Linear Regression Assignment

- Ajit M

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 : Plot of categorical variables against the target variables we derive the following,

- Fall has highest count in the Seasons.
- Demand for the 2nd year has grown.
- Middle of the Year demand is high (Month-wise).
- Mean of holidays is high while the non-holidays have less mean which states that on holidays generally the mean is high.
- On weekday basis there is not much difference, each day the demand is approximately same.
- On excellent weather the demand is pretty high than on other weather conditions.

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

Answer:

Dummy variable creation is done for categorical variables where each category is treated as a separate column and is used in binary type. So, for 2 power n categories we first need n variables.

But if we take as n then it will be a redundant column producing multicollinearity because in n columns say 1st column can be easily represented by the rest n-1 columns, thus to reduce columns and to escape from multicollinearity issue we need to drop one column of the n columns. World wide its commonly used to drop first row and hence drop_first = True is required.

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

Answer:

The variable temp or atemp both have maximum correlation with target variable(cnt).

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

Answer:

Post checking the below items, concluded that this is linear regression model,

a. Error terms belong to normal distribution with mean at 0.
b. Error does not have any specific patten.
c. The model seems to be linearly related (target variables and the other variables).
d. R-squared and adjusted R squared are nearly same and no overfitting.
e. Multicollinearity check using VIF method.

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:

Top 3 variables highly related with target column are

a. Atemp
b. Year
c. Weathersit_excellent

General Subjective Questions:

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

Answer:

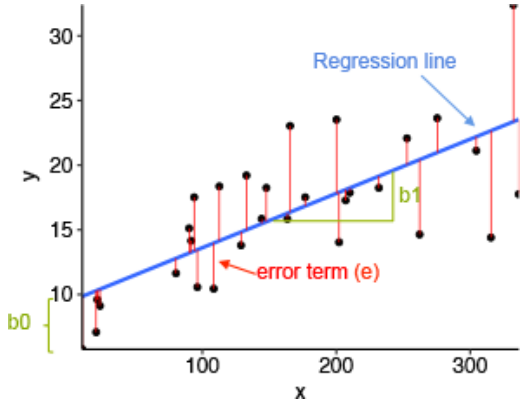
Linear regression, part of regression analysis, is a machine learning algorithm based on supervised learning. It's a predictive modelling technique, through which we find the relationship between input and the target variable.

Linear regression is one of the very basic forms of machine learning in the field of data science where we train a model to predict the behaviour of data based on some selected variables. The name linear states that always the data in x axis and y axis are linearly related to each other.

Consider an example where the sales of electric bike increase as the fuels price increases and as the maintenance cost of e-bikes goes down. So here the count of units sold directly depends on the maintenance cost of e-bikes. Thus, it has a linear relationship but a negative relationship. But if you consider the count of units sold vs the fuel price it has a positive linear relationship. Using this relationship and having in hand the previous data, we can assume how the fuel price increases and how far we can reduce the maintenance cost so that we will be able to increase the sales, and get the teams be prepared to supply the demand. This is to predict the value with historical data patterns behaviour.

Mathematically we can write as y = b0 + b1*x

Where y is the predicted variable, b1 is the slope of the line, x is independent variable, b0 is intercept or constant. It is the cost function which helps to find the best value for m & c providing best fit line for the data points.

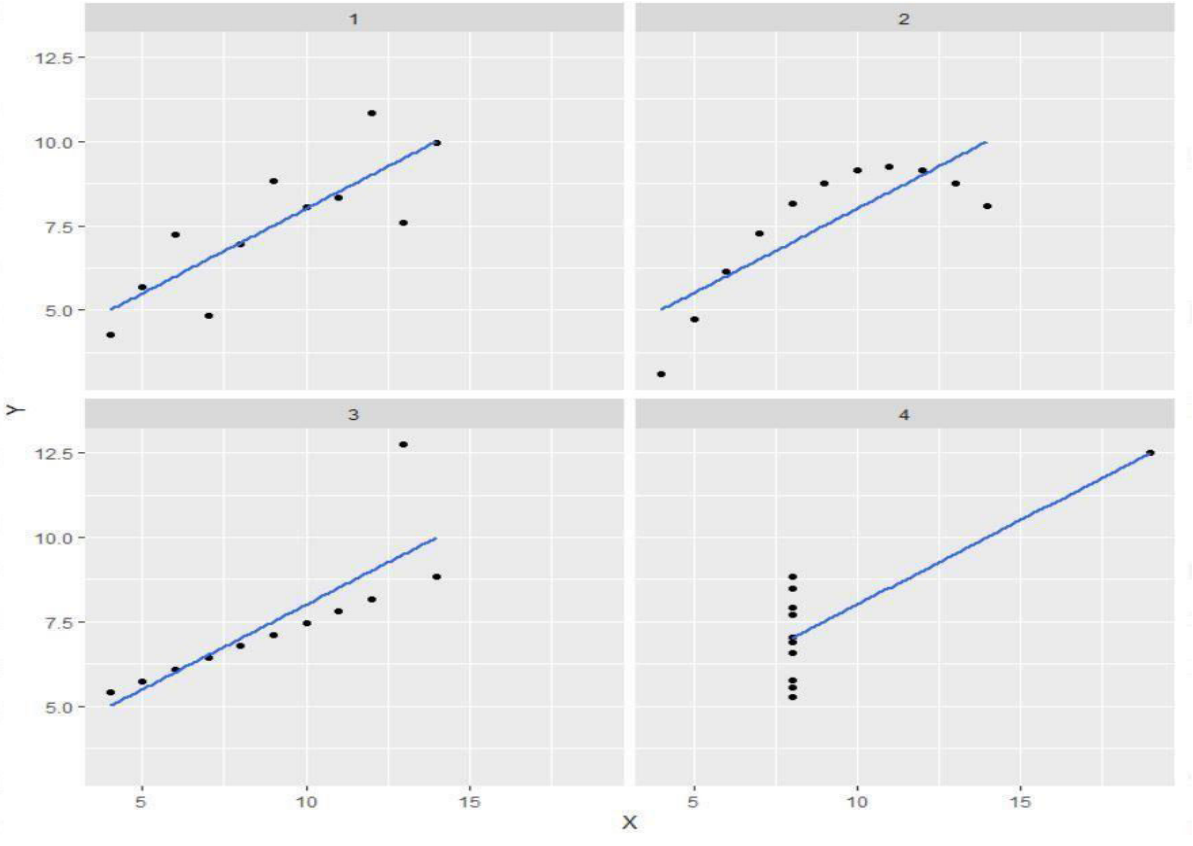


Here, x and y are two variables on the regression line.
b1 = Slope of the line.
b0 = y-intercept of the line.
x = Independent variable from dataset
y = Dependent variable from dataset

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

Answer:

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet 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 before analysing it and the effect of outliers on statistical properties.



- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

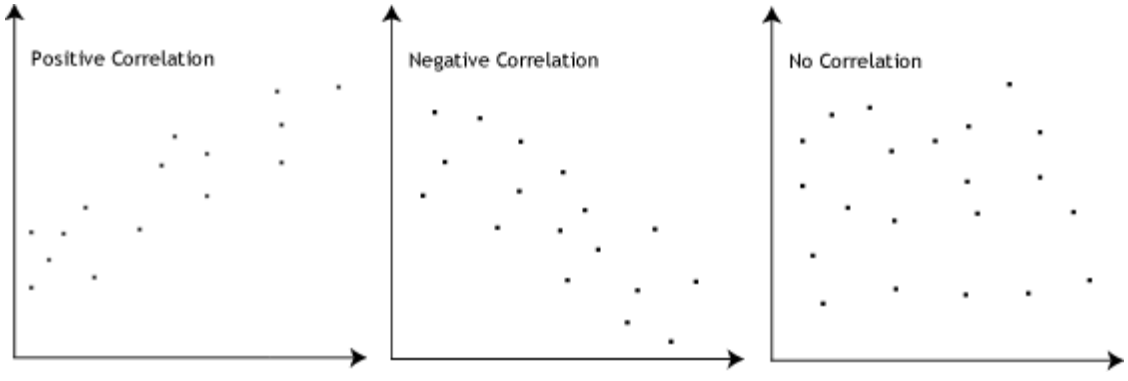
3. What is Pearson's R? (3 marks)

Answer:

Pearson's R is the most common way of measuring a linear correlation. It is a number between -1 and 1 that measures the strength and direction od relationship between two variables. If data lie on a perfect straight line with negative slope, then r = -1.

It is also known as Bivariate corelation, PPMC (Pearson product-moment correlation coefficient), Correlation coefficient.

It summarises the statistics of dataset, specifically, it describes the strength and direction of the linear relationship between two quotative variables.



Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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

Answer:

Scaling is a method of standardization that's most useful when working with a dataset that contains continuous features that are on different scales, and you're using a model that operates in some sort of linear space. Feature scaling transforms the features in your dataset so they have a mean of zero and a variance of one This will make it easier to linearly compare features. Also, this a requirement for many models in scikit-learn.

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.

Example Weight of a device = 500 grams, and weight of another device is 5 kg. In this example machine learning algorithm will consider 500 as greater value which is not the case. And it will do wrong prediction. Machine Learning algorithm works on numbers not units. So, before regression on a dataset it is a necessary step to perform.

Scaling can be performed in two ways: Normalization: It scale a variable in range 0 and 1. Standardization: It transforms data to have a mean of 0 and standard deviation of 1.

Normalization is good to use when distribution does not follow gaussian distribution, while on other hand the standardization is used when it is gaussian distributed data. Unlike normalization, standardization does not have bounding range so even if we have outliers, it will not be affected by standardization.

Normalised scaling (or) Min-Max Scaling

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

Standardized scaling

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation

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

Answer:

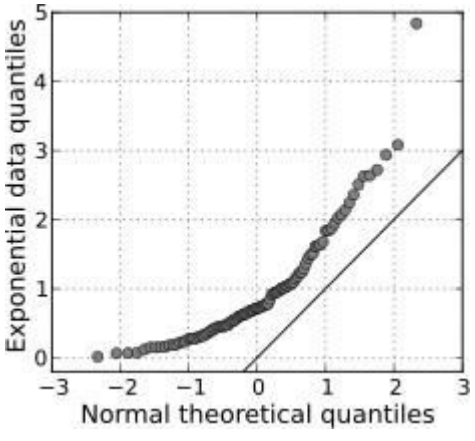
If there exist a excellent correlation, then VIF is infinite. Here we get R2 = 1, which leads to 1/(1-R2) infinity. This is always an issue and so we need to solve this by dropping one variable from the dataset which is causing multicollinearity.

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

Answer:

Quantile – Quantile (Q-Q) plot is a method to analyse the set of data if it theoretically contributes to normal, exponential or uniform distribution. Also, it helps to determine if those sets represent the same population of common distribution.

if the two data sets come from a common distribution, the points will fall on that reference line.
A Q-Q plot showing the 45-degree reference line:



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.
A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Few advantages:

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:
If two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour.