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. Our dataset had 5 categorical variables, namely: yr (year), mnth(month), weathersit(weather situation), season, holiday, while our dependent variable was 'cnt'. To infer the impact of our categorical variables on 'cnt', we took help of boxplot representation of the data.

The deduction we made were like as follows:

- in terms of most profitable year for the company was 2019

- September was when people liked to go for bike sharing more than the other months

- holidays were not profitable for the bike sharing service

- people preferred calm, clear, cloudy weather over all other situations

- seasonal profits were more during fall and lowest in spring

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

Ans. Dummy variable creation will add a column, which can increase the correlation between the variables. If we have to make sure that the algorithm fits appropriately, we need to reduce the extra column by dropping it.

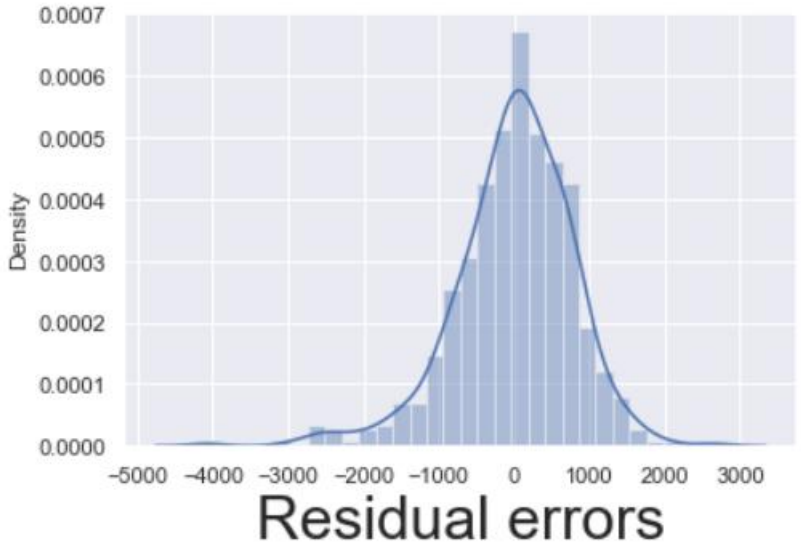
For example, if we are trying to analyse a categorical column, and the dummy variable for that variable would be able to give the same idea as the already present categorical variable, it will make the algorithm think the variables are correlated more than the actual scenario. So as a rule of thumb if there are n number of categorical variable columns, we should use n-1 columns when we are creating dummy variables.

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

Ans. The highest correlation was between 'temp' and 'atemp' and our target variable 'cnt'.

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

We prefer the residual distribution to be centred around 0 and also follow normal distribution, which we did check using a distplot . we saw that the residuals were centres around 0, therefore proving our assumptions.



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. The top 3 features are: 'atemp', 'yr', and 'season_spring'.

General Subjective Questions

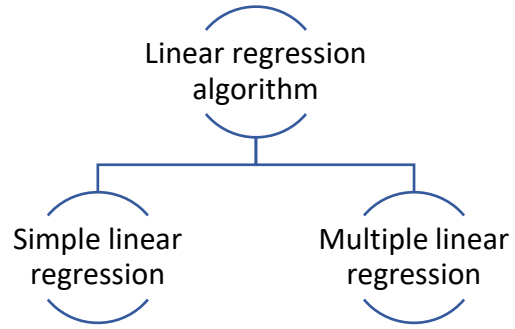
1. Explain the linear regression algorithm in detail. (4 marks)

Ans. Linear regression is one of the most popular Machine Learning algorithms, that makes predictions based on numerical variables. It basically shows a linear relation between an independent variable (x) and a dependent variable (y), hence the name. So, it is measure of how the independent variable is affecting the values of the dependent variable. Mathematically the representation of linear regression:

$$y = a_0 + a_1 x + \varepsilon$$

Here,

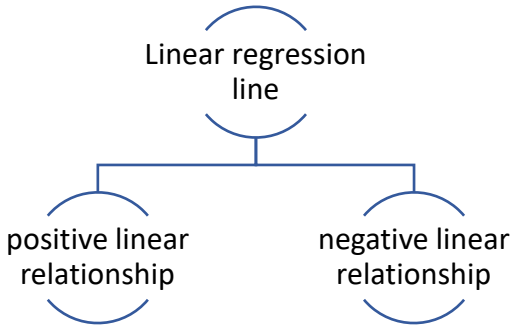
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.



SLR: there is one independent variable which is used to predict the value of the dependent variable

MLR: there is more than one independent variable which is used to predict the value of the dependent variable

Linear regression line

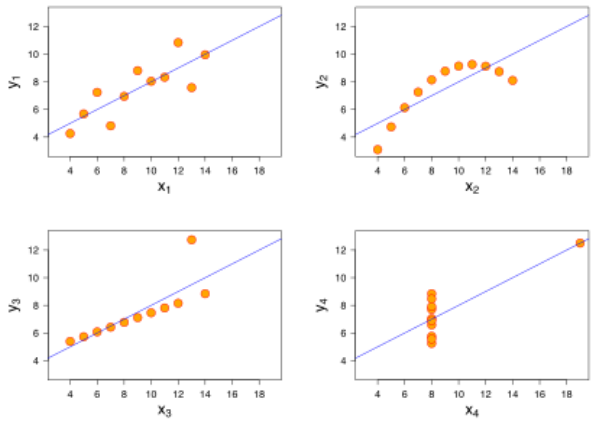


Positive linear relationship: When the dependent variable increases on Y-axis, the independent variable will also increase on X-axis.

Negative linear relationship: When the dependent variable decreases on Y-axis, the independent variable will increase on X-axis.

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

Ans. Anscombe's Quartet is a group of four datasets that appear similar when analysed by just looking at summary statistics, but when these datasets are plotted, a clear distinction between these can be seen.



As can be seen from the diagram above, there is a clear distinction in the distribution of these sets.

These datasets have 11 (x,y) values , created by statistician Francis Anscombe. And using these datasets, the following properties (that prove to be equal):

- Mean x value for all these sets is 9
- Mean y value for all these sets is 7.50
- Sample variance for all these sets,
    - for x is 11
    - for y is 4.125
- The correlation for all these sets between x and y is 0.816
- A linear regression (line of best fit) for each dataset follows the equation y = 0.500x + 3.00
- Coefficient of determination is 0.67

The first scatter plot on top left can be modelled as gaussian, the next plot on top right doesn't follow a normal distribution, the bottom left plot although has a linear distribution, a different regression line should be chosen for it, and the last graph has x as a constant apart from that one outlier.

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

Ans. Also known as Pearson correlation coefficient/correlation coefficient states the linear correlation between datasets, which is the ratio of covariance of two variables and product of standard deviations. More simply it is a normalized value of covariance which can only be between -1 and 1.

Pearson's R is calculated by this:

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Where,

r_{xy} = Pearson's R

n = sample size

x_i = sample points (i); similarly, for y_i

$\overline{x}$= sample mean; similarly, for $\overline{y}$

S_x = sample standard deviation; similarly, for S_y

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

Ans. Scaling is used to normalize the range of features in dataset. It is usually performed for normalize data during data pre-processing.

Raw data varies from dataset to dataset. This can impact some machine learning algorithms such as linear regression by hampering objective functions. It is also applied so that gradient descent converges faster.

Normalized scaling: values are moved while rescaling them making sure that the range ends up to be between 0 and 1. This is also known as min-max scaling. Normalization formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Xmin and Xmax is the minimum and maximum values respectively.

Standardized scaling: Values are usually centred around the mean with a unit standard deviation; mean of a feature becomes 0. Standardization formula:

$$X' = \frac{X - \mu}{\sigma}$$

μ Is the mean and sigma is the standard deviation.

Data with Gaussian distribution should be subjected to standardized scaling.

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

(3 marks)

Ans. VIF is a measure of how much variance of regression coefficient increases due collinearity. VIF is determined by fitting regression model between independent variables.

Larger VIF values indicate that the independent variables are correlated. It is basically inflation of coefficient of the model. If all the independent variables uncorrelated or orthogonal, VIF would be 1. In the case where R_2 score is 1, making VIF ($1/1-R_2$) or infinity. This means the independent are perfectly correlated and prompts us to drop one of the variables that 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)

Ans. Quantile-quantile plot/q-q plot is to graphically represent if the datasets have a common distribution. It is a plot of quantiles from the first dataset and another from second dataset. Quantile means a fraction/percentage of points below a threshold. For an example, 0.4/40% quantile means that 40% of the data falls under the threshold and the remaining 60% falls above the threshold.

Importance of q-q plot: datasets with differing sample sizes can be analysed with the help of q-q plot. Distributional aspects such as shifts in location, changes in symmetry etc. can be simultaneously tested.

Q-Q plots share some similarity with probability plots, just that in probability plots instead quantiles of the sample dataset is replaced with quantiles of theoretical distribution.