Linear Regression Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer:

Linear regression is the basic model and has been used since almost around 200 years, hence it is also called a linear model.
Example: This model assumes a linear relationship between the Input Variables (X) and the Single Output Variable (Y). That is, Y can be calculated from a linear combination of the input variables (X).

There are two kinds of Linear Regression:
One in which there is only one Input Variable which is known as Simple Linear Regression.
And
Other is which there are more than one Input Variables hence also referred to as Multiple Linear Regression.
Equation for a Simple Linear Regression:

$$\mathbf{Y = B0 + B1*X}$$

From the above equation has the below terms.

- Capital Greek letter Beta (B) is the coefficient factor.
- Intercept is denoted by B0
- Slope of the Equation is denoted by B1
- X is the Input Variable
- Y is the Output Variable

Linear Regression is used in finding a relationship between 2 continuous variables. The main objective of this model is to find a line that best best fits the data.

Question 2: What are the assumptions of linear regression regarding residuals?

Answer:

Residual plot helps in analyzing the model using the values of residues. It is plotted between predicted values and residue. Their values are standardized. The distance of the point from 0 specifies how bad the prediction was for that value. If the value is positive, then the prediction is low. If the value is negative, then the prediction is high. 0 value indicates prefect prediction. Detecting residual pattern can improve the model.

Characteristics of a residue:
Residuals do not exhibit any pattern
Adjacent residuals should not be the same as they indicate that there is some information missed by the system.

The Main Assumptions that are made by Linear Regression are:

- Linear Relationship
- Multivariate Normality
- No or Little Multicollinearity
- No Auto-Correlation
- Homoscedasticity

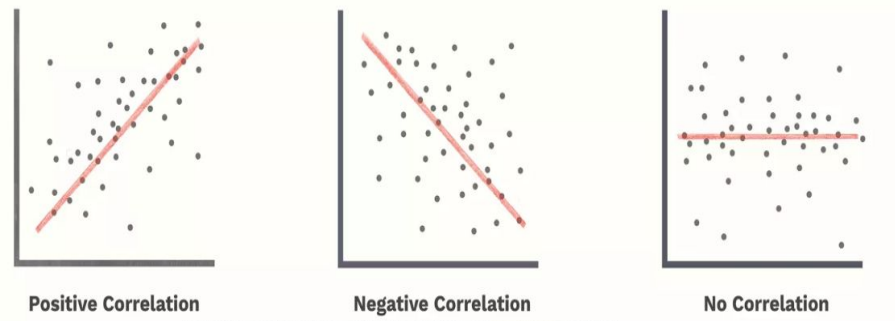
Question 3: What is the coefficient of correlation and the coefficient of determination?

Answer:

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression.

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

1 indicates -> strong positive relationship.
-1 indicates -> strong negative relationship.
0 indicates -> no relationship at all.



The **coefficient of determination(R2)** is key output of regression analysis. It is interpreted as the proportion of the variance in the dependent variable.

The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
In case of linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
If:
R2 = 0 -> means that the dependent variable cannot be predicted from the independent variable.
R2 of 1 -> means the dependent variable can be predicted without error from the independent variable.
R2 between 0 and 1 -> indicates the extent to which the dependent variable is predictable. An R2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R2 of 0.20 means that 20 percent is predictable; and so on.

$$R2 = \{ (1 / N) * \Sigma [(xi - x) * (yi - y)] / (\sigma x * \sigma y) \}2$$

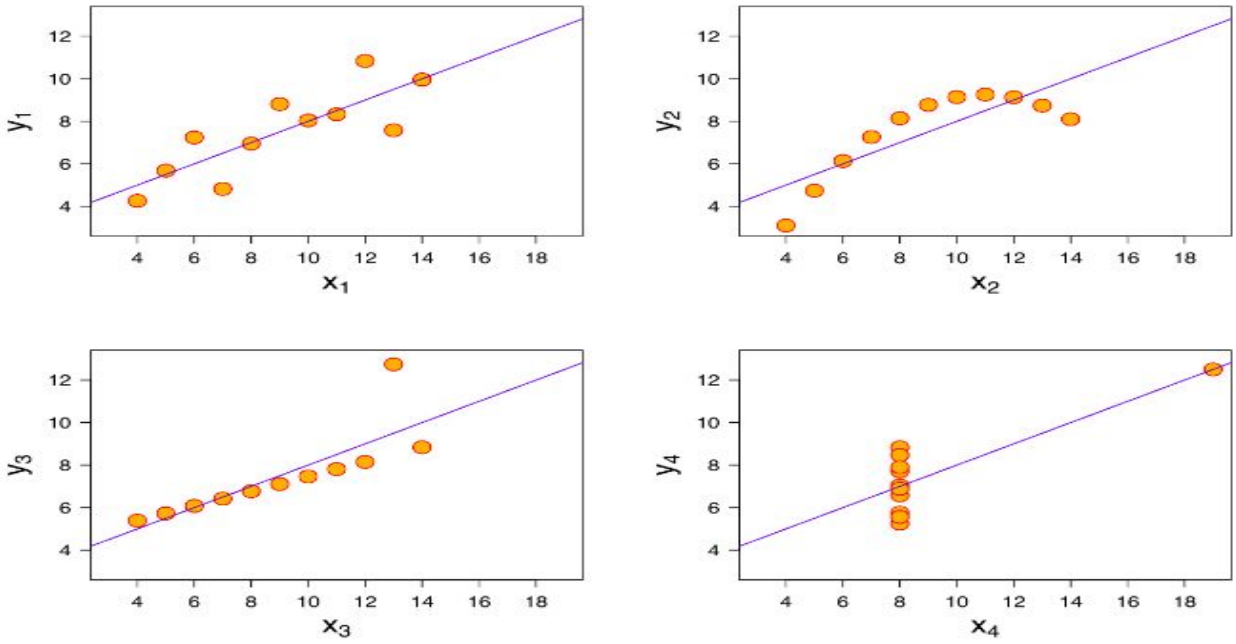
Question 4: Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet comprises of 4 datasets. Each containing eleven (x,y) pairs.
All these datasets share the same descriptive statistics. But, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups :
Mean of x is 9 and mean of y is 7.50 for each dataset.
Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset
When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I has clean and well-fitting linear models.
- Dataset II is not distributed normally.
- Dataset III the distribution is linear, but Outlier is present
- Dataset IV has one outlier

Question 5: What is Pearson's R?

Answer:

Pearson correlation coefficient or Pearson's R, is a measure of the **linear correlation between two variables X and Y.**

a Pearson product-moment correlation **attempts to draw a line of best fit through the data of two variables**, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:

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

Answer:

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.

It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

Formula for Normalization:
$X_{changed}=(X-X_{min})/(X_{max}-X_{min})$

Formula for Standardization:
$X_{changed}=(X-\mu)/\sigma$

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

Answer:

If there is perfect correlation, then **VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.** If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

If VIF is large and multicollinearity affects our analysis results, then we would need to take some corrective actions before using multiple regression.

Question 8: What is the Gauss-Markov theorem?

Answer:
the Gauss–Markov theorem states that in a **linear regression model in which the errors are uncorrelated,** have **equal variances** and **expectation value of zero**, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists.

Question 9: Explain the gradient descent algorithm in detail.

Answer:

There are three kinds of optimization algorithms:

- Optimization algorithm that is not iterative and simply solves for one point.
- Optimization algorithm that is iterative in nature and converges to acceptable solution regardless of the parameters initialization such as gradient descent applied to logistic regression.
- Optimization algorithm that is iterative in nature and applied to a set of problems that have non-convex cost functions such as neural networks. Therefore, parameters' initialization plays a critical role in speeding up convergence and achieving lower error rates.

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).

The **procedure** starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

Answer:
A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions

QQ-plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don’t, our residuals aren’t Gaussian and thus our errors aren’t either. This implies that

for small sample sizes, can't assume our estimator \hat{\beta} is Gaussian either, so the standard confidence intervals and significance tests are invalid.

Let's fit OLS on an R datasets and then analyze the resulting QQ plots.

plot(lm(dist~speed,data=cars))