Assignment-based Subjective Questions

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

Ans: Through the analysis we can say different things for the different categorical variables.

1. season - for the season the median of the fall is better and can be the reason for the high count of bike rentals.
2. year- for the year 2019 is the year where the maximum bikes have been rented out when it is compared to the 2018
3. month- for the month of july the median is on the top hence july saw the maximum rentals for the bike.
4. weekday- sunday shows the highest bike rentals out of all the weekdays.
5. weathersit- on clear whether people tend to rent bike more instead of mist and light snow.

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

Ans 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.

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

Ans: looking at the pairplot the Temp is having the best positive correlation with the cnt dependent variable.

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

Ans: Taking a more statistical view:

● Linear regression, at each X, finds the best estimate for Y

● At each X, there is a distribution on the values of Y Model predicts a single value, therefore there is a distribution of error terms at each of these values as can be seen from the figure below.the assumptions of simple linear regression were:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity) With these assumptions we can go ahead and make inferences about the model which, otherwise, we wouldn't have been able to. Also note that,there is NO assumption on the distribution of Xand Y, just that the error terms have to have a normal distribution

How we check these assumptions are:

- The normal distribution of the residual terms is a very crucial assumption when it comes to making inferences from a linear regression model. Hence, itis very important that you analyses these residual terms before you can move forward. The simplest method to check for the normality is to plot a histogram of the error terms and check whether the error terms are normal.

- Apart from this, you also need to check for visible patterns in the error terms in order to determine that these terms have a constant variance.

----

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

Ans: the top three features affecting the rents of the bicycle:

- Temperature
- Season(Summer/Winter)
- Year(2019)

General Subjective Questions

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

Ans: Regression:The output variable to be predicted is a continuous variable, e.g. scores of a student.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 straightline. The straight line is plotted on the scatter plot of these two points.

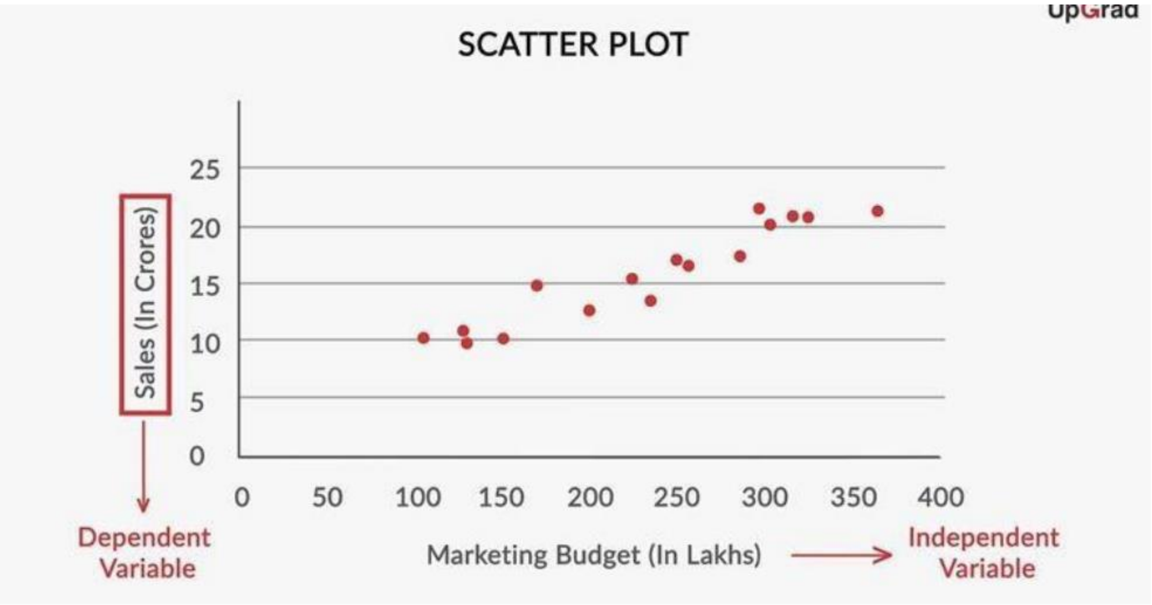


Figure 2 - Scatter plot

The standard equation of the regression line is given by the following expression: Y = $\beta_0 + \beta_1 X$

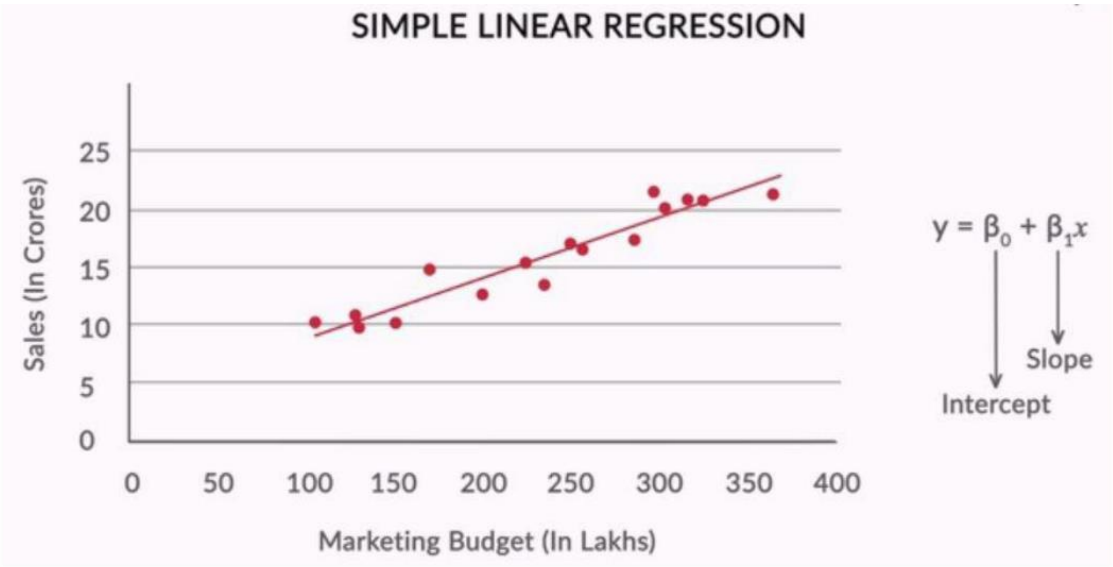
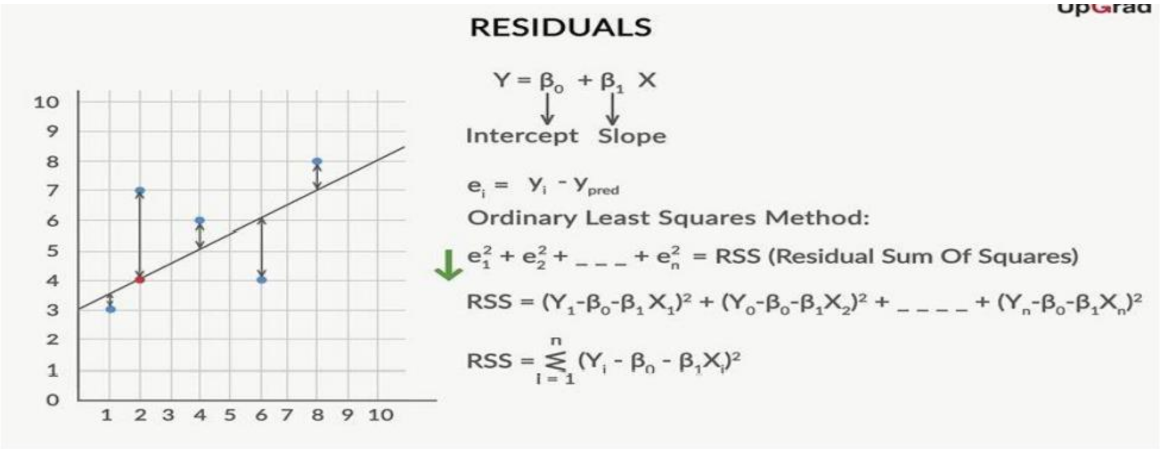


Figure 3 - Regression Line

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



RESIDUALS

$$Y = \underset{\text{Intercept}}{\underset{\downarrow}{\beta_o}} + \underset{\text{Slope}}{\underset{\downarrow}{\beta_1}} X$$

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$\downarrow$ $e_1^2 + e_2^2 + \_\,\_\,\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$$RSS = (Y_1 - \beta_o - \beta_1 X_1)^2 + (Y_o - \beta_o - \beta_1 X_2)^2 + \_\,\_\,\_\,\_ + (Y_n - \beta_o - \beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

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

Ans: Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

They were constructed in 1973 by the statistician Francis

Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

The 4 Quarters specifies that:

- In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Applications:

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.

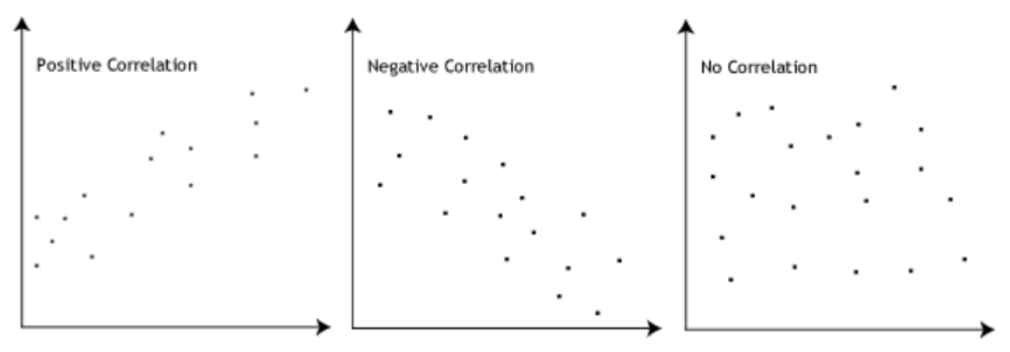
----

3. What is Pearson’s R? (3 marks)

Ans : In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear 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
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

----

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

Ans: Feature Scaling: Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. 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. **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

----

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

Ans: If there is perfect correlation, then VIF = infinity. 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.

----

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

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. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.