UPGRAD Subjective Questions

Name: Sankar Thulasimani

Email: sankarthulasimani@yahoo.com

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 variable in the dataset were season, weathersit, holiday, mnth, yr, weekday, and workingday. These were visualized using a boxplot and had the following effect on our dependent variable,

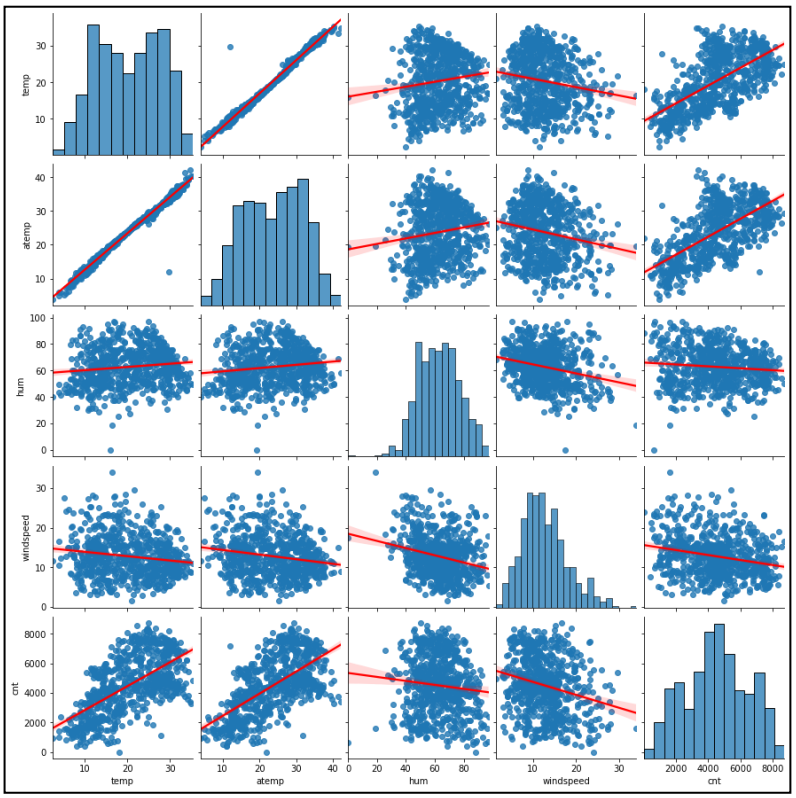
- **Season –** The boxplot depicts that fall season has the greatest number of counts followed by summer, whereas spring has very a smaller number of counts.
- **Weathersit** - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable since it is biking. Highest count was seen when the weathersit was Clear, Partly Cloudy
- **Holiday –** Count reduces slightly during holidays compared with working days
- **Mnth -** September saw highest number of counts whereas December has the lease, it is almost like the season, and they are correlated.
- **Yr** - The number of rentals in 2019 was more than 2018, showing an increasing trend.

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

If you don't drop the first column then your dummy variables will be highly correlated (redundant). This may affect some models adversely and the effect is stronger when thecardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance may be distorted. Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column.

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

“temp” and “atemp” are the two numerical variables which were highly correlated with the target variable (cnt)



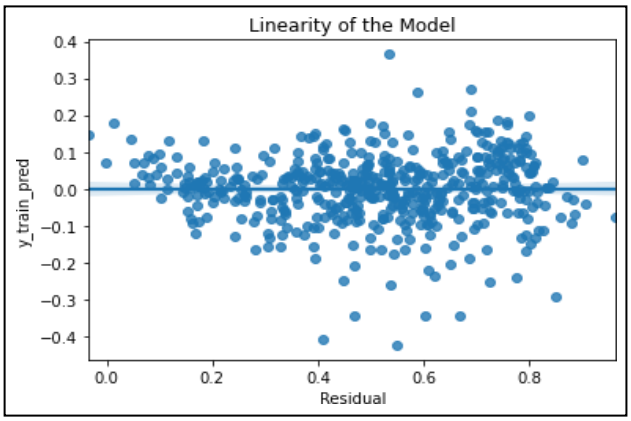
4. How did you validate the assumptions of Linear Regression after building themodel on the training set?

The following assumptions of linear regression were validated.

1. Linearity of the model
2. Homoscedacity
3. Error terms are independent of each other
4. Error terms normal distribution

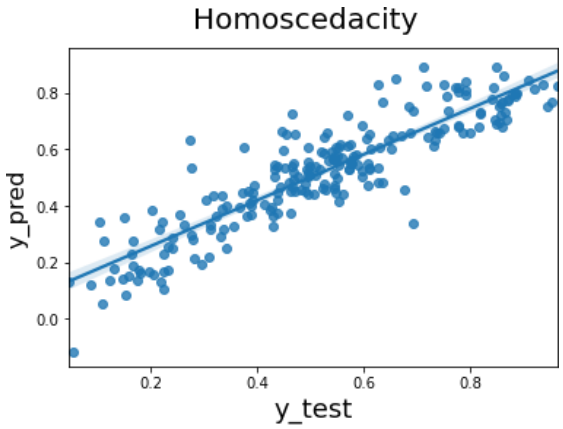
Linearity of the model:

The dependent variable (y) is assumed to be a linear function of theindependent variables (X, features) specified in the model. To detect nonlinearity one can, inspect plots of observed vs. predicted values or residuals vs.predicted values.



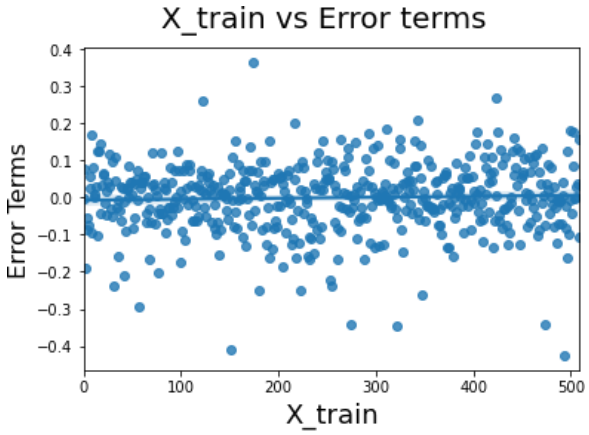
Homoscedacity:

The variance should not increase (or decrease) as the error values change. Also, the varianceshould not follow any pattern as the error terms change. It is validated in the below graph.



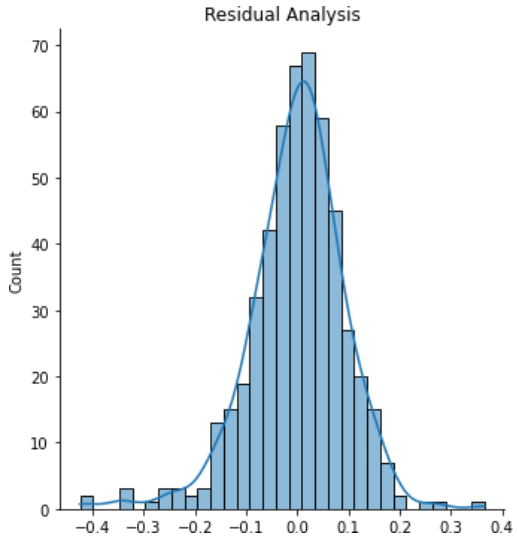
Error terms are *independent* of each other:

- The error terms should not be dependent on one another, it is validated below with residual vs x_train.



Error terms are normally distributed:

Residuals distribution should follow normal distribution and centered around 0 (mean =0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows thatthe residuals are distributed about mean = 0.



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 features are

1. **temp - coefficient: 0.539793** – it is positively correlated
2. **yr - coefficient: 0.231247** - it is positively correlated
3. **weathersit_Light Snow & Rain - coefficient: (- 0.301727)** – it is negatively correlated

General Subjective Questions:

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

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most used predictive analysis model.

Linear regression is based on the popular equation **"y = mx + c".**

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s)/independent variable(x). In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable.

Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc. Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression: SLR** is used when the dependent variable is predicted using only **one** independent variable.
2. **Multiple Linear Regression: MLR i**s used when the dependent variable is predicted using multiple independent variables.

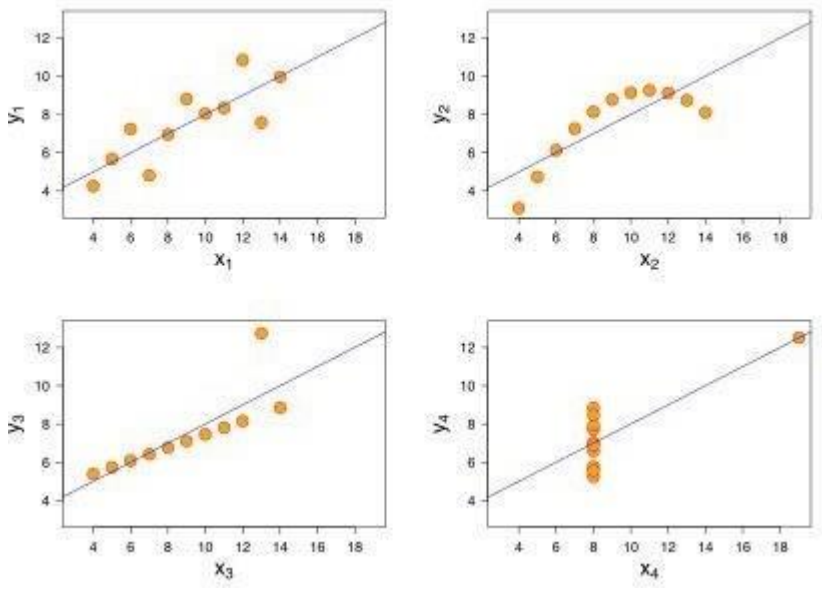
The equation for MLR will be:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\epsilon$$

β1 = coefficient for X1 variable β2 = coefficient for X2 variable and so on… **β0 is the intercept (constant term).**

2. Explain Anscombe's quartet in detail.

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distributionand look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them, it's not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line 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.

3. What is Pearson's R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. It's value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us can we draw a line graph to represent the data? r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 meansthere is no linear association.

The Pearson coefficient correlation has a high statistical significance. It looks at the relationshipbetween two variables. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help Pearson correlation coefficient calculator. This linear relationship can be positive or negative.

Example:

- **Positive linear relationship:** In most cases, universally, the income of a person increases as his/her age increases.
- **Negative linear relationship:** If the vehicle increases its speed, the time taken to travel decreases, and vice versa.

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

Feature **scaling** is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset because of different range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

- Normalization (Min max scaling) is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks. **It ranges from 0 to 1.**

  Minmaxscaler = (x - xmin)/(xmax-xmin)

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

  Standardisation = (x-mu)/sigma

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

VIF - the variance inflation factor -The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. (VIF) =1/(1-Ri^2). If there is perfect correlation, then VIF = infinity. Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables- If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1.So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" **.**

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come frompopulations 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. It is used to compare the shapes of distributions. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

The q-q plot is used to answer the following questions:

1. two data sets come from populations with a common distribution.
2. two data sets have common location and scale
3. two data sets have similar distributional shapes.
4. two data sets have similar tail behavior

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