

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)

Ans:

Following are the observation about the dependent Variable 'cnt' in relation to Categorical variables:

- i) 'Season' - There is a hike in good demand in summer and fall season
Bike demand takes a dip in spring season.
- ii) 'Yr' - Bike demand in year 2019 is higher as compared to 2018.
- iii) 'Month' - Increase in demand from month of May till the month October
- iv) 'Weathersit' - we can clearly see a drop in demand in the snow weather.
- v) 'Holiday' - There is increase in demand on holidays
- vi) 'Weekday' - The demand of bike is almost similar throughout the weekdays.
- vii) 'Workingday' - Bike demand doesn't change whether day is working day or not

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

Ans:

It is important to use drop_first=True during dummy variable creation as it helps in reducing the extra column created during dummy variable creation.

Hence it reduces the correlations created among dummy variables.

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

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

Ans:

atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables

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

Ans:

The linearity assumption can best be tested **with scatter plots** between dependent variable and Predictors. The linear Regression equation would be in format - $E[Y] = \beta_0 + \beta_1 X$

There are four assumptions associated with a linear regression model:

Linearity, Homoscedasticity, Independence, Normality.

In Bike sharing model following points we have checked on linear regression assumptions:

- i) We have drawn a scatter plot between y_test vs y_Pred.
- ii) Derived Model formula:
$$\text{Count} = 0.07887 + (\text{temp} * 0.5515) + (\text{workingday} * 0.0482) - (\text{windspeed} * 0.1839) - (\text{weathersit_moderate} * 0.0677) + (\text{Season_summer} * 0.0875) + (\text{Season_winter} * 0.1174) + (\text{Year} * 0.2389) + (\text{Month_Sept} * 0.0874) + (\text{Weekday_sat} * 0.0593)$$
- iii) Derived Residual Analysis on Test and Train data sets.
Test data r^2 : 76.92
Train data r^2 : 79.1

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

Ans:

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- i) Temp (coefficient='0.5515'): Indicate a unit increase in temp variable, increases the bike demand by .5515 units
- ii) Yr (2019) (coefficient='0.2389'): Indicate a unit increase in Yr variable, increases the bike demand by 0.2389 units.
- iii) Wind speed (coefficient='-0.1839') : Indicate a unit increase in wind speed variable, decreases bike demand by 0.1839 units

General Subjective Questions

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

Ans:

Linear regression is the most basic and commonly used predictive analysis. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

As the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y=a+bx$$

Where a and b given by the formulas:

$$a(\text{intercept}) = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

$$b(\text{slope}) = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = 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)

Ans:

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 analyzing it and the effect of outliers on statistical properties.

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

| I | | II | | III | | IV | |
|------|-------|------|------|------|-------|------|-------|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

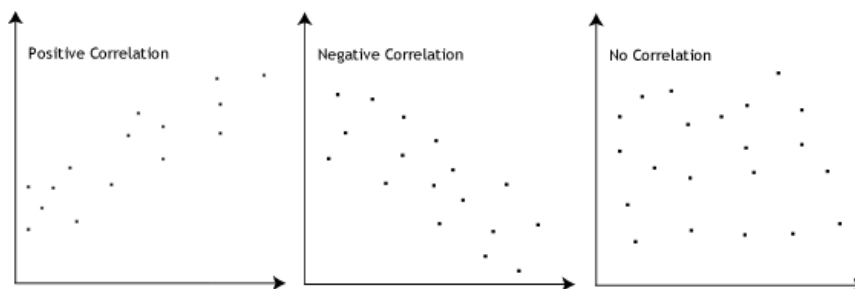
3. What is Pearson's R? (3 marks)

Ans:

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



Pearson r Formula

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

Here,

- 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

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

Ans:

It 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 hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

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.

Difference between normalized scaling and standardized scaling:

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

| S.NO. | Normalisation | Standardisation |
|-------|-------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 6. | This transformation squishes the n-dimensional data into an n-dimensional unit hypercube. | It translates the data to the mean vector of original data to the origin and squishes or expands. |
| 7. | It is useful when we don't know about the distribution | It is useful when the feature distribution is Normal or Gaussian. |
| 8. | It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |

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

Ans:

If there is perfect correlation, then $VIF = \text{infinity}$. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

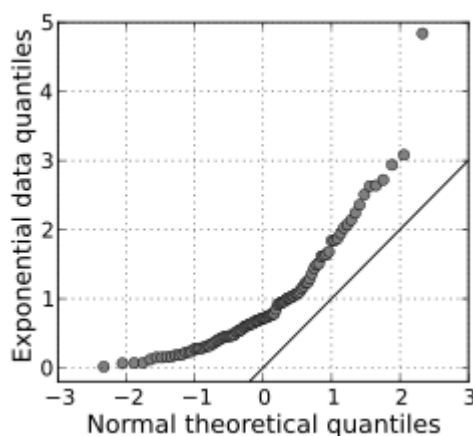
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans:

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.

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.

