Linear Regression - Subjective Q & A

Q 1: **What are the assumptions of linear regression regarding residuals?**

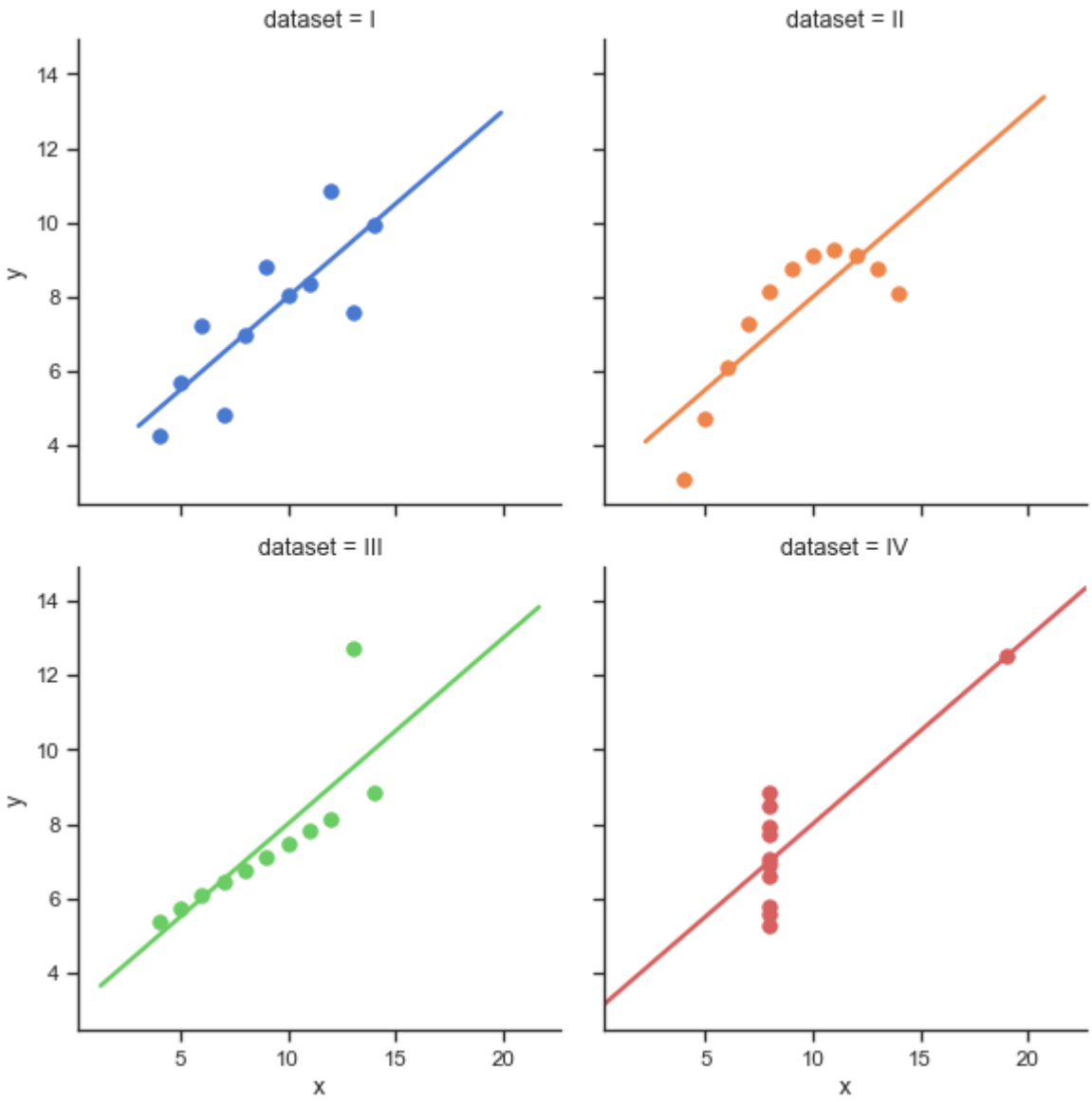
Answer: One of the major & crucial assumptions of Linear regression is that the error terms i.e. residuals should be normally distributed irrespective of X & Y variables.
To check the normal distribution of residuals we can plot the histogram/distplot of the residual.

Q 2: **What is the coefficient of correlation and the coefficient of determination?**

Ans : Coefficient of correlation is the measure of how strong two variables are related to each other. It is a statistical parameter which indicates an association between independent variable and dependent variable. It lies in the range of -1.0 to +1.0 and is denoted by r. -1.0 denotes strong inversely correlated variable and +1.0 denotes strong positive correlation between variables.

The coefficient of determination also know as R-Squared value is a measure used to assess and predict future outcome of a model.
It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

Q 3: **Explain the Anscombe's quartet in detail.**



Enscombe Quartet represents four datasets having nearly identical statistical measures (mean, standard deviation & correlation) however when plotted they appear very different.
It was represented by statistician Francis Anscombe in 1973 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

	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

Q 4: **What is Pearson's R?**

Ans: Pearson R or Pearson product-momentum correlation coefficient measures strength of linear relationship between two variables. It is denoted by r. Pearson product momentum correlation tries to draw a best fit line when the data of both the variables are plotted as scatterplot and the coefficient shows how far these points are from the best fit line. It varies from -1.0 to +1.0 and zero value denotes no association between variable, +1 denote strong positive correlation and -1 denotes strong negative correlation.

Q 4: **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

In most cases, dataset contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations. To supress this effect, we need to bring all features to the same level of magnitudes we use scaling of variables.

This can be done in two ways:

1. standardisation

Standardisation replaces the values by their Z scores.

$$x' = \frac{x - \bar{x}}{\sigma}$$

$$x' = \frac{x - \bar{x}}{\sigma}$$

This redistributes the features with their mean **μ = 0** and standard deviation **σ =1**

2. Normalisation:

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

This distribution will have values between **-1 and 1**with **μ=0**.

Q5: **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

An infinite VIF value mean a R square value of 1 between the variables. This is a case of extreme correlated between independent variables.