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

   Answer:

   Residual is defined as the difference between the actual value and the value predicted *via the model. For linear regression model to be a best fir model for a given data set, the following needs to hold true –*

   1. Residuals will to be ***normally distributed***

   2. The mean of this distribution should be 0

   3. Constant variance of residuals – Homoscedasticity

   4. Residuals are independent of each other

   5. No dependence between predictor variables and the residuals

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

   Answer:

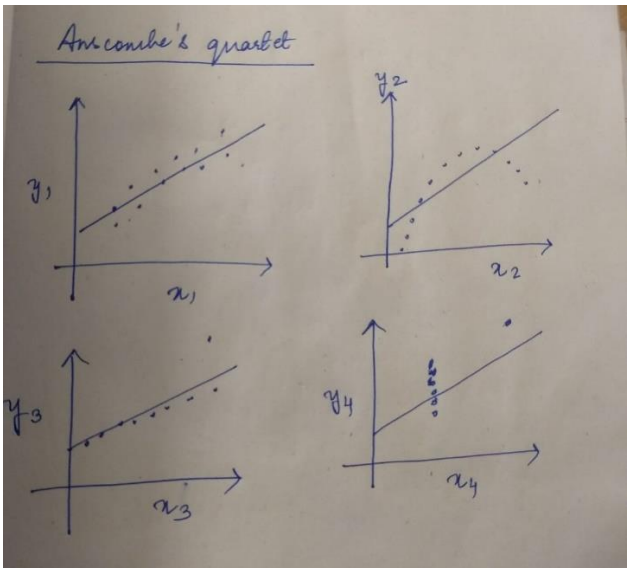
   Coefficient of correlation, R *is the* ***measure of the correlation between 2 variables****. It explains the* ***interdependence between 2 variables*** *and how the change in once affects the to other. The value for R can range between* ***-1 to 1****. 1 implies a perfect positive or proportional correlation, 0 implies no correlation and -1 implies a negative or inverse correlation. R is a good measure between any 2 variables. When we need to explain dependence between multiple variables and a target variable, R is inadequate and hence we have the* ***Coefficient of determination****, which is nothing but the square of R,* ***R^2****.* ***R^2*** ***explains the variability in data. It is a very useful metric in explaining the multiple linear regression and in evaluating machine learning models.*** ***R^2*** *basically shows how the target variable changes when the predictors change in a combined manner.* ***R^2*** ***is in the range 0 to 1.***

3. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a ***collection of 4 datasets*** *which* ***have same descriptive statistic values*** *but when plotted on a graph,* ***the plots are very different****. Each data set contains* ***11 datapoints****. This basically demonstrates that studying the mean, variance and other such statistics does not provide the complete understanding of the data. It highlights the importance of graphing the data in order to study the outlier and other influential factors. That said, the graphs cannot be looked at in isolation as well. The basic idea here is to ensure that the underlying data is investigated and not just the summary statistics.*

The quartet looks like below –



4. What is Pearson's R?

Answer

Pearson's R is the same as the coefficient of correlation explained above. *This is a specific method of quantifying the linear correlation between 2 variables. The value can range from -1 to 1.*

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

Scaling is a process of transforming the data to a given scale/range. **Scaling is required when analyzing variables which have different units**. For example, while analyzing the car price assignment, the variable price is in ten thousand whereas the car width and height are in tens. It is difficult to compare the coefficients.

Advantages of scaling

1. **Ease of interpretation**
2. **Faster Convergence** when using the gradient descent method
3. Since nothing else changes in the model, it is convenient to use scaled values

Normalized scaling rescales the values in the range of 0 and 1, whereas, in standardized scaling the data is rescaled to have a mean of 0 and standard deviation of 1

6. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer
   VIF defines the ***severity of multi-collinearity of a variable****. If the VIF for a variable is INF, that means that the* ***variable is severely collinear and hence that can be removed from the model****.* ***Mathematically, this will happen when R^2 is 1****.*