Linear Regression Subjective Q & A

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. From the model analysis I find the following,
   a) Bike sales are growing yearly.
   b) Higher temperatures influence rentals positively.
   c) Snow, windspeed and mist and snow have negative effects on rentals, in that order
   d) The months, Nov to Feb, have a negative effect on rentals

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

   Ans. When we use drop_first=True, the number of columns generated is k-1 where k is the number of values in the original column. This helps in reducing 1 column and that helps in keeping the complexity lower.

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

   Ans. As per the pair-plot the highest correlation between the variables with the target variable is with the column name, **registered**.

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

   Ans. I validated the assumptions of Linear regression by looking at the model parameters after building it. Those were the following,

   a) Identifying correlated variables and removing them from the model
   b) Keeping variables with VIF scores around 5 or below.
   c) Checking that P value of the coefficients are around 0.05 or below.
   d) A good P squared score. In our case it is .784.
   e) Checking the residuals to check the distribution and see it is a normal distribution with the mean around 0.

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. As per the final model the top 3 features are,

   a) Year
   b) Workingday
   c) Weathersit

General Subjective Questions

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

Ans. Linear Regression algorithm is a Machine Learning Algorithm. It is a type of supervised Machine Learning, i.e. the target variable is known or available and it is also continuous. Linear Regression algorithm takes the input dataset and tries to fit the input into a single line, i.e. the model, that closely resembles the relationship of the input variables to the target variable. This technique provides us with clear insights into the effect of the independent variables on the target variable. The coefficients of each variable gives us the impact of each independent variable.
There are 2 types of Linear Regression,

a) Simple Linear Regression: 1 independent variable/feature relating to the target variable.
The equation of Simple Linear Regression is,

$Y = \beta_0 + \beta_1 X$

- Y is the target variable
- β_0 is the intercept
- X is the independent variable
- β_1 is the coefficient of X, i.e the slope

b) Multiple Linear Regression: More than 1 independent variable/feature relating to the target variable.
The equation of Multiple Linear Regression is,

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + ... + \beta_n X_n$

- Y is the target variable
- β_0 is the intercept
- X_1 to X_n are the independent variables
- β_1 to β_n are the coefficients of X_1 to X_n

Using Linear Regression the objective is find the best fit line that best represents the data points. The best fit line is the one that has minimum errors.

There are a number of ways errors are calculated,

a) Mean Square Error (MSE)
b) Mean Absolute Error (MAE)
c) Root Mean Squared Error (RMSE)
d) R squared error
e) Adjusted R squared error

The most common method of checking for the best fit line is to have a high R squared error/Adjusted R squared error, as close to 1 as possible.

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

Ans. Anscombe's quartet is a set of 4 datasets with almost identical summary statistics yet completely different representations on graphs. The datasets have similar statistical properties like, means, variances, R-squared, correlations, linear regression lines.

The datasets were created by Francis Anscombe to illustrate the importance of visualizing data and to show solely relying on summary statistics can be misleading.
The 4 datasets of Ansombe's quartet has a set of 11 rows, i.e. unique pairs of x and y. On a graph each dataset has a unique pattern, whereas the summary statistics are almost matching with each other.
The purpose of this quartet is to highlight the importance of EDA and visualize the data to spot trends, outliers and other important details that otherwise might be not be spotted just by looking at the summary data.

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

Ans. Pearson's R is the Pearson correlation coefficient. It is a measure of strength and direction of the linear relationship between x and y variables.

The formula of Pearson's R is,

$$r = \frac{n(\sum xy) - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Here,

- x is the independent variable
- y is the dependent variable
- n is the sample size
- Σ is the summation of all the values

The value of r ranges from -1 to 1.

- 1 means that there is a perfectly positive linear relationship between x and y
- -1 means that there is a perfectly negative linear relationship.
- 0 means that there is no relationship between x and y

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

Ans. When we try to fit a model there are a number of variables we need to consider, including the dependent and the independent ones. Each variable could on a very different scale. Say, x_1 could be between 0 and 9 whereas x_2 could be in the thousands range and y_2 could be in lakhs. In this scenario there will be a very wide scale of the coefficients.

- Thus the coefficients could be imprecise as the effect of one coefficient might look insignificant to the other. Model will have to assign larger weights to one variable over the other.
- Computational time will be higher as the computer will have to compute wide variance in scales across variables.

Scaling is a technique to bring data points closer to each other, reducing the distance between them and reduce the time to build a more reliable model.

There are 2 main methods for scaling features,

a) Normalization: It is a method to bring all the values on a common scale, e.g. Adjusting all values to range between 0 and 1.

The formula for Normalisation is as follows,

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

So for each column, the values are computed as a normalised between the maximum and minimum. This ends up with a column value set between 0 and 1.

b) Standardization: This is another method for converting the values in columns with a range, i.e. mean and standard deviation. However the technique aims to keep the values centred around mean and go up to the standard deviation 1.

The formula for Standardisation is,

$$x_{stand} = \frac{x - mean(x)}{standard\ deviation(x)}$$

Since the results provided by the standardization are not bounded by a fixed range as we have seen in normalization, it can be used with the data where the distribution is following the Gaussian distribution.

Standardization is not affected by outliers wherein normalization does, as it captures all the data points in their ranges.

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

Ans. When 2 or more variables are exactly equal then they are perfectly correlated and their R square value is 1. This results in an infinite VIF. The formula for VIF is,

$$VIF = \frac{1}{1 - R^2}$$

As you can see when R square equals 1, the denominator becomes 0 and hence the VIF value becomes infinite.
This happens when the variables are exactly the same or there are dataset issues like there are more variables than observations, e.g. 1000 variables for 40 observations.

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

Ans. A Q-Q plot or Quantile-Quantile plot is a graphical method for determining if a dataset follows particular probability distribution or whether 2 data samples came from the same population or not.
Q-Q plots are helpful in determining if a dataset is normally distributed or if it follows some other type of distribution.

Quantiles are points in a dataset that divide the data into intervals containing equal probabilities or proportions of the total distribution, e.g. 25th, 50th , 75th percentiles.

The way a Q-Q plot is generated is as follows,

a) The data is sorted and the percentiles are demarcated
b) Then we take a theoretical distribution, say a normal distribution, and demarcate the same percentiles on it.
c) We then plot the values from point (a) against (b) and draw a straight line to match the plotted points as closely as possible.

If the straight line matches very closely then the assumed theoretical distribution was correct. If not, we have to take up another kind of distribution and redo the steps.

Q-Q plots are used to validate predictive models ad so are very useful in linear regression as it shows how well the model fits the data.