Linear Regression Subjective Questions

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?**

Answer:

The dependent variable in the data set is Cnt (Renamed to totalCount in the notebook)

The categorical variables in the dataset are:

- Year
    1) The demand for bike increases year on year
    2) Demand increased by 24% in 2019 compared to 2018
    3) This could be related to increasing popularity
    4) It could also be a result of more conscious effort and investment made by Boom Bikes to increase demand for bike sharing
- Month
    1) The demand for bikes is generally higher in the months from May to Nov
    2) This could be linked to Summer season and Fall during the period from May to Nov
    3) Demand is lowest in Jan and Feb which could be linked to winter season and snowfall in these months
- Weekday
    1) The demand for bikes is not impacted much by days of the week
    2) Demand is almost similar on most days and there is no significant increase in demand on week days or weekends (13-15%)
    3) The average demand is the lowest on Sundays(13%) and highest on Thursday, Friday and Saturday(15%) on all these days
- Weather situation
    1) Weather conditions significantly impact demand for bike rentals
    2) We see more rentals in clear weather followed by misty weather
    3) There are low rentals in light snow and rain and absolutely no rentals during heavy snow fall and rain
- Season
    1) Seasons also have significant impact on bike rentals
    2) There are highest rentals in the fall season followed by summer and moderate rentals in winter
    3) There are lowest rentals in spring season
- Working day
    1) There is no significant impact of working days on rentals

2) Overall working days sum up to 70% of the rentals and non-working days sum up to the remaining 30% of the rentals

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

Answer:

1) Dummy variables are created for categorical variables. For a categorical variable with n levels, the data can be represented using n-1 dummy variables.
2) Consider an example of a categorical variable 'Season' which could take 4 values or has 4 levels – Spring, Summer, Rainy and Winter
3) This could be represented using 3 dummy variables as shown in the below table:

Season	Season_Summer	Season_Rainy	Season_Winter
Spring	0	0	0
Summer	1	0	0
Rainy	0	1	0
Winter	0	0	1

   - The get_dummies function in pandas creates Summer takes value 1 in summer season and 0 in others
   - Rainy takes value 1 in rainy season and 0 in others
   - Winter takes value 1 in winter season and 0 in others
   - If all the 3 variables are 0, it is Spring so we do not need any fourth dummy variable to represent Spring
4) Dummy variables equal to number of levels in the categorical variables.
5) The drop_first attribute when set to True drops the first dummy variable created to have n-1 dummy variables

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

Answer:

- Looking at the pair plot temp/atemp has the highest correlation with the target variable.
- The heatmap gives the correlation value which is 0.6

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

Answer:

The assumptions of linear regression are:

1) There should be linear relationship among the variables.
2) Independent variables should not be correlated with each other. Multi-collinearity should not be observed.
3) Error terms are normally distributed with mean zero(not X, Y)
4) Error terms are independent of each other
5) Error terms have constant variance (homoscedasticity):
    a. The variance should not increase (or decrease) as the error values change.
    b. Also, the variance should not follow any pattern as the error terms change.

Assumptions validated in the assignment are:

1) Linear relationship is observed among the variables
2) No multi-collinearity observed among the independent variables in the final model
3) Error terms are normally distributed around mean zero
4) The error terms are independent of each other with no pattern observed in their distribution is validated by using plot on the training model and scatter plot on the test model

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

Answer:

The top 3 features contributing significantly towards explaining demand of shared bikes are:

1) Year increases demand
2) Temperature increases demand
3) Weather Situation – LightSnowRain – Reduces demand

General Subjective Questions

1) Explain the linear regression algorithm in detail

Answer:

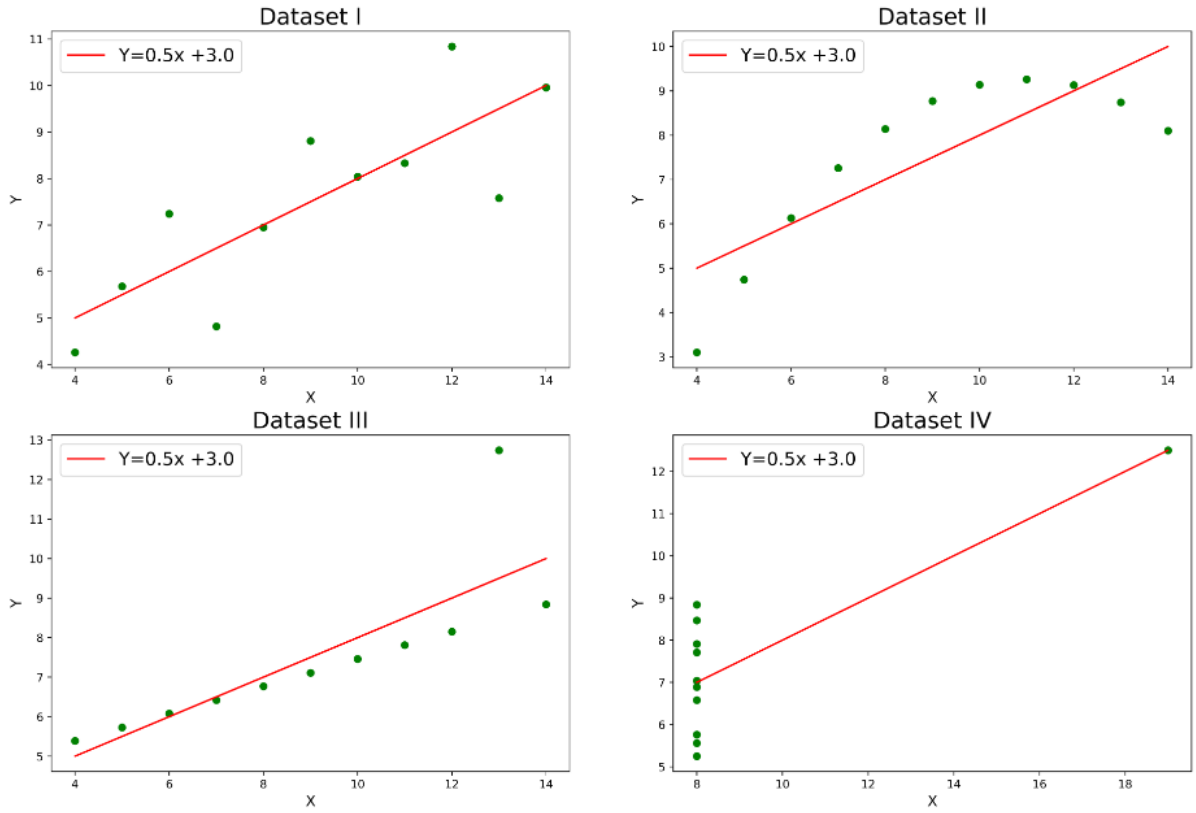
1) Linear Regression is a form of predictive modelling technique which tells us the relationship between the dependent (target variable) and independent variables (predictors).
2) The relationship between target and predictors is described using a straight line or regression line
3) The relation between target and dependent variables could be of 2 types:
    a. Increase in target variable with increase in dependent variable (Positive regression)
    b. Decrease in target variable with increase in dependent variable (Negative Regression)
4) Regression could also be of 2 types:
    a. Simple Linear Regression:
        a) Explains the relationship between a dependent variable and one independent variable using a straight line
        b) The equation is of the form $Y = \beta_o + \beta_1.X$
    b. Multiple Linear Regression:
        a. It is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables)
        b. The equation is of the form $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + ... + B_nX_n$
6) The steps involved in the algorithm are:
    a. Understand the data and perform EDA
    b. Split the data into test and training data set
    c. Split the training and test data into X and Y data frames where Y is a series containing the target variable and X is data frame containing all the predictors
    d. A linear model is fitted on the training data set using libraries from python or manually
    e. The library finds co-efficients for the best fit line by minimizing errors
    f. The effectiveness of the model can be measured using R^2 and Adjusted R^2
    g. Collinearity in the model is measured using VIF metric and a threshold of 5 and below is recommended
    h. This is then evaluated to verify the assumptions of linear regression (Linear relation, No multi-collinearity, Normal distribution of errors with zero mean, No pattern among the errors and homoscedasticity)

i. The test data set is also split to have the same features in Y and X as the training data set
j. The model is then applied on the test data
k. The R^2 and Adjusted R^2 of the test data should be close to that of the training data to indicate a good model
l. All the assumptions of linear regression must hold true on the test data also for a good regression model.

2) Explain the Anscombe's quartet in detail.

Answer:

1. Anscombe's quartet comprises of four data sets that have nearly identical simple descriptive statistics, but have very different distributions and appear very different when plotted on graph.
2. Each dataset consists of eleven (x, y) points.
3. The Anscombe's quartet were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.
4. These were created to prove the impression among statisticians incorrect that "numerical calculations are exact, but graphs are rough".
5. The statistics used for all the data sets are as below:
   - Mean
   - Sample variance
   - Correlation between x and y
   - Linear regression line
   - Coefficient of determination of the linear regression



6) The four data sets comprising the Anscombe's quartet have identical statistical parameters, but the graphs show them to be considerably different.

- In the scatter plot for Dataset 1 we see that there is a linear relationship between x and y.
- In the plot for Dataset 2 we see that there is a non-linear relationship between x and y.
- In the plot for Dataset 3 there is a linear relationship for all the data points except the outlier which is present far away from the line.
- The plot for Dataset 4 shows an example when one high-leverage point is enough to produce a high correlation coefficient.

7) While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

3) What is Pearson's R?

Answer:

1) Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables.
2) It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

3) It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.
4) Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.
5) It is independent of the unit of measurement. For example, if a variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
6) Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

3) Feature scaling is the process of normalizing the range of features in a dataset.
4) Real-world datasets often contain features that are varying in degrees of magnitude, range, and units.
5) To interpret these features on the same scale, we need to perform feature scaling.
6) Not scaling can cause models to make predictions that are inaccurate and difficult to interpret.
7) There are two major methods to scale the variables -
    a. Standardisation
    b. MinMax scaling.

Standardisation	MinMax scaling
Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.	MinMax scaling, on the other hand, brings all of the data in the range of 0 and 1.
Formula for Standardization is: X = X – mean(X) / SD(X)	Formula for MinMax scaling is: X = X-min(X)/Max(X) – Min(X)

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

Answer:

1) VIF is calculated using formula –

$$VIF_i = 1/1-R_i^2$$

2) If R^2 for any feature is 1, we get 0 in the denominator for division which gives VIF as infinity
3) This happens when there is perfect correlation between features
4) Removing this multi-collinearity from the model helps to address this

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

Answer:

- Q-Q plots also known as Quantile-Quantile plots, plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
- This helps to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.
- QQ plots is very useful to determine:
    - If two populations are of the same distribution
    - If residuals follow a normal distribution.
    - Skewness of distribution
- If the datasets we are comparing are of the same type of distribution, we would get a roughly straight line using the QQ plot
- One of the Linear regression assumption states that error terms are normally distributed with mean zero