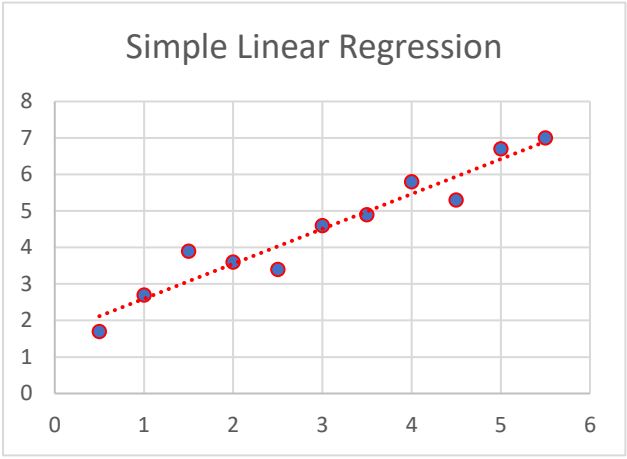
1. **Explain the linear regression algorithm in detail.**

Linear Regression is a kind of supervised machine learning model which attempts to explain the relationships between set of dependent/predictor variables and an independent/output variable using a straight line.



There are two kind of linear regression model present.

- Simple Linear Regression

  $Y = \beta_0 + \beta_1 X$, where
  β_0 -> Intercept
  β_1 -> Slope

- Multiple Linear Regression

  $Y = \beta_0 + \beta_1 X + \beta_2 X + \beta_3 X + \cdots + \beta_n X$
  β_0 -> Intercept
  $\beta_1 \ldots \beta_n$ -> slope

The objective of a linear regression model is to find the best fit line of a given dataset where we need to predict/project the output variable. Ordinary Least Squares Method is used to determine the best-fitting regression line which states that the sum of squares of residuals (RSS) which is also called Cost Function, should be minimum.

Residuals are defined as the difference between the Y of actual and Y of predicted data.

The equation of RSS for simple linear regression is defined as:

$$RSS = \sum_{k=0}^{n} (Y_k - \beta_0 - \beta_1 X_k)^2$$

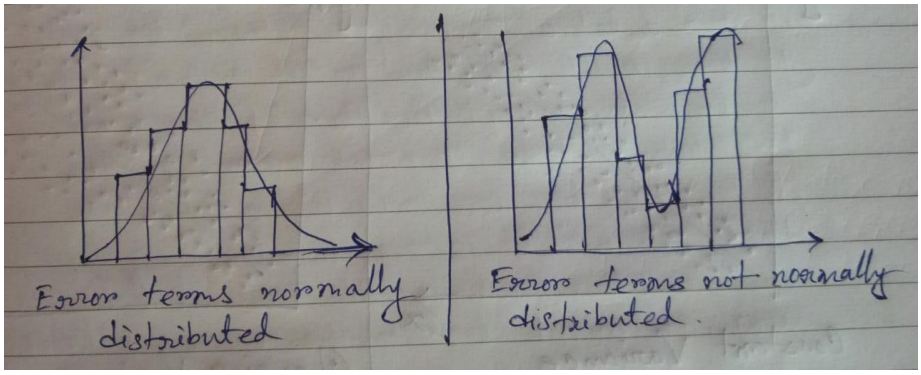
There are two methods popular methods which are used to optimize the cost function.

a. Closed form method: using differentiation find the slope and intercept.
b. Gradient Descent method: using iterative minimizations with learning rate.

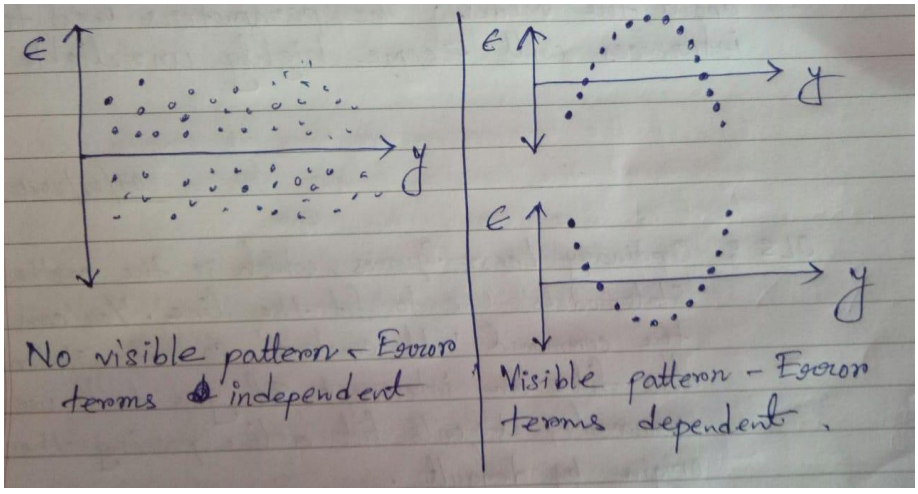
2. What are the assumptions of linear regression regarding residuals?

We know that Residuals are defined as the difference between the Y of actual and Y of predicted data. Now in order to fit a dataset in linear regression model there are some assumptions that need to be considered.

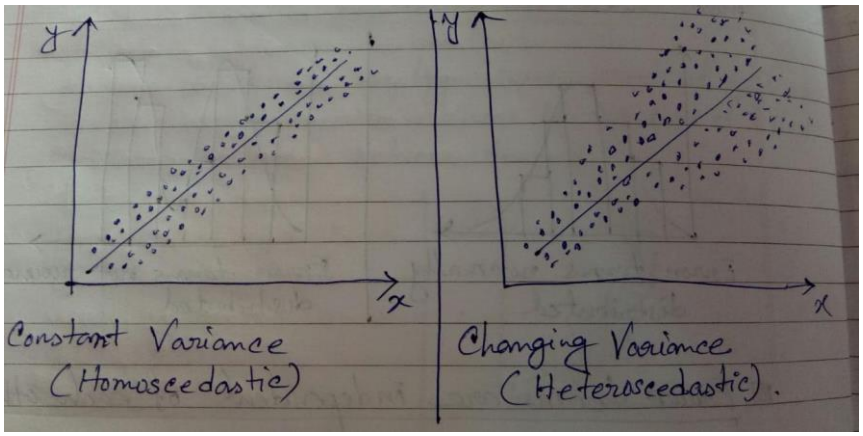
a) Error terms (residuals) are normally distributed with mean zero.
We can always fit a straight line and not make any prediction out of it. But if we need to use the regression line to make some inferences on the model then we need to have a notion of the distribution of the residuals.
If the error terms are not normally distributed the p-values obtained during the hypothesis testing which is used to determine the significance of the coefficients become unreliable. In case of the residuals being normally distributed the mean will equal to zero in most of the cases.



b) Error terms are independent of each other.
There should not be any kind of pattern in the error terms which means the error terms should not be dependent on each other.



c) Error terms have constant variance (homoscedasticity).
There should not be any pattern in variance as the error terms changes. In case of the error terms not being homoscedastic in nature, we cannot not rely on the prediction made by the model.



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

a) Coefficient of correlation: Correlation coefficient is a statistical value that defines the degree of relationship between two variables in dataset. It is a number between -1 to 1 which quantifies the extent to which two variables "correlate" with each other.
- If one increases as the other increases the correlation is positive.
- If one increases as the other decreases the correlation in negative.
- If one stays constant as the other varies, the correlation is zero.

b) Coefficient of determination: This is also referred as "R squared" value. The strength of a linear regression model is mainly explained by R squared value. The is explained by the below mathematical equation:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where,

RSS (Residual sum of squares)

RSS = $\sum_{k=1}^{n}(Y_k - \beta_0 - \beta_1 X_k)^2$, $\beta_0 \rightarrow intercept, \beta_1 \rightarrow slope$

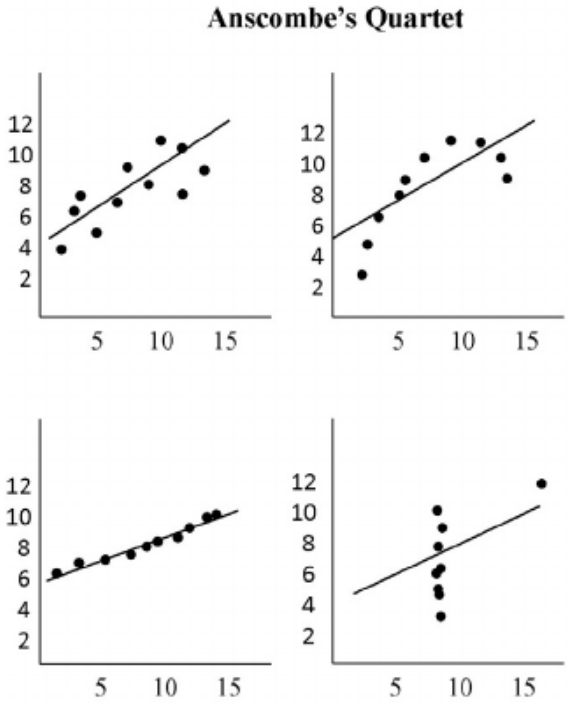
TSS (Total sum of squares)

$TSS = \sum_{k=1}^{n}(Y_k - \bar{y})^2$, $\bar{y}$ = mean of the variable

R squared value provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The R squared value is the square of the correlation coefficient between 2 variables, where the value of R squared varies between 0 and 1.

4. Explain the Anscombe's quartet in detail.

Anscombe's quartet was introduced by statistician Francis Anscombe. It demonstrates the importance of data visualization in data analysis. Anscombe's quartet contains a set of 4 datasets where all the datasets have identical descriptive statistics.



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

The datasets were prepared by statistician Francis Anscombe. If we visualize the datasets, we can clearly observe different kind of patterns in all the four datasets. This emphasizes the importance of visualization in Data Analysis.

5. **What is Pearson's R?**

Correlation between set of data is measures how well they behave. Pearson's R is a one such method which measures the correlation. The Pearson correlation coefficient or Pearson's r, can have a range of values from -1 to 1. A value of 0 signifies that there is no association between the two variables. A value greater than 0 signifies a positive correlation. A value less than 0 signifies a negative correlation.

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

The above equation calculates the Pearson's R between two sets of variables.

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

Scaling is a method which transforms a range of numerical values in a particular range.

In machine learning model building purposes scaling is performed in numeric variables which helps normalize the data as well as it helps in speeding up the calculation in algorithm.

Standardized Scaling: It brings all the data into a standard normal distribution with mean zero and standard deviation one.

$$x = \frac{x - mean(x)}{sd(x)}$$

Normalized Scaling: It brings all of the data in the range of 0 and 1.

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}$$

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

VIF or variance inflation factor of value 'x' signifies that the variance of the model coefficient is inflated by a factor of 'x' due to the presence of multicollinearity in the variables. Now if VIF value is high then we can state that there is a correlation present between the variables. The VIF of a given variable is represented by the below equation:

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

Where 'I' refers to the i-th variable which is being represented as the linear combination of rest of the independent variables.

If VIF value is more that 10 then the variable should be eliminated.

8. **What is the Gauss-Markov theorem?**

Gauss-Markov theorem states that if certain assumptions are considered as true then ordinary least square estimate for regression produces unbiased estimates that have the smallest variance of all possible linear estimators. The assumptions are below:

a) There has to be some linear relationship between dependent and independent variables.
b) The error terms have to be normally distributed with mean zero.
c) The error terms are independent of each other.
d) The error terms have constant variance i.e. homoscedasticity.

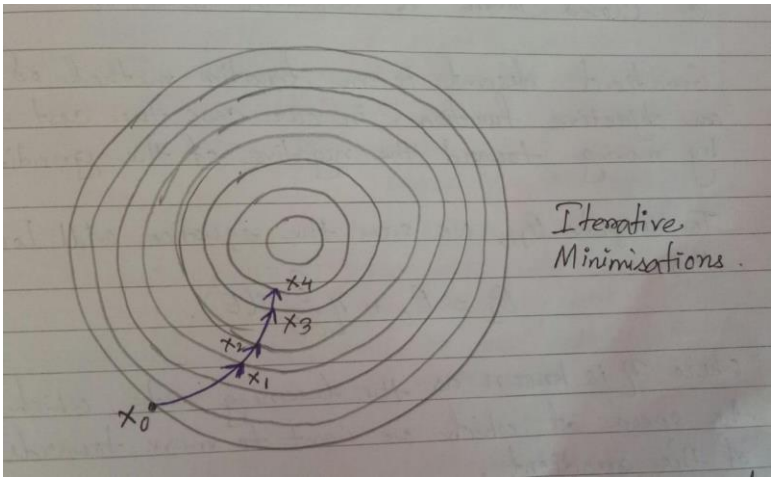
9. **Explain the gradient descent algorithm in detail.**

Gradient descent is a kind of unconstrained minimization algorithm which is used to minimize the cost function. The algorithm starts by assuming an initial value of the parameter and setting a learning rate. At initial value we calculate the output of the differentiated cost function and assign the same back to the parameter. For an example if we have the initial value as x_0, learning rate as α and the cost function as f(x) then the new parameter becomes,

$$x_1 = x_0 - f'(x_0)\alpha$$

We continue the process until the algorithm reaches an optimal point where the parameter value does not change effectively.

The parameter α is the learning rate and its magnitude decides the iterative steps of the algorithm. The range of α is (0,1].



The gradient descent algorithm is an iterative method of optimizing the cost function.

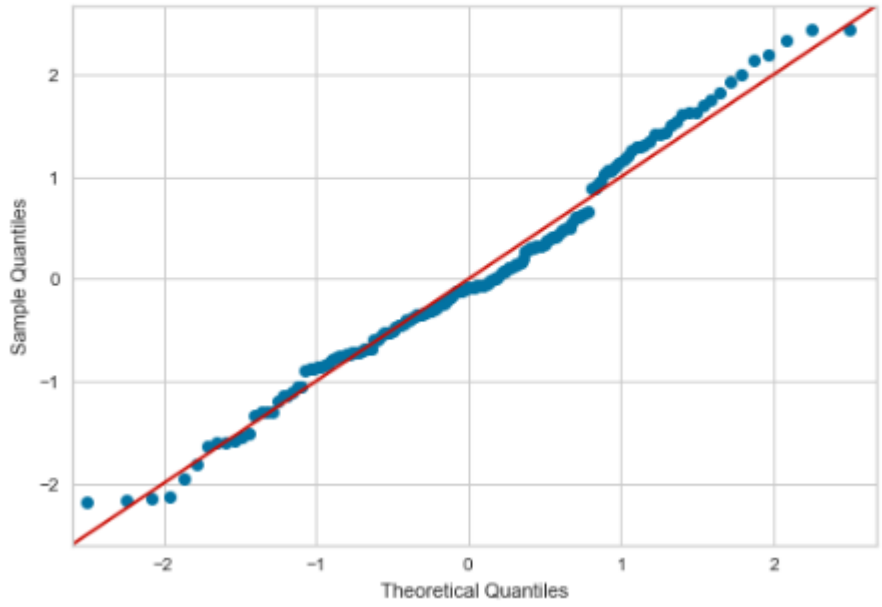
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot or Quartile-Quartile plot is used to compare two datasets and their distributions. If two datasets come from a population with similar distribution then the all the points should fall on the 45-degree reference line. As the plots depart from the reference line, we can conclude that the two datasets have come from a different distribution.

In linear regression the Q-Q plot signifies the measure of the relative location of the intercept and slope between the quantile. The intercept of a regression line is a measure of location, and the slope is a measure of scale. The distance between medians is another measure of relative location reflected in a Q–Q plot.

In linear regression this plot can be used to check the distribution of errors or residuals.

```
import statsmodels.api as sm
mod_fit = sm.OLS(y_train,x_train).fit()
res = mod_fit.resid # residuals
fig = sm.qqplot(res,fit=True,line='45')
plt.show()
```



The above example shows a fair distribution of errors / residuals along the reference 45-degree line.