Multiple Linear Regression question/Answers - Vimala Subramanian

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)

1. season
    a. There seems to be a outlier in spring
    b. No of bikes used in fall is high followed by summer and winter. Surprisingly spring is less than winter

2. weathersit

a. As expected the usage of bike is when weather is clear followed by misty_cloudy. b. There is no bike usage during heavyrain_thunder

3. holiday

The usage of bikes is low on a holiday, so on working days the average is more

4. weekday

The usage of bikes on most of the weekdays is same.

5. workingday

Whether it is a working day or not, the bike usage is more or less same

6. yr

The bike usage increases from 2018 to 2019 almost 1.5 times up.

7. mnth

the usage of bikes is high in jul, sep, jun

8. As temp and humidity increases the usage of bike is also increasing, and when windspeed is low

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

As we have observed the first dummy variable column can be represented by other dummy variables for the same categorical variable and as it is redundant, we drop the first dummy variable column. It is also used to reduce the collinearity between the dummy variables .

For ex: for a categorical variable with values like low, medium, high = the value, low can be represented in the absence of medium and high.

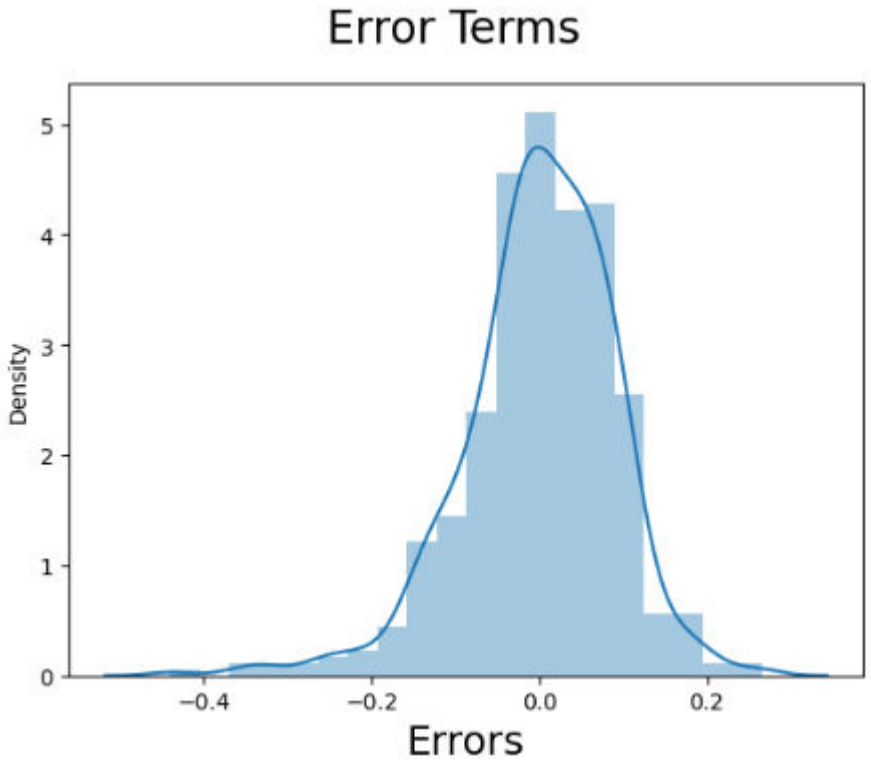
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

temp (independent variable) has high correlation with the target variable cnt

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

set? (3 marks)

I performed a residual analysis on the train data set. Plotted a distplot to see the histogram which turned out to be with mean=0, normally distributed



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

a. Positive influencing features are temp = 0.5640 - indicates that a unit increase in temp variable, increased the bike sales by 0.5640 units ,

b. yr_2019 = 0.2377 - indicates that a unit increase in yr_2019 variable, increased the bike sales by 0.2377 units followed by

c. weathersit_LightSnow_Rain = -0.2305 which is a negative influencing feature. - indicates that a unit increase in weathersit_LightSnow_Rain variable, decreased the bike sales by 0.2305 units.

General Subjective Questions

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

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

This form of analysis estimates the coefficients of the linear equation, involving one or more

independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. We can then estimate the value of X (dependent variable) from Y (independent variable).

Linear regression models are relatively simple and provide an easy-to-interpret mathematical formula to generate predictions. Linear regression is an established statistical technique and applies easily to software and computing. Businesses use it to reliably and predictably convert raw data into business intelligence and actionable insights. Scientists in many fields, including biology and the behavioral, environmental, and social sciences, use linear regression to conduct preliminary data analysis and predict future trends. Many data science methods, such as machine learning and artificial intelligence, use linear regression to solve complex problems.

Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages – (1) analyzing the correlation and directionality of the data, (2) estimating the model, i.e., fitting the line, and (3) evaluating the validity and usefulness of the model.

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

Developed by F.J. Anscombe in 1973, Anscombe's Quartet is a set of four datasets, where each produces the same summary statistics (mean, standard deviation, and correlation), which could lead one to believe the datasets are quire similar.

It enhances our view in data visualization by recommeding us to look closely at our data, question our assumptions, and use a variety of analytical tools to get a full picture. The key takeaway from Anscombe's quartet is that summary statistics alone may not be sufficient to capture the essence of a dataset, and visual exploration through graphs can reveal important insights into the nature of the data.

3. What is Pearson’s R? (3 marks)

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables.

Pearson correlation draws a line of best fit through two variables, indicating the distance of data points from this line. A 'r' value near +1 or -1 implies all data points are close to the line. An 'r' value close to '0' suggests data points are scattered around the line.

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features with highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling is necessary for a model to be functional with the appropriate range of coefficients. For e.g., if there were two independent variables named price and months on which the sale of car depended, the price range would be far too high because there are only 12 months in a year. In that case, scaling the variable price appropriately won't allow decimal errors to happen in the model.

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-squared, etc.

There are two types of scaling:

1. Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing. MinMaxScaler helps to implement normalization in python. This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in Neural networks broadly.

2. Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/(1-R^2)) turns out to approach infinity.

The greater the VIF, the higher the degree of multicollinearity. In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

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

(3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure

from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.

2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

3. The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.