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

Linear regression:

- Supervised learning.
- It's one of the regression techniques.
- Main intention is to predict a dependent variable using independent variables provided the variables exhibit a linear relationship.

While building a linear regression model, following steps are performed:

- Data understanding and exploration
    - In this step, one usually analyzes the provided data to get statistical information (descriptive statistics) such how many columns and rows, what's the distribution of data etc.
    - Data is visualized through different kinds of plots such as histogram, box plots, bar charts and correlation matrix
    - While exploring, one will come to know about the problems such as missing values, null values, outliers etc. in the data.
- Data cleaning
    - The problems observed (such as missing values, duplicate values, outliers) during the data exploration are taken care in this step.
    - Operations such as null values removal, imputation of values, treating outliers are usually performed in this step.
- Data preparation
    - Handling categorical variables.
        - Linear regression can be performed only with numerical predictor variables.
        - So, all the categorical variables need to be converted to numerical variables. This can be done in the following ways.
            - Creating dummy variables (one hot key encoding)
                - This method can be used for un-ordered categorical variables such as color of the car, company of the car, gender of a person etc.
            - Label encoding
                - This can be used for ordered categorical variables such as rating of a restaurant, employee designation, car engine cylinder number etc.
    - Train-test split of the data
        - Data is split into train set and test set.
        - Train set is known to the model. It's used to learn and generalize the other (unknown) data set.
        - After generalizing the model using the train data, it is tested on test data to see how well it can predict values on an unknown data.
    - Feature Scaling

        - To interpret the coefficients better it's recommended to have all the variables on the same scale.
        - Also, it makes the Gradient Descent process much faster.
        - Scaling can be done in the following ways:
            - Standardization – make mean as 0 and standard deviation 1.
                X = (X – Mean) / standard deviation
            - MinMax Scaling / Normalization – scale between 0 and 1.
                X = (X-Xmin) / (Xmax-Xmin)
            - Usually categorical variables are not scaled.
- Model building and evaluation
    - Data preparation is done and now model can be built.
    - There are many libraries in Python/R to build the model.
    - Statsmodel and sklearn are the popular python libraries to build model.

    - Steps usually followed:
        - Create X and Y set using the train set
        - Feature Selection.
            - Can be done manually, if the columns are less (Forward/Backward/Stepwise selection).
            - Can be done using Recursive Feature Elimination (RFE).
            - Recommended approach is to use the combination of automated and manual selection.
        - Fit a model with the selected features.
        - Measure the statistical significance of the coefficients (by checking the p-values) and collinearity (using the VIF values)
        - Remove the features which are not significant (p-values > 0.05) or highly correlated (VIF > 5).
        - Repeat the above three steps until a model is generated with acceptable adjusted r-squared, p-values, F-stat and VIF.
    - Check assumptions for Residual Analysis.
        - Assumptions are
            - Error terms are normally distributed with mean zero
            - Error terms are independent of each other.
            - Error terms have constant variance (homoscedasticity).
        - Steps:
            - Predict the values on train data and test data sets
            - Calculate the residuals. (yi – ypred)
            - Plot the histogram – Check the trend
            - Check the adjusted r-squared of train set and test set. It should be comparable; difference should not be more/less than 5%.
            - If it's more than 5%, add or remove features to the model and evaluate further.

If all steps are performed correctly, the final model will be the best fit line for the given data.

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

Assumptions are:

- Linear relationship between dependent and independent variable.
- Error terms are normally distributed.
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity).

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

Coefficient of correlation, R:

- It quantifies the extent to which two variables correlate with each other.
- Range of R is -1 to +1.
    - Where +1/-1 indicates a strong relationship between two variables.
    - 0 indicates that variables are not correlated
- if one variable increase as other increases, then the correlation is Positive.
- if one variable decrease as other increases, then the correlation is Negative.
- One of the popular measures is Pearson correlation

Coefficient of determination, R-squared:

- It is the measure of variance of a variable explained by a model. Higher the -squared better the model.
- For example, if R-squared of a model is 0.8, it means that the model can explain the 80% of variance of the data points.
- Coefficient of determination, R-squared is the square of Coefficient of correlation, R.
- Formula,

    R-squared = 1 – (Residual sum of squares, RSS / Total sum of squares, TSS)

    Where,

    RSS = sum of squares of errors

    Where error, ei = yi – ypred

    TSS = sum of (yi – y mean) ^ 2,

    = (y1 – y mean) ^ 2 + (y2 – y mean) ^ 2 + … + (yn – y mean) ^ 2

- In other words, R-squared tells how good the model is.

4. Explain the Anscombe's quartet in detail.

Anscombe's quartet demonstrate how important is to plot the data rather than relying only on summary statistics. It comprises four data set that have nearly identical descriptive statistics yet have very different distributions and appear very different when graphed. These data sets were constructed in 1973 by a 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.

Data set:

The average x value is 9 for each dataset

The average y value is 7.50 for each dataset

The variance for x is 11 and the variance for y is 4.12

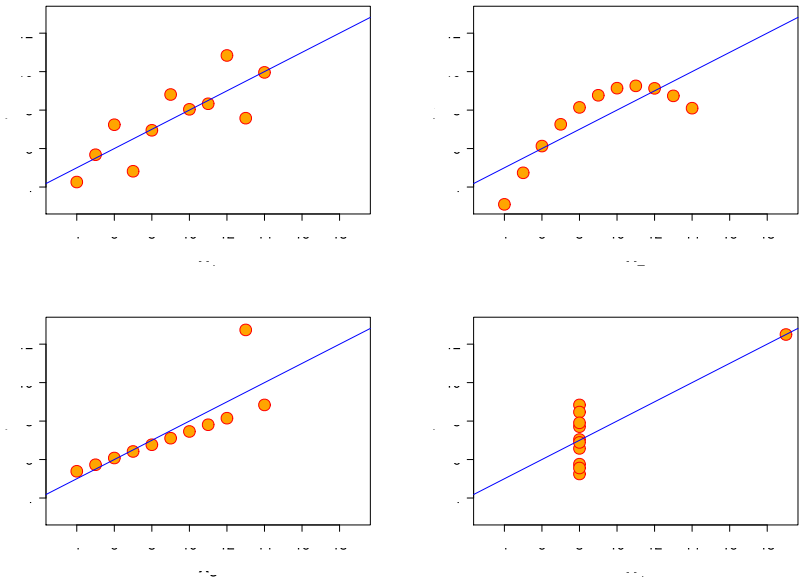
The correlation between x and y is 0.816 for each dataset

A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

Anscombe's quartet

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

The graphs for these data sets are given below:



Reference: Wikipedia

5. What is Pearson's R?

Pearson's R is a type of correlation coefficient. It's the most common measure of correlation. R can have any value in the range -1 to 1. The absolute value indicates how strong the relation between the two variables. 1 indicates very strong relationship and 0 indicates no relationship. A positive sign indicates that two variables increase together, and negative sign indicates that they reduce together.

One important point is that Pearson's R is the measure of linear association between two variables. That is, it's a quantification of how well the association is represented by a straight line. For instance, two variables may be highly related to each other, but Pearson's R may be zero. This could mean that the variables may not be linearly related. In this case, it's recommended to use other correlation coefficient measures such as Spearman's R.

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

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

If a feature's variance is orders of magnitude more than the variance of other features, that feature might dominate other features in the data set. This make the interpretation difficult. Another reason is that some libraries require all the variables in the same scale. Also, Gradient Descent algorithm performs faster on rescaled values.

Standard scaling:

The result of standardization (or Z-score normalization) is that the features will be rescaled so that they'll have the properties of a standard normal distribution with

$\mu=0$ and $\sigma=1$

where μ is the mean (average) and σ is the standard deviation from the mean; standard scores (also called z scores) of the samples are calculated as follows:

$z=x-\mu\sigma$

MinMax scaling:

Min-Max scaling often also simply called "normalization". In this approach, the data is scaled to a fixed range - usually 0 to 1.

A Min-Max scaling is typically done via the following equation:

$Xnorm=X-XminXmax-Xmin$

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

Variance Inflation Factor (VIF) gives a basic quantitative idea about how much the feature variables are correlated with each other. In other words, it tells us the collinearity exist in the model. It is an extremely important parameter to test our linear model. Higher the VIF, higher the collinearity.

VIF = 1 / (1-R-squared)

So,

If VIF = Infinite, then (1 – R-squared) = 0,

R-squared = 1, R = 1 => this indicates that the variables are perfectly correlated (strong correlation). In other words, the value of that independent variable can be predicted by other independent variables. For instance, two columns with exactly same values can lead to infinite VIF.

8. What is the Gauss-Markov theorem?

The Gauess-Markov theorem states that if certain assumptions are met, the ordinary least squares estimate for regression coefficients gives the best linear unbiased estimate possible.

Assumptions:

Gauss-Markov assumptions guarantee the validity of OLS for estimating regression coefficients.

Linearity

- The parameters that are being estimated using OLS must have a linear relationship.

Random

- Data is randomly sampled from the population.

Non-Collinearity

- The regressors being calculated aren't perfectly correlated with each other.

Exogeneity

- The regressors aren't correlated with the error term.

Homoscedasticity

- The error variance is constant

Validating these assumptions is one of the important steps in estimating the coefficients. When the assumptions are violated, one may have to tweak the model/rebuild the model to fit the ideal Gauss-Markov assumption more closely.

In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they indicate what 'ideal' conditions would be. They also help pinpoint problem areas that might cause estimated regression coefficients to be inaccurate or even unusable.

It's represented in algebra, by saying that a linear regression model $yi = xi' \beta + \varepsilon i$ and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

$E\{\varepsilon i\} = 0, i = 1, . . ., N$

$\{\varepsilon 1 \ldots\ldots \varepsilon n\}$ and $\{x1, \ldots., xN\}$ are independent

$cov\{\varepsilon i, \varepsilon j\} = 0, i, j = 1,\ldots., N\ I \neq j.$

$V\{\varepsilon 1 = \sigma 2, i= 1, \ldots.N$

9. Explain the gradient descent algorithm in detail.

Gradient descent is one of the popular optimization algorithms in machine learning. It's a first order optimization algorithm (iterative form). It means that the algorithm calculates the first order derivative of the cost function while updating the parameters. On each iteration, the parameters get updated in the opposite direction of the gradient of the function J(w) with respect to the parameters where the gradient gives the direction of the steepest ascent. The rate of the step we take on each iteration to reach the minima is called learning rate.

Learning rate:

- If learning rate is very small, it would take a long time to converge.
- If the learning rate is large, it may fail to converge and overshoot the minimum (or oscillates)

Formula:

Cost function, J(Theta),

Theta(t) = Theta(t-1) – learning rate * (partial derivative cost function at t-1), where Theta(t-1) =>

value of theta at t-1 step

Let's understand the algorithm using an example.

Assume that a logical regression model has two parameters: weight (w) and bias (b)

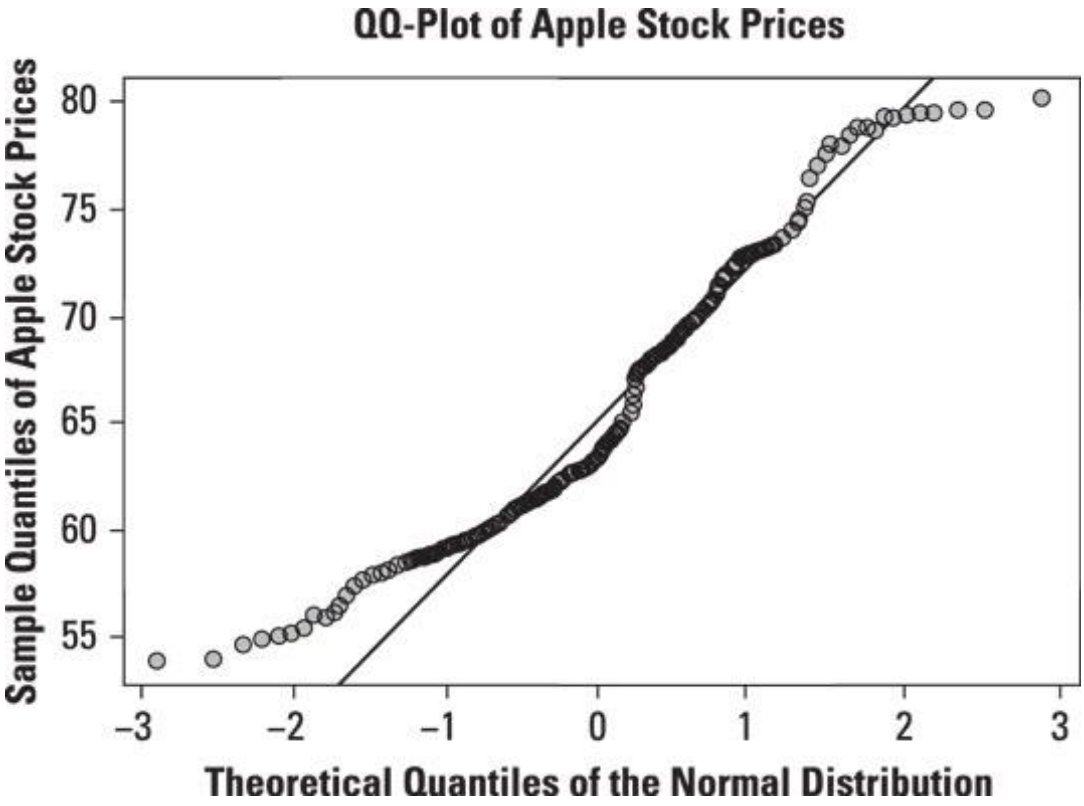
1. Initialize weight w and bias b to any random numbers.
2. Pick a value for the learning rate, α. The learning rate determines how big the step would be on each iteration.
3. Scale the data if necessary.
4. On each iteration, take the partial derivative of the cost function J(w) w.r.t each parameter (gradient). Then update the parameters,
   w = w – learning rate * partial derivative of cost function w.r.t w
   b = b – learning rate * partial derivative of cost function w.r.t b
5. Continue the process until the cost function converges. That is, the partial derivatives are zero and w/b values don't change further.

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

Quantile-Quantile (QQ) plots are used to determine if data can be approximated by a statistical distribution. For example, you might collect some data and wonder if it is normally distributed. A QQ plot will help answer that question. QQ plots are also used to compare to different datasets to determine if their distributions are comparable.

The basic idea 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.

For example, this figure shows a normal QQ-plot for the price of Apple stock from January 1, 2013 to December 31, 2013.



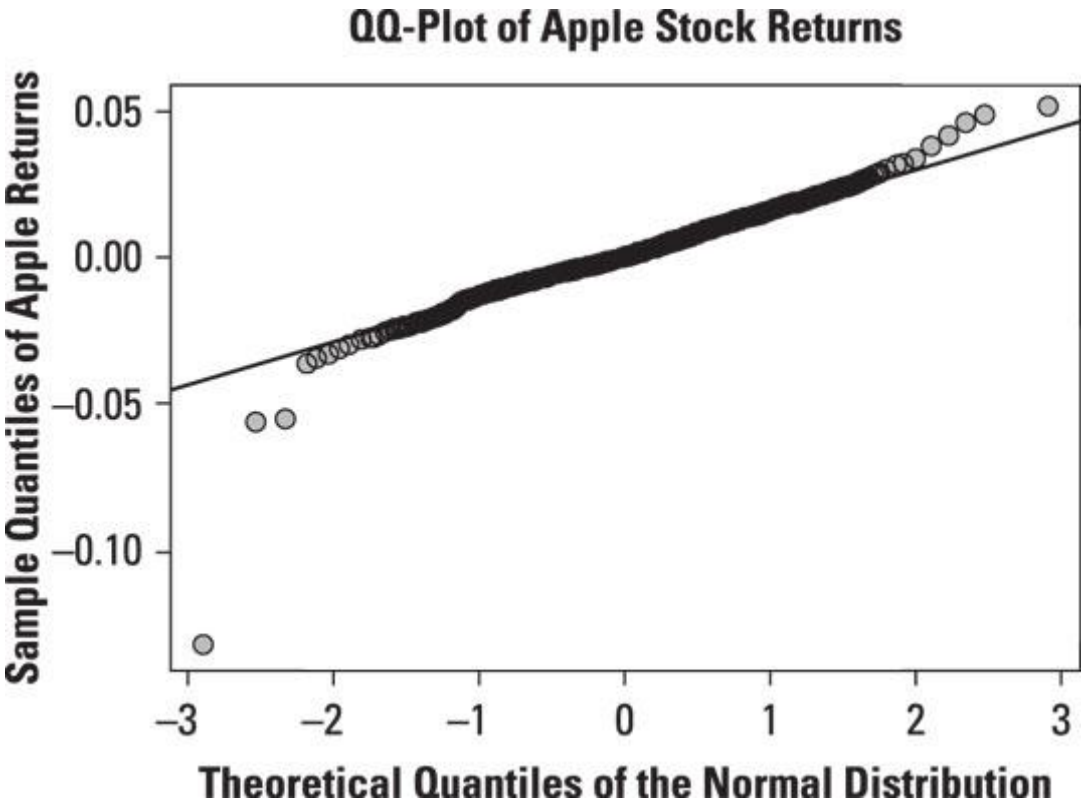
Normal QQ-plot of daily prices for Apple stock.

The QQ-plot shows that the prices of Apple stock do not conform very well to the normal distribution. In particular, the deviation between Apple stock prices and the normal distribution seems to be greatest in the lower left-hand corner of the graph, which corresponds to the left tail of the normal distribution. The discrepancy is also noticeable in the upper right-hand corner of the graph, which corresponds to the right tail of the normal distribution.

The graph shows that the smallest prices of Apple stock are not small enough to be consistent with the normal distribution; similarly, the largest prices of Apple stock are not large enough to be consistent with

the normal distribution. This shows that the tails of the Apple stock price distribution are too “thin” or “skinny” compared with the normal distribution. The conclusion to be drawn from this is that the Apple stock prices are not normally distributed.

This figure shows a normal QQ-plot for the daily returns to Apple stock from January 1, 2013 to December 31, 2013:



Normal QQ-plot of daily returns to Apple stock.

The QQ-plot shows that the returns to Apple stock do not conform to the normal distribution, either. In this case, the smallest returns to Apple stock are too small to be consistent with the normal distribution. Similarly, the largest returns to Apple stock are too large to be consistent with the normal distribution. This shows that the tails of the Apple return distribution are too “thick” or “fat” compared with the normal distribution. Therefore, Apple returns are not normally distributed.

In linear regression, Q-Q plots (along with residual plots) are used to visually check that data meets the homoscedasticity and normality assumptions. It compares the distribution of the data to a normal distribution by plotting the quartiles of the data against the quartiles of a normal distribution. If the data is normally distributed, then they should form an approximately straight line.