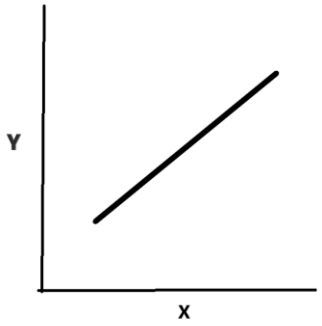
Linear Regression Assignment – Subjective Questions

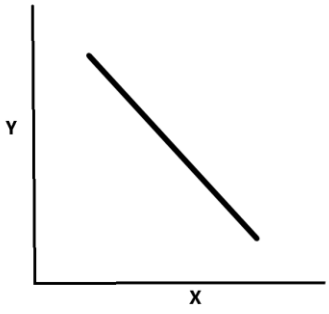
Q1. What are the assumptions of linear regression regarding residuals?

Linear regression has overall 3 kind of assumptions. These are mentioned below:

a. Assumption on Linearity: It is assumed that there is a linear relationship between X and Y. Here Y is the dependent variable that is dependent on X and X is the independent variable or predictor that has an impact on the value of Y. It can be explained with positive or negative relationship. With the increase of X, if Y also increases we can say that both have positive linear relationship. Similarly, if with increase of X, is Y decreases or vice versa, we can call it as negative linear relationship.

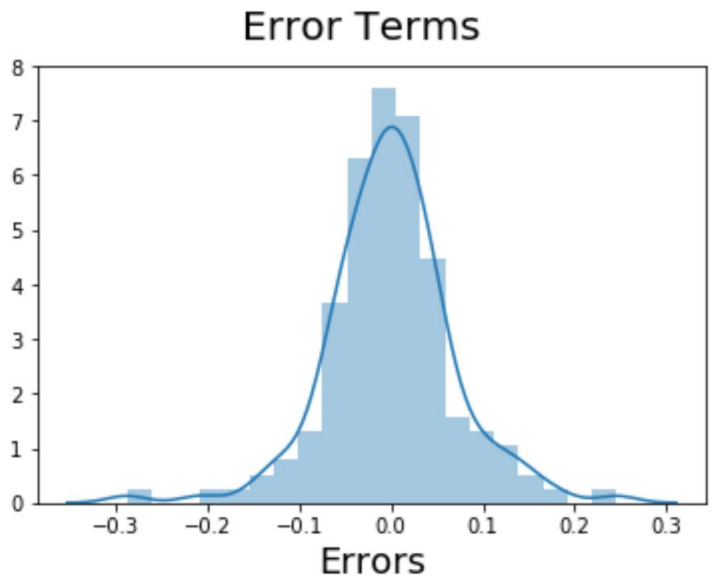


Positive Linear Relationship



Negative Linear Relationship

b. Assumption on Error terms:
   i. The error terms are normally distributed with a mean value of zero.
   ii. The residual terms also has constant similar variance, which is termed as Homoscedasticity.
   iii. The residual terms are independent on each other.



c. Assumption on the predictors:
   i. There is no multi-collinearity between the estimators i.e. they are linearly independent of each other.
   ii. The estimators are measured without any error.

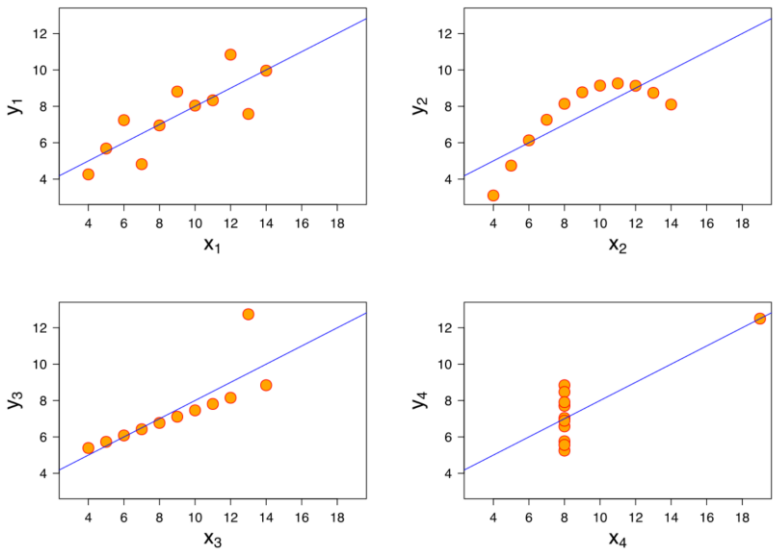
Q2. What is the coefficient of correlation and the coefficient of determination?

The coefficient of correlation is defined as the degree of relationship between two variable here one can be the dependent variable i.e. Y and another can be the estimator i.e. X. It is normally symbolized as R. It can in range of -1 to 1. The negative correlation means increase of one variable cause the decrease of others and vice versa. For example, R=-1 means one variable is completely opposite to other whereas R=1 means both the variable are almost similar and R=0 indicates there is no relation at all.

The coefficient of determination is nothing but the square of coefficient of correlation, symbolized as R^2. It shows the percentage variation in Y which is explained by all of the variables in X. Its value lies in between 0 to 1. It can never be negative as it is a square value. It is very useful in multiple linear regression as it is used to assess the model and tells the extent of the fit.

Q3. Explain the Anscombe's quartet in detail.

Anscombe's quartet is an important example, which shows the importance of the data visualization before running any regression. It is explained with four dataset on an X/Y plane. The mean, variance and correlation are same of X and Y for all the four data set, but the visualization reveals a different story. Now, let us have a look at the diagram below:



Here the first visualization shows a clean and well-fitting line. The second one does not seem to be distributed normally. Whereas the third and fourth one are highly influenced with the outliers. So if those can be removed, it could be a nice fit as well. It shows how important is it to take a look into the data before performing any regression.

Q4. What is Pearson's R?

Pearson's R is the most reliable coefficient in linear regression. This correlation coefficient is the measure of strength of association between two variables in linear regression. The value of this coefficient lies in between -1 to 1. It can be visualized below.

R=-1 R=0 R=1

Here the first plot shows a negative correlation with R=-1, the second one does not show and correlation with a coefficient of 0 and third one shows a positive relation with a coefficient of 1.

In non-linear relationship also, the Pearson's R can be calculated and it can be high value, but it is not reliable. With increase of power of estimator, the value of the coefficient increase the variation as well.

Q5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is an important procedure in linear regression, which helps to bring all the independent variables in a stipulated range. When there are many independent variables in a model and they belong to different magnitudes, units and ranges. It actually confuses the model with wired coefficients when it only takes the magnitude neglecting the units. For example, from magnitude point of view 500 is always greater than 5, but if we consider 500 gm and 5 Kg, then the comparison and difference is changed. One important aspect of scaling is it just influences the coefficients, but it does not alter other parameters like p-value, R-squared, t-statistic, F-statistic etc.

Normalized scaling is done is such a way that the value will be in range of 0 to 1 using the maximum and minimum of the date whereas the standardized scaling is done in a way to keep the mean at 0 and standard deviation as 1.

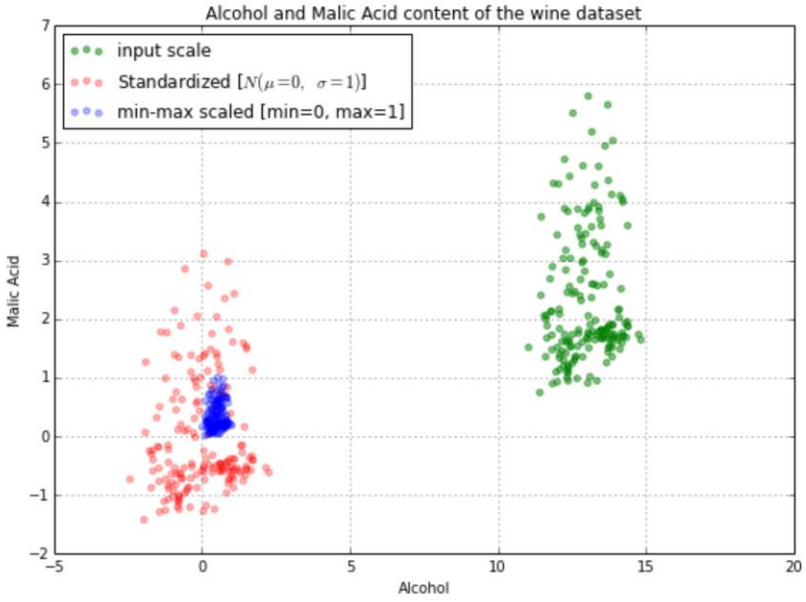
$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Normalised Scaling

$$x = \frac{x - mean(x)}{sd(x)}$$

Standardised Scaling

A comparison example of these two is shown in below self-explanatory diagram,



Q6. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The variance inflation factor (VIF) of a variable is defined as $1/1\text{-}R^2$. The value of VIF is infinite indicates it has strong correlation with other variable. Therefore, when the VIF is infinite, it indicates $1\text{-}R^2$ is 0.

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor with other reaches to unity, the $1\text{-}R^2$ reaches to 0 and the VIF results to infinite.