Subjective Question and Answers

Ques-1: Explain the linear regression algorithm in detail.

Answer: Linear regression is one of the simplest machine learning algorithms. Linear regression is a type of statistical modelling that permits you to explore whether one a variable is reliant on others. The connection between variables is justified by a trend-line which is overlaid on your data and can be used for predicting numerous unique things.

There are two types of regression i.e. Linear regression and Multiple linear regression.

When you're discovering the relationship between two variables, an independent variable and dependent variable (x and y), the method of regression used to question this is described to as a simple linear regression.

If you're comparing one dependent and multiple independent variables, then you're doing multiple linear regression.

The equation for simple linear regressions is:

$$y = b_0 + b_1 * x_1$$

Let's break this equation down.

y is the dependent variable.

x is the independent variable.

b_1 is what's known as the coefficient for the independent variable.

b_0 is the constant term,

If you have multiple variables the equation looks almost the same, except you add more independent variables into the mix:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

Question: What are the assumptions of linear regression regarding residuals?

Answer: The Regression has five key assumptions:

1.Normality Assumption: it is considered that error terms are normally distributed.

2.Zero mean assumption: It is considered that the residuals have mean value of zero i.e. as already stated the error terms are normally distributed around zero.

3.Constant Variance Assumption: it is considered that the residuals have the same variance. This assumption is also known as Homogeneity or Homoscedasticity.

4.Independent error assumption: It is assumed that residual terms are independent of each other i.e. pairwise correlation is zero.

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

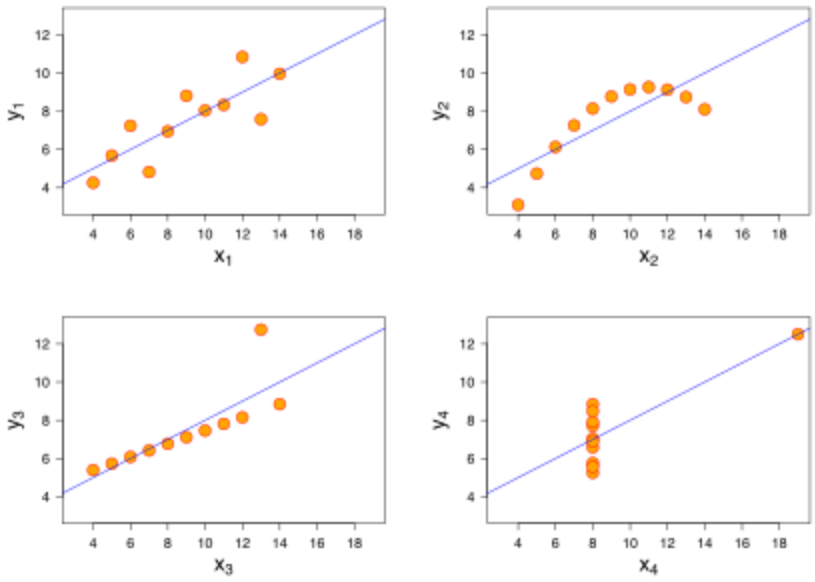
Answer: Coefficient of Determination is the square of Coefficient of Correlation.

R square or coefficient of determination shows percentage difference in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

Coefficient of Correlation is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in agreement. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated, then the correlation value can still be computed which would be 0. You can explain R square for both simple linear regressions and for multiple linear regressions.

Question:4 Explain the Anscombe's quartet in detail.

Answer: **Anscombe's quartet** contains four data sets that have nearly the same simple descriptive statistics yet have very diverse allocations and show very dissimilar when graphed. Each dataset involves of eleven (*x*, *y*) points. They were formed in 1973 by the statistician Francis Anscombe to explain both the significance of graphing data before analysing it and the effect of outliers and other powerful reflections on statistical properties. He termed the paragraph as being planned to counter the notion among statisticians that "numerical calculations are exact, but graphs are rough.



As we can see , all the four linear regression are the same. But there are some peculiarities in the data sets that have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression in model's sensitivity to outliers. Had the outlier not been present, we could have got a great line fitted through the data points. So, we should never run a regression without having a good luck at out data.

Question: What is Pearson's R?

Answer: Correlation coefficient methods are used to find how solid a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

One of the most used formulas in stats is Pearson's correlation coefficient formula. This correlation coefficient is designed for linear relationships and it might not be a good measure for a nonlinear relationship between the variables

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

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling means that you're converting your data so that it matches within a certain scale, like 0-100 or 0-1. You want to scale data when you're using techniques established on methods of how much apart data points like support vector machines or k-nearest neighbours. With these processes, a change of "1" in any numeric feature is given the same importance.

By scaling your variables, you can help compare different variables on equal balance.

Normalization rescales the values into a range of [0,1]. This might be beneficial in some situations where all factors need to have the same positive scale. Though, the outliers from the data set are lost.

$$Xchanged=X-XminXmax-XminXchanged=X-XminXmax-Xmin$$

Standardization rescales data to have a mean of 0 and standard deviation of 1 (unit variance).

$$Xchanged=X-\mu\sigma$$

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

Answer: Variance inflation factor (VIF) is used to check the presence of multicollinearity in a dataset.

If the value of VIF is high for a variable, it means that R2 value of corresponding model is high i.e. we already have the variable which is correlated with this variable in our model.

VIF can be calculated through many means some of them are mathematical formula or through python code.

If there is perfect correlation, then VIF = infinity. A huge value of VIF suggests that there is a correlation amongst the variables.

Question: What is the Gauss-Markov theorem?

Answer: The Gauss Markov theorem informs us that if a specific set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.

There are five Gauss Markov assumptions (also called *conditions*):

1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.

2. **Random**: our data must have been randomly sampled from the population.
3. **Non-Collinearity**: the regressors being calculated aren't perfectly correlated with each other.
4. **Exogeneity**: the regressors aren't correlated with the error term.
5. **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

We can review the Gauss-Markov Assumptions concisely in algebra, by saying that a linear regression model represented by
$y_i = x_i \text{ 'β} + \varepsilon_i$

Question: Explain the gradient descent algorithm in detail?

Answer: Gradient descent is a very simple optimization algorithm. It makes iterative movements in the direction opposite to the gradient of a function at a point.

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

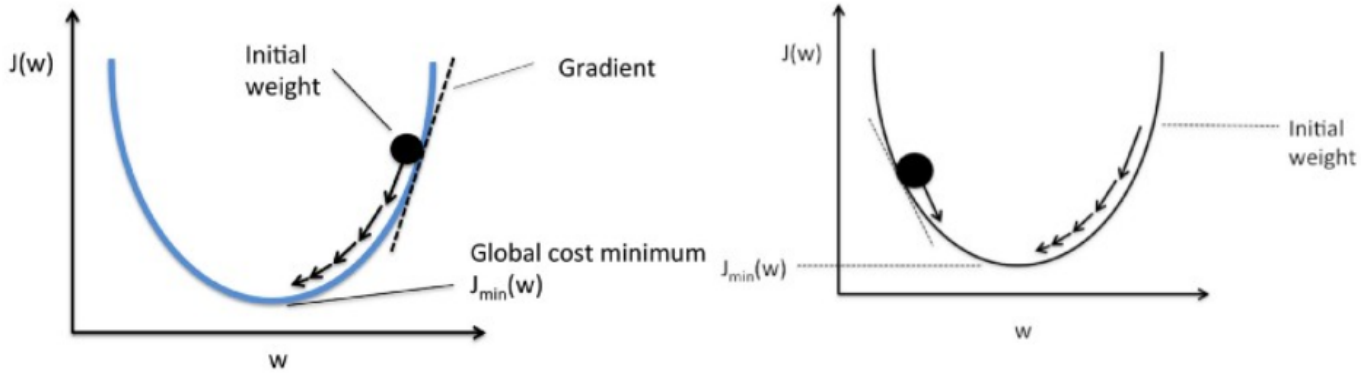
$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)}$$

$$\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_2^{(i)}$$

It is easy to understand if we visualize the procedure.

Assume we are trying to find the minimum of the function f(x)=(x−1)2f(x)=(x−1)2
(we know the minimum is at x=1. Let's try to apply gradient descent)

We must make an initial guess for the minimum. Let our initial guess be x (0) = -1



According to Gradient descent algorithm, our new point x (1) =x (0) −ηf'(x(0)) x(1)=x(0)−ηf'(x(0))

where ηηis the adjustable step size (let's fix it as 0.6) and f'(x)f'(x)denotes the derivative of f(x)f(x) at point x.x. The derivative of (x−1)2(x−1)2is 2(x−1).2(x−1).
So, x (1) =−1−x (1) =−1−0.6∗2∗−2=0.6∗2∗−2= 1.4

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

Answer: The **Q-Q plot**, or **quantile-quantile plot**, is a graphical tool to help us assess if a set of data reasonably came from some hypothetical distribution such as a Normal or exponential. .. If both sets of quantiles came from the same distribution, we should see the points forming a line that's almost straight.

The Q-Q Plots or Quantile-Quantile Plots exceeds all the limits of the Histogram plot. Let's take an example:

Consider a sample data from a population. Building of Q-Q plots starts with the order the sample data from smallest to largest. Let k denote the ranking or order number. Therefore k=1 for the smallest and k = n for the largest. The q-q plot is because the ordered value of k is an estimate of (k-1/2)/n quantile of the sample data. In other words, the ordered values are close to inverse of cumulative distribution of (k-0.5)/n, where n is the sample size. If the cumulative distribution function belongs to an appropriate known distribution, then the plot of ordered values and the known cumulative distributional values will approximately form a straight line. On the other hand, if the assumed distribution is inappropriate, the points will deviate from the straight line in a systematic manner. Therefore, if we assume that the cdf is from a normal distribution, then obtaining a straight line after the plot confirms that the sample data indeed belongs to the normal population.