1. Explain the linear regression algorithm in detail.

Solution: Regression helps us examine two things. Firstly, to determine whether a set of predictor/independent variables perform well in predicting an outcome variable/dependent variable or not. Secondly, the linear regression algorithm helps us to determine the variables that are significant predictors of the outcome variable and in what manner do they impact the outcome variable. This is indicated by the magnitude and sign of the beta/regression coefficient estimations. These regression coefficients are estimates of the unknown population parameters (independent/predictor variables) and describe the relationship between a predictor variable and the response.

In simple terms, linear regression is linear approach to modelling the relationship between a dependent/response variable and one or more independent/predictor variables. It is mostly done using Sum of Squares Method. The objective is to obtain a line that best fits the data and minimizes a quantity called Residual Sum of Squares (RSS). The best fit line is the one for which total prediction error are as small as possible and best expresses the linear relationship between dependent and independent variable/s. The error encountered is the absolute measure of the typical distance that the data points fall from the regression line.

Suppose Y is a dependent variable, and X are independent variables. The regression equation would be:

$$\hat{\mathbf{Y}}_t = \mathbf{b}_0 + \mathbf{b}_1\mathbf{X}_{1t} + \mathbf{b}_2\mathbf{X}_{2t} + \ldots + \mathbf{b}_k\mathbf{X}_{kt}$$

Where b_o is the y-intercept and $b_1, b_2 \ldots b_k$ are the regression coefficients of the features.

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

Solution: Assumptions of Linear Regression regarding residuals are as follows:

- *Normality assumption*: The first assumption is that the error terms, ε(i), are normally distributed.
- *Zero mean assumption*: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
- Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance. This assumption is also known as the assumption of homogeneity or homoscedasticity. The data is said to homoscedastic when the residuals are equal across the line of regression. In other words, the variance is constant.
- *Independent error assumption*: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

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

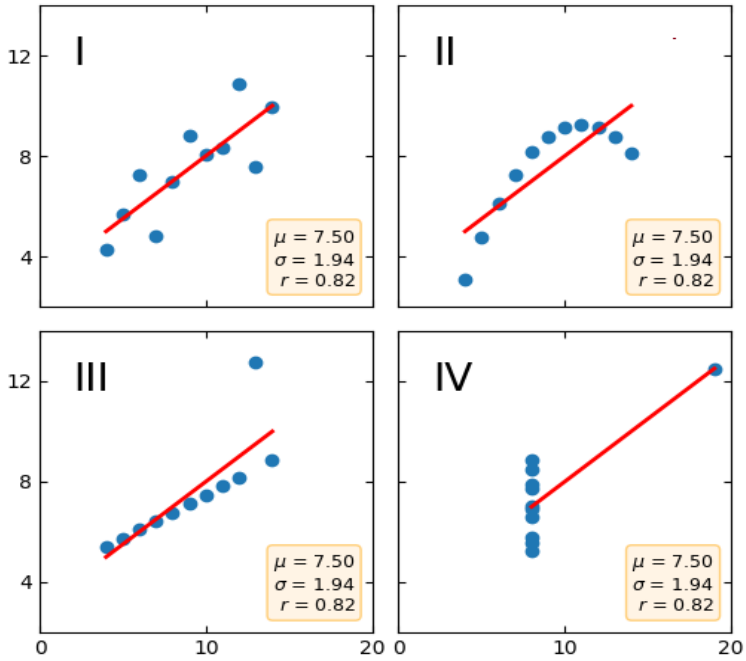
Solution: *(i) The coefficient of correlation* is used in statistics to measure of the strength of the relationship between the relative movements of two variables. The values always range between '-1' (strong negative relationship) and '+1' (strong positive relationship). The negative value of '-1' implies that the variables move in opposite directions i.e. for a positive increase in one variable, there is a decrease in the second variable. On a similar note, positive value of '1' implies that the variables move in same directions i.e. for a positive increase in one variable, there is an increase in the second variable. Values at or close to zero imply weak or no linear relationship.

(ii) The coefficient of determination, "R squared", is the ratio of the explained variation to the total variation. It represents the percent of the data that is the closest to the line of best fit; it explains all the variations. Hence, its value ranges from 0 to 1. In general, a high R2 value indicates that the model is a good fit for the data. An R2 of '0' means that the dependent variable cannot be predicted from the independent variable, and '1' means it can be predicted without error from the independent variable. Any value between 0 and 1 corresponds to the extent to which the dependent variable is predictable.

4. Explain the Anscombe's quartet in detail.

Solution: Anscombe's Quartet was developed by statistician, Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. Statistically, all the four linear regression are exactly the same, but the graphical representation shows the hidden peculiarities of the dataset. Let the decision not be made only on the basis of regression line.

- (I) image shows that the linear regression is quite good.
- (II) image clearly shows that linear regression can only model linear relationships. It is incapable of deal any other kind of data.
- (III) and (IV) images show the linear regression model's sensitivity to outliers. In the absence of the outliers, we could easily get a best-fit line through the datapoints.

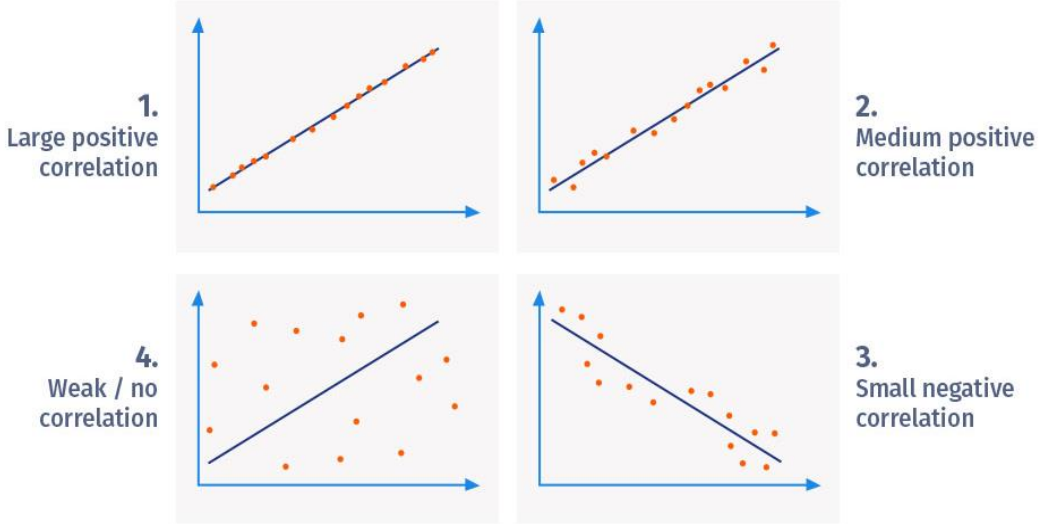


Anscombe's Quartet

5. What is Pearson's R?

Solution: *Pearson's correlation coefficient* is a statistical measure of the strength of a linear relationship between paired data. In a sample, it is denoted by 'r' and it lies between -1 and 1.The closer the value is to 1 or –1, the stronger the linear correlation.

- Positive values denote positive linear correlation, (increase in one variable, increases another)
- Negative values denote negative linear correlation; (increase in one variable, decreases another)
- A value of 0 denotes no linear correlation.



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

Solution: Scaling is a method used to formalize the range of independent variables or features of data. Although scaling is not mandatory, it helps handle disparities in units and helps reduce computational expenses during long processes. This method helps improve the performance of the model and reduces the values from varying widely.

- Normalized scaling rescales the values into a range of [0,1]. This is a good technique to use when the distribution is not Gaussian or standard deviation is very small.
- Standardized scaling is a technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation This scaling assumes that the data has a Gaussian distribution, however, this does not have to be true but the technique is more effective if the attribute distribution is Gaussian.

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

Solution: Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset.

$$VIF_i = \frac{1}{1-R_i^2}$$

Where VIF: Variable Inflation Factor of i^{th} variable, R_i^2: R2 value of model when that variable is regressed against all other independent/predictor variables.

An infinite VIF is an indication that there is perfect correlation between the independent variables; meaning the corresponding variable may be expressed exactly by a linear combination of other variables. A high value of VIF implies that R2 score is high. Note that, in such cases, the other independent variable/s should also have infinite VIFs.

8. What is the Gauss-Markov theorem?

Solution: Gauss-Markov theorem states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, provided

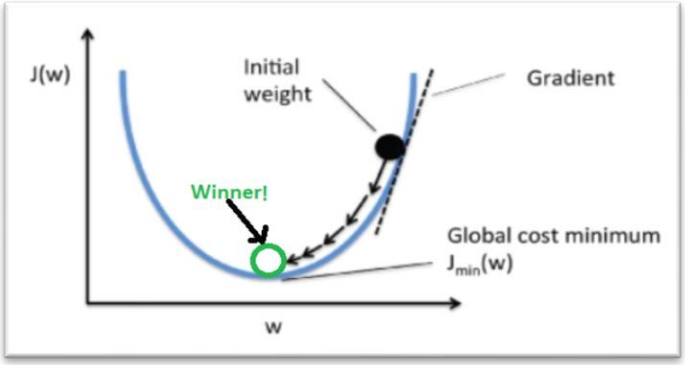
- The errors in the linear regression model aren't correlated i.e. uncorrelated.
- They have equal variances and expectation value of 0.

In short, Gauss-Markov Theorem states that OLS is BLUE (Best Linear Unbiased Estimator).

The errors specifically don't need to be normal. Also, they don't need to be independent and identically distributed. Meaning, they should only be uncorrelated with mean zero and must be homoscedastic with finite variance. The requirement that the estimator has to be unbiased cannot be dropped, since biased estimators exist with lower variance.

9. Explain the gradient descent algorithm in detail.

Solution: Gradient descent is an optimization algorithm. In linear regression, it is used to optimize the cost function and find the values of the βs (estimators) corresponding to the optimized value of the cost function. The working of this algorithm is like a ball rolling down a graph (ignoring the inertia). Please refer the figure below.



From the above figure, we see Gradient Descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima). In mathematical terms, the aim of gradient descent for linear regression is to find the solution of ArgMin J(Θ0,Θ1), where J(Θ0,Θ1) is the cost function of the linear regression.

In the equations (on-right), h is the linear hypothesis model, h=Θ0 + Θ1x, y is the true output, and m is the number of data points in the training set. This algorithm starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value. At every step, repetition is performed until convergence.

Cost Function

$$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y_i]^2$$

Predicted Value; True Value

Gradient Descent

$$\Theta_j = \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta_0, \Theta_1)$$

Learning Rate

Now,

$$\frac{\partial}{\partial\Theta}J_\Theta = \frac{\partial}{\partial\Theta}\frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y]^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x_i) - y)\frac{\partial}{\partial\Theta_j}(\Theta x_i - y)$$

$$= \frac{1}{m}(h_\Theta(x_i) - y)x_i$$

Therefore,

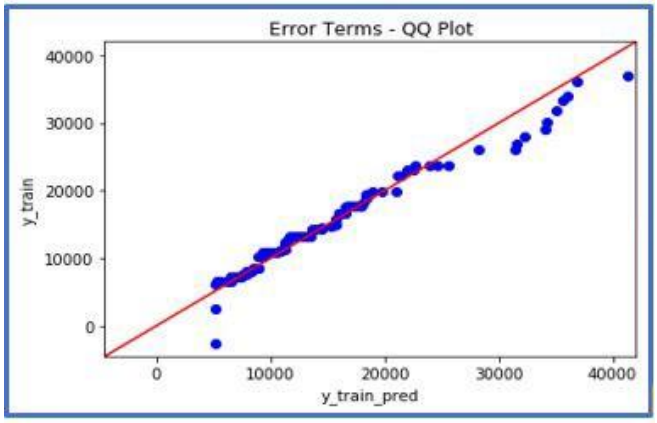
$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

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

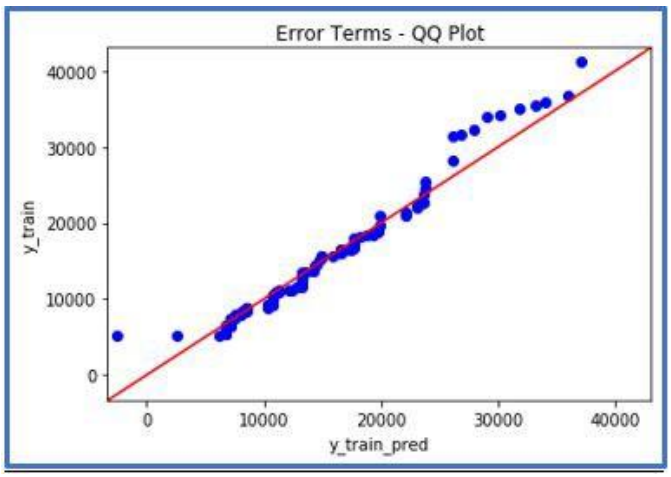
Solution: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly that come from some theoretical distribution such as a Normal, Exponential or Uniform distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It helps us determine whether two data sets come from populations with a common distribution or not. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Its advantages include:
(i) Q-Q plot can be used with sample sizes, can detect many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers, etc.
(ii) It is used to check the scenarios. If two data sets - (a) come from populations with a common distribution (b)have common location and scale (c)have similar distributional shapes and (d)have similar tail behavior.

Interpretations: Possible interpretations for two data sets:
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis