Subjective Questions_ Sukanya Roy Sikdar

1. Explain the linear regression algorithm in detail.

Answer : Linear regression is typically a machine learning algorithm based on supervised learning. It is used to find the relationship between variables of a data set and forecast. To begin with linear regression we have to go through the following steps :

I. Data collating, cleaning and manipulation.

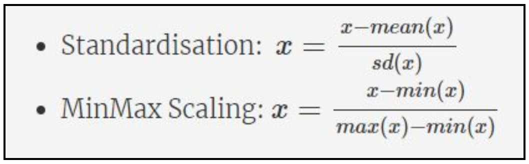
II. Identifying the dependent and independent variable.

III. Exploration of data using EDA techniques like univariate and bivariate analysis, data visualization, etc. Checking the correlation of the dependent variable against all the independent variable and see how one affects the other.

IV. Finalizing the dataset, either using RFE Technique, or manual or mixed techniques

V. Creating dummy variables across the categorical values. Dummy for nominal and Encoding for ordinal variables.

VI. Dividing the data into train and test data and feature scaling the data (numerical variables) using techniques like below :



- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

VII. Perform linear regression modeling on the train data and find the significance of the data using the p value (at a 0.05 level of significance). We typically use Hypothesis testing technique where H0: beta = 0, and Ha: beta !=0 (beta being coefficient of an independent variable). So, when p value is <0.05, then we reject the null hypothesis and claim that the independent does have some predictive relationship with the dependent variable. We need to perform a VIF analysis to track the correlation among the independent variables as well and remove those with high VIF value (typically >5)

VIII. We need to check the residual values and see if they have any correlation among themselves and also check if it follows a normal distribution.

IX. We then check the model in the test data as well and check the significance and Adj. R Sqr value before finalizing it.

X. We also need to evaluate the model which the tested and the predicted variable and once we finalize the model we can verify the value of the Adj. R Sqr, F-statistic and Prob(F-statistic) and see if the model fitted is good.

XI. Typically the model would be in a format like y = a0 + a1x1 + a2x2+...+anxn

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

Answer : The typical assumptions of a linear regression regarding residuals / errors are -

I. Errors / residuals are normally distributed

II. There should be homoscedasticity of error (equal variance around the line)

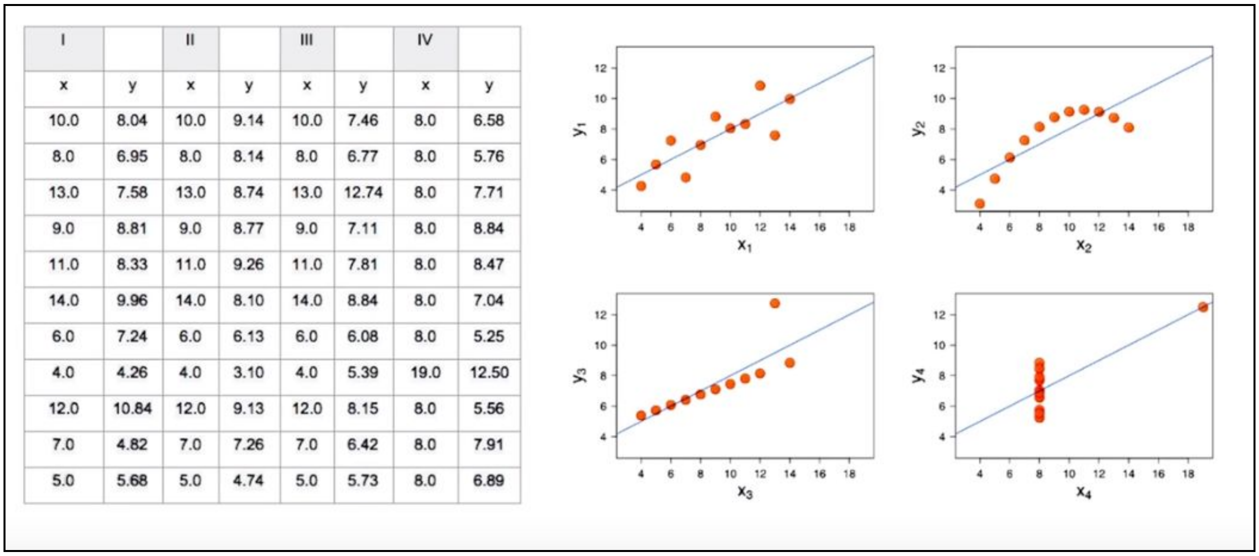
3. What is the coefficient of correlation and the coefficient of determination?

Answer : Coefficient of correlation (also known as Correlation Coefficient) is a statistical measure which indicates how strong a relationship is between two variables. The relationship can be positive, negative or no correlation and it lies between -1 to +1. Positive correlation is represented by +1 while -1 represents negative correlation.

Coefficient of determination is a statistical measure with interprets the proportion of the variance in the dependent variable that is predictable from the independent variable. It is the square of correlation (r) and lies between 0 and 1. When a dependent variable cannot be predicted from the independent variable R-sqr is 0, and when a dependent variable can be predicted without error from the independent variable then R-sqr is 1. The values in between 0 and 1 indicates the extent or % to which the dependent variable is predicted by the independent variable,i.e if R-sqr is 0.20 then 20 % of the variance in the dependent variable is predictable by the independent variable.

4. **Explain the Anscombe's quartet in detail.**

Answer : There may be scenarios when few data sets are nearly identical based on descriptive statistics but different when visualized. Anscombe's quartet comprises of four data sets that have nearly identical data, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. It was constructed by the statistician Francis Anscombe in 1973 to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. Figure below shows an example of Anscombe's quartet. The 4 data sets may have similar statistical values (mean, median, etc), however when we plot the graph we can observe how different the data sets are from each other.



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

5. **What is Pearson's R?**

Answer : Pearson's R, also known as Pearson's correlation Coefficient is a test statistic that measures the statistical relationship or association between two continuous variables. It is one of the most preferable methods of measuring the association between variables of interest since it is based on the method of covariance. It typically indicated the association, magnitude as well as the direction of the relationship.

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

Answer : Scaling is a step of data pre-processing which is applied to independent variables or quantitative features of data. Scaling in statistics usually means a linear transformation

Scaling is performed because it helps to normalize / standardize the data within a particular range and speeds up the calculation in an algorithm.

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.

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

Answer : VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. Hence, if value of VIF is infinite then it indicates, perfect correlation.

8. **What is the Gauss-Markov theorem?**

Answer : The Gauss–Markov theorem is an important theorem for linear regression in statistics. It states that in a linear regression model in which the errors are uncorrelated, and have equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists. Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance.

9. **Explain the gradient descent algorithm in detail.**

Answer : Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model.

I. The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

II. The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

III. The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

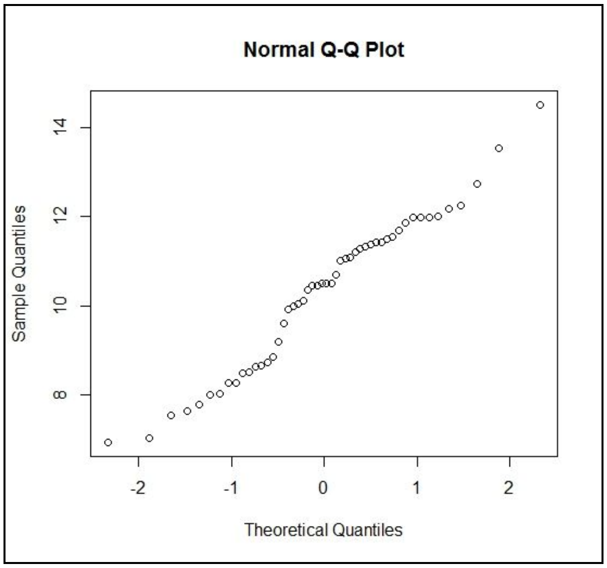
IV. Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

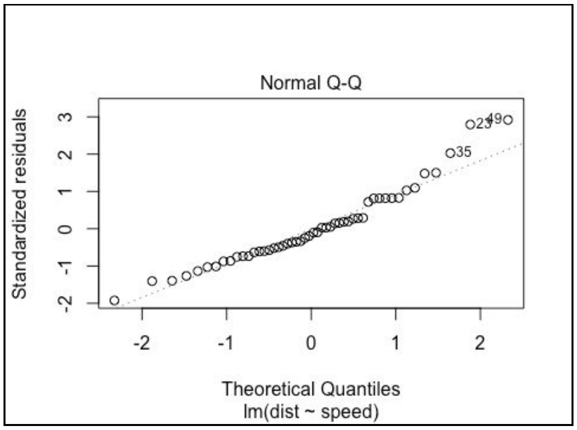
V. This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

Answer : A Q-Q plot, also known as quantile - quantile plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight.



Q-Q plots are important in linear regression because we can use them in a single, simple way when fitting a linear regression mode. We can check if the points lie approximately on the line, and if they don't, we can say that the residuals aren't Gaussian and thus the errors aren't either.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)