General Subjective Questions

Question – 1. Explain the linear regression algorithm in detail.

Linear regression is a data analysis technique that is used to predict the value of an unknown variable by analysing another related and known data variables. It mathematically models the unknown or dependent variable (aka explanatory or predictor variables) and the known or independent variable (aka response variables or predicted variables) as a linear equation.

Based on the number of dependent variables, there are 2 types of linear regressions:

Simple Linear Regression – Involves only one independent variable and one dependent variable. Simple linear regression is defined by the linear function **Y = β0 + β1X**.

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression – Involves more than one independent variable and one dependent variable. Multiple linear regression is defined by the linear function **Y = β0 + β1X1 + β2X2 + ….. + + βnXn.**

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

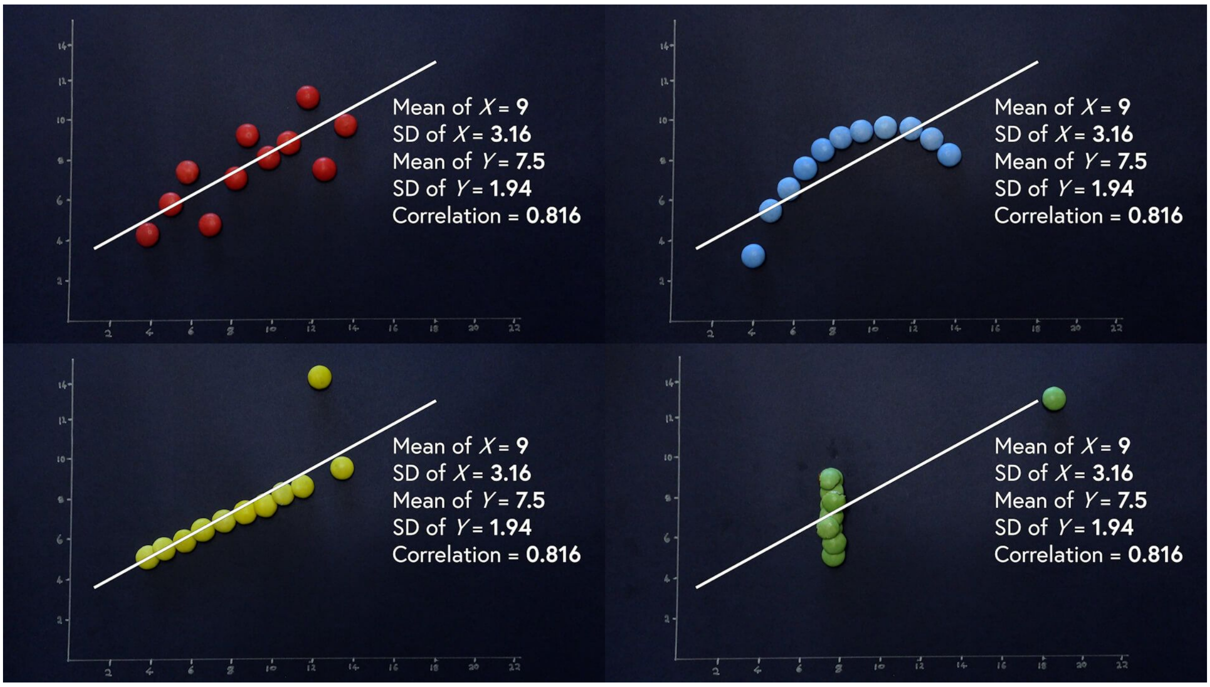
A simple linear regression technique attempts to plot a line graph between two data variables, x and y. As the independent variable, x is plotted along the horizontal axis. The dependent variable, y, is plotted on the vertical axis. The goal of the algorithm is to find the best Fit Line equation that can predict the values of dependent variable based on the independent variables. The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables.

Question – 2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises **four** datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of **eleven** (*x*, *y*) points.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. The set of four datasets have have identical descriptive statistical properties in terms of **means**, **variance, R-**

squared, **correlations**, and **linear regression lines** but having different representations when we scatter plots on a graph.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where *y* could be modelled as gaussian with mean linearly dependent on *x*.
- The second scatter plot (top right), shows an obvious but non-linear relationship between the variables x and y.
- The third graph (bottom left) shows that the modelled relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- The fourth graph (bottom right) shows an example when one data point with heavy weightage is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship. It also reveals the inadequacy of basic statistic properties for describing realistic datasets.

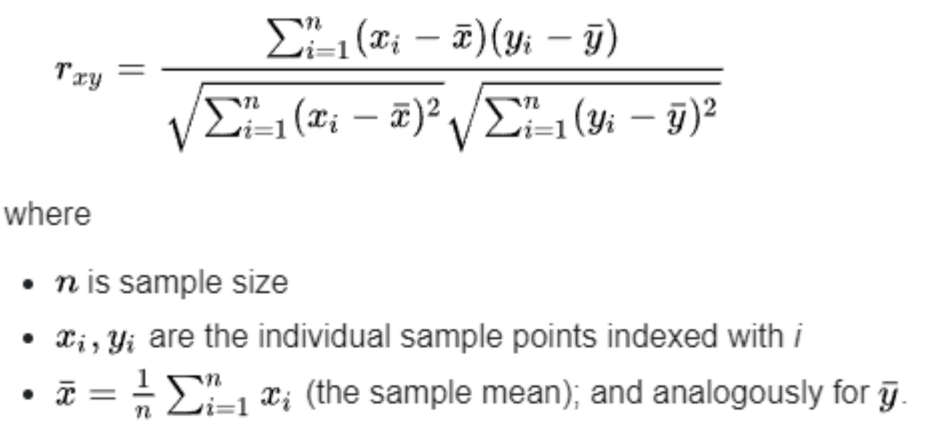
Question – 3 What is Pearson's R

The **Pearson correlation coefficient** (**PCC**)is a correlation coefficient that measures linear correlation between two sets of data.

The correlation coefficient ranges from −1 to 1. An absolute value of exactly 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope.

- a value of +1 implies that all data points lie on a line for which *Y* increases as *X* increases
- a value of -1 implies that all data points lie on a line for which *Y* increases while *X* decreases.
- a value of 0 implies that there is no linear dependency between the variables.

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r(XY) and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*.



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

$(X_i - X)$ $(Y_i - Y)$ is positive if and only if X_i and Y_i lie on the same side of their respective means. Thus, the correlation coefficient is positive if X_i and Y_i tend to be simultaneously greater than, or simultaneously less than, their respective means.

Question – 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. For example, if we have multiple independent variables like age (18-100 years), salary (25000 – 75000) and height (1-3 Meters); feature scaling would scale all of them to be in the same range i.e. (0-1). If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

There are 2 types of Scaling:

- Normalization – Also known as min-max scaling. The method performs rescaling the features in the range of [0, 1] or [−1, 1]. **Normalization** is good to use when the distribution of data does not follow a Gaussian distribution.

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

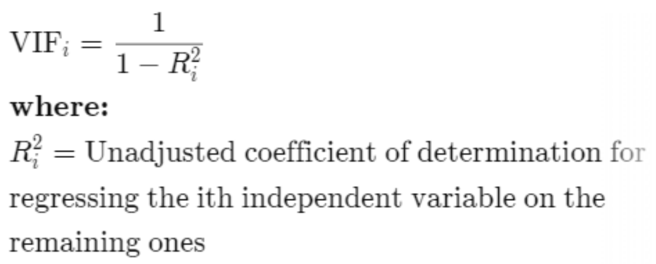
- Standardization - The method makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. In the equation below **x̄** is the average of the feature and **σ** is the standard deviation of the feature.
  Standardization can be helpful in cases where the data follows a Gaussian distribution. Since standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization.

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

Question – 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. The VIF is an estimate of how much the variance of a regression coefficient is inflated due to multicollinearity. A multiple regression model where there is a high multicollinearity makes it more difficult to estimate the relationship between each of the independent variables and the dependent variable.

In the equation below: VIF ∞ is arrived when R-squared is 1. The value of 1 for r-squared indicates that the model predicts 100% of the relationship and perfectly collinear relationship



$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Question – 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q (quantile-quantile) plots play a vital role in graphically analysing and comparing two probability distributions by plotting their quantiles against each other. If the two distributions that we are comparing are exactly equal, then the points on the Q-Q plot will perfectly lie on a straight line y = x.

- We plot the theoretical quantiles, basically known as the standard normal variate (a normal distribution with mean of zero and a standard deviation of one) on the x-axis and the ordered values for the random variable, which we want to determine whether or not is a Gaussian distribution, on the y-axis.
- If the points at the ends of the curve formed from the points are not falling on a straight line but are scattered significantly from these positions, then we cannot conclude a relationship between the x- and y-axes. This result clearly signifies that the ordered values that we wanted to calculate are not normally distributed.
- When the data points are few, the Q-Q plot does not perform very precisely, and it fails to give a conclusive answer. When we have an ample amount of data points and plot a Q-Q plot using a large data set, however, then it gives a result significant enough to draw clear conclusions about the type of distribution.

Assignment-based Subjective Questions

Question – 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Categorical Variables showing linear relationship

- Relation with season:
  - The mean count is **lowest** for season **SPRING**
  - The mean count is **highest** for season **FALL**
- Relation with year:
  - The mean count is significantly lower for Year: **2018**
- Relation with month of year:
  - The mean count increases linearly with month peaking at **September** and reducing thereafter. So, September month sees the maximum users.
- Relation with holiday:
  - The mean count is **less** on a holiday
- Relation with weather situation:
  - The mean count is **low** for days with **Light Rain**
  - The mean count is **high** for days with **Clear** weather

Categorical Variables **NOT** showing linear relationship

- Relation with day of week:
  - The mean count **NOT** seem to be affected much by the day of the week
- Relation with working day:
  - The mean count **NOT** seem to be affected whether the days is working or not
- Relation with day of month:
  - The mean count **NOT** seem to be affected by the day of the month

Question – 2. Why is it important to use drop_first=True during dummy variable creation?

- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Question – 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- Relation with temp/atemp:
  - The count increases with increase in temperature
  - The variables temp and atemp is highly correlated. Therefore, for simplicity we can ignore one of the two.

- Relation with humidity

    - The counts are very rare for **humidity < 40**

- Relation with windspeed
    - The counts are very rare for **windspeed > 25**

Question – 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

- Checking the **VIF factors** of various Independent/Predictor Variables to ascertain the non-existence of collinearity. All the Independent Variables in the final model have the **VIF < 5**.

- Checking the **Adjusted R-squared** value of the model to model is able to predict the variance in the data. Both the models (with constant and without constant) seem to be predicting **~80%** of variance in the data

- Checking the **p-value** of the Independent/Predictor Variables to ascertain that the null hypothesis is successfully rejected with the coefficient of the variable. All the Independent Variables in the final model have **p-value = 0**.

- Checking the **F-Statistic** of the final model to ascertain the high degree of fitment. Both the models have significantly high value of **F-Statistics ~ 1000.**

- Checking the **distribution of error terms** reveals normal distribution of errors centered around 0.

Question – 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The final equation for the model is:

total_count = 0.125926 + (0.232861 * Year) + (-0.098685 * Is_Holiday) + (0.548008 * Temperature) + (-0.153246 * Windspeed) + (0.088080 * Summer) + (0.129345 * Winter) + (-0.282869 * Lite Rain) + (-0.078375 * Mist) + (0.101195 * September)

Based on the equation, the features that contribute to the demand of the shared bikes are tabulated below. So while temper

Feature Name	Correlation	Coefficient
Temperature	**Positive**	**0.548**
Year	**Positive**	**0.232**
Winter	**Positive**	**0.129**
Lite Rain	**Negative**	**0.282**
Windspeed	**Negative**	**0.153**
Holiday	**Negative**	**0.100**