ASSIGNEMENT: LINEAR REGRESSION

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

Solution:

Following are the assumptions to be made regarding the residuals while building a linear regression model:

a.) The mean of residuals is zero:

We should always assume that the mean of the residuals will be equal to zero for a perfect model.
However we won't be able to get exactly equal to zero, however we always should try to get the least value of it.

b.) Homoscedasticity of residuals or equal variance:

It means that there should be no clear pattern in the distribution of the values.
If there is a cone shape based pattern then it leads to heteroscedastic.

c.) No autocorrelation of residuals:

The residuals should always be unique, which means that if the new residual value is dependent on the previous value then it leads to autocorrelation, which is no acceptable.

d.) The X variables and residuals are uncorrelated:

There should not be any relativity between the residual values and the X variables, else it would lead to complexity in the model which cannot be accepted.

e.) Normality of residuals:

The residual values should be distributed normally, which would also reflect that the X and Y variable values will also be distributed normally

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

Solution:

A. Correlation coefficient (r):

Correlation coefficient is the value that measures the strength and the direction of a linear relationship between two variables.

It is also referred to as the Pearson correlation coefficient.

- The formula is given by:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

where n is the number of pairs of data.

- The value of 'r' lies between -1 < r < +1. Positive correlation: If x and y have a strong positive linear correlation, r is nearer to +1.

- Positive value indicates that as the X variable value increases even the Y variable values increases.

- Negative correlation: If x and y have a strong negative linear correlation, r nearer to -1. It indicates that as the value of X variable increases, the value of Y variable decreases.

- No correlation: If there is no linear, r value is equivalent to zero. These values can vary based upon the "type" of data being examined.

B. Coefficient of determination (R2):

- It is a measure that explains us the how an individual variable is influenced by the other variables.

- The coefficient of determination is the ratio of the explained variation to the total variation.

- It is useful because it gives the proportion of the influence (variance) of one variable that is predictable from the other variable.

- The value of R2 lies between 0 and 1, and denotes the strength of the linearity between x and y.

- It represents the percent of the data that is the nearest to the line of best fit. For example, if r = 0.889, then r2 = 0.81, which means that 81% of the total variation in y can be explained by the linear relationship between x and y

- It shows how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, It means it is able to reflect all of the variation in the data.

3. Explain the Anscombe's quartet in detail?

Solution:

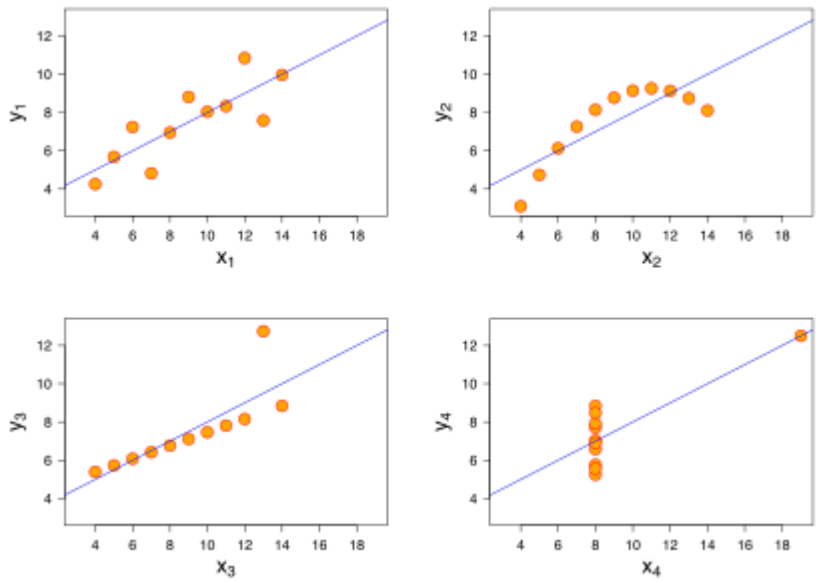
- **Anscombe's quartet** is a concept which explains the importance of Graphs along with numerical calculations. It comprises of four data sets that have two variables X and Y which are taken in such a way that they are identical/ similar ones, like having same mean value of both the variables X and Y

- Below are the samples used to demonstrate it :

Anscombe's quartet							
I		**II**		**III**		**IV**	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91

5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

- The above samples have same mean value of X and Y with similar statistical properties.



(Source: Google.com)

- The first plot appears to be having a linear relation with both the variables correlating.
- Second one appears to be a nonlinear relation model.
- The third one is almost similar but with very evident outliers present in the data set.
- In the fourth one there is no relationship between two variables and high correlation coefficient in the data point.

4. What is Pearson's R?

Solution:

- Pearson's R is a correlation coefficient which is a measure of linear relation between two variables.
- It is denotes as "ρ" when it is measured in the population and "r" when it is measured in a sample.
- The value of relation lies between -1 and +1.
- If we obtain a positive correlation coefficient value, then it says that the as the value of one variable increases the other one increases.
- If we obtain a negative correlation coefficient value, then we ca conclude that the values of one variable decreases as the other variable increases.
- If we do not obtain a zero equivalent value then we can say that there is no linear relation between both the variables

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

Solution:

- **Scaling** is a method of normalizing the range of an individual variable in a data set.
- Since the range of data varies from one unit of range to other with the individual variables
- When we do analysis due to a high variant values will impact the outcome, which would mislead the interpret of the analysis
- Hence Scaling of a data is performed to understand the analysis in a better way.

- Normalized scaling V/s Standardized scaling:
- **Normalized scaling** is a method of scaling in which the range of rescaling is in the range of
- [0, 1] or [−1, 1].
- This usually means dividing each component of feature vector by Euclidean length
- We have min- max normalization and mean normalization approaches:

- **Min- max normalization:**
  $x' = (x - x_{min}) / (x_{max} - x_{min})$

  Where; x' – normalized value
  x – Original value

- **Mean normalization:**
  $x' = (x - x\ mean) / (x_{max} - x_{min})$

  Where; x' – normalized value
  x – Original value

- **Standardized scaling** considers the Gaussian distribution for the input values and standardizes for obtaining mean of 0 and standard deviation value equal to 1.
- This method is mostly widely used in logistic regression, linear regression, vector machines etc.
- The formula used to perform standardized scaling is given by;
  $X' = (X - \mu) / \sigma$
  Where; X' – standardized value
  X – Original value
  μ – Mean value
  σ – Standard deviation

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

Solution:

- **VIF**: Variance inflation factor is the value which explains or provides the ratio of variance in a model with multiple variables with reference to an individual variable alone.
- It helps us to understand the multicollinearity
- It is given by the formula;
- $$VIF_j = \frac{1}{1 - R_j^2}$$

  Where;
  (Rj)^2 – is multiple R2 for the regression
- Higher the VIF value, higher is the collinearity
- So we have to try to keep the VIF value as low as possible.
- Usually value of VIF between 5 -10 is considered to be good. However it varies from application to application.
- Ideally we shouldn't get the VIF value as infinite
- If we get an infinite value it means that the two variables are exactly collinear with each other
- If we same variables are used then we would get an infinite value as it would be the exact data points involved in it.
- And if the (Rj) ^2 value of any predictor variable with the other variable approaches unity, that will correspond to infinite VIF value.

Submitted by

NAME: Shivaprasad A

ROLL NO: DDS1910473