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? (3 marks)

Answer:

Month:
The lowest booking counts happens in start of year during JAN which starts increases to max in July and Sept and then again decrease till December

Season: Maximum Bookings happened in Summer and during Fall which gradually start decreasing from Winter to Spring
Weekday: there is no significant changes can be observed across the working
Weathersit: Less bookins happen in light_snow, Most of the bookings happen in Misty and Clear weather
Holiday: on holidays the bookings were not like as expected

workingday: More than 4000+ booking happened on working day. workingday can be a good predictor for the dependent variable

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

Answer:

- 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.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the targetvariable? (1 mark)

Answer:
temp and atemp
temp and count
atemp and count

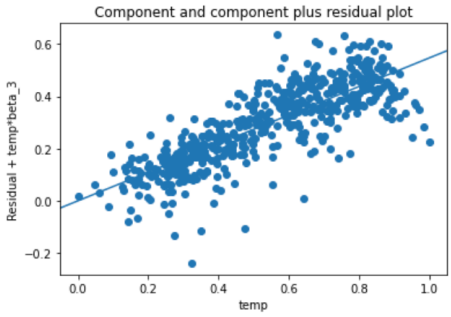
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

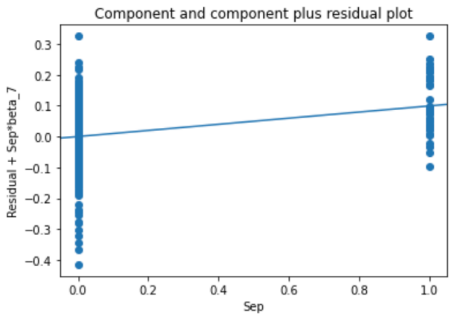
Answer:

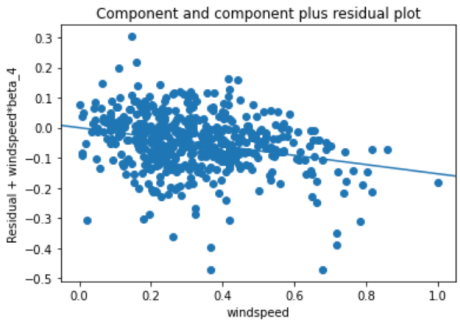
The way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds

We can check the linearity of the data by looking at the Residual vs Fitted plot. Ideally, this plot would not have a pattern where the red line (lowes smoother) is approximately horizontal at zero.
We can check this assumption using the Scale-Location plot.

In this plot we can see the fitted values vs the square root of the standardized residuals. Ideally, we would want to see the residual points equally spread around the red line, which would indicate constant variance.







5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:

- Weathershit
- Month
- Season

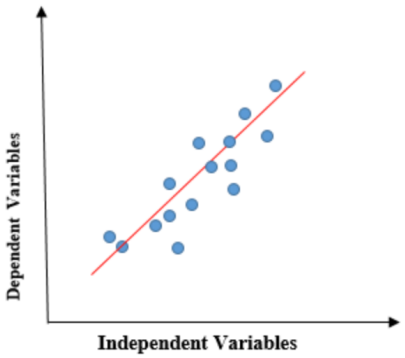
General Subjective Questions

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

   It is one of the machine learning techniques that fall under supervised learning.

Autocorrelation
This assumption made by linear regression indicates little to no autocorrelation in data. Autocorrelation takes place when residual errors are dependent on each other in one or the other way.
2. Multi-collinearity:
This assumption says that data multi-collinearity either doesn't exist at all or is present scarcely. Multi-collinearity happens when independent features or variables show some dependency.
3. Variable relationship:
The model has an assumption that there is a linear relationship between feature and response variables

linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.



To calculate best-fit line linear regression uses a traditional slope-intercept form.

$Y = \beta 0 + \beta 1 x$

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression

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

Answer:

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set. As you can see, the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plot

The purpose and importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

The four datasets can be described as:

Dataset 1: this fits the linear regression model 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

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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

Machine learning algorithms are sensitive when the data is not scaled.

There are various machine learning algorithms that use the same kind of basic strategies as their base concept under the algorithm.

These base concepts are totally based on the mapping of the distance between data points.

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
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, regardless of the unit of the values.
In the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.

So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.
Normalising typically means to transform your observations xx into f(x)f(x) (where ff is a measurable, typically continuous, function) such that they look normally distributed. Some examples of transformations for normalising data are power transformations.

Scaling simply means f(x)=cxf(x)=cx, c∈Rc∈R, this is, multiplying your observations by a constant cc which changes the scale (for example from nanometers to kilometers).

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

This shows a perfect correlation between two independent variables.

If there is perfect correlation, then VIF = infinity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile

For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
The purpose of Q Q plots is to find out if two sets of data come from the same distribution.
A 45 degree angle is plotted on the Q Q plot;
if the two data sets come from a common distribution, the points will fall on that reference line.

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

Normal Distribution:

The normal distribution (aka Gaussian Distribution/ Bell curve) is a continuous probability distribution representing distribution obtained from the randomly generated real values.
Usage: The Quantile-Quantile plot is used for the following purpose:

• Determine whether two samples are from the same population.

• Whether two samples have the same tail •

Whether two samples have the same distribution shape.

• Whether two samples have common location behavior.

Advantages of Q-Q plot:

• Since Q-Q plot is like probability plot.

So, while comparing two datasets the sample size need not to be equal. • Since we need to normalize the dataset, so we don't need to care about the dimensions of values.