LINEAR REGRESSION SUBJECTIVE QUESTIONS

A. ASSIGNMENT BASED SUBJECTIVE QUESTIONS

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: Analysis for the categorical variable is done using boxplot diagrams. Below are the few inferences from the visualisations.

- The boxplots show no outliers or very minimal outliers hence the probability of measurement errors is minimal or zero.
- Fall Season has the highest amount of count for the rented bikes, hence can even see that September is the month with the highest number of sales.
- It's clearly visible that the count has seen a downfall for 3 months - July, November, and December.
- For weathersit, we see that "clear, few clouds, partly cloudy" i.e., clear weather have the highest number of registered and casual users.
- People tend to hire the BoomBikes on a weekend or a holiday.
- We see an increase in the count for the future year i.e, 2019, showing a good progress in the business.

Q2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

Ans: drop_first=True is important during dummy variable creation because if we do not use this, then n dummy variables will be created and because of this it will lead to multicollinearity as these dummy variables are corelated within themselves and hence avoiding Dummy Variable Trap.

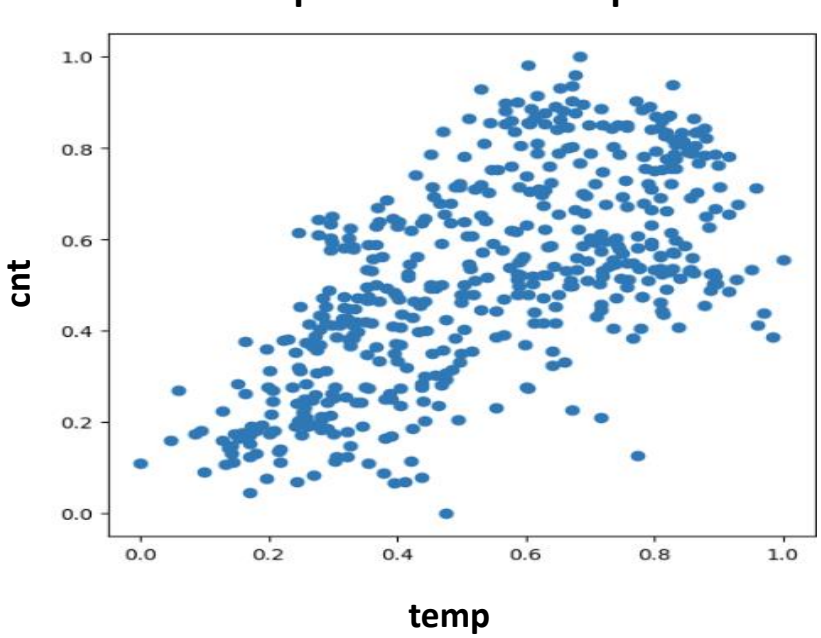
Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: '**temp'** variable has the highest correlation with the target variable (0.63).

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

1. **Linear relationship between dependent and independent variables:** The linearity is valid if there is a linear relationship i.e., a straight-line relation between the independent and dependent variables, here, in this example, target variable being 'cnt' and independent variable being 'temp'.



2. **Little or no Multicollinearity:** There should be insignificant multicollinearity between the variables. If the multicollinearity is high, then one of those variables should be dropped.

3. **Normality of error terms**: Error terms should be distributed normally for the purpose of reliability of tests used to build the model.

4. **Homoscedasticity:** There should be no visible pattern in residual values i.e., the error should be constant along the values of the dependent variable.

Q5. 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 contributing significantly explaining the demand is –

- temp
- weathersit - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- year

Hence, the focus should be on these 3 variables when planning to achieve maximum demand.

B. GENERAL SUBJECTIVE QUESTIONS

Q1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear regression is a statistical model which analyses the relationship between a dependent variable with a given set of independent variables. In simple terms, it means that if the value of any independent variable increases or decreases (basically changes), then the value of dependent variable also changes accordingly i.e., increases, or decreases.

Mathematically, linear regression is represented by the equation:

$$y = mx + c$$

where,

y = dependent variable which is being predicted

m = slope of the regression line

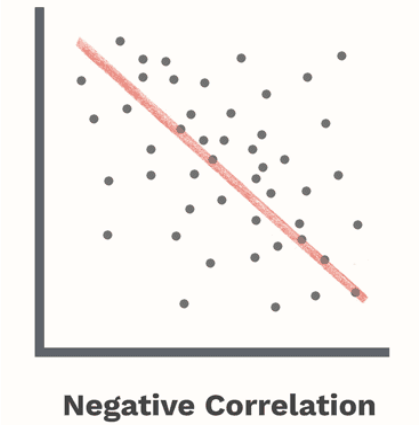
x = independent variable which is used to predict dependent variable

c = constant (known as y-intercept)

Linear Regression can be positive or negative in nature:

- Positive Linear Relationship - There is a positive linear correlation when the variable on x – axis (independent variable) increases as the variable on the y – axis increases (dependent variable). It is shown by an upward sloping straight regression line.

- Negative Linear Relationship - There is a negative linear correlation when either one of the variable increases and the other variable decreases. It is shown by a downward sloping straight regression line.



A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-values.

The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. It ranges between 0 and 1. Using a p-value, one can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet comprises of four datasets (x,y) which have the same mean, standard deviation, and regression line but are qualitatively different. Basically, they share same descriptive statistics. However, when plotted they do not resemble same characteristics.

Here are the characteristics of the four datasets in Anscombe's quartet:

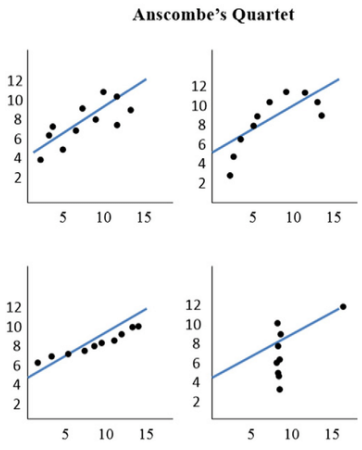
A. **Dataset I**:
- The data points form a perfect linear relationship.
- This dataset has a strong positive linear correlation.
- The summary statistics (mean, variance, correlation) would accurately reflect this linear relationship.

B. Dataset II:
- The data points form a non-linear relationship, but it is still relatively strong.
- There is a clear quadratic pattern in the data.
- While the linear correlation coefficient is still relatively high, it fails to capture the true nature of the relationship.

C. Dataset III:
- There is no apparent relationship between the variables.
- The data points are scattered randomly without any discernible pattern.
- Summary statistics might suggest no relationship between the variables, but visually, it's evident that there's no linear relationship.

D. Dataset IV:
- There is a strong linear relationship between most of the data points, except for one outlier.
- The linear correlation coefficient would suggest a strong relationship, but the presence of the outlier significantly affects the relationship when visually examined.

Anscombe's Quartet of Different XY Plots of Four Data Sets Having Identical Averages, Variances, and Correlations



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

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	5.76
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	8.81	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	7.26	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

Source: Adapted from Anscombe (1973, pp. 19-20)

The quartet shows that summary statistics alone can be misleading, as they might suggest similarities or patterns in the data that are not actually present. When visualized the data through plots (such as scatter plots), it reveals important patterns or outliers that summary statistics might overlook.

Q3. What is Pearson's R? (3 marks)

Ans: Pearson's R developed by Karl Pearson is a statistical measure which quantifies the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction. It is the most common way of measuring a linear correlation between two variables.

Pearson's correlation takes values between -1 and 1:

- A value of 1 shows a perfect positive linear relationship i.e., with increase in one variable there is an increase in another variable.
- A value of -1 shows a perfect negative linear relationship i.e., with increase in one variable there is a decrease in another variable.

Mathematically, it is the covariance of the two variables by the product of their standard deviations i.e.,

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

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

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling is a process where the value of a dataset is transformed to a specific range or distribution. It is a data preparation step for regression model.

Performing scaling is important as it ensures that no single feature is dominating the distance calculations in an algorithm and can even help to improve the performance of the algorithm.

Normalized Scaling: This scales the values of variables to a specific range, typically between 0 and 1. Mathematically,

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

where x is the original value, and $\min(x)$ and $\max(x)$ are the minimum and maximum values of the variable, respectively.

Standardized Scaling: This scales the values of variables to have a mean of 0 and a standard deviation of 1. Mathematically,

$$x_{\text{ standarized}} = \frac{x - \text{mean}\,(x)}{standard\ deviation(x)}$$

where x is the original value, and mean(x) is the mean of the variable and standard deviation(x) is the standard deviation of the variable.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in regression analysis. The formula for VIF for an independent variable x_i is –

$$\text{VIF}_i = \frac{1}{1-R_i^2}$$

where, R_i^2 and R^2 value is obtained by regressing x_i against all other independent variables in the model.

VIF is infinite when R_i^2 = 1, which means that the independent variable x_i is perfectly linearly related to the other independent variables in the model. In these situations, the regression model fails to compute accurate estimates of the regression coefficients due to the perfect multicollinearity and hence are unreliable.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans: A Q-Q plot i.e., quantile-quantile plot is a graphical tool which is used to assess whether a given sample of data follows a particular theoretical distribution. It compares the quantiles of the sample data to the quantiles of the theoretical distribution being tested.

It is used to compare the distribution of the two sets of data. It pinpoints deviations between distributions and identifies the data points responsible for them.

In linear regression, Q-Q plot is mainly used to know if the assumption of a common distribution is satisfied for two datasets provided. If yes, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If not, then, it makes one understand the difference between two datasets.

The Q-Q plot provides insights into the nature of the difference than analytical methods with the help of various tests such as chi square tests.

By Aashna Mehta