Assignment Based Subjective Questions

Q1.) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1.)

- 1.)Bike rental demand has gone up from 2018 to 2019.Thus, year is a key parameter in the model.
- 2.)Highest bike booking were happening in Fall with a median of over 5000 bookings (for two years). It is followed by Summer & Winter. It indicates that the season can be a good predictor of the dependent variable.
- 3.)large no. of of the bike booking were happening in the months' May to Sep with a median of over 4000 bookings per month(for two years). It indicates that the mnth has some trend for bookings and can be a good predictor for the dependent variable.
- 4.)Most of the bike booking was happening during Clear weather with a median of close to 5000 bookings (for two years). This was followed by Misty weather. It indicates that the weathersit does show some trend towards the bike bookings, and it can be a good predictor for the dependent variable.
- 5.)Further analysis would be needed to determine whether working day and weekday attributes need to be included in the model parameter selection.they seem not so significant.
- 6.)Most of the bike rentals are happening during non-holiday time.

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

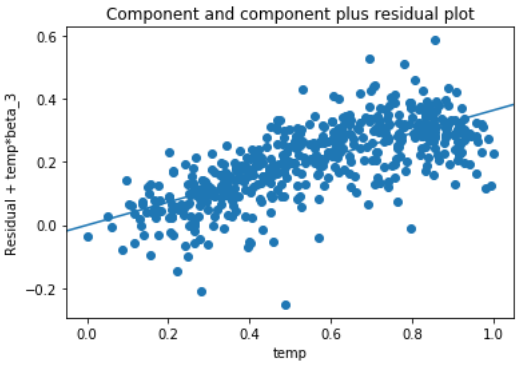
A2.) Drop_first =True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

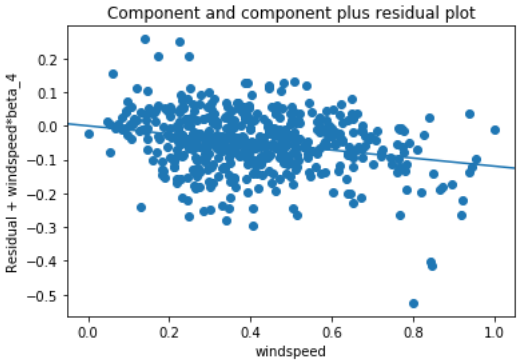
Q3.) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3.) 'Registered' which denotes the no of bike registrations is the highest correlated variable to the target variable

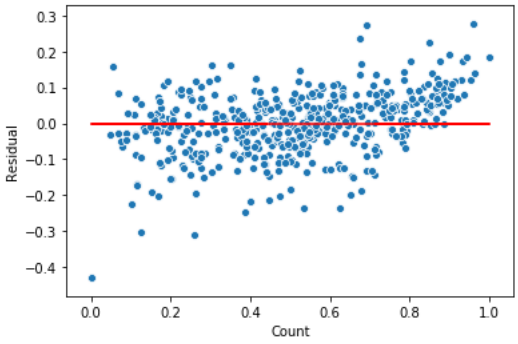
Q 4) How did you validate the assumptions of Linear Regression after building the model on the training set?

A4.)





1.)The above plots represents the linear relationship between the model and the predictor variables.

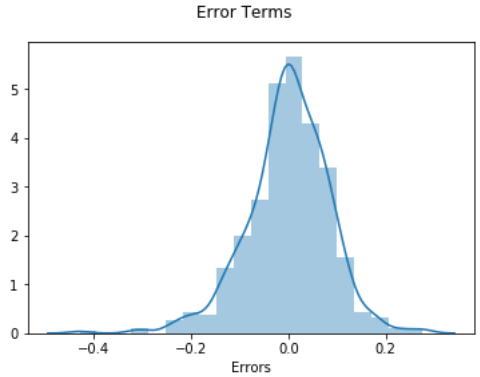


2.) From the above plot we observe that there is no visible pattern in residual values, thus homoscedasticity is well preserved.

```
In [40]: print('The Durbin-Watson value for Final Model lr is',round(sm.stats.stattools.durbin_watson((y_train - y_train_pred)),4))

The Durbin-Watson value for Final Model lr is 2.0632
```

3.) Autocorrelation refers to the fact that observations' errors are correlated. To verify that the observations are not auto-correlated, I used the Durbin-Watson test. The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables. As the value comes out to be 2.0632 there is almost no autocorrelation.



4.) Based on the histogram, we can conclude that error terms are following a normal distribution with mean at zero.

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

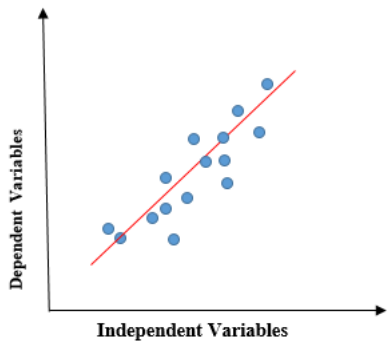
A 5.) As per final model the top 3 predictor variables contributing significantly towards explaining the demand of the bikes are:

- Temp(temperature): temp variable has a coefficient value of 0.3639 which makes it one of the top 3 predictor variables.
- RainSnow: With a negative coefficient of 0.2723 Rainsnow becomes one the significant variables in our model as it tells us that rain and snow deter people from renting bikes.
- Year: Year having a positive coefficient of 0.2367 becomes the third most significant variable as in 2019 there were more sales compared to previous year.

General Subjective Questions

Q1.) Explain the linear regression algorithm in detail.

A1.) Linear regression is a simple and quiet statistical regression method for determining the relationship between continuous variables. Linear regression, as the name implies, depicts the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis). When there is only one input variable (x), the linear regression is referred to as simple linear regression. When there is more than one input variable, this type of linear regression is referred to as multiple linear regression. The linear regression model produces a sloped straight line that describes the relationship between variables.



The graph above depicts the linear relationship between the dependent and independent variables. When the value of x (independent variable) rises, so does the value of y

(dependent variable). The red line is known as the best-fit straight line. Based on the data points provided, we attempt to plot a line that best models the points.

A traditional slope-intercept form is used to calculate best-fit line linear regression.

$$y = mx+b \implies y = a_0+a_1x$$

y= Dependent Variable.

x= Independent Variable.

a0= intercept of the line.

a1 = Linear regression coefficient.

Need of a Linear regression

As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable. Let's understand this with an easy example:

Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

Q2.) Explain the Anscombe's quartet in detail.

A2.) Anscombe's Quartet is a collection of four data sets that are nearly identical in simple descriptive statistics, but have some peculiarities that fool the regression model if built. They have very different distributions and show up differently on scatter plots.

It was built in 1973 by statistician Francis Anscombe to demonstrate the significance of plotting graphs before analysing and modelling, as well as the impact of other observations on statistical properties.

There are four data set plots that have nearly identical statistical observations and provide the same statistical information, which includes the variance and mean of all x,y points in all four datasets.

This emphasises the importance of visualising the data before applying various algorithms to build models out of it, implying that the data features must be plotted in order to see the distribution of the samples, which can help you identify the various anomalies present in the data such as outliers, diversity of the data, linear separability of the data, and so on. Furthermore, Linear Regression can only be considered a fit for data with linear relationships and is incapable of dealing with any other type of dataset.

This demonstrates the significance of visualising data before applying various algorithms available.

The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

Q3.) What is Pearson's R?

A3.) The Pearson's Correlation Coefficient is also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation in statistics. It's a statistic

that calculates the linear relationship between two variables. It, like all correlations, has a numerical value between -1.0 and +1.0.

When we talk about correlation in statistics, we usually refer to Pearson's correlation coefficient. However, it is incapable of capturing nonlinear relationships between two variables and of distinguishing between dependent and independent variables.

Pearson's correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

Q4.) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4.) It is a data Pre-Processing step that is applied to independent variables in order to normalise the data within a specific range. It also aids in the speeding up of algorithm calculations.

Most of the time, the collected data set contains features with widely disparate magnitudes, units, and ranges. If scaling is not performed, the algorithm only considers magnitude rather than units, resulting in incorrect modelling. To solve this problem, we must scale all of the variables to the same magnitude level.

It is important to note that scaling has no effect on the other parameters such as t-statistic, F-statistic, p-values, R-squared, and so on.

Normalization/Min-Max Scaling:

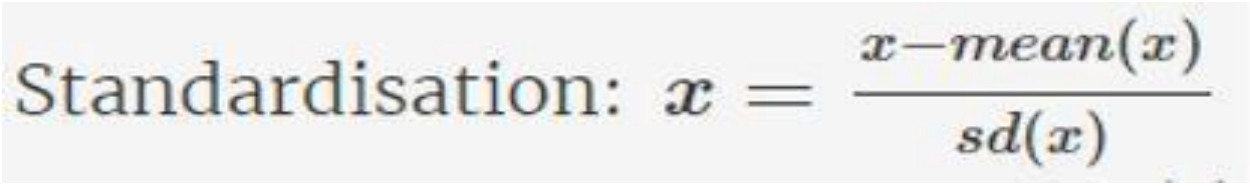
- It brings all of the data in the range of 0 and

1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

- 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**).



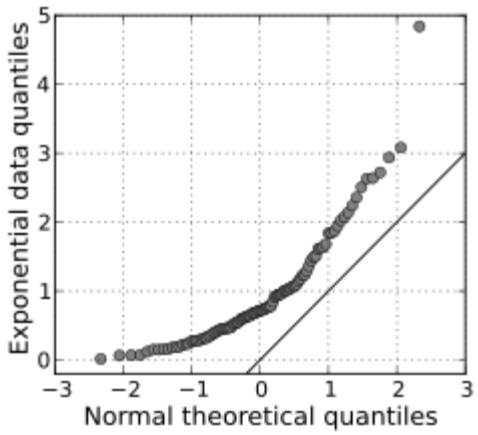
- sklearn.preprocessing.scale helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Q5.) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5.) VIF = infinity if there is perfect correlation. This demonstrates that two independent variables have a perfect correlation. We get R2 = 1 in the event of perfect correlation, which leads to 1/(1-R2) infinite. To overcome this issue, we must remove one of the factors that is producing the perfect multicollinearity from the dataset. An infinite VIF value suggests that a linear combination of other variables may exactly express the related variable (which show an infinite VIF as well).

Q6.) . What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6.) Q-Q plots (Quantile-Quantile plots) are plots that compare two quantiles. A quantile is a percentage of the population in which specific values fall below it. The median, for example, is a quantile where 50% of the data falls below it and 50% of the data falls above it. Q Q plots are used to determine whether two sets of data are from the same distribution. On the Q Q plot, a 45 degree angle is drawn; if the two data sets are from the same distribution, the points will fall on that reference line.
A Q Q plot showing the 45 degree reference line:



The points in the Q–Q plot will roughly lie on the line y = x if the two distributions being compared are similar. The points in the Q–Q plot will roughly lie on a line if the distributions are linearly connected, but not necessarily on the line y = x. Q–Q plots can also be used to estimate parameters in a location-scale family of distributions graphically.A Q–Q plot is used to compare the morphologies of two distributions, offering a graphical representation of how features like location, scale, and skewness differ in the two distributions.