increases.
- **No correlation**: it indicates that there is non-linear relationship between x and y.
- r is dimensionless quantity.
- **Perfect correlation** exists when r = ±1

Coefficient of determination (r^2 or R^2):

- It defines the variance between 2 variables.
- The value of r^2 lies between 0 and 1 which determines the linear relationship between x and y
- It determines what percent of data is close to the best fit line.

$$r^2 = 1 - (RSS / TSS)$$

Where, RSS = Residual Squared Error

TSS = Total Sum of Square

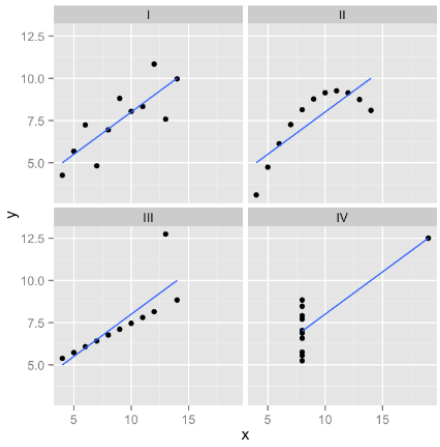
3. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet consists of four datasets having identical statistical properties but appear different when graphed.

Each dataset consists of eleven sets of x and y points.

It is used to demonstrate the importance of graph before analyzing it.

```
## Source: local data frame [4 x 6]
##
##   set mean(x) sd(x) mean(y) sd(y) cor(x, y)
## 1   I       9  3.32     7.5  2.03     0.816
## 2  II       9  3.32     7.5  2.03     0.816
## 3 III       9  3.32     7.5  2.03     0.816
## 4  IV       9  3.32     7.5  2.03     0.817
```



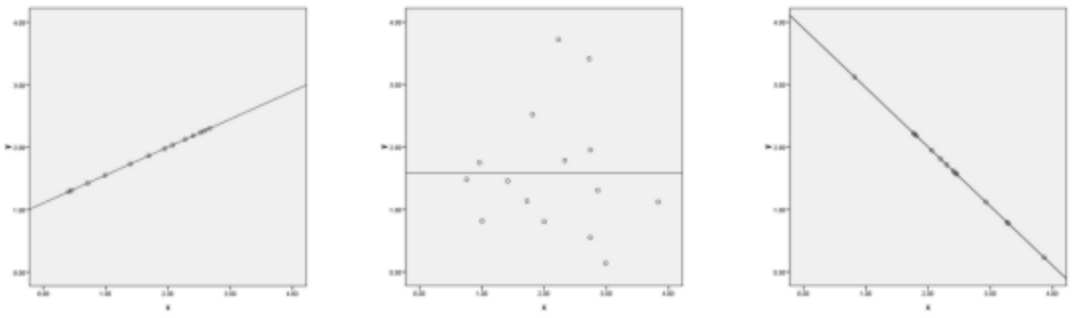
The above example shows how sets 1 to 4 statistical property of x and y are identical but are represented different while visualizing it.

4. What is Pearson's R?

Ans. Pearson's R determines the relationship between two variables by measuring the strength of the association between two variables.

It is useful when the relationship is linear between the variables.

$$-1 \leq r \leq 1$$



$r = -1$
perfect -ve correlation

$r = 0$
no correlation

$r = 1$
perfect +ve correlation

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

Ans. **Scaling** is the method to normalize the features of the data.

It is generally during data preprocessing step.

Reason behind scaling the features:

- Dataset might contain highly varying magnitude, units or range.
- And while calculating Euclidean distance between 2 data points, the features with high magnitude will have more weightage than the features with low magnitude.
- To avoid this, features need to be brought on the same level of magnitude.
- It also cause ease of interpretation and faster convergence for gradient descent method.

Two scaling method includes:

- Standardization
- Normalization

Standardization: it brings all the data into standard normal distribution with mean = 0 and standard deviation = 1

$$x' = \frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

Normalization (MinMax scaling): it brings all the data in the range 0 and 1

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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

Ans. Variance inflation factor (VIF) quantifies the correlation between two predictor.

It is used to diagnose multicollinearity.

The formula for calculating VIF is given by:

$$VIF = \frac{1}{1 - R^2}$$

Where, R^2 is the extent to which a predictor is correlated with other variables in linear regression.

The value of R^2 lies between 0 and 1. The higher the value of R^2, the better the fit.

And when R^2 is 1, then VIF tends to infinity. i.e., the model has learned all the dataset which lead to model overfitting.