Assignment by - Shivani Bhatia

Submitted as a part of Multiple Linear Regression Assignment for IIIT B and Upgrad Executive PG in AI and ML Programme.

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)

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

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

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

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

General Subjective Questions

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

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

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

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

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

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

General Subjective Questions

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

Linear regression is a basic and commonly used type of predictive analysis. The idea is to model the relationship between a dependent variable (what you want to predict) and one or more independent variables (the features or inputs).

The Basic Idea

1. Draw a Line: Imagine you have a scatter plot of data points. Linear regression tries to draw the best straight line through these points. This line is called the "line of best fit."

2. Equation of a Line: The line of best fit can be described by the equation:

$$y = mx + c$$

- y is the dependent variable (what you want to predict).
- x is the independent variable (the feature/input).
- mis the slope of the line (how much y changes for a one-unit change in x).
- cis the intercept (the value of y when x is 0).

In the case of multiple features, the equation becomes:

- y = b0 + b1x1 + b2x2 +……+ bnxn

- y is the dependent variable.
- x1, x2, ..., xn are the independent variables.
- b0 is the intercept.
- b1, b2, ..., bn are the coefficients for each independent variable.

How It Works

1. Fit the Line: The algorithm tries to find the best values for the slope(s) and intercept that minimize the difference between the actual data points and the predictions made by the line. This difference is called the "error."

2. Minimize Error: Linear regression uses a method called "least squares" to minimize the sum of the squared differences between the actual data points and the predicted points on the line.

Steps in Linear Regression

1. Collect Data: Gather data with the dependent variable and one or more independent variables.

2. Split Data: Divide the data into training and test sets.

3. Train Model: Use the training data to find the best line of fit by calculating the slope(s) and intercept that minimize the error.

4. Make Predictions: Use the line of best fit to make predictions on new data.

5. Evaluate Model: Check how well the model performs by comparing predictions to actual values using metrics like R^2 score or Mean Squared Error (MSE).

Example

Imagine we want to predict the price of a house based on its size. So, our data might look like this:

- Size of house (in square feet) [x]
- Price of house (in rupees) [y]

Linear regression would help us draw a line that best predicts the house price based on its size. Once you have the line, you can use it to estimate the price of a house given its size.

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

Anscombe's quartet consists of four datasets, each containing 11 pairs of x and y values. When graphed, each dataset reveals a distinct relationship between x and y, characterized by unique patterns of variability and different strengths of correlation. Despite these visual differences, all four datasets share identical summary statistics, including the same means and variances for x and y, the same correlation coefficient between x and y, and the same linear regression line.

It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analysing it and the limitations of relying solely on statistical properties.

Importance and Lessons

Visual Inspection: Anscombe's Quartet illustrates that relying solely on statistical summaries (mean, variance, correlation) can be misleading. Visual inspection through graphing is crucial for understanding the true nature of data.

Graphing Data: Before performing any statistical analysis, it's important to plot the data. Different datasets can have the same statistical properties but very different distributions and relationships.

Outliers and Patterns: The datasets show how outliers and different patterns in data can drastically change the interpretation, even if the statistical properties are the same.

Conclusion

Anscombe's Quartet serves as a powerful reminder that statistics alone do not tell the whole story. Visualizing data can reveal underlying patterns, relationships, and anomalies that are not apparent from statistical summaries alone. Always visualize your data to gain a complete understanding before drawing conclusions.

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

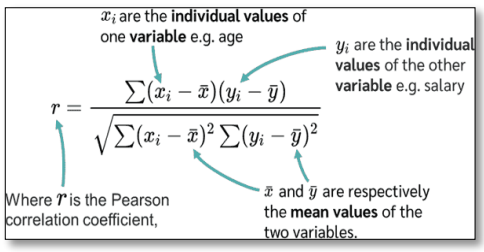
Pearson's R, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is widely used in statistics to determine how well one variable can predict another.

Key Points

Range: Pearson's R ranges from -1 to 1.

- **+1** indicates a perfect positive linear relationship: as one variable increases, the other variable increases proportionally.
- **-1** indicates a perfect negative linear relationship: as one variable increases, the other variable decreases proportionally.
- **0** indicates no linear relationship: the variables do not exhibit any linear association.

Calculation: Pearson's R is calculated using the formula:



Interpretation:

- **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.

Usage:

- Pearson's R is commonly used in fields such as social sciences, natural sciences, and economics to measure the degree of relationship between variables.
- It helps in understanding the strength and direction of the linear association, which can inform predictions and decisions based on the data.

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

When a model includes many independent variables, they can be on very different scales, leading to coefficients that are hard to interpret. Therefore, scaling features is necessary for two main reasons:

1. **Ease of Interpretation**: Scaling ensures that all features contribute equally, making the model's coefficients more meaningful and easier to understand.
2. **Faster Convergence for Gradient Descent Methods**: Scaling improves the efficiency of gradient descent algorithms, allowing them to converge more quickly.

Scaling is particularly important in machine learning for algorithms that rely on the distance between data points, such as k-nearest neighbour (KNN), support vector machines (SVM), and gradient descent-based algorithm.

Normalized Scaling (Min-Max Scaling):

- **Definition**: Transforms the features to a fixed range, usually [0, 1] or [-1, 1].
- **Use Case**: Useful when you know the minimum and maximum values of your data and want to bring all features to the same range.
- **Formula**: $(x - x\_min)/x\_max - \mathrm{x\_min})$
- **Example**: If you have a dataset with features that range from 0 to 1000, normalization would scale these features to the range 0 to 1.

Standardized Scaling (Z-score Standardization):

- **Definition**: Transforms the features to have a mean of 0 and a standard deviation of 1.
- **Use Case**: Useful when the data follows a Gaussian distribution (normal distribution) and you want to center the data around 0 with a unit variance.
- **Formula**: $\frac{x-Mean}{Standard\ Deviation}$
- **Example**: If you have a dataset with features that have different means and standard deviations, standardization would transform these features to have a mean of 0 and a standard deviation of 1, making them comparable.

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

Variance Inflation Factor (VIF) measures the degree of multicollinearity among predictor variables in a regression model. VIF can become infinite due to perfect multicollinearity. This occurs when one predictor variable is an exact linear combination of others, leading to the following:

1. Perfect Multicollinearity:

- Definition: When a predictor is perfectly predicted by other predictors.
- Formula: **VIF (x_i)= (1/(1-R_i^2))**
- Result: If ((R_i^2) = 1), then (1 - R_i^2) = 0), making VIF infinite.

2. Linear Dependence:

- Impact: Redundancy among predictors, making the regression matrix singular or nearly singular.

Practical Implications:

- Infinite VIF indicates perfect multicollinearity.

Solution:

- Remove one of the collinear variables.
- Combine collinear variables.
- Use regularization techniques like Ridge Regression.

Conclusion:

Infinite VIF values occur due to perfect multicollinearity, where one predictor is an exact linear combination of others. This results in an undefined or infinite VIF, signalling the need for corrective measures to ensure reliable regression models.

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

A **Q-Q plot** (quantile-quantile plot) is a graphical tool to compare the distribution of a dataset with a theoretical distribution, typically the normal distribution. It helps to visually assess whether the data follows a specific distribution.

How a Q-Q Plot Works

1. **Quantiles**: The data points are plotted against the quantiles of the theoretical distribution.
2. **Line of Fit**: If the data follows the theoretical distribution, the points will approximately lie on a straight line (the 45-degree reference line).

Use and Importance of a Q-Q Plot in Linear Regression

1. **Normality Check**:
    - **Assumption**: Linear regression assumes that the residuals (errors) are normally distributed.
    - **Q-Q Plot**: Helps in assessing whether the residuals follow a normal distribution. If they do, the points will align closely with the reference line.
2. **Detection of Deviations**:
    - **Outliers**: Points that deviate significantly from the reference line indicate outliers or non-normality.
    - **Skewness and Kurtosis**: Patterns in the Q-Q plot can reveal skewness (asymmetry) or kurtosis (tailedness) in the data.
3. **Model Validation**:
    - **Fit Assessment**: Ensures that the assumptions of linear regression are met, improving the reliability and validity of the model.
    - **Decision Making**: Helps in deciding whether transformations or alternative models are needed.

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)

Following are the inferences made from analysis of categorical variables:

- Bike rentals were more in the year 2019 compared to 2018.
- Bike rentals increased during good weather situation.
- Bike rentals were more in the fall season.
- Bike rentals particularly in September month were high.
- It was more popular on working days and non-holidays.
- Year and Count are positively correlated.

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

A dummy variable is a binary variable created to represent categorical data, often used in regression models and other statistical analyses. Each category of a categorical variable is converted into a separate dummy variable, taking the value 1 if the observation belongs to that category and 0 otherwise.

When creating dummy variables, using the drop_first=True option is important for the following reasons:

1. **Avoiding Multicollinearity**: In regression analysis, multicollinearity occurs when independent variables are highly correlated. If dummy variables are created for all categories of a categorical variable without dropping one, the sum of these dummy variables will always equal one, leading to perfect multicollinearity. Dropping the first category prevents this issue by ensuring that the dummy variables are linearly independent.

2. **Providing a Reference Category**: Dropping the first dummy variable creates a baseline or reference category against which the effects of other categories can be compared. This makes the model interpretable, as the coefficients of the remaining dummy variables indicate the effect of each category relative to the reference category.

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

Temperature has highest correlation with target variable.

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

Residual Analysis: Residuals in the obtained model were approximately normally distributed, and there were no discernible patterns in the residual plot.

Homoscedasicity (Constant Variance): The spread of residuals was roughly constant across all levels of the predicted values.

Linearity: The points fell approximately along a diagonal line on scatterplot of observed vs. predicted values., indicating a linear relationship.

Independence of Residuals: There were no discernible pattern in the residuals when plotted against time or other relevant variables.

Multicollinearity: VIF values for final selected features should be below a certain threshold (commonly 5 or 10) to ensure no problematic multicollinearity.

Cross-Validation: Assessed the model's performance on new data to ensure generalizability and consistency.

Check for Overfitting: The model generalizes well to new, unseen data without overfitting the training set.

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 following three features significantly contribute to explaining the demand for shared bikes:

- Temperature (temp)
- Winter season (winter)
- Calendar year (year)