Linear Regression Subjective Questions

Submitted by:

Shravan Kumar Yadav

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)

1. Yr: in 2019 the demand for shared bike is more than 2018.
2. Workingday:- The data for working day is more than that of not working day. The mean is same for cnt weather it is working day or not.
3. Weathersit:- The demand for bikes decreased as the weather become bad.
4. Holiday:- During holidays we have minimum demand and when there are no holidays then demand is higher. The mean for non-holiday is higher than holiday one.
5. Weekday:- The 25^{th} percentile and 75^{th} percentile differs for each day but the mean is same across all weekdays. The demand increase as we move ahead in a week except on 5^{th} day of week.
6. Season:- Spring has the lowest demand for bikes as compared to other seasons and fall has the highest.
7. Month:- From Jan to Oct the demand for shared bikes increases. And from Nov to Jan the demand decreases.

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

The drop_first=True syntax will drop the first dummy variable after creation. By default get_dummies will create dummy variable for each unique value of data in a column. If there are n unique values then get_dummies will create n dummy columns. But to represent n unique values in a column , we need n-1 dummy variables. So to get n-1 dummy variables we are using drop_first=True in get_dummies function.

For examples we have

Gender
Male
Female
Other

get_dummies without drop_first=True

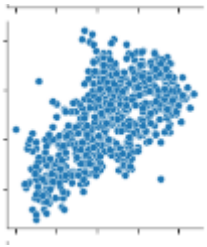
Male	Female	Other
1	0	0
0	1	0
0	0	1

get_dummies with drop_first=True. When Female and other are 0 then it is Male.

Female	Other
0	0
1	0
0	1

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

atemp has highest correlation with cnt (target variable). Below is the graph showing correlation.



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

We will observer the residual or errors. Residual or errors tells us deviation of predicted value (obtained from model) from actual value in data set. We plot the residuals/errors on distribution plot using sns.distplot() and observe that residual are normally distributed around 0 and sum of all residuals or errors is 0. For this bike sharing data we got below residual plot.

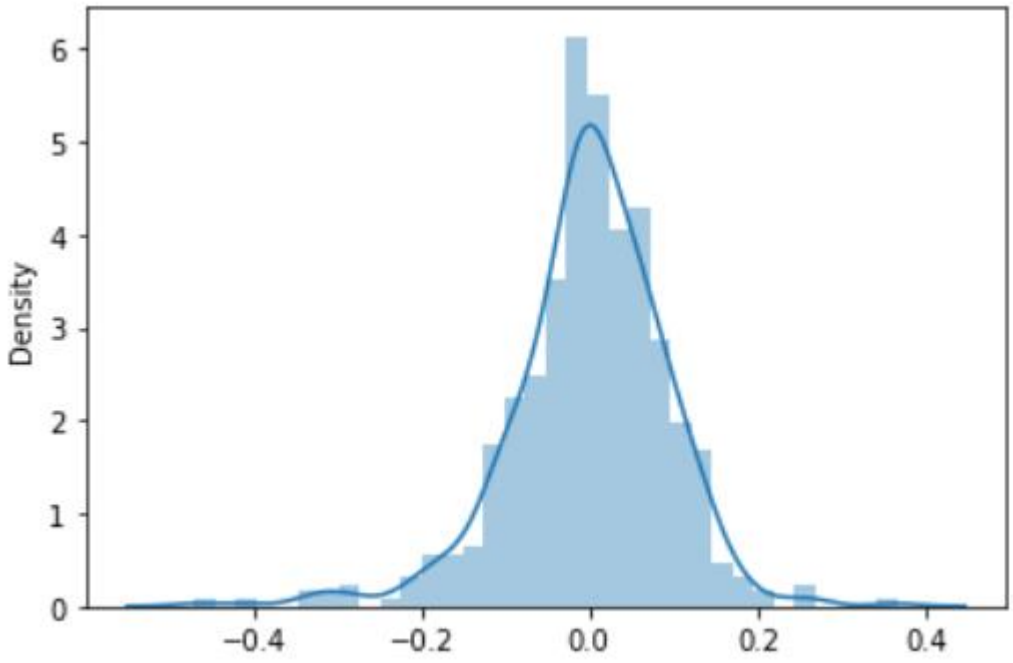
```
X_train_rfe1=sm.add_constant(X_train_rfe1)

y_train_pred=lr.predict(X_train_rfe1)

res=y_train-y_train_pred

sns.distplot(res)

plt.show()
```



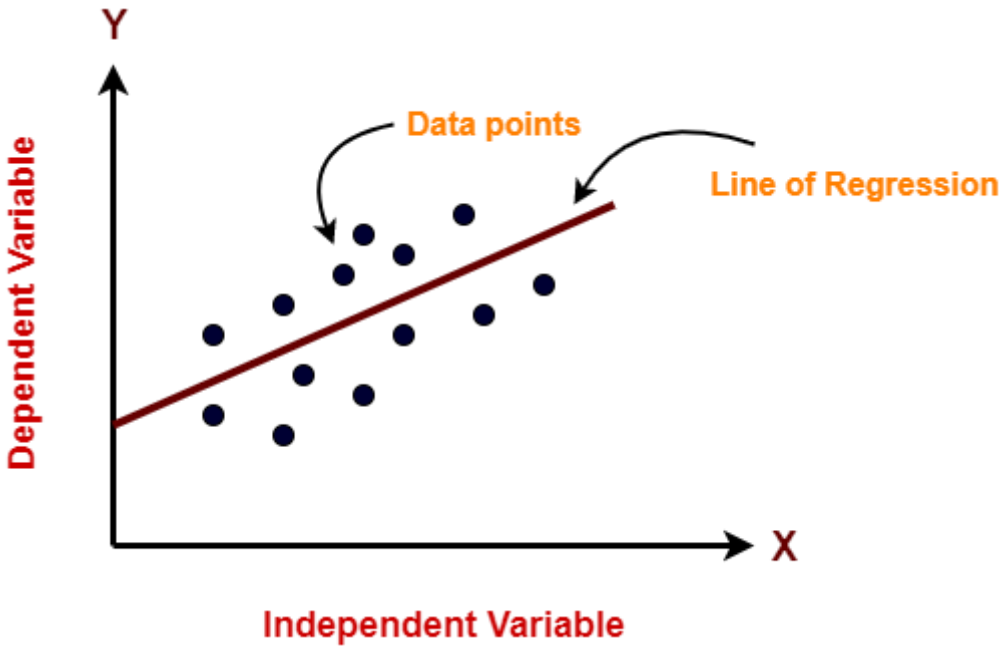
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. atemp: feeling temperature in Celsius. With one unit increase in atemp , the demand of bikes will increase by 0.36.
2. weathersit: Weather has negative impact on demand of bikes specially for below 2.
   1. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
   2. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered cloud
3. yr: year has positive correlation demand of bikes.

General Subjective Questions

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

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables.



Based on the number of independent variables, Linear regression algorithm divided into

1. Simple Linear Regression
   When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression/ Univariate Linear regression.
   The mathematical equation can be given as

   $$y = \beta_0 + \beta_1 * X$$

   Where y is the response or the target variable , X is the independent feature

β1 is the coefficient of X

β0 is the intercept

β0 and β1 are the model coefficients (or weights). To create a model, we must "learn" the values of these coefficients. And once we have the value of these coefficients, we can use the model to predict the target variable.

Let's suppose we have a dataset which contains information about the relationship between 'a number of hours studied' and 'marks obtained'. Many students have been observed and their hours of study and grade are recorded. This will be our training data. The goal is to design a model that can predict marks if given the number of hours studied. Using the training data, a regression line is obtained which will give the minimum error. This linear equation is then used for any new data. That is, if we give the number of hours studied by a student as an input, our model should predict their mark with minimum error.

2. Multi Linear Regression
   Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
   The ideal equation of Multi linear regression is

   $$y= \beta_0+\beta_1*X_1 + \beta_2*X_2 + \beta_3*X_3 ..... + \beta_n*X_n + €$$

   Model now fits a hyperplane instead of line.

Measuring strength of Linear regression

1. Coefficient of determination or R^2
   R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.
   Mathematically R^2 is represented as :-

   $$R^2 = 1 - \frac{RSS}{TSS}$$

   Where

   RSS=residual sum of square=$\sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$

   TSS=Sum of errors of the data from mean=$\sum_{i=1}^{n} \left(y_j - \bar{y}\right)^2$
2. Residual Standard Error (RSE)

   $$RSE = \sqrt{\frac{RSS}{df}}$$

   Where df=n-2 where n=number of data-points

Assumption in Linear regression

1. Linear relationship between dependent and independent variables.
2. Error terms are normally distributed
3. Error terms are independent of each other
4. Error terms have constant variance

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

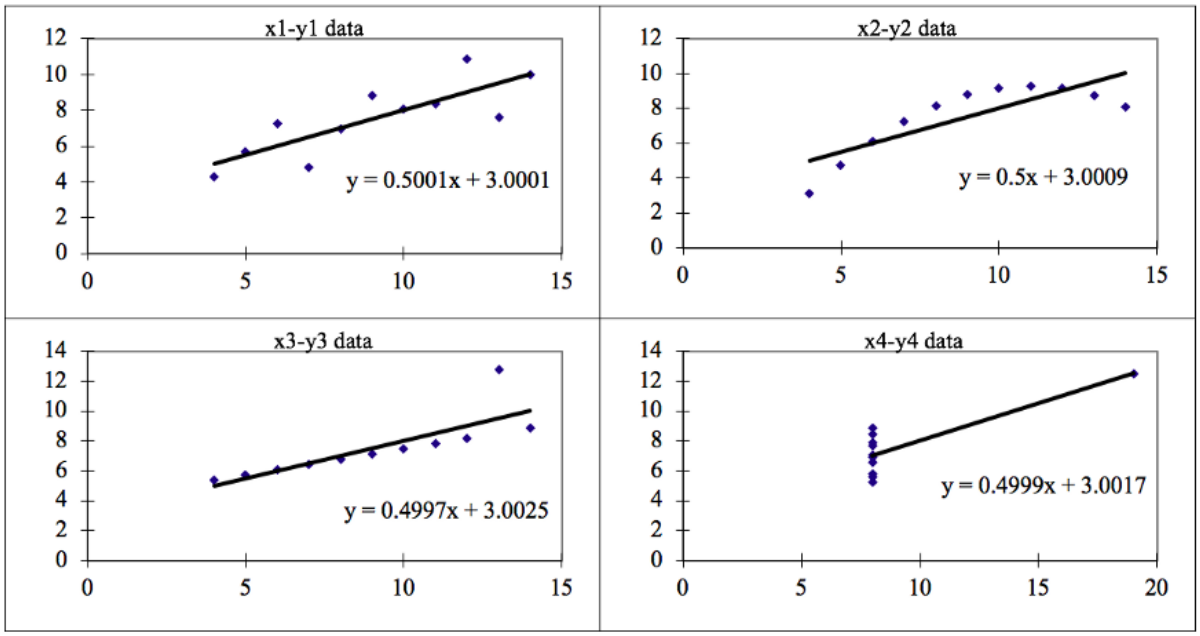
The four datasets can be described as:

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)

It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviation. Thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\Sigma(y_i-\bar{y})^2}}$$

r = correlation coefficient

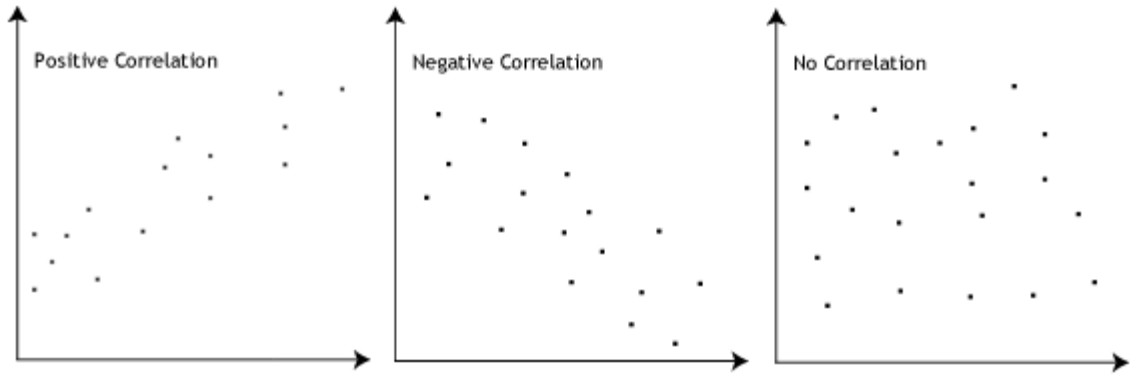
x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units. This results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. That's why scaling is performed.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

1. Normalisation/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.

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

2. Standardisation 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).

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

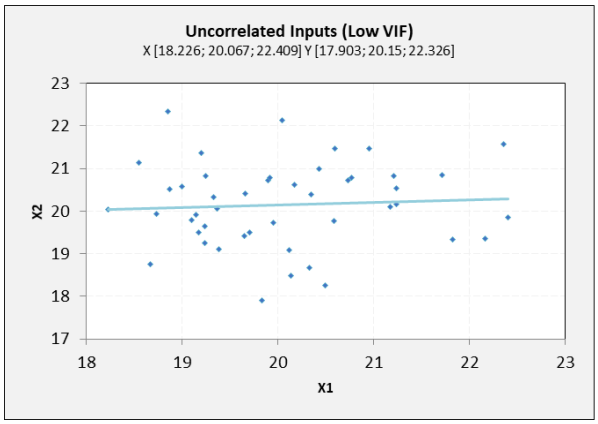
    - sklearn.preprocessing.scale helps to implement standardization in python.
    - One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

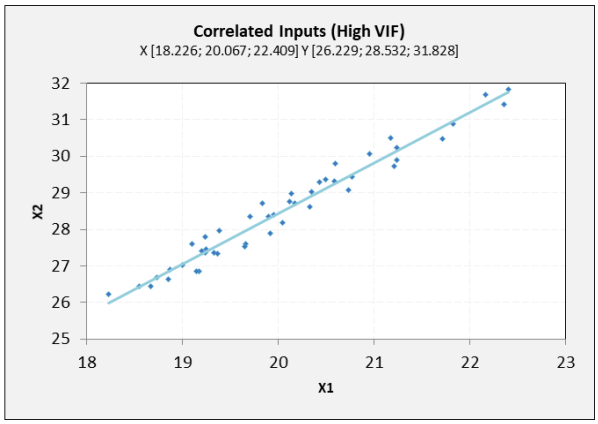
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well)





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

(3 marks)

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It 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. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.