LinearRegression 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? (3 marks)

Answer: Here are some inferences made by analysing categorical variables on dependant variable (count)

- **Season** - Approx 32% of bookings were made in season 3(fall), 27% and 25% in summer and winter respectively
- **year** - 62% of bookings were made in 2019
- **month** - 10% of bookings were made in the month 5,6,7,8,9 (i.e. May, June, July, August, September)
- **holiday** - Most of the bookings happened on non-holiday
- **weekday** - Bookings made on weekdays were almost same across
- **workingday** - 69% of bookings were made on working day
- **weather** - 68% of bookings were made on weathersit 1: Clear, Few clouds, Partly cloudy, Partly cloudy

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

Answer: When creating dummy variables for categorical variables with k categories, the parameter drop_first=True is used to create k-1 dummy variables to represent all the information.
Dropping one of the dummy variables serves as a reference category.
If you create dummy variables for all categories of a categorical variable, it introduces multicollinearity because the information about one category can be derived from the others.

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

Answer: From the pair plot , it can be determined that temp, atemp has correlation with cnt. From the model it is evident that atemp is not considered and hence temp has highest correlation with cnt

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

Answer: From the distplot of residuals it is observed that Error terms are normally distributed with mean 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)

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- **Temp (Temperature):** It has coefficient of **0.509903**
- **Yr (year):** It has coefficient of **0.232930**
- **Weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds):** It has coefficient of **-0.294243**

General Subjective Questions

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

Answer:
Linear regression is a type of supervised learning algorithm used for predicting a continous output variable(dependant varaible) based on one or more predictor variable (independant variable)

There are 2 types of linear regression

- Simple Linear regression
- Multiple linear regression

Simple Linear regression: A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line. The independent variable is also known as the predictor variable. And the dependent variables are also known as the output variables.

It is based on the equation y=mx+c or (y(x) = B0 + B1(x))
y = independent variable
x = predictor variable
m = slope
c = intercept

Multiple linear regression: A multiple linear regression model attempts to explain the relationship between a dependent and multiple independent variable..

The new aspects to consider in multiple linear regression are:

- *Overfitting*: As you keep adding the variables, the model may become far too complex
  It may end up memorising the training data and will fail to generalise
  A model is generally said to overfit when the training accuracy is high while the test accuracy is very low
- *Multicollinearity:*Associations between predictor variables (VIF can be used to determine this)

- *Feature selection*: Selecting the optimal set from a pool of given features, many of which might be redundant becomes an important task

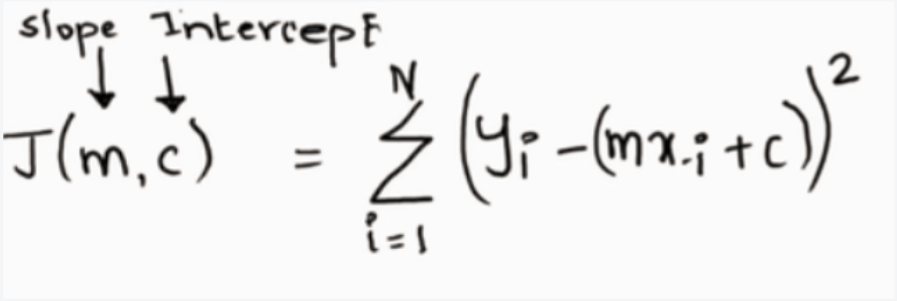
Equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Y is the dependent variable.
x1, x2, .. are the independent variables.
b0 is y intercept
b1, b2, b3...are the coefficients representing the impact of each independent variable.
Coefficients are estimated using methods like Ordinary Least Squares (OLS).

cost function: Every machine learning model there exists a term called cost function which needs to be optimized. Performance of ML model depends on cost function.
Objective is to minimize the sum of squared differences between the predicted and actual values.
Loss function is represented as Mean Squared Error (MSE)

Note : The right side of equation is divided by N or multiplied by 1/N .



Cost Function

Assumptions of linear regression (Simple & Multiple regession):

- There is a linear relationship between X and Y:
- Error terms are normally distributed with mean zero(not X, Y): error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable.
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity): The variance should not increase (or decrease) as the error values change.
  Also, the variance should not follow any pattern as the error terms change

Evaluation: R2, Adjusted R2 shall be used to determine the model performance.

Note: Some important points on linear regression

Linear regression:

- parametric model
- Uses: Forecasting, prediction
- Does interpolation not extrapolation
- Shows correlation not causation

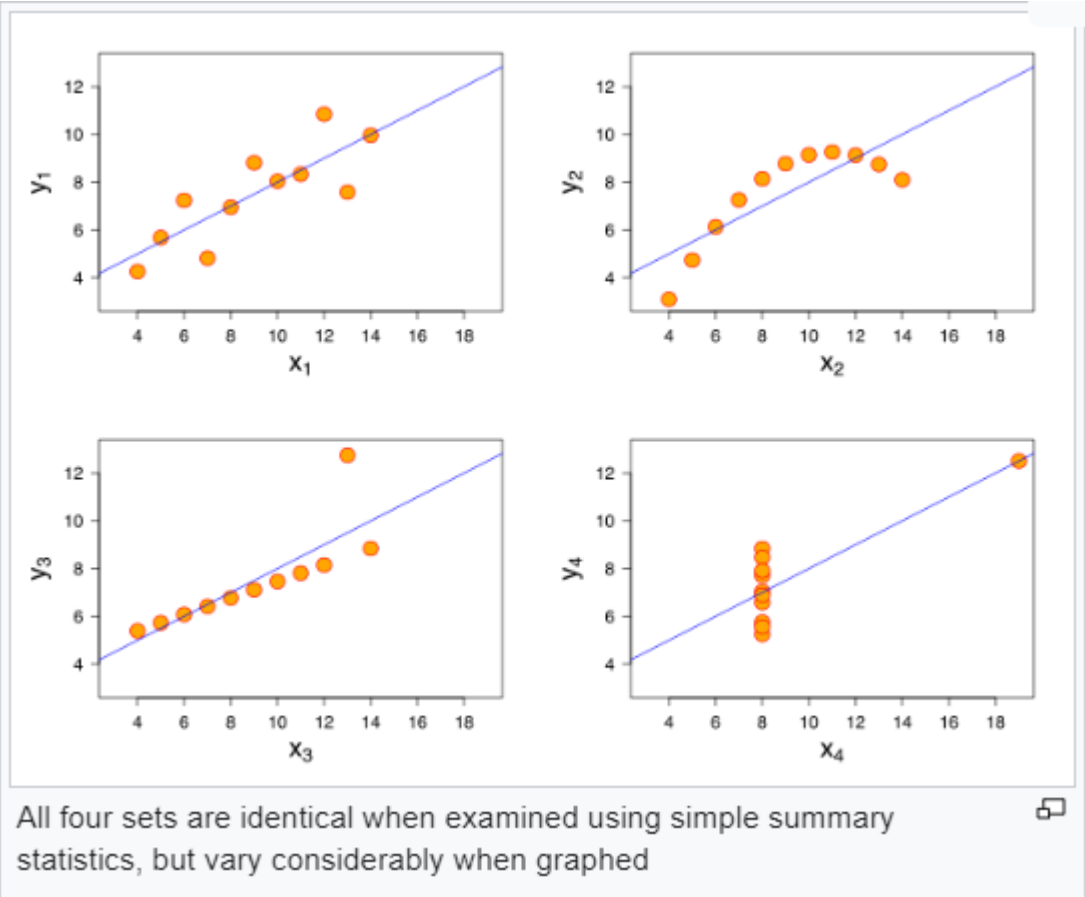
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

Anscombe's quartet serves as a cautionary example against overreliance on summary statistics and underscores the value of graphical analysis in understanding datasets.



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

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

Answer:

Pearson's correlation coefficient (r) ,is a measure of the strength and direction of a linear relationship between two continuous variables**.**

Pearson's correlation ranges from -1 to 1

- −1 indicates a perfect negative linear relationship.
- 1 indicates a perfect positive linear relationship.
- 0 indicates no linear relationship

Direction of correlation:

- r>0: Positive correlation (as one variable increases, the other tends to increase).
- r<0: Negative correlation (as one variable increases, the other tends to decrease).
- r=0: No linear correlation.

Strength of correlation:

- The absolute value of r indicates the strength of the correlation.
- |r| close to 1 suggests a strong linear relationship.
- |r| close to 0 suggests a weak or no linear relationship.

Formula:

Formula

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

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

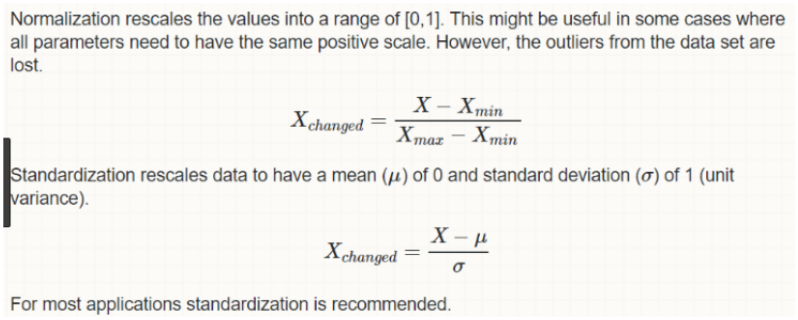
$\bar{y}$ = mean of the values of the y-variable

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

Answer:

- Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step
- Scaling is performed for - *feature interpretation, faster convergence of gradient descent*
- There are two major methods to scale the variables: **Standardisation and 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.



Example below:

	Original	Standardization	Max-Min Scaler
2	6.9314183	-0.2244971	0.0000003
3	2.6674115	-0.2244979	0.0000001
4	7.7248183	-0.2244970	0.0000003
5	5.7388433	-0.2244973	0.0000002
6	0.8965615	-0.2244982	0.0000000
7	4.5147618	-0.2244975	0.0000002
8	2.9934144	-0.2244978	0.0000001
9	4.8708377	-0.2244975	0.0000002
10	4.2797819	-0.2244976	0.0000002

- scaling just affects the coefficients and none of the other parameters like

  t-statistic, F-statistic, p-values, R-squared, etc.

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

Answer:
A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.
Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

Reason for encountering infinite VIF values is perfect multicollinearity, and it typically happens due to one of the following scenarios:

- one predictor variable can be expressed as a perfect linear combination of other predictor variables.

- If you include a dummy variable for each category and one of them can be exactly predicted from the others, it leads to perfect multicollinearity.

When R2 =1 , VIF becomes infinity

Formula

$$VIF_i = \frac{1}{1-R_i^2}$$

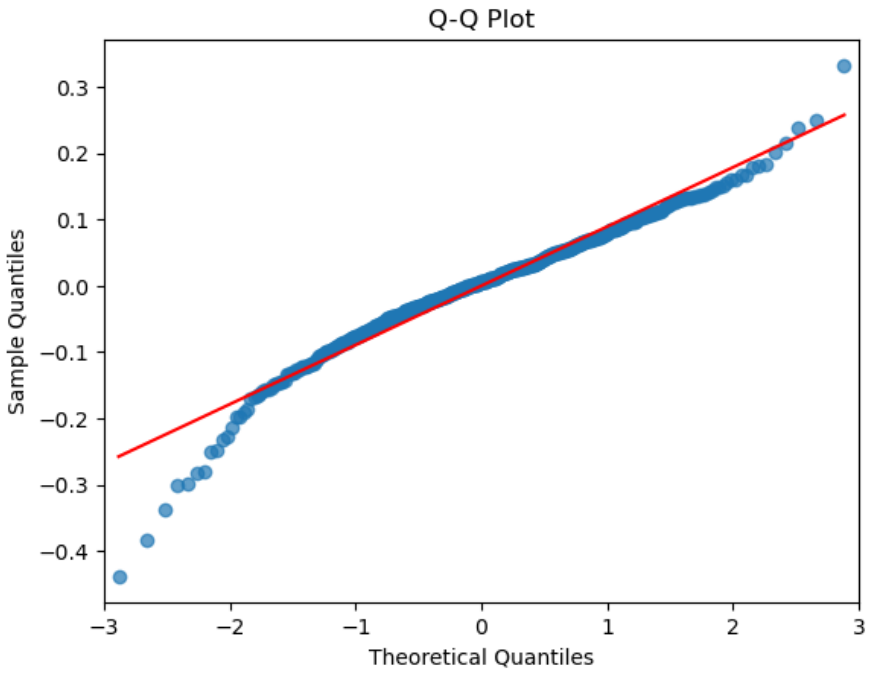
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer:

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set

In the context of linear regression, Q-Q plots are often used to check the normality assumption of residuals. Here are few points:

- Q-Q plots provide a visual means to compare the distribution of a sample to a theoretical distribution (usually normal distribution).
- A straight line on the Q-Q plot suggests that the residuals are normally distributed.
- Departure from the straight line indicates deviations from normality.
- The x-axis of the Q-Q plot represents the quantiles expected from a theoretical distribution (e.g., normal distribution).
- The y-axis represents the quantiles observed in the actual dataset (residuals).


Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−3
−2
−1
0
1
2
3