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?

- The categorical variables available in the assignment are “season”, “workingday”, “weathersit”, “weekday”, “yr”, “holiday”, and “mnth”.
- “season” –
  - Based on the data available, the most favourable seasons for biking are summer and fall followed by Winter and very less bike rentals happened in Spring.
  - Spring has significant low consumption ratio.
- “workingday” –
  - Working day represents weekday and weekend/holiday information.
  - The count of the Bike rentals is more on weekdays as compared to weekends. So some offers can be given on weekends to attract more customers.
- “weathersit” –
  - Most bike rentals took place in the clear weather.
  - Very less Bike rentals happened during Light snow and Mist + Cloudy weather.
- “weekday” –
  - Saturday, Wednesday and Thursday are the days where more bikes are rented
- “yr” –
  - 2 years data is available and the increase in the bikes has increased from 2018 to 2019.
- “holiday” –
  - More Bikes were rented on holidays
- “mnth” –
  - Bike rental rates are the most in September and October

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

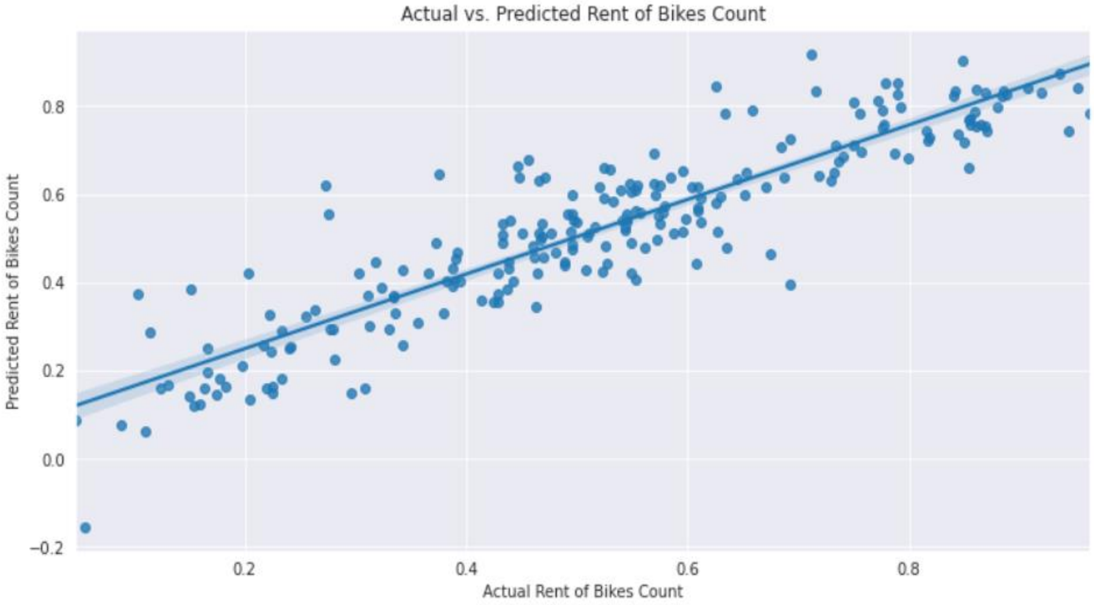
Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and other one is semi_furnished, then It is obvious that the 3rd one is unfurnished. So we do not need 3rd variable to identify the unfurnished.

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

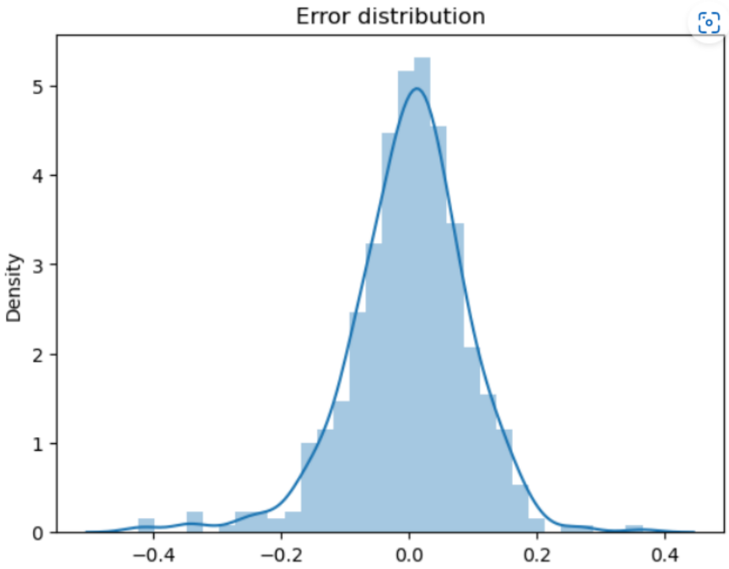
- We can see that temperature variable is having the highest coefficient 0.5398, which means if the temperature increases by one unit the number of bike rentals increases by 0.5398 units. “atemp” is the derived parameter from temp hence not considering it as it is eliminated in the model preparation.

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

1. Linear relationship between independent and dependent variables – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



2. Error terms are normally distributed: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.



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

The top 3 variables are:

‘temp’ :

Temperature is the Most Significant Feature which affects the Business positively, Whereas the other Environmental condition such as Light snow or Mist + Cloudy affects the Business negatively**.**

‘Yr’:

More Bike rentals were seen in 2019 as compared to 2018.

‘season’:

Winter season is playing the crucial role in the demand of shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.

The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

Machine learning models can be classified into following two categories on the basis of learning algorithm:

Supervised learning method: Past data with labels is available to build the model

- Regression: The output variable is continuous in nature
- Classification: The output variable is categorical in nature

Unsupervised learning method: Past data with labels are not available

Clustering: No pre-defined notion of labels is there

Past data set is divided into two parts during supervised learning method:

Training data is used for the model to learn during modelling

Testing data is used by the trained model for prediction and model evaluation

Linear regression models can be classified into two types depending upon the number of independent variables:

Simple linear regression: When the number of independent variables is 1

Multiple linear regression: When the number of independent variables is more than 1

The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:

->Differentiation

->Gradient descent method

->The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

->RSS: Residual Sum of Squares

->TSS: Total Sum of Squares

When one variable might not be enough

A lot of variance isn't explained by just one feature

Inaccurate predictions

Formulation of MLR: MLR helps us to understand how much will the dependent variable change when we change the independent variables.

New considerations to be made when moving from SLR to MLR

Overfitting - When the model becomes complex and gives very good results in training data and fails in the testing data.

Multicollinearity - To identify if there is any dependency within the pool of independent variables to remove redundancy.

Feature selection - Out of the pool of many features what features are considered to be most important. We drop the redundant features and those features that are not helpful in prediction.

Dealing with categorical variables

Dummy variables - USed when there are fewer levels. You learnt about it using the marital status example.

Feature Scaling

Standardisation - Method used to make sure that data is internally consistent.

MinMax scaling - Method used to make sure that data is internally consistent.

Scaling for categorical variables - Categorical variables cannot used as they are, so they are converted to numeric format.

Model Assessment and Comparison

Adjusted R-squared - The adjusted R-squared value increases only if the new term improves the model more than would be expected by chance.

AIC, BIC - Various types of criteria used for automatic feature selection

Feature Selection

Manual feature selection - A very tedious task in order to select the correct set of features.

Automated feature selection - The three step process is involved.

Select top 'n' features

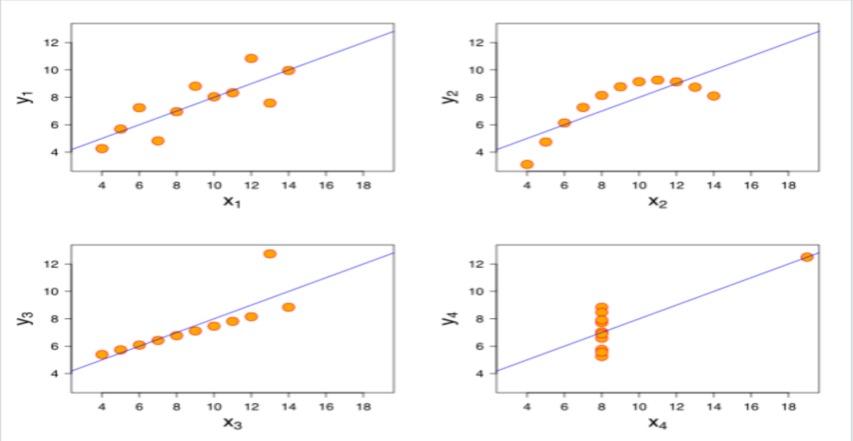
Forward/backward/Stepwise selection based on AIC

Regularization

Finding a balance between the two - A balance of both manual and automatic feature selection is required to attain the features.

Explain the Anscombe’s quartet in detail.

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

2. What is Pearson’s R?

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.

The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.. The Pearon’s R returns values between -1 and 1. The interpretation of the coefficients are:

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

$$r = \frac{n(\Sigma \mathrm{x} * \mathrm{y}) - (\Sigma \mathrm{x}) * (\Sigma \mathrm{y})}{\sqrt{[n\Sigma \mathrm{x}^2 - (\Sigma \mathrm{x})^2] * [n\Sigma \mathrm{y}^2 - (\Sigma \mathrm{y})^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx^2 = the sum of squared x scores

Σy^2 = the sum of squared y scores

3. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

- The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range. All the data should be in same scale. For eg:, we have one variable as CGPA (Values ranging from 7-10) and other variable as Salary (values ranging from 100K$-150K$). As you can see the scale of the these variables are not same. To make the variables in same scale, so that plot of the graphs will be relevant, scaling is required. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

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

The VIF formula clearly signifies when the VIF will be infinite. If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

- The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. Whenever we are interpreting a Q-Q plot, we shall concentrate on the 'y = x' line. We also call it the 45-degree line in statistics. It entails that each of our distributions has the same quantiles. In case if we witness a deviation from this line, one of the distributions could be skewed when compared to the other.
- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.