Linear Regression Questions and Answers

By - Sangram Sinha

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

Answer: **Assumption I:** The error terms should be normally distributed with mean equal to zero.

So, once we build the model, we need to check that it is not violating this assumption. We need to check this by plotting histogram of the error terms to find the model is normally distributed.

Assumption II: The error terms must be independent to each other, which means while plotting with X or y we need to check for any patterns. If there is no visible pattern in the error terms, then we can say our model fit has not violated this assumption and we are good to go.

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

Answer: **Coefficient of correlation(r)**: It tells us the direction and strength of a linear relationship between two variables. If we have positive correlation, then if one variable increases the other one will also increase and if we have negative correlation, then if one variable increases the other one will decrease. If we have no correlation then there is a random, non-linear relationship with both variables.

Coefficient of determination(r^2): It tells us the variance of one variable from the other variable. It also tells us the percentage of the data that is closest to best line fit in regression. Example: If we have r= 0.872 and r^2 =0.834 then we can say that 83% of the variance in y from x can be explained in a linear regression equation. So higher the r^2 better it is as it covered most of the data points.

3. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet tells us why visualization is important. It has four datasets of identical descriptive statistics but different distribution which can be shown in a graph and contain 11 datasets.

The first graph shows there is a linear relationship between x and y and there maybe a normal distribution but not the same for second one. It is non-linear and not normally distribution, but it can tell the variance. It tells than even though we have non-linear we can have high correlation between x and y. The third one is having linear relationship with x and y but have outlier. The correlation can go up to 80%. The last graph shows high correlation coefficient even though it is non-linear.

4. What is Pearson's R?

Answer: It is a measure of a statistical relationship between two continuous variables. -1 and -1 shows strong correlation. -1 tells strong negative relationship while +1 tells strong positive relationship.

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

Answer: Scaling: It is a pre-processing step in data processing helps us to normalize the data.

It is important to have all the features in same scale for better interpretation.

Normalized scaling: In this we rescale the values from 0 to 1. This is very useful when all the data are positive. But if we have outliers they are lost.

The basic formula is X – Xmax/Xmax – Xmin

Standardized scaling: In this we rescale the data to have mean value 0 and standard deviation is 1. In this we can see the outlier.

The basic formula is X – mean /standard deviation

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

Answer: VIF is $1/1-R^2$, so when R-squared is close to 1 the denominator value tends to zero which means the equation become 1/0 which is infinity. Now when does R-squared tends to 1, it is when the data is overfitting, or having large variance.