1 Explain linear regression in detail

A regression analysis is a very popular technique used in analytical problem solving and many business problems typically will lend themselves to a regression-based approach, especially when we are trying to predict or understand future outcomes using data that already exists on consumer behaviour.
Linear regression is an equation for finding relation between one dependent variable and one or more independent variable. When there is only one variable it is known as linear regression however when there are more than one variable it is multiple regression. A regression model is used to understand and quantify cause-effect relationships.

For example rltn between height and weight. The main idea behind this is to find the best fitted line. The method used is ordinary least square equation.

2What are the assumptions of linear regression regarding residuals?

Residuals are the difference between the observed value of the dv and the predicted value.

Both the sum and mean of the residuals are equal to zero.

The error terms should be normally distributed with the mean equal to zero.

Residual plot should show heteroscedasticity

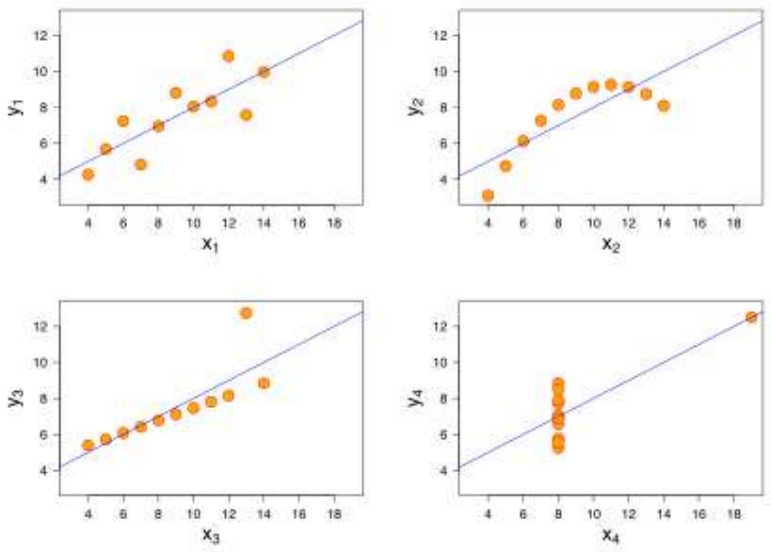
3What is the coefficient of correlation and the coefficient of determination?

Coefficient of correlation is “R” value which is obtained in the linear regression summary table and coefficient of determination is the square of the “R” value which is square of the coefficient of correlation value. R square tells us the variance in the dependent variable explained by all the independent variable. Higher the value better the model is performing. The correlation value lies between -1 to +1.

4Explain the Anscombe’s quartet in detail.

Anscombe’s quartet consists of four data sets and each data set have eleven data points. These data sets almost have same descriptive statistics but when these data is plotted there graphs shows a very different distributions. This theory was demonstrated by a famous statistician Francis Anscombe to demonstrate how important graphing and visualizing data is to analyse the data.

Below is the graph:



The first shows a simple linear relationship of two variables.

The second is a non linear relationship between two vars

The third is also shows a linear distribution but more of a robust regression.

The third shows when a high leverage point is enough for a high correlation coefficient. Even though the data is not have any relationship among the variables.

5What is Pearson's R?

Pearson's correlation coefficient r is the measure of the strength of the association among two variables. The value ranges between +1 to -1. Where +1 tells a strong positive correlation and vice versa. And zero suggests no relation.

Formula attached.

6What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

In a data usually the data is in different scales to bring all the data into one scale and build the model we need to bring all the data to one scale hence scaling is the first step before model building. This process brings all the features to same level of magnitudes. There are four types of scaling:

1. Standardisation
2. Mean Normalisation
3. Min Max Scaling
4. Unit Vector

Formula attached:

Normalization scales the value between into the range (0,1) however standardisation means the value is rescaled to get a mean of 0 and standard deviation of 1.

7You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Vif is infinity if the correlation between variables is high. Vif =1/(1-R2)=1/0=infinity. The variables with inf Vif should be dropped. Higher the value the greater the correlation with other variables. The variables with value anything less than 5 is consider good predictor.

8What is the Gauss-Markov theorem?

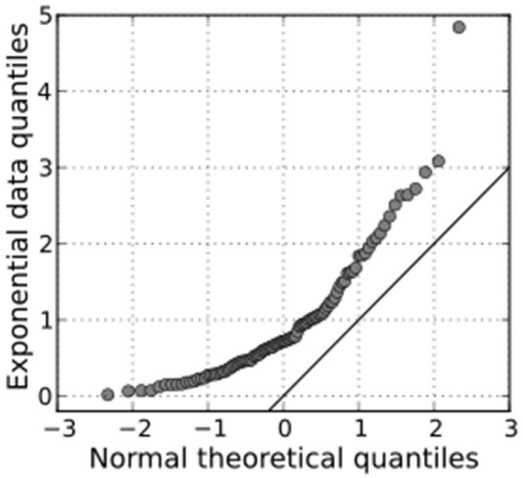
The theorem explains that if the model follows linear regression then the ordinary least square is best linear Unbiased estimator or BLUE. Which means minimum variance. The theorem have six assumptions and they are: Linearity, Random, Non Collinearity, Exogeneity, Homoscedasticity. If these assumptions are met the model follows BLUE in ideal condition. Usually all these assumptions are not 100 percent met.

9Explain the gradient descent algorithm in detail.

Gradient descent algorithm is an iterative optimization algorithm which is used to move from higher region (high cost) to lower region (low cost). This algorithm is used to update the parameters of the model. Parameters are the coefficients in Linear regression. Gradient descent is used to minimize the cost function.

10What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot or quantile – quantile plot are plot of two quantiles against each other. The main reason of this plot is to find out if the two sets of data are coming from same distribution or not. A 45 degree angle is plotted if the two data sets falls in this reference line then are considered of same distribution. If the plots are not aligned with the reference line are considered not to be normally distributed. It is used to validate the distribution of a dataset.



$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

cov = covariance
σ_x = Standard deviation of X
σ_y = S.D. of Y.

In terms of mean and Expectation

$$cov(x,y) = E[(x-\mu_x)(y-\mu_y)]$$

$$\rho_{x,y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

μ_x = mean of X.
μ_y = mean of Y.
E = Expectation.

1. Standardisation

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

mean $\mu = 0$ & S.D. $\sigma = 1$.

2. Mean normalisation

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

distribution will have value b/w -1 & 1 with $\mu = 0$.

3. Min - Max Scaling:

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

This brings the value b/w 0 & 1.

4. Unit vector:

$$x' = \frac{x}{\|x\|}$$