LINEAR REGRESSION ASSIGNMENT

Submitted By:

Abhishek Kumar Goyal (APFE21709647)

ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

Ques 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 1) In the bike sharing dataset provided for the assignment, there are the following categorical variables that were analyzed. These variables individually had an impact on the target variable 'cnt' which is described as follows:

i. *'season' :* It was observed that the count of the bike rentals is highest in the Fall season followed by Summer and then Winter season. The count is very less for Sprnig season as compared to the other seasons.

ii. *'mnth' :* The maximum number of bike rentals were observed in the month of September with the least number being in January. Generally observed, the bike count increased as we moved from January till September after which the count started decreasing till December.

iii. *'weekday' :* The count of bike rentals is very high on Thursdays and Fridays while it is the lowest on Sundays.

iv. *'weathersit':* It was observed that most users rented bikes when the weather was clear or partly cloudy. In the case here were heavy rainfall or snowfall, no user rented a bike as the weather is not suitable for biking.

v. *'yr' :* The count of bikes rented increased from 2018 to 2019.

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

Ans 2) When we create dummy variables for a particular variable, we use drop_first=True, in order to remove an extra column from being created. It helps in reducing the multicollinearity between the independent variables as the *nth* column that is dropped could be explained using the other *n-1* variables. Hence, it helps in reducing the number of independent variables, reduces multicollinearity and simplifies the model.

For e.g : If we have 4 values for the 'season' column namely 'Fall', 'Spring', 'Summer' and 'Winter', then while creating dummy variables 4 new columns will be created. But, we can drop the 'Fall' column as it can be easily derived from other 3 columns when their values are 0.

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

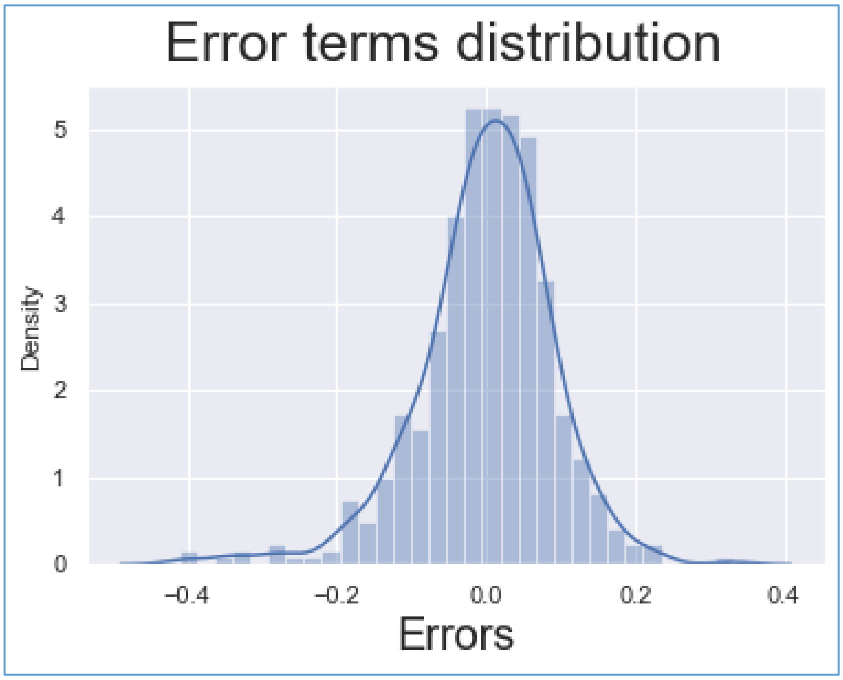
Ans 3) After plotting the pair-plot, it was observed that out of all the variables, ‘temp’ and ‘atemp’ are the two relevant variables that had the maximum correlation with the target variable ‘cnt’.

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

Ans 4) In order to validate the following assumptions of linear regression, I did the below steps :

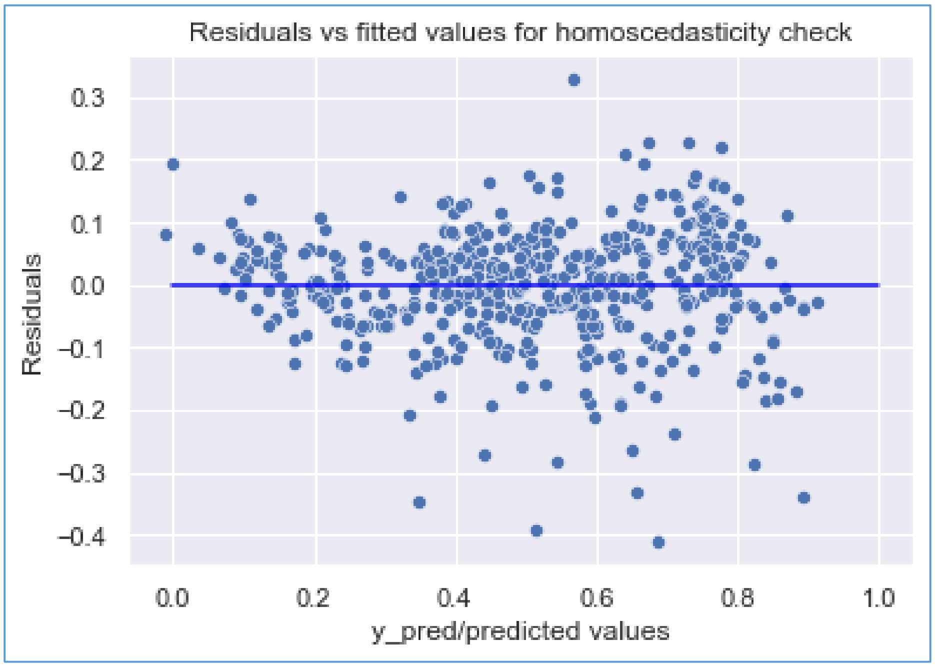
i. *The error terms should be normally distributed*

Plotted a distribution plot of the residuals (Computed using the actual values and the predicted values from the model).
The below plot shows that the error terms are normally distributed with mean centered around 0.



ii. *Error terms have constant variance homoscedasticity*

Plotted a scatterplot of the residual values vs the fitted/predicted values.
The below plot shows that the error terms have no discernible pattern and hence have constant variance.



Ques 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 5) The top 3 features from the final model which contribute significantly towards the demand of shared bikes are :

- temp
- yr
- weathersit_light_snow_rain

GENERAL SUBJECTIVE QUESTIONS

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

Ans 1) The Linear Regression is a Machine Learning algorithm which falls under the category of Supervised Learning algorithm.

It is used to explain the relationship between dependent (output variable) and independent variable using a straight line. In such models, the output variable to be predicted is a continuous numerical variable. The Linear Regression model looks to find the best-fitted line to explain the effect of the independent variables (predictor variables) on the dependent variable (target variable). It is mostly done by the Residual Sum of Squares method.

There are two types of Linear Regression model:

a. Simple Linear Regression (1 predictor variable)

Equation of the best fit line is represented by: $y = \beta_0 + \beta_1 * x$

Here, y represents the target variable and x represents the predictor variable.

b. Multiple Linear Regression (Multiple predictor variables)

Equation of the best fit line is represented by: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots. + \beta_n * x_n + \varepsilon$

Here, y represents the target variable and $x_1, x_2, \ldots, x_n$ represents the n predictor variables.

The strength of a linear regression algorithm can be identified by using the R-squared value which should lie in the range of [0, 1] with 1 being perfect-fit case.

Assumptions:

There are some assumptions which are made by the linear regression model which are:

i. Linearity: It assumes that the relationship between the dependent and independent variables is linear.

ii. Assumptions about residuals :

a. Error terms are normally distributed.

b. The residuals have a mean value of 0.

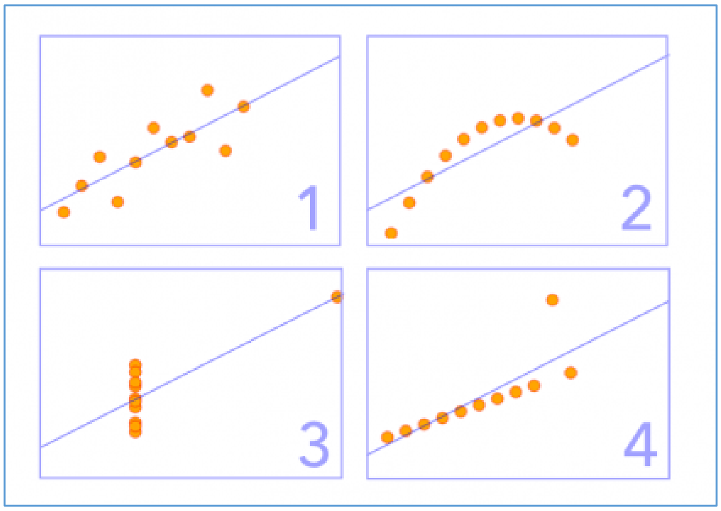
c. Error terms are independent of each other.

d. Error terms have constant variance (homoscedasticity).

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

Ans 2) Anscombe's quartet is a set of 4 datasets that was created to illustrate the importance of plotting and visualizing data before applying Linear Regression model. The four datasets used in this quartet are very much similar when compared on the basic of summary statistics but portray a completely different picture when they are graphed.

The Anscombe's quartet graphically looks like:



The summary statistics of these four datasets show that the means and variances for these datasets were same.

However, when plotted graphically, the results look quite different as portrayed above.

Here,

→ The **first** dataset seems to be well explained and has a good fit on the regression line with some variance.

→ The **second** dataset shows that Linear Regression is not suitable for modelling relationships other than linear as the data does not appear to be normally distributed.

→ The **third** dataset shows that the value of 'x' remains constant except for one outlier.

→ The **fourth** dataset looks to be very well explained apart from the one outlier which makes the line look not well fitted.

Ques 3) **What is Pearson's R? (3 marks)**

Ans 3) Pearson's R is a correlation coefficient which is used to measure the strength of the association of two variables and is widely used in Linear Regression. The value of the coefficient is numeric and lies in the range of [-1, 1]. When the value is :

a. Negative (Less than 0): It indicates that there is a negative correlation between the variables i.e. if value of one variable increases, the value of other variable increases. A coefficient of -1 indicates a perfect negative correlation.

b. Zero (0): It indicates that there is no correlation between the variables.

c. Positive (Grater than 0): It indicates that there is a positive correlation between the variables i.e. if value of one variable increases/decreases, the value of other variable also moves in the same direction. A coefficient of 1 indicates a perfect positive correlation.

It is calculated by the covariance of the two variables divided by the product of their standard deviations.

It is mainly used for variables which have a linear relationship between them and might not give good results for those with a non-linear relationship.

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

Ans 4) Scaling is a method which is used to align the values of the features between a specific range for e.g [0, 1]. It is performed to bring the different features into the same range so that the model can be effective without any bias towards the higher values. It eases the understanding of the effect of various features and their implementation as change by same factor for each feature will be interpreted in the same manner depending only on its coefficient value.

a. Normalized Scaling:

It is also known as Min-Max scaling and is one of the simplest methods used for scaling. It rescales the features in the range of [0, 1].

General Formula is given by:

$$x' = x - \min(x) / \max(x) - \min(x)$$

where,

x' → Scaled value

x → Actual value

max (x) → Maximum value of feature x

min (x) → Minimum value of feature x

b. Standardized Scaling:

It bring s the values of all the variables into a standard normal distribution with mean 0 and standard deviation 1.

General formula is given by:

$$x' = x - \bar{x} / \sigma$$

where,

x' → Scaled value

x → Actual value

$\bar{x}$ → Mean value of feature x

σ → Standard deviation of feature x

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

Ans 5) The VIF i.e Variance Inflation Factor is used to identify the multicollinearity in a dataset. VIF is used to calculate how well one independent variable is explained by all the other independent variables combined.

It is calculated as:

$$VIF = 1/ 1 - R_i^2$$

The value of VIF for a variable will be infinite when the value of R_i^2 is 1 which implies that the other variables are very well able to explain that variable effectively meaning that the variable is linearly dependent on other variables.

So, if there is a perfect correlation among variables, the VIF value will be infinite.

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

Ans 6) Q-Q plot or the Quantile-Quantile plot is a plot which is used to graphically plot the quantities of a sample distribution with those of a theoretical distribution.

Use:

It is used to identify if there is any sort of probability distribution followed by the dataset. It helps to determine if two datasets which are being compared come from populations which have a common distribution.

Importance:

It is useful for determining various distributional aspects such as shifts in location and scale, changes in the symmetry and also the presence of outliers