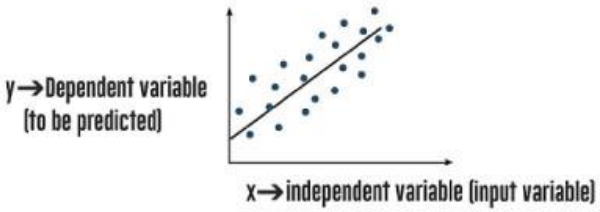
1. Explain the linear regression algorithm in detail.

Ans Linear regression algorithm is a machine learning algorithm based on supervised leaning. In this algorithm we find relationship between input and output considering that they are linearly dependent.

Simple Linear regression: Speaking about simple linear regression which has one depended and one independent variable, we try to predict depend variable by fitting a straight line through the 2D dataset.



Let the equation of line: $y=a+b$

So, In order to find the equation of straight line, we need to calculate intercept and slope value. Since there can be lot of lines passing through the dataset, best fit line is calculated using ordinary least square method. The line for which sum of errors is least is known as best fit.

After solving the equation for least sum of error squares(also known as cost function), value of a and be comes out to be

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Assumptions of Linear Regression: Linear regression model is based on below assumption:

a. Linear relationship between X and Y

b. Error terms are normally distributed (not X, Y)

c. Error terms are independent of each other

d. Error terms have constant variance (homoscedasticity)

Multiple Linear regression : But the real world liner regression problems are having much more independent terms. Therefore multiple linear regression model is used.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

Feature selection: As there are lots of variables, it comes with complexity like multicollinearity. So all the feature are generally not taken into final model. Some of the features which are not of much importance or are dependent on others are deleted using various methods: filter, wrapper and embedded methods. Iterative methods are used to remove the unnecessary variables.

Either analytical of numerical methods are used to minimize the cost function.

Residual Analysis: Simple linear regression models the relationship between the magnitude of one variable and that of a second—for example, as x increases, y also increases. Or as x increases, y decreases. Correlation is another way to measure how two variables are related. The models done by simple linear regression estimate or try to predict the actual result but most often they deviate from the actual result. Residual analysis is used to calculate by how much the estimated value has deviated from the actual result.

***Null Hypothesis and p-value*:** During feature selection, null hypothesis is used to find which attributes will not affect the result of the model. Hypothesis tests are used to test the validity of a claim that is made about a particular attribute of the model. This claim that's on trial, in essence, is called the null hypothesis. A p-value helps to determine the significance of the results. p-value is a number between 0 and 1 and is interpreted in the following way:

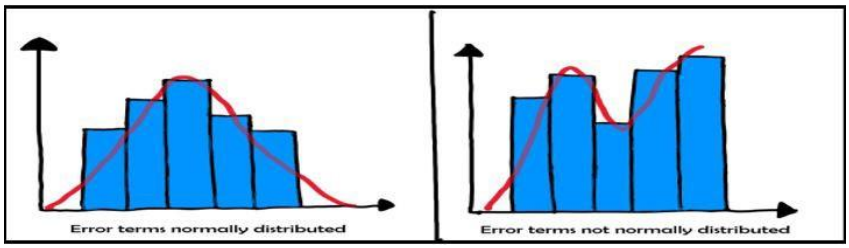
- A small p-value (less than 0.05) indicates a strong evidence against the null hypothesis, so the null hypothesis is to be rejected.
- A large p-value (greater than 0.05) indicates weak evidence against the null hypothesis, so the null hypothesis is to be considered.
- p-value very close to the cut-off (equal to 0.05) is considered to be marginal (could go either way). In this case, the p-value should be provided to the readers so that they can draw their own conclusions

Model Assessment: Overall model fitting is also checked using probability of F-statistics. R-squared value and adjusted R-squared values are calculated to assess the model. R-squared value denotes the extent of variance which can be explained by linear regression model. Root mean square error (RMSE) is also used to assess model. A low RMSE shows a better model.

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

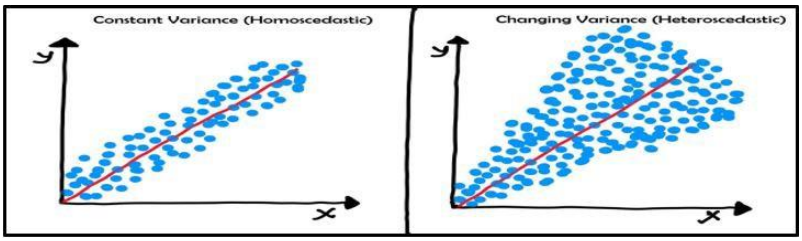
Ans. There are four assumptions in linear regression regarding residuals:

a) Normality assumption: The residuals are normally distributed. This assumption assures randomness in residuals. In the below figure, second figure is not valid for linear regression

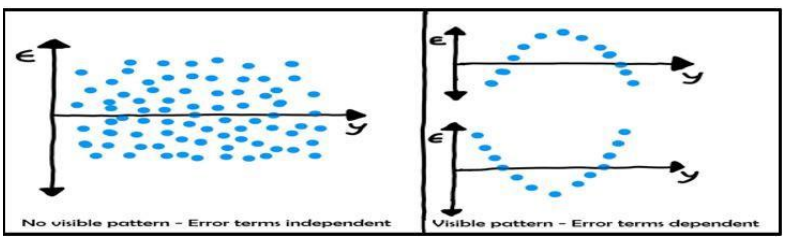


b) Zero mean of normal distribution: The residual are supposed to be normally distributed around zero mean.

c) Homoscedasticity: The residual should have same variance across complete distribution. Here also for second figure linear regression can not be used as the variance is large in latter part



d) Residual Independence: The residuals should be independent of each other i.e. there should be not pattern in the distribution. Presence of pattern shows that the model has missed some correlation between varaibles. First figure is okay but for second there is pairwise covariance.



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

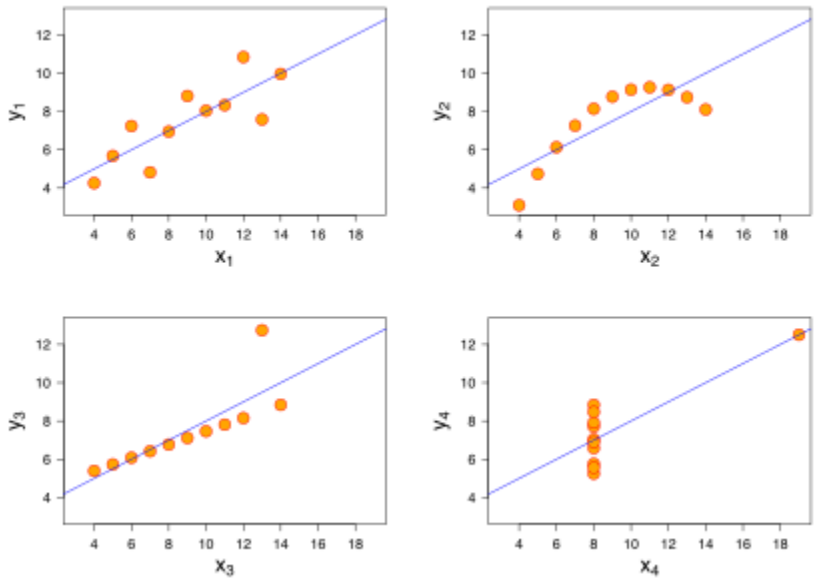
Ans. Coefficient of correlation: Coefficient of correlation is a statistic which shows the direction and strength of linear association of two variables. It is denoted by r. Its values lie between -1 to +1.

Negative value shows that when one variable increases, the other one decrease and vice versa. Positive correlation shows that when one variable increases or decreases so does other. Values of -1 denoted a perfectly negative fit linear association and +1 shows a perfectly positive fit linear association of variables. A value of zero shows that there is no correlation between variables.

Coefficient of determination: When coefficient of linear regression is squared, we get coefficient of determination. It is denoted by r^2. It gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. Therefore it is helpful determine how certain one can be in making predictions from a certain model/graph. A value of 1 shows that all points fall on the linear model. E.g. If r is 0.8, then r^2 will be 0.64. It indicates that 64% variance can be explained by our model, rest 36% remains unexplained.

4. *Explain the Anscombe's quartet in detail.*

Ans. Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises of four datasets containing 11 points each. The datasets have same descriptive summary: mean of x and y, variance, correlation, fitting line and coefficient of determination. But while plotting them on graph, they come to be totally different. All the graphs show different stories.



a) Dataset 1 shows a well-defined linear model.

b) Dataset 2 shows distribution is following a particular pattern and linear model is not able to explain it.
c) Dataset 3 shows all the points are following same line except one point and the the line is shifted because of that point.
d) Dataset 4 shows that linear models are highly sensitive to outliers.

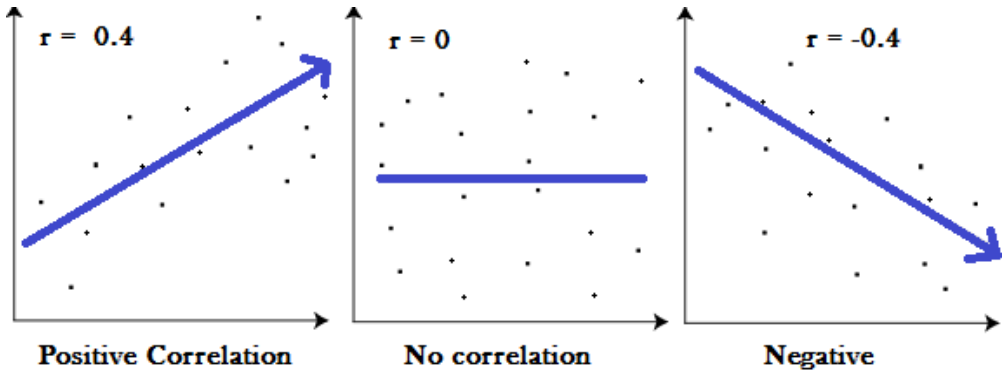
So, the dataset reject the general belief: 'numerical calculations are exact, but graphs are rough". It shows the importance of visualization.

5. What is Pearson's R?

Ans. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. It is given by formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2]\,[\, n\Sigma y^2 - (\Sigma y)^2]}}$$

Where x and y denotes the independent and dependent variables values and n denotes total number of observations. Value of Pearson's R varies between +1 and -1.



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

Ans. Scaling is a method of modifying the interval of features by stretching or shrinking.

In Linear regression, the general solution is denoted by:

$$y = X_1\beta_1 + X_2\beta_2 + \ldots + X_k\beta_k + \varepsilon.$$

Independent variables have different range of values eg: one variable can be in thousands while other can be between 0 and 1. So the coefficient will also vary according to the range. The higher the range, lower will be coefficient of variable and vice versa. So while defining the model, some of the coefficient will be very high and others may be too small and there is high possibility that they may be ignored. Interpretation becomes difficult in such situation. So, it is required to give equal importance to all coefficients and to achieve this scaling is done. Scaling also helps numerical methods to converge faster

Normalization: Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization: Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. Formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

If there are outliers in data, normalization shrinks it between 0 and 1, so it will shrink the data more compactly while standardization will show outliers in scaling.

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

Ans. VIF for any variable is calculated from formula:

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

An infinite VIF means that R_i^2 is 1 which means that the variable is perfectly dependents on other variables. So the variable must be deleted from final model.

8. What is the Gauss-Markov theorem?

Ans. Gauss-Markov theorem states that Ordinary Least Square (OLS) is the best estimators among unbiased estimators given that the model satisfies all the assumptions of linear regression. OLS will give the smallest variance among all other estimators. In other words, OLS is BLUE (Best Linear Unbiased Estimator)

9. Explain the gradient descent algorithm in detail.

Ans. Our aim is to find the best fit regression line in linear regression in order to predict value of dependent variable on the basis of dependent variable. We need to calculate the cost function and then minimize it.

Cost function

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

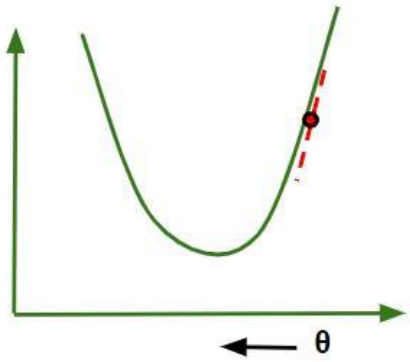
For a simple linear regression we have two values of θ_1 and θ_2. In Gradient Descent method, randomly two values of θ_1 and θ_2 are selected and then the values are updated iteratively to minimize the cost function until it reaches a minimum value. Using these finally updated values of θ_1 and θ_2 in the hypothesis equation of linear equation, model predicts the value of x in the best manner it can. The formula used in this method is:

Gradient descent formula:

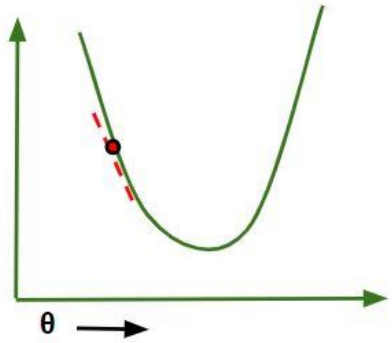
$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Where α is known as learning rate. Our convergence rate will depend on it. We graph cost function as a function of parameter estimates i.e. parameter range of our hypothesis function and the cost resulting from selecting a particular set of parameters. We move towards downward direction in the graph in order to find the minimum value. Way to do this is taking derivative of cost function as explained in the above formula. Gradient Descent step downs the cost function in the direction of the steepest descent. Size of steps strongly depend on learning rate

In the Gradient Descent algorithm, one can infer two points :



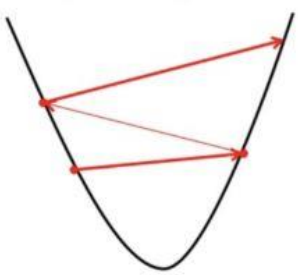
1. If slope is +ve : $\theta_j = \theta_j$ – (+ve value). Hence value of θ_j decreases.



2. If slope is -ve : $\theta_j = \theta_j - (-\text{ve value})$. Hence value of θ_j increases.

So learning rate is the deciding factor here:

1) If learning rate is too much, it may overshoot the minimum value



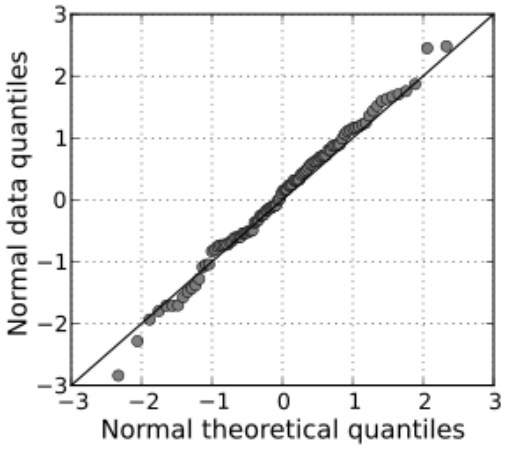
2) If learning rate is too small, convergence rate will be slow.

So, learning rate should be optimum one.

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

Ans. A quantile-quantile (Q-Q) plot is a graphical method in which quantiles for two probability distributions are plotted against each other. It helps in comparing two distributions in statistics. Also it can be used to check the validity of a distributional assumption for a data set. In general, the basic methodology is to compute the theoretically expected value for each data point based on the distribution in question. If the data indeed follow the assumed distribution, then the points on the q-q plot will fall approximately on a straight line (45^0 line). For example in the below figure the data points of sample ans standard normal are plotted with respected quantile values. As the

approximation can be done by a straight line, so we can say that the sample data is normally distributed.



When we have training and test data sets separately, then it is necessary that both the data sets should follow same distribution. This can be done by Q-Q plot. If we are able to draw a line (45 degree) then it means that both data sets are following the same distribution. If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis it means the distributions are different. So, the comparison can be made on unequal data sizes.