Linear Regression Subjective Questions

Q.1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: There are 6 categorical variables in the dataset, namely - season ('season'), month ('mnth'), weather situation ('weathersit'), holiday ('holiday'), weekday ('weekday') and working day ('workingday'). We used Box plot for visual analysis of their effect on the target variable – count ('cnt').

Let us look at the box plots of independent categorical variables and then state all the inferences.

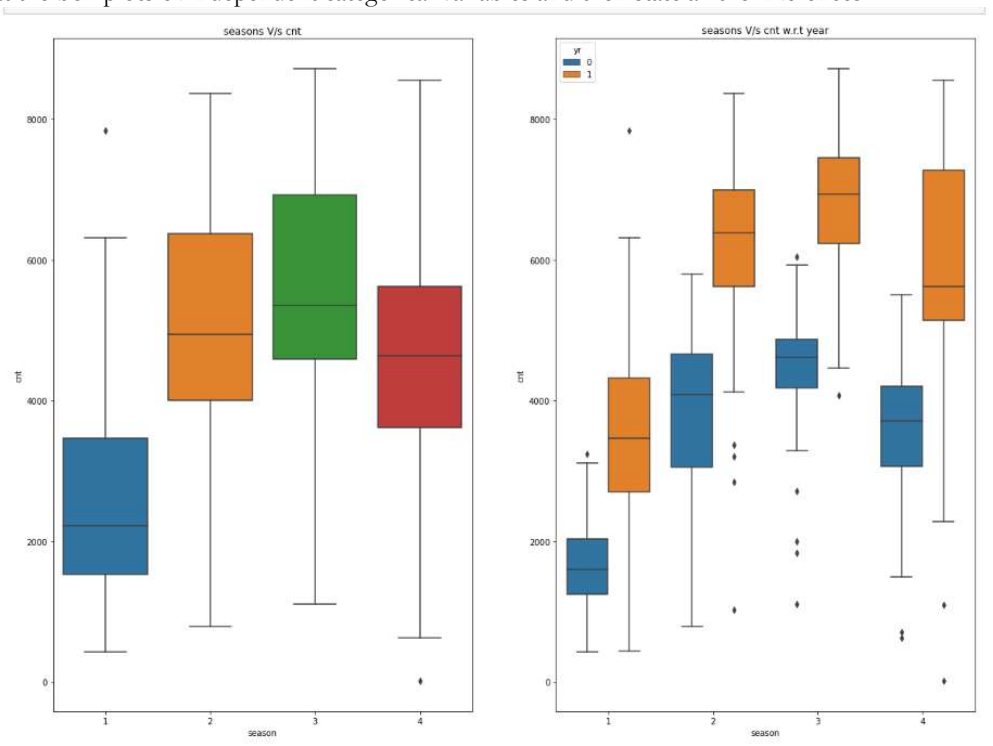


FIGURE 1

The inference derived:

1. season v/s cnt:

Refer to Figure 1, left plot

- Bike Demand is high in Summer and Fall and it decreases in Winter. Highest bike booking are happening in fall (season3) with a median of over 5000 booking. This is followed by summer (season2) & winter (season4).
- Demand for bike in spring is less, which is a little surprising, as it should be actually high during that time as weather is favourable. To understand the reason behind this we also plotted "Seasons V/s cnt w.r.t Year" and we did get some answers.

Refer to Figure 1, right plot

- In 2018-Spring, when the bike sharing venture was launched, it was still a new concept and may be that is why demand was less. But gradually it increased from 2018-summer onwards and now if you look at spring 2019, overall demand and median has gone higher compared to 2018- spring.
- This boxplot also shows how the cnt got increased from the 2018 to 2019. That indicates that the whole concept was received well and was gaining popularity with time. But unfortunately after that Covid period started, which obviously was not a usual circumstances.
- now, once covid period is over and situation is getting back to normal, thanks to already registered customers, the initial demand pattern may follow and demand will increase with time once again
  This indicates, season can be a good predictor for the dependent variable.

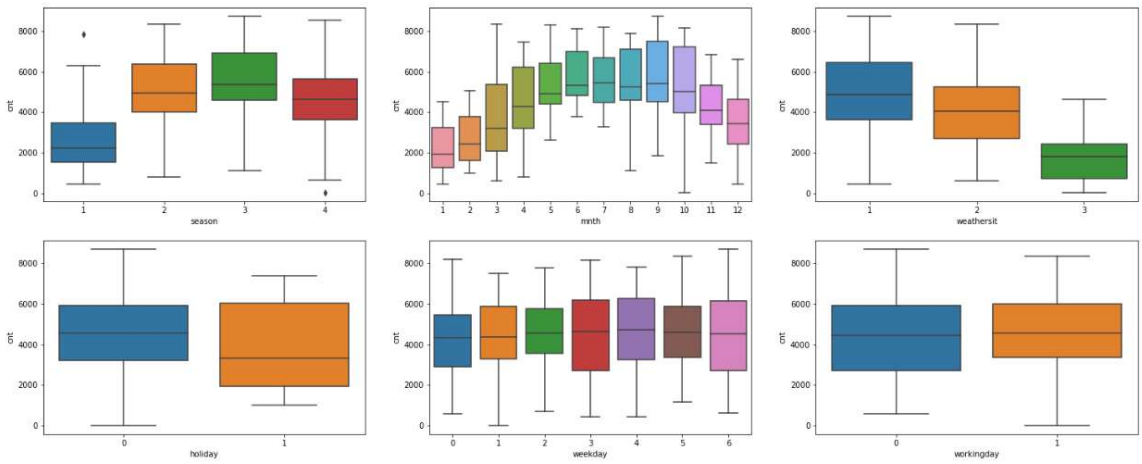


FIGURE 2

2. mnth v/s cnt:

High amount of the bike bookings are happening in the months 5, 6, 7, 8, 9 and 10 (May to Oct) with a median of over 4000 booking per month. This indicates, month has some trend for bookings and can be a good predictor for the dependent variable.

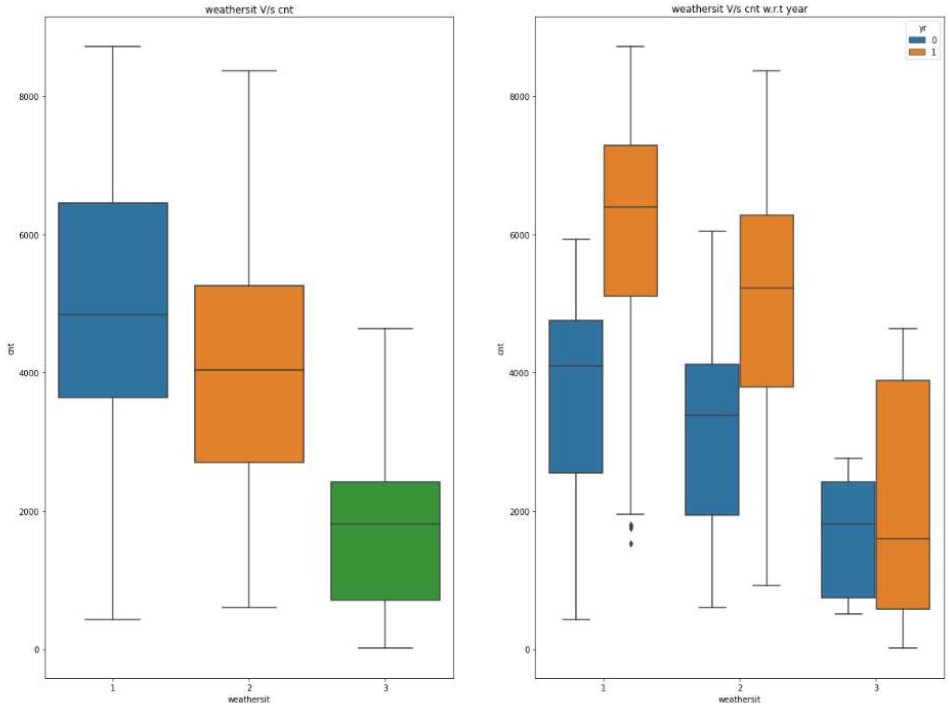


FIGURE 3

3. weathersit v/s cnt:

Refer to Figure 3

- To start with, we only see 3 categorical variables (clear, mist and light snow). Fourth weather situation, namely heavy rain-snow-fog is completely missing in the plot. Which means zero demand in case of heavy rains or snow or fog.
- Secondly, the median cnt is decreasing from clear to misty to light snow situation, i.e. clear > misty > light snow.
- Highest bike bookings are happening during 'weathersit 1' (clear) with a median of close to 5000 booking. This also makes sense logically as weathersit 1 is - Clear, Few clouds, Partly cloudy. This is followed by weathersit 2 and then weathersit 3.
- Once again we see that in 2018 the cnt (casual + registered) was less and it gradually increased in 2019

This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
4. holiday: Median as well as maximum value of count is higher when it is not a holiday. We need to look at what weekday plot is showing as oliday and workingday will be eventually giving the similar information.
5. weekday: On first look, Median is almost similar for all week days (little above 4000). Which means bike demand on all days are more or less similar. Hence, this variable can have some or no influence towards the predictor. Hence we let the model decide if this needs to be added or not.
6. workingday: Median of bike demand cnt is more or less similar for working (1) and not-a-working-day (0), a little above 4000. This indicates, workingday may or may not be a good predictor for the dependent variable. And so we let the model decide if this needs to be added or not.

Q.2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: Categorical variables cannot be used directly in the model creation and hence it has to be converted into meaningful numeric values which is done through dummy variables encoding for each Categorical variables.

During the dummy variable creation, each categorical variable with say N levels, we create 'N-1' new indicator variables for each of these levels by droping one level using drop_first=True.

It is important to use drop_first=True for mainly two reasons:

- as it helps in reducing the extra column created during dummy variable creation
- and hence it reduces the correlations created among dummy variables and avoids multi-collinearity. Multi-collinearity can impact model training and prediction.

Let's understand this better with categorical variable 'weekday'. There are 7 weekdays, Sunday to Saturday (num 0 to num 7). So we have 7 types of values in Categorical column and we want to create dummy variable for that column. We can do this with just 6 types of values, e.g. say weekday_1 to weekday_7. Whichever weekday it is, that column will be 1 and rest will be zero. And when all columns are zero, it indicates it is weekday_0. So we do not need 7th variable to identify weekday_0.

Hence if we have categorical variable with N-levels, we need to use only N-1 columns to represent the dummy variables.

Q.3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Let us look at the pair-plot and heatmap of correlation to answer this question.

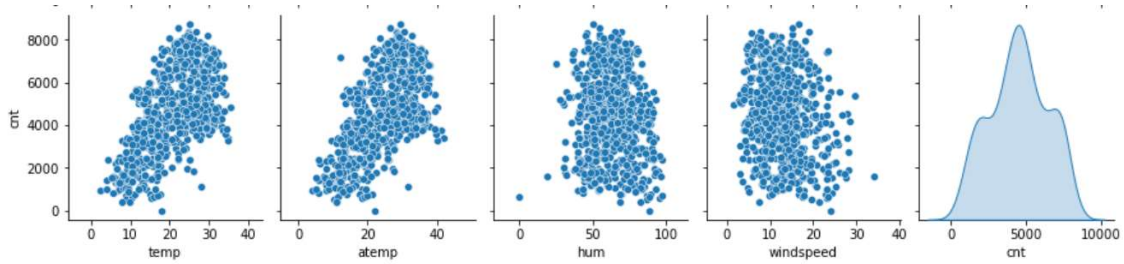


FIGURE 1

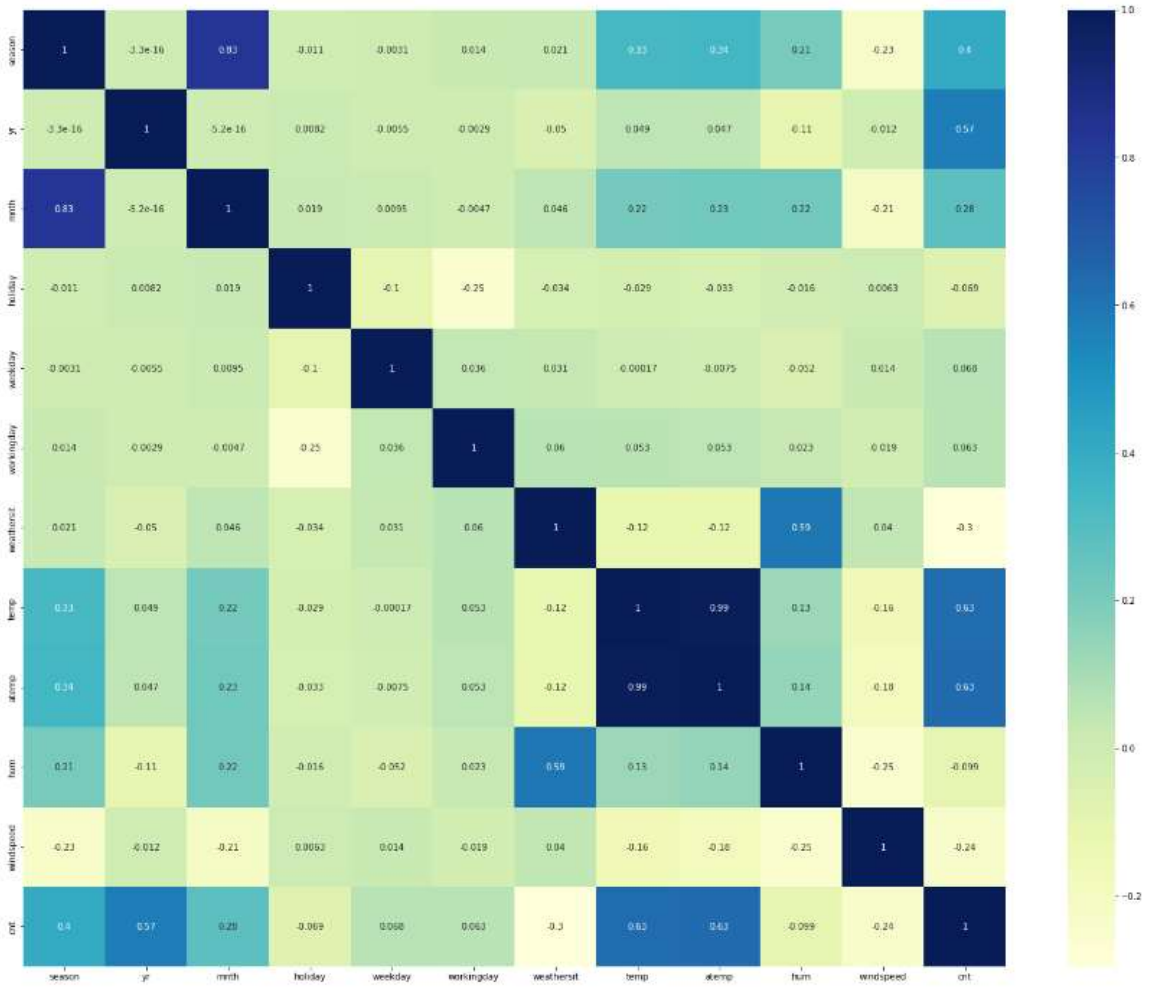
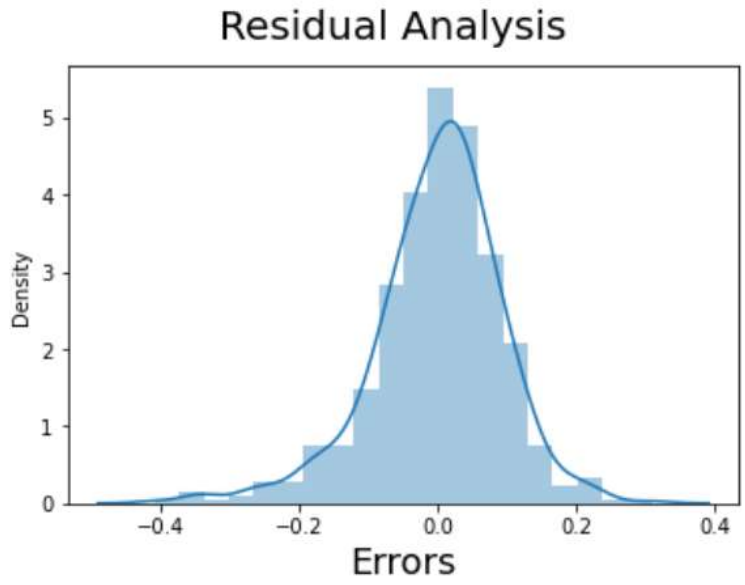


FIGURE 2

After looking at both the figures, we can conclude that count is highly correlated with variable 'temp' with correlation value of 0.63. We are not considering 'atemp' as correlation value of 'temp' and 'atemp' is 0.99 indicating 'atemp' is redundant.
The next highest correlation is with variable year, 0.57. Then season with correlation of 0.4, then month with correlation of 0.28 and then weathersit with negative correlation of -0.3 and then windspeed with negative correlation of -0.24.

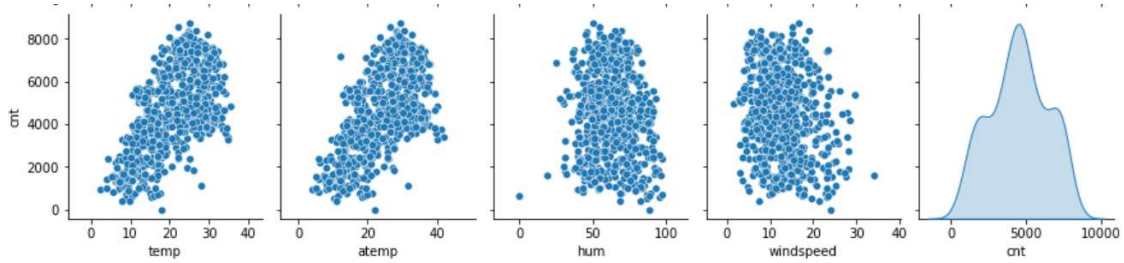
Q.4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: The assumptions of linear regression are as follows:
1. Normality: The residuals of the model (error terms) are normally distributed with mean zero (not X, Y)
Residual Analysis was performed on the training data set and predicted target values to check the Distribution of Error Terms. The graph below shows that Error Terms are Normally Distributed around mean value 0 as assumed.



2. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y

As discussed in answer to Q.3. , 'temp' is linearly related to count and so are other categorical variables that our model is using for prediction.



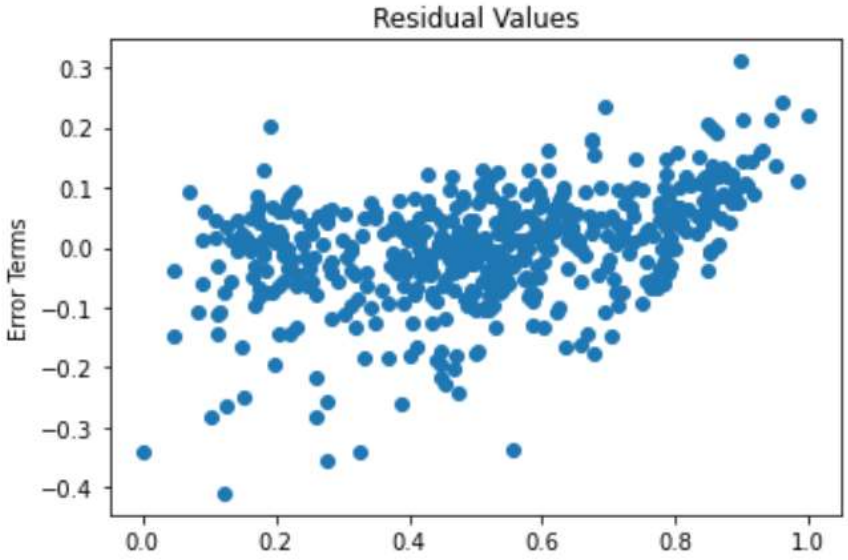
---

3. Multicollinearity: There is No Multicollinearity between the predictor variables i.e. predictor variables are not correlated to each other. And this is clear from the VIF values of all selected variables, all values under 5.

This aspect too was taken care of by removing variables like 'atemp' that had high correlation with other variable. Also, while creating dummy variables, we used N-1 levels instead of N levels using drop_first=True and hence avoiding multu-collinearity.

	Features	VIF
2	temp	4.72
3	windspeed	4.02
1	workingday	4.01
0	yr	2.00
7	weekday_6	1.65
4	season_2	1.56
8	weathersit_2	1.52
5	season_4	1.38
6	mnth_9	1.20
9	weathersit_3	1.07

4. Independent Error terms: Error terms are independent of each other as clear from following scatter plot. They do not follow any patterns.



5. Homoscedasticity: Error terms have constant variance.
The plot also confirms Homoscedasticity. As can be seen the variance does not increase (or decrease) as the error values change. Also, the variance does not follow any pattern as the error terms change.

Q.5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: Let us look at the coefficient values of each predictor variable used in te final model to answer this question.

const	0.084143
yr	0.230846
workingday	0.043203
temp	0.563615
windspeed	-0.155191
season_2	0.082706
season_4	0.128744
mnth_9	0.094743
weekday_6	0.056909
weathersit_2	-0.074807
weathersit_3	-0.306992

The top 3 features are:

(1) temperature with coeff 0.563615 which indicates that a unit increase in temp variable, increases the bike hire numbers by 0.5636 units (almost 56% hike in count which makes sense logically too)

(2) weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) with a coefficient value of '-0.3070' indicates that, a unit increase in Weathersit3 variable, decreases the bike hire numbers by 0.3070 units (30% reduction in count)

(3) year with a coefficient value of '0.2308' indicates that a unit increase in year variable, increases the bike hire numbers by 0.2308 units (which is almost 23% hike in count). This indicates that with time, popularity of bike is increasing which in turn increases the count

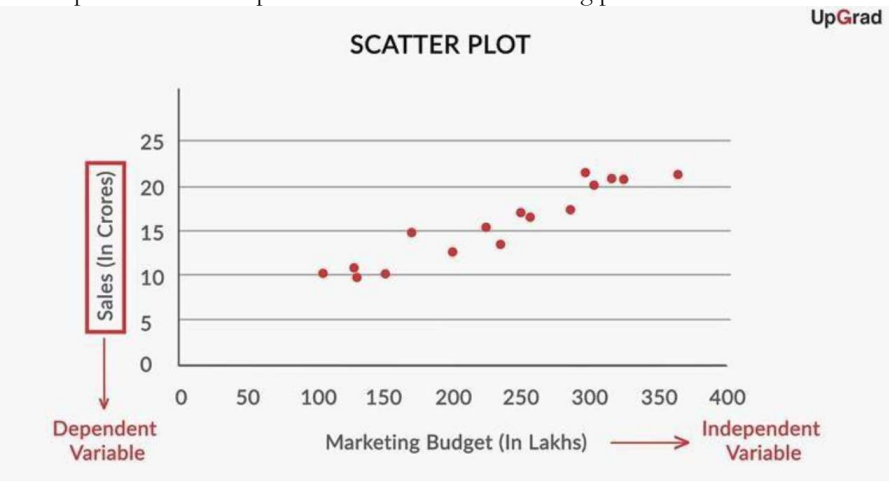
General Subjective Questions

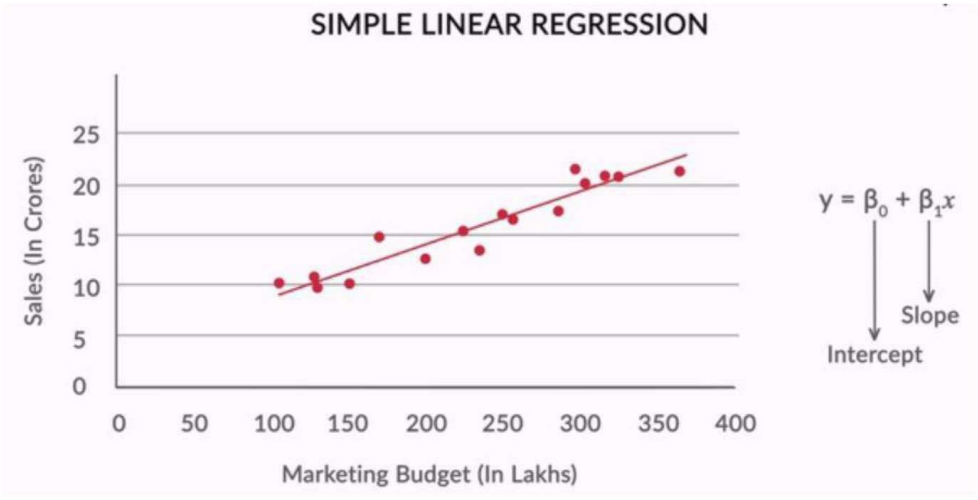
Q.1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Regression is the most commonly used supervised predictive analysis model.
Broadly speaking, it is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).
There are two types of linear regression:

- Simple linear regression - When one predictor variable is used
- Multiple linear regression - When multiple predictors are used

Simple Linear Regression
The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points as shown in the following plots.



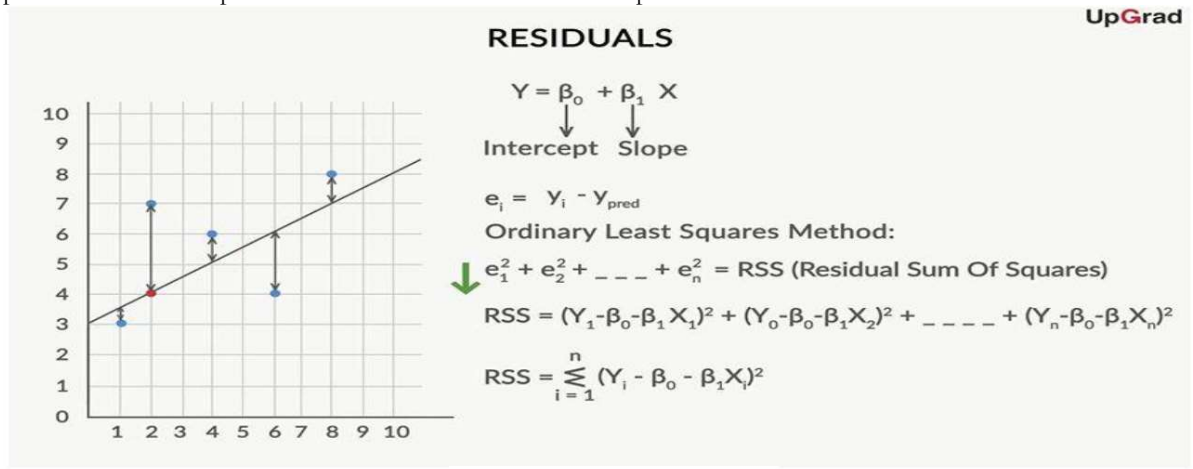


- Regression Line

In a Simple Linear Regression, the regression line is given by
y = B0 + B1 * X
Where
B0 : Intercept means Value of Y when X=0
B1 : Slope of the fitted line
Once we find the best B1 and B2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of X.
Similarly in Multiple Linear Regression, the Regression line is given by
y = B0 + B1 * x1 + B2 * x2 .... + Bi * xi
where x1, x2, x3, … , xi are multiple data points

Best Fit Line

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



Basically, we find the Best Fit line by minimizing the cost function which is RSS (Residual Sum of Squares).
RSS = sum of squares of residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

Drawbacks of RSS is it's value will change based on the unit so need to define alternative measure which should be more "Relative" and Not absolute so TSS (Total Sum of Squares) is defined.

Using RSS and TSS, R-Squared or Coefficient of Determination is defined as R-Squared = 1 -(RSS/TSS) .
R-Squared statistics provides measures of how well actual data points are replicated by model based on the total variations of outcomes as explained by the model i.e. expected data points. Basically, Higher R-Squared value mean Model fits the data.

Once the Simple or Multiple linear regression Model is built, we need to do the residual analysis and see if it meets the Assumptions of Linear Regression to confirm the model fit

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

In Multiple linear regression Model, need to check for

- Overfit
    - Model is complex and provides a near perfect linearity for Actual and predicted values
    - Basically this show High R2 values for Training set while low R2 for test set
- Multicolinearity
    - This scenario happens when there is co-relation b/w variables used in model
    - This needs to be avoided for a decent model
    - Variance Inflation Factors values can be used to eliminate Multicolinearity withing variables
        - VIF < 5 - acceptable and no need to drop variable
        - VIF > 5 - Ok, but need to rechecked
        - VIF > 10 - Drop the variable

Q.2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that can impact the regression model and rather fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."
This can be used to demonstrate both, the importance of graphing data when analysis to identify any anomalies and effect of outliers and other parameters on statistical properties.
There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

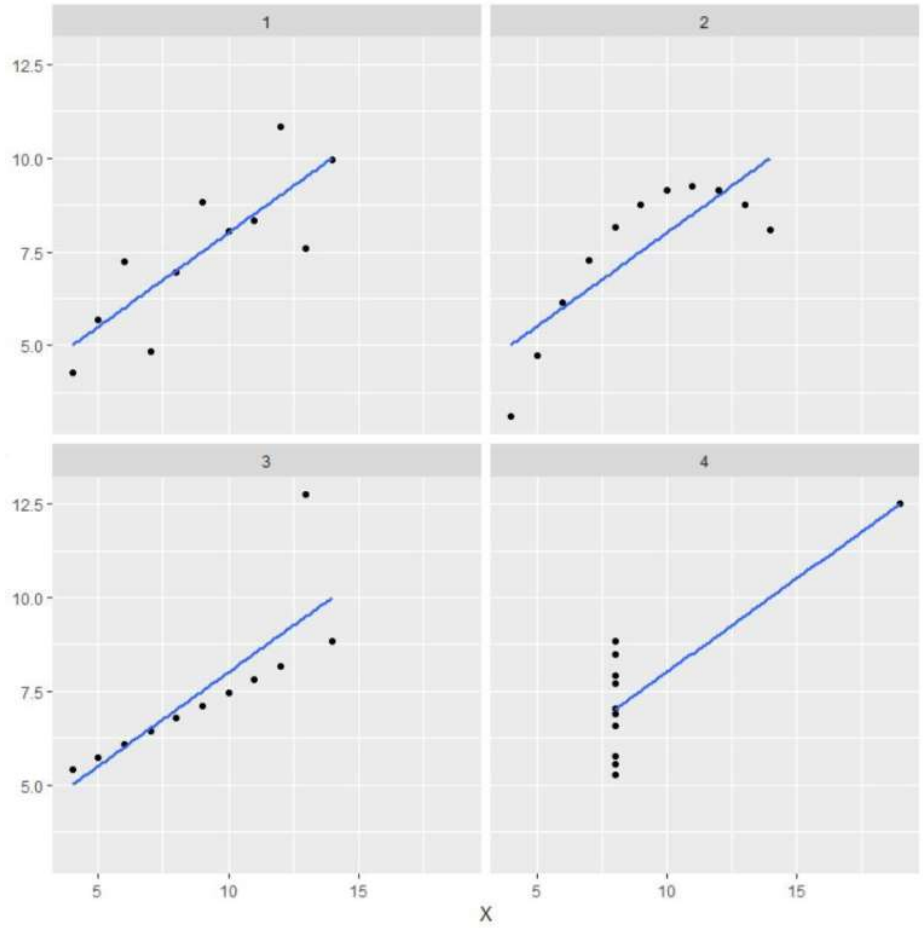
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Statistical Information for all above datasets is approximately similar.

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But when these models are plotted they generate totally different kind of scatter plots as below:



This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data etc.

The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is nonlinear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression mode
-

Summary:
The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Q.3. What is Pearson's R? (3 marks)

Answer: Pearson's R is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrates.

Pearson's R or Bivariate correlation, is a statistics that measures the linear correlation between 2 variables and like other correlations it's numerical value lies between -1.0 and +1.0.

R = 1 means the data is perfectly linearly correlated with a positive slope (i.e., both variables tend to change in the same direction)
R = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
R = 0 means there is no linear association
0< R < 5 means there is a weak association
5 < R < 8 means there is a moderate association
R > 8 means there is a strong association

Peasrson's correlation coefficient is the Covariance of 2 variables divided by the product of their standard deviations.

Pearson's correlation coefficient, when applied to population (rho) can be terms as Population correlation coefficient or Population Pearson correlation coefficient.

For Random variables X , Y
Population(X , Y) = Covariance(X,Y) / (Std dev of X * Std dev of Y)

A key mathematical property of the Pearson correlation coefficient is that it is invariant under separate changes in location and scale in the two variables. That is, we may transform X to a + bX and transform Y to c + dY, where a, b, c, and d are constants with b, d > 0, without changing the correlation coefficient.

Q.4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Feature Scaling is an important step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units, it will lead to a model with very weird coefficients that might be difficult to interpret. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It also helps in speeding up the calculations in an algorithm.

So we need to scale features because of two reasons:
1. Ease of interpretation
2. Faster convergence for gradient descent methods

You can scale the features using two very popular method:
1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

2. Normalization or MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

Q.5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity / multicollinearity.

Based on the VIF Formula as below, VIF can be infinite if R^2 is 1. If R^2 is 1, this means R= 1 and there is a perfect correlation between 2 independent Variables.

VIF = 1/ (1-R^2)

A large value of VIF indicates that there is a strong correlation between the variables. To solve this, we have to drop one of the two variables from the dataset that is causing this perfect Multi-collinearity.

Q.6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: Q-Q plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q to find out if two sets of data come from the same distribution or not. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, then points will fall on or very close to that 45-degree reference line.

The slope tells us whether the steps in our data are too big or too small. For example, if we have N observations, then each step traverses 1/(N-1) of the data. So we are seeing how the step sizes (a.k.a. quantiles) compare between our data and the normal distribution. A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed. One example cause of this would be an unusually large number of outliers (like in the QQ plot we drew with our code previously).

In Linear Regression this is helpful 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 or not.

Some Advantages:
- it can be used with sample sizes
- Distributional aspects like scale shifts, changes in symmetry, presence of outliers can be detected

It can be used to check below scenarios,
- Population data with a common distribution
- Have Common scale
- Have Similar distributional shapes
- Have similar tail behaviour