# Linear Regression Assignment

## A. Assignment-based Subjective Questions

### Que 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

**Ans.**

Categorical variables have a value which can be separated into different distinct groups based on nature or characteristics. In the given 'BoomBikes' dataset four categorical variables are present, which is season, month, weekday and weathersit.

The season variable explains, season of corresponding city which could be spring, summer, fall or wither. The weathersit variable explains the weather condition on a particular day, which can be divided into clear weather, cloudy, snowy or rainy. Whereas, month and day variables explain weekday and month of the year. For the ease of the calculation and building effective model these variables are converted into dummy variables.

From the obtained model, it is observed that summer and winter seasons are positively associated with the total bike rent with the coefficient of 0.0751 and 0.1267 respectively. In January, September and July months, the bike rental is significant with the coefficient of -0.0379, -0.0473 and 0.0937 respectively.

In case of weather parameters, temperature has positive influence on the overall bike rentals with the coefficient of 0.5703, indicating clear weather is helpful for bike rantings. Whereas, cloudy, snowy, hum and windspeed is negatively associated with the bike rentals, with the coefficient of 0.0543, 0.2385, 0.1708 and 0.1935 respectively.

### Que 2. Why is it important to use drop_first=True during dummy variable creation?

**Ans.**

The dummy variables are created over the categorical variable for better performance and ease to understand data. If a particular categorical variable has 'n' number of levels, then the result of dummy variables divides it into 'n' columns with values 0 and 1. If a particular level is true then it would return 1 else 0.

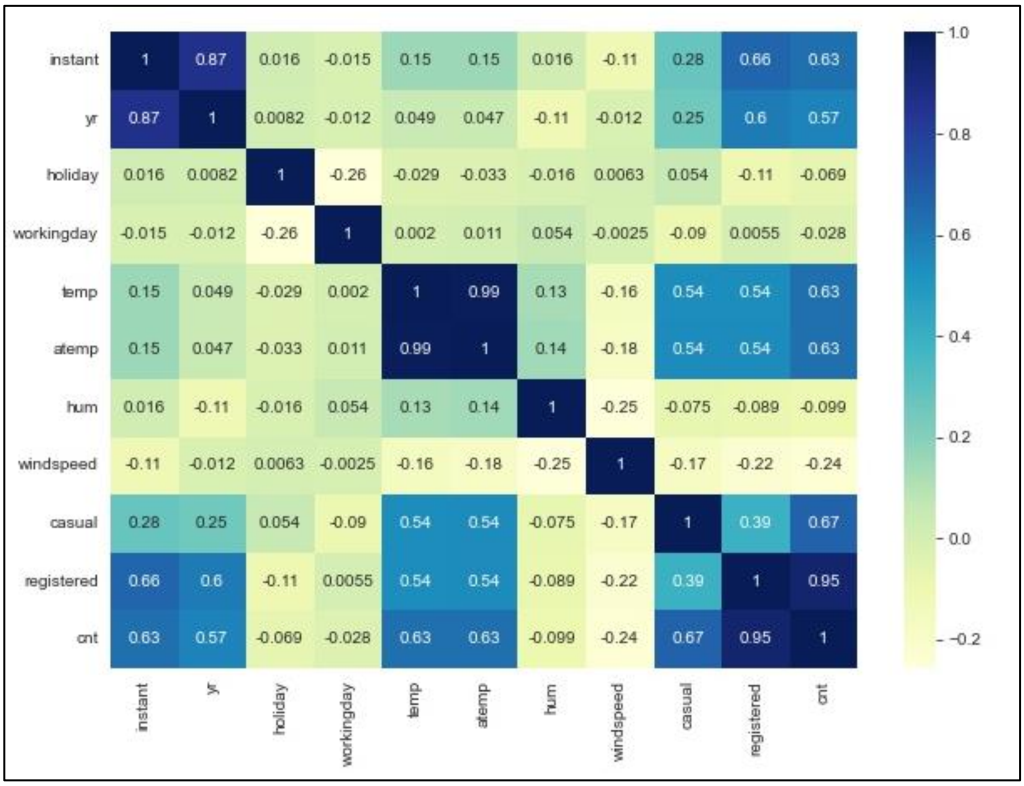
The dummy variables for particular variables can also be understand using n-1 columns. The once column in dataset is becomes redundant as it contains same level of information from data.

### Que 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

**Ans.**

In ‘BoomBikes’ dataset, weather parameters such as temp, atemp, hum, windspeed are numerical variables along with business variables which is casual, registered, cnt. The correlation between these numerical variables is shown in Figure 1. In business parameters, cnt is the summation of casual and registered variables hence it showing higher correlation of 0.28 and 0.66 respectively with target variable ‘cnt’, which in the later stage of analysis might induce the multicollinearity hence need to drop it.

In case of weather parameters, humidity and windspeed is negatively correlated with target variable with the correlation of -0.099 and -0.24. This indicates in windy weather people have not rented the bike because of excessive wind pressure. The temp and atemp variables are almost similar and it showing equal correlation with target variable cnt which is 0.63.



|  | instant | yr | holiday | workingday | temp | atemp | hum | windspeed | casual | registered | cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|
| instant | 1 | 0.87 | 0.016 | -0.015 | 0.15 | 0.15 | 0.016 | -0.11 | 0.28 | 0.66 | 0.63 |
| yr | 0.87 | 1 | 0.0082 | -0.012 | 0.049 | 0.047 | -0.11 | -0.012 | 0.25 | 0.6 | 0.57 |
| holiday | 0.016 | 0.0082 | 1 | -0.26 | -0.029 | -0.033 | -0.016 | 0.0063 | 0.054 | -0.11 | -0.069 |
| workingday | -0.015 | -0.012 | -0.26 | 1 | 0.002 | 0.011 | 0.054 | -0.0025 | -0.09 | 0.0055 | -0.028 |
| temp | 0.15 | 0.049 | -0.029 | 0.002 | 1 | 0.99 | 0.13 | -0.16 | 0.54 | 0.54 | 0.63 |
| atemp | 0.15 | 0.047 | -0.033 | 0.011 | 0.99 | 1 | 0.14 | -0.18 | 0.54 | 0.54 | 0.63 |
| hum | 0.016 | -0.11 | -0.016 | 0.054 | 0.13 | 0.14 | 1 | -0.25 | -0.075 | -0.089 | -0.099 |
| windspeed | -0.11 | -0.012 | 0.0063 | -0.0025 | -0.16 | -0.18 | -0.25 | 1 | -0.17 | -0.22 | -0.24 |
| casual | 0.28 | 0.25 | 0.054 | -0.09 | 0.54 | 0.54 | -0.075 | -0.17 | 1 | 0.39 | 0.67 |
| registered | 0.66 | 0.6 | -0.11 | 0.0055 | 0.54 | 0.54 | -0.089 | -0.22 | 0.39 | 1 | 0.95 |
| cnt | 0.63 | 0.57 | -0.069 | -0.028 | 0.63 | 0.63 | -0.099 | -0.24 | 0.67 | 0.95 | 1 |

Figure 1: Correlation between variables in dataset

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

**Ans:**

The assumptions of linear regression are as follows:

**a.** Linear regression between dependent and independent variables.

Int the given data set the dependent variable is ‘cnt’ and independent variables are weather parameters. By plotting scatter plot between them one can identify the relation between these variables. The relation could be linear or non-linear with positive or negative slope. Figure 2 shows the relation between variables with the help of scatter plot.

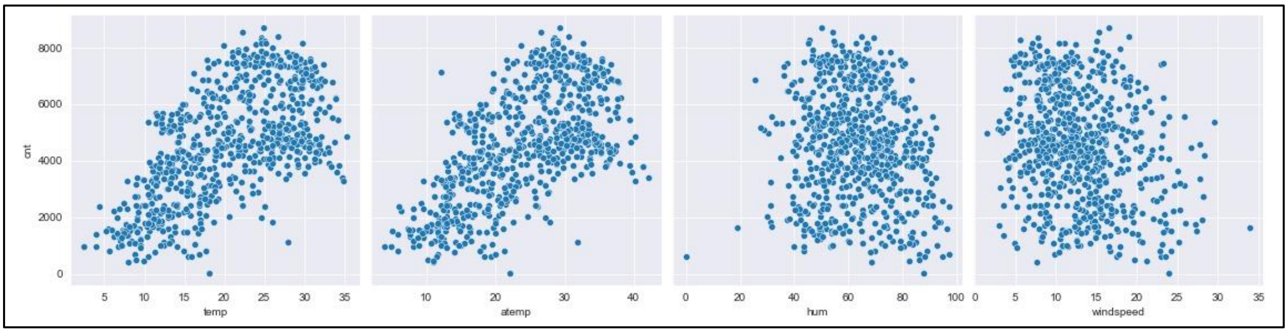


Figure 2: Scatter plot between dependent and independent variables

**b.** Error between actual and predicted dependent variable should be normally distributed

The mean of error term between actual and predicted dependent variables should equal to zero, also it should be normally distributed on both sides of mean value of error term. Figure 3 shows the normal distribution of error terms.

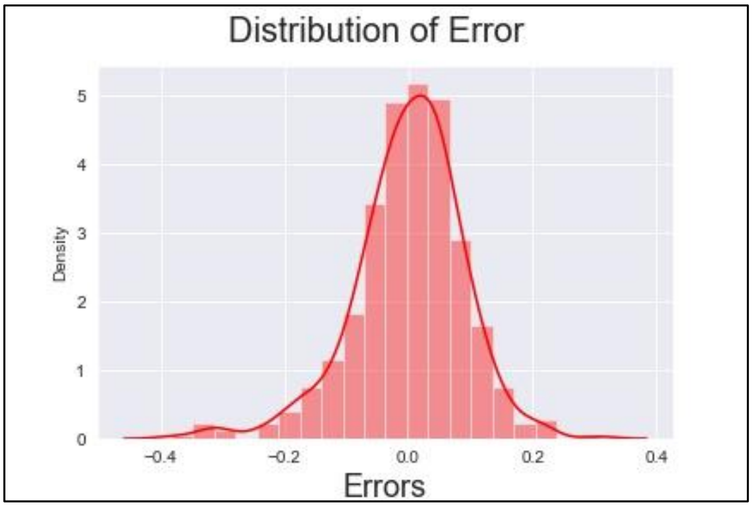


Figure 3: Shows the normal distribution between errors

**c.** Error terms should have the constant variance of error terms are similar across the values of the independent variables (Homoscedasticity)

The error term between train and predicted dataset should have linear relationship with train data also the variance should be constant these is called as Homoscedasticity. Figure 4 shows the relationship between error and predictor variable.

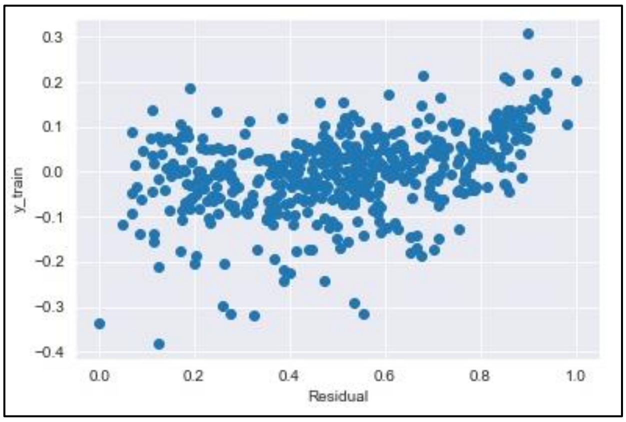


Figure 4: Relation between residuals and predictor variables

**d.** Multicollinearity

The independent variables in the dataset should not be highly correlated with each other, this assumption is tested with the help of Variance Inflation Factor (VIF). The higher the VIF value more will be multicollinearity among the variables, generally it should be kept below 5.

**Que 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

**Ans.**

The final predictive model is as follows:

**Cnt =** 0.2556 + 0.2289 * **yr -** 0.1127 * holiday **-** 0.0189 * workingday + 0.5703 * **temp** – 0.1708 * hum **-** 0.1935 * windspeed + 0.0751 *summer + 0.1267 * **winter -** 0.0379 * jan – 0.0473 * Jul + 0.0937 * sept – 0.0543 * cloudy -0.2385 * snow

In the above model, variable such as yr, temp and winter is highly associated with the bike rental with the coefficient of 0.2289, 0.5703 and 0.1267 respectively.

## B. General Subjective Questions

**Que 1. Explain the linear regression algorithm in detail.**

**Ans:**

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between continuous variables. The term linear regression shows the linear relationship between the independent variables and dependent variables. If the linear regression model contains more than one independent variable then these models is called as multiple linear regression model.

The assumptions of single linear regression are as follows:

a. Linear relationship between X and Y
b. Error term are normally distributed
c. Error term are independent of each other
d. Error term have constant variance.

And in multiple regression analysis the multicollinearity and overfitting concept is added.

Steps involved in multiple linear regression analysis is as follows:

**A.** Reading, understanding and visualizing the data.

Performing the exploratory data analysis over the datasets, numerical variables visualised sung pairplot and categorical variables visualised using the boxplots.

**B.** Preparing of data for machine learning models.

In this stage the scaling, creation of dummy variables and splitting data into train-test sets.

The dataset with different scale is brought into particular range by applying standardization techniques, min-max scaling etc. In dummy variables, the categorical variables with ‘n’ levels is converted into the ‘n-1’ variables for further analysis. For training and testing the model dataset need to divide into train and test sets by using train_test_split with 30-70 or 20-80 ratio. On the basis of train dataset, we build the ML model and we validate the model with the help of test datasets.

**C.** Training the model

In this stage we build the machine learning model with two approaches which is bottom-up, top-down approach and feature selection technique.

In bottom-up approach, model will be built by taking one variable at a time hence choice of the variable becomes very crucial. In top-down approach we consider all the variables at a time and eliminating the unwanted variables on the basis of p-value and VIF.

**D.** Residual analysis:

In residual analysis we analyse the difference between y_train and the predicted model, the residual term should be normally distributed along mean equal to zero. Also, we check for pattern in the error terms by plotting scatter plot between X_train dataset and error term, there should not be any pattern with independent variables.

**E.** Prediction and evaluation on test sets:

In this section we test the model with the obtained model of test datasets. The accuracy of the model is evaluated based on r2_score i,e coefficient of determination and mean squared error. for best suitable model, the r2_score should be higher and mean squared error should be on lower side.

**Que 2. Explain the Anscombe’s quartet in detail.**

**Ans:**

Anscombe’s quartet comprise of four datasets, which contains x and y the descriptive statistics of these 4 datasets are appears to be identical. But there graphical distribution is different because of their values. This dataset was developed by Fransis Anscombe in 1973 who was the statistician by profession.

In the result of descriptive statistics these four different datasets contains equal mean, standard deviation and correlations. The Anscombe’s quartet is used to illustrate the importance of looking at a data graphically before starting to analyse according to particular type of relationship.

**Que 3. What is Pearson’s R?**

**Ans:**

Pearson’s correlation coefficient is one of the most common ways to represent and measure a linear correlation between dependent and independent variables, which is denoted by ‘r’. it is a number between -1 to 1 which measures the strength as well as the direction of the relationship between two variables.

If the value of r is between 0 and 1, this indicates a positive correlation between two variables, indicating one variable changes and the other variable changes in the same direction. If the value of r is equal to 0, which means that there is no correlation between these variables. The negative correlation appears when the value of ‘r’ lies between 0 to -1, this happens when one variable changes the other variable in opposite direction.

**Que 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

**Ans:**

Scaling is the process of normalizing the data within a particular range, the scaling is performed during the pre-processing stage of the data analysis to the variables.

The given dataset might consist of multiple variables, sometimes they are not associated with each other and for their analysis it must be brought into same scale. Hence the scaling is performed on dataset. The scaling on data only affects the coefficient of variables it does not affect the t-statistic, F-statistic, p-values, R-squared parameters. There are many Standardization, Normalization/Min-Max Scaling, Binarizing, etc.

In the normalization scaling, data is converted in range of 0 to 1, with respect to the minimum and maximum values. The normalization is calculated based on the following formula:

$$\text{MinMax Scaling (x)} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The standardization replaces the values by their z score. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardization (x)} = \frac{x - \text{mean}(x)}{\text{std. dev}(x)}$$

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

**Ans:**

The variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. VIF measures the correlation between one predictor and other predictor variable in model, which can be determined by following formula:

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

In mathematics, any division is said to be infinity when its denominator is equal to zero. In above equation the denominator will be zero when the value of $R^2$ is equals to 1, which means the dependent and independent variable in the model is identical and perfect correlation exist between them. So the main reason for VIF is infinity is the strong correlation between the variables.

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

**Ans:**

The quantile-quantile (Q-Q) plot is a graphical method to determine whether the two-sample dataset belongs to the same population dataset or not. In the Q-Q plot the quantile means the fraction of the dataset given below the particular values. For the reference purpose and line with the $45^O$ is plotted against the quantile datasets, if these two samples are from same population then these points are lies along with this line.

Q-Q plot is used to determine the two samples are from the same population or not, also to check whether two samples have the same distribution shape. Whether two samples have common location behaviour.

Q-Q plot is drawn to get the general understanding of the data, if the datapoints on the graph is lies exactly on $45^O$ line which indicates data points are from same population. If the upper end of the datapoints in QQ plot is deviates from line in leftward direction which indicates the dataset is skewed to the right side. Similarly, if the bottom end of the dataset is lies in right side of the dataset and top portion follows the $45^O$ line which indicates dataset is skewed to left.