Shayamon Bastakoti

Assignment-based Subjective Questions

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

I analyzed the categorical variables against the target variable using boxplots and observed the following impacts on demand:

- **Season:** Fall (season 3) shows the highest rental bike demand.
- **Year:** There is a noticeable growth in demand in the following year.
- **Month:** Demand increases steadily each month until June, with September experiencing the peak. After September, there is a decline in demand.
- **Holiday:** Demand tends to decrease on holidays.
- **Weekday:** The data does not reveal a clear pattern of demand based on weekdays.
- **Weather Situation:** Clear weather conditions are associated with the highest demand.

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

drop_first=True is important to use, as it helps in reducing the extra column created during 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 multicollinearity issue.

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

The feature “temp” has highest correlation. It is very well linearly related with target “cnt”

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

I have checked the following assumptions:

- Error terms are normally distributed with mean 0.
- Error Terms do not follow any patterns.
- Checking multicollinearity using VIF(s).
- Linearity Check.
- Ensured the overfitting by looking the R2 value and Adjusted R2.

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

Features like “holiday”, “temp” and season “hum” are highly related with target column, so these are top 3 contributing features in model building.

General Subjective Questions:

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

Linear Regression is a supervised machine learning algorithm that is a core part of regression analysis. Regression analysis involves predictive modeling techniques used to establish the relationship between an input variable and a target variable. Linear regression, being one of the simplest machine learning techniques, helps in training a model to predict outcomes based on observed data. The term "linear" signifies that the variables on the x-axis (independent variable) and y-axis (dependent variable) are expected to have a linear relationship.

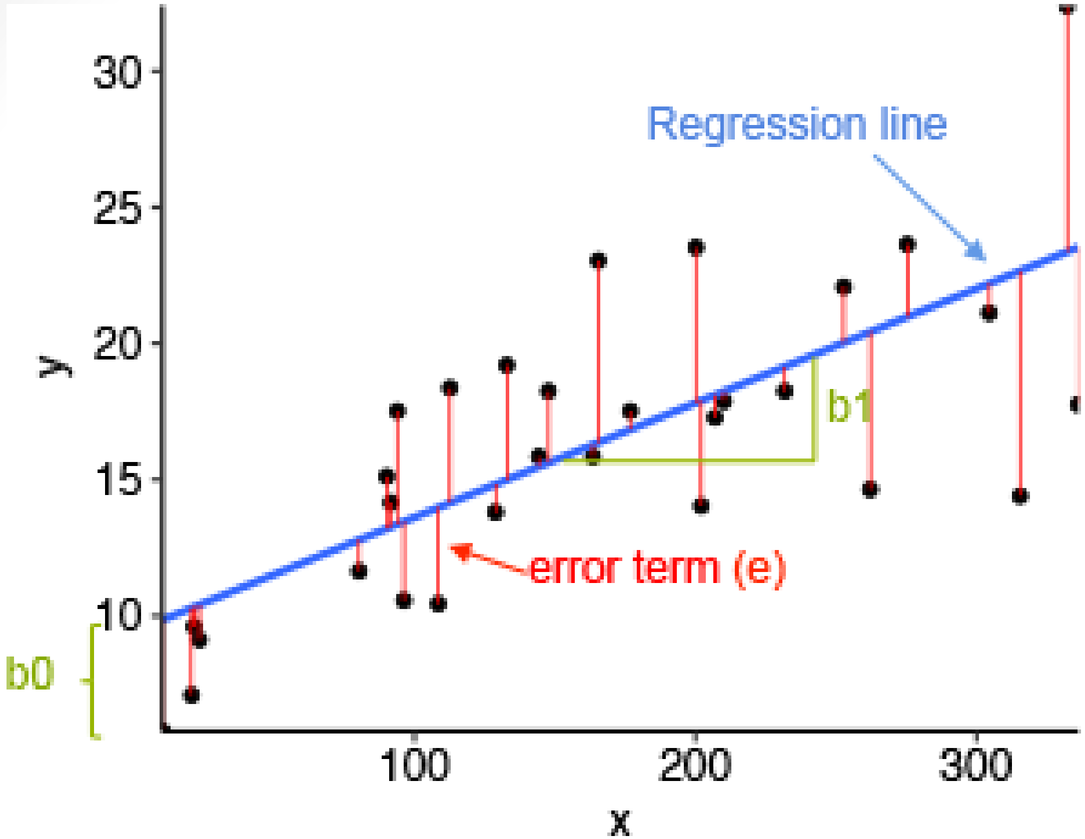
For instance, consider running a sales promotion where you anticipate an increase in customer numbers. By analyzing historical data from past promotions, you can plot this data to observe trends and correlations. Using these insights, you can predict customer counts for the current promotion. This prediction enables better planning, such as deciding on the number of staff required to accommodate the expected customer surge. Linear regression thus helps estimate future values by identifying patterns in historical data.

Linear relationships can be upward, where both variables increase together, or downward, where an increase in one variable corresponds to a decrease in another. For example, a downward trend could represent a scenario like a campaign by law enforcement to reduce crime rates, where increased efforts result in fewer robberies.

Linear regression aims to predict a numeric outcome (Y) from an independent variable (X). The equation for simple linear regression is:

y = b0 + b1*x

Here, **y** is the dependent variable, **x** is the independent variable, **b0** represents the intercept, and **b1** is the slope of the line. A cost function is used to determine the optimal values of **b0** and **b1**, providing the best-fit line accurately.



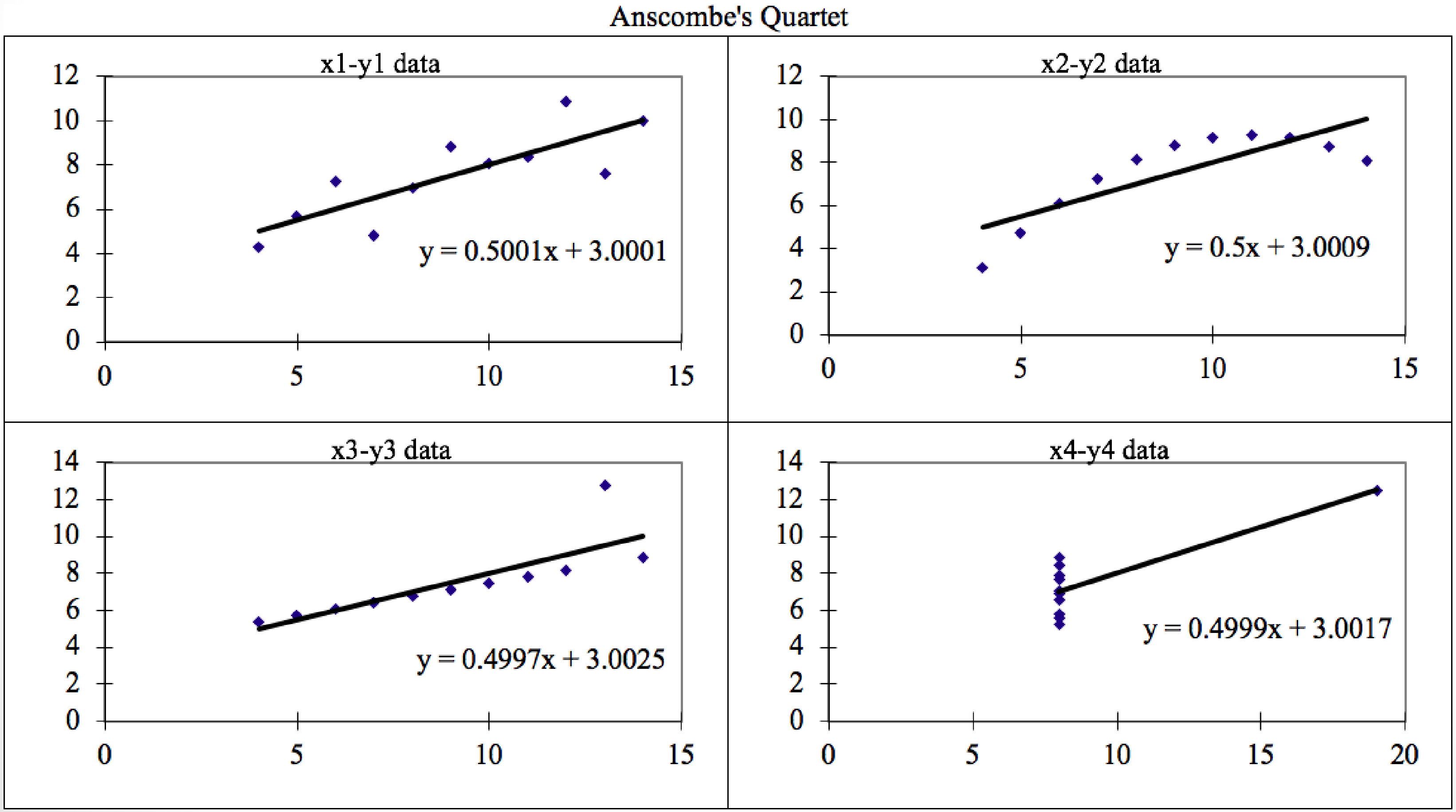
Here, x and y are two variables on the regression line. b1 = Slope of the line.

b0 = y-intercept of the line.

x = Independent variable from dataset y = Dependent variable from dataset

- **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.



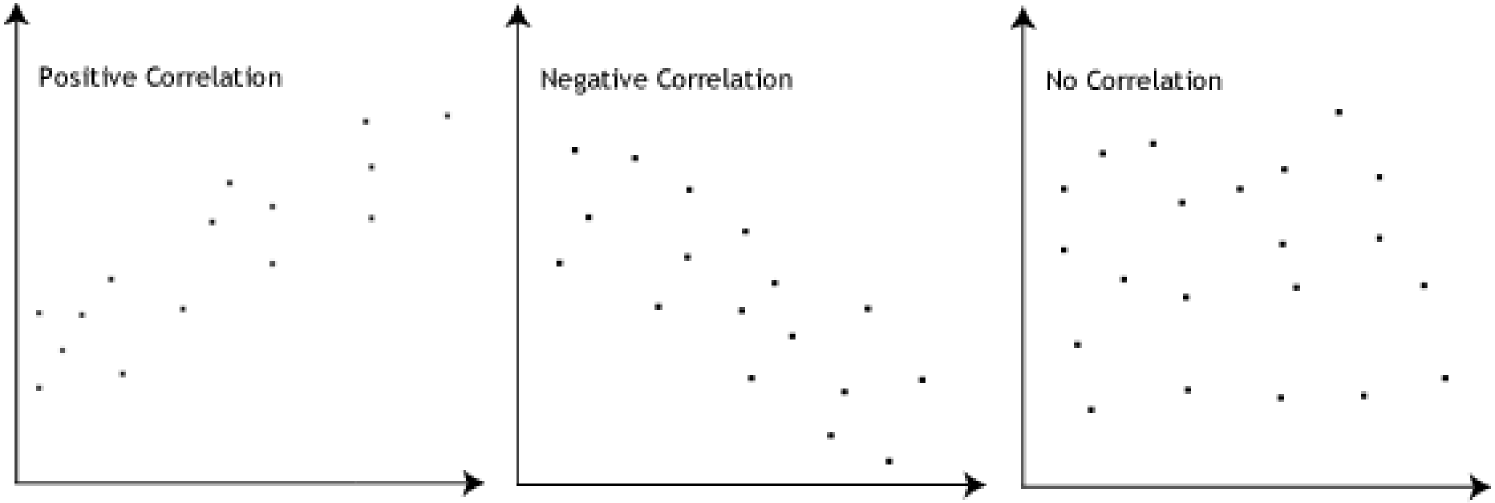
- The four datasets can be described as:
- **Dataset 1:** this **fits** the linear regression model pretty well.
- **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the **outliers** involved in the dataset which **cannot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

- **What is Pearson’s R? (3 marks)**

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive.

Pearson's r measures the strength of the linear relationship between two variables. It lies between -1 and 1.

If data lie on a perfect straight line with negative slope, then r = -1.



Positive correlation indicates the both the variable increase and decrease together. Negative correlation indicates the one the variable increase and the other variable decrease and vice versa.

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

Scaling is a technique used to normalize the range of independent variables, ensuring that all features are on the same scale in regression analysis. It helps to prevent larger numerical values from being mistakenly interpreted as more significant by the regression algorithm, which could skew the results. Without scaling, variables with higher magnitudes may disproportionately influence the model's performance.

It's crucial to understand that scaling only impacts the coefficients and does not alter other parameters such as the t-statistic, F-statistic, p-values, or R-squared values.

Example:

Consider a scenario where the weight of one device is 500 grams, while another device weighs 5 kilograms. Without scaling, the algorithm might incorrectly perceive 500 as the larger value, leading to inaccurate predictions. Since machine learning models operate on numerical values rather than units, scaling is an essential preprocessing step for regression.

Methods of Scaling:

Normalization: Rescales variables to fall within the range of 0 to 1.

Standardization: Adjusts data to have a mean of 0 and a standard deviation of 1, centering and scaling the features.

Normalized scaling rescales data to a specific range, typically between 0 and 1, and is useful when the data has known bounds. **Standardized scaling**, on the other hand, transforms data to have a mean of 0 and a standard deviation of 1, making it ideal for data that doesn't have fixed limits. Both methods are used to bring features to a common scale but are suited for different types of data and algorithms.

Normalized scaling rescales data to a specific range, typically between 0 and 1, and is useful when the data has known bounds. **Standardized scaling**, on the other hand, transforms data to have a mean of 0 and a standard deviation of 1, making it ideal for data that doesn't have fixed limits. Both methods are used to bring features to a common scale but are suited for different types of data and algorithms.

- **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.

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

The Quantile-Quantile (Q-Q) plot is a graphical tool used to evaluate whether a dataset aligns with a specific theoretical distribution, such as Normal, Exponential, or Uniform distributions. Additionally, it can determine if two datasets originate from populations with the same distribution. A Q-Q plot visualizes the quantiles of one dataset plotted against the quantiles of another dataset, allowing comparisons across various distributions, including Gaussian, Uniform, Exponential, or Pareto distributions.

Key Advantages:

- **Versatile with Sample Sizes:** It works effectively even with small sample sizes.
- **Detects Distributional Features:** The plot can reveal shifts in location and scale, changes in symmetry, and the presence of outliers.

Common Applications:

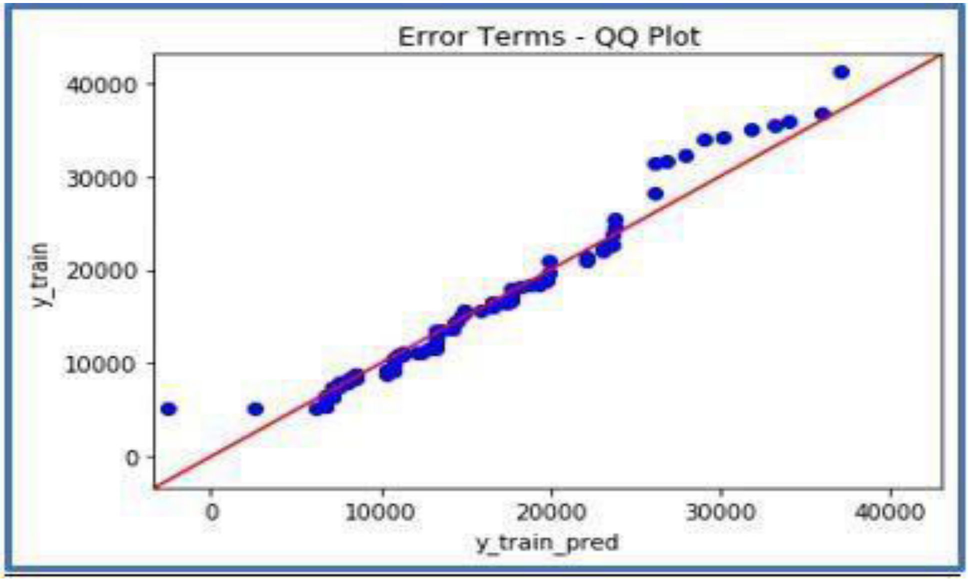
- The Q-Q plot helps verify if two datasets:
- Share the same population distribution.
- Exhibit a common location and scale.
- Have similar distributional patterns.
- Display comparable tail behavior.

Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets:

- Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
- Y-values < X-values: If y-quantiles are lower than the x-quantiles.
- X-values < Y-values: If x-quantiles are lower than the y-quantiles.



- **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis