LINEAR REGRESSION ASSIGNMENT

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)

I have plotted the categorical variables with the target variables on bar chart and has inferred following effect on target:

1. Season: A large number of clients are coming during 'Fall' season.
2. Weather situation: large number of clients are coming during ' Clear to Partly Cloudy'.
3. Week day: Majority of clients come on 'Friday'.
4. Month: Largest number of clients were during 'September'.
5. Year: The demand of bikes has grown in the next year.

2. Why is it important to use drop first=True during dummy variable creation? (2 marks)

Drop first=True is important in order to reduce the number of columns generated due to the dummy variable creation. Hence it reduces the correlations created among dummy variables. If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with dataset as we will have constant variable(intercept) which will create issue of multicollinearity.

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

“temp” has the highest correlation with the target variable “cnt”.

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

The assumptions of the classical linear regression model are:

1. There is a linear relationship between the dependent and the independent variables.
2. There should not be any relationship among the independent variables.
3. The variance of error term is same regardless of the value of independent variable. That is there should be homoscedasticity.
4. The mean value of the error term must be zero.

To validate these assumptions first of all we will check the value of R square, adjusted R square which should be greater than 80% and the value of p-value should be less than 5%. Then we can check multicollinearity by the variance inflation factor method. Lastly, we can plot the distplot of the error term to see whether it is a bell-shaped curve of a normal distribution with mean value being zero.

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

   The regression equation obtained is:

   cnt =1328.10+ 1949.18*yr+5146.58*temp-1640.54hum-1249.41*winter+ 577.67*September-1232.35*Light Rain or Snow

   Thus, we can see that the predictor variables with highest coefficient values are ‘mnth’, ‘temp’ and ‘yr’. That is the top 3 variables contributing significantly to the target variable are the temperature, month and year.

General Subjective Questions

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

The term regression was introduced by Francis Galton. Linear regression algorithm is a supervised machine learning algorithm. Regression is a technique to measure the relationship between a dependent and independent variable. Linear regression is the regression analysis where the relation between the dependent and the independent variable is linear. The equation of a linear regression is:

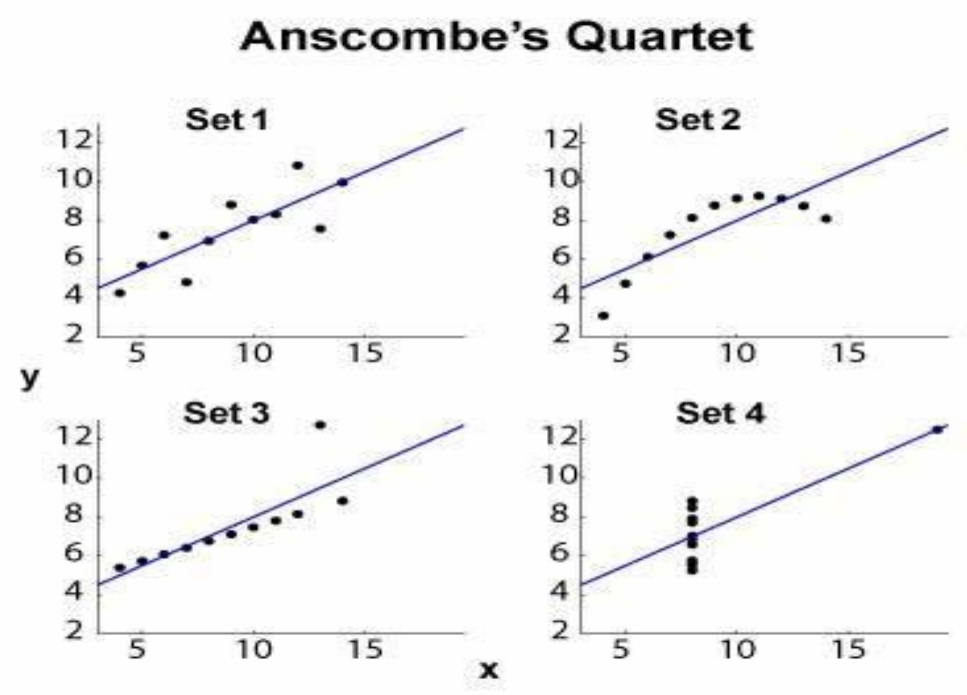
y =b0 + b1*x+u

Where y is the predicted variable (dependent variable), b1 is slope of the line, x is independent variable, b0 is intercept(constant) and u is the error term or the residual.

In python we can perform a linear regression analysis with the help of scikit library.

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. Each dataset consists of eleven (x,y) points.



The four datasets can be described as: Set 1: this fits the linear regression model pretty well. Set 2: this could not fit linear regression model on the data quite well as the data is non-linear. Set 3: shows the outliers involved in the dataset which cannot be handled by linear regression model Set 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

3. What is Pearson’s R? (3 marks)

Pearson's r is the correlation coefficient developed by Karl Pearson to find the correlation between a dependent and independent variable. The value of Pearson’s R always lies between -1 to +1. he Pearson’s R can be obtained by the formulae:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

1.If r = -1 then there is a perfect negative correlation

2.If r = 0 there is no correlation

3.If r= +1 then there is a perfect correlation

Positive correlation indicates the both the variable increase and decrease in the same direction whereas negative correlation indicates the variable increase and decrease in the opposite direction.

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

Scaling is a method to normalize the range of values of the variables. It is performed to bring all the variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values. 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. Machine Learning algorithm works on numbers not units. So, before regression is done on a dataset it is a necessary step to perform scaling. Scaling can be performed in two ways: normalization and standardisation.

Normalization: It scales a variable in range of 0 and 1.

Standardization: It transforms data by standardising the z value in order to have the features of a normal distribution that is a mean of 0 and standard deviation of 1.

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

When there is a perfect relationship then VIF = Infinity whereas if all the independent variables are orthogonal then to each other then VIF = 1.0. Means if a variable is expressed exactly by a linear combination of other variable then it is said that VIF is infinite.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution It is used for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out. Few advantages are:

a) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry.

b) presence of outliers can all be detected from this plot. It can be used with sample sizes also. 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.

Interpretation: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. Below are the possible interpretations for two data sets. a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45degree from x -axis b) Y-values < X-values: If y-quantiles are lower than the x-quantiles. c)X-values < Y-values: If x-quantiles are lower than the y-quantiles. d)Different distribution: If all point of quantiles lies away from the straight line at an angle of45 degree from x -axis.