Linear Regression Assignment

upGrad & IIITB | Data Science Program - January 2023

Saba Afreen

DSC-52

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

The following are some of the inferences of the Categorical variable that effect on the dependent variables.

1. Season: fall has the highest demand for bikes However, spring has counted as the lowest demand in bikes.
2. In 2019 the demand was at peak.(1:2019, 0:2018)
3. In June the demand recorded as 7000 which is highest among other months.
4. The maximum demand for holiday or no holiday is same and the demand decreased when there is a holiday.
5. There is no major difference between weekdays.
6. A good weather sit has highest demand.

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

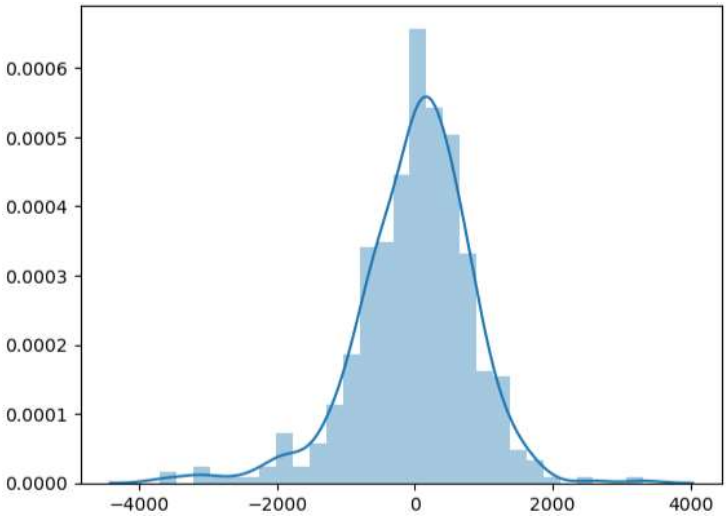
- It is important to use drop_first=True as it helps in reducing the extra column which is created when we perform created_dummies,
- Example: we have t
  hree variable of winter i.e. summer, winter, spring, rainy.
- When we perform dummies all four columns are created separately with placing 1 on preferred place.
- Assuming 0 in three columns will automatically takes as summer, which is removed by the function drop_first=True

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

- The numerical variable ‘atemp’ has the highest correlation with target variable ‘cnt’ with a value of ‘0.65’ followed by a ‘temp’ with a value of ‘0.64’.

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

- We validate the assumptions of the linear Regression by plotting a distplot of the residual and analysing it to see if it is a normal distribution or not and if it has a mean=0. The diagram below shows that it is normally distributed with mean=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)

The following are the top3 features contributing significantly towards explaining the demand of the shared bikes:

1. Temp: A coefficient value 4763.6123 indicates that a unit increase in temp variable, increases the bike hire numbers by 4763.6123 units.
2. Light Snow(Weathersit): A coefficient value of -2467.1107 indicates that, a unit increase of this variable, decreases the bike hire numbers by -2467.1107 units.
3. Yr: A coefficient value of 2023.3801 indicates that, a unit increase of this variable, increase the bike hire numbers by 2023.3801 units.

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

- Linear regression is a supervised machine learning method that basically uses for analysis and modeling purpose.
- This method is mostly used by the autoML tool for describing the linearity or correlation between the independent and dependent variable.
- This can be achieve by the visual representation by fitting a line on scatter plot, this can be done by Least Square method.
- Finally, models are tested on various combinations unless the R-squared recorded as approximate to 70 or 80.
- Then the prepared model is applied on the test data to predict the target set, here it is compulsory that the R-squared should be equal to R-squared of the final model.

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

- Anscombe's quartet is a group of the plot that appears to be same when perform statistics but there are drastic difference when the scatter plots are plotted.
- 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. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

3. What is Pearson’s R? (3 marks)

- Pearson’s R is nothing but Pearson's correlation coefficient (r).
- It is used to measure the linear correlation.
- The range is from -1 to 1 that measures the strength and relationship between two variables.
- If the correlation of two variables is 1 it is said that this has highest linearity However, if the number is in negative e.g. -0.87 this means that the variables are strong negatively correlated.

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

- It is one of the compulsory step of linear regression which is applied to independent variables to normalize the data to a particular rage. Also helps in speedup the calculation.
- The collected data may contain many types & they may have highly varying magnitude, units and range. If scaling is not done then algorithm take the units which leads to incorrect modelling. To avoid from this issue, scaling are perform.
- Normalization/min-max scaling brings all the data in the range 0 and 1 whereas, 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**).

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

- A variance inflation factor gives a measure of multicollinearity among the independent variable in a multiple regression model.
- If there are any highly correlated variable then the VIF of that variable will be infinite.

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

- Q-Q plot is known as quantile-quantile plots. The use of this is to plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
- With the use of this we can observe the particular type probablity distrubution like normal, uniform, exponential.