25 September 2024 07:12 PM

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: most of the variables do not have any impact on the dependent variable except a few which are linearly related to the dependent variable.

Temp/Atemp have a big impact on the dependent variable. We have seen in the data that the increase in the temperature cause more bike rides bookings to take place.

Most of the bike rides are booked during Aug-September month by the customers.

Seasons where Bike rides are preferred most are summer and fall. Spring sees the least bookings in the all the seasons.

There is no such preference for the day of the week. All the days see almost similar kind of booking requests.

We see more booking requests on a working day compared to a holiday. A day being a working

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

Answer: drop_first=True is used creating dummy variables to avoid multicollinearity, enhance model performance, and simplify the interpretation of results in regression analysis.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

Answer: Temp and Atemp both having the 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)

Answer:

Below is the list of assumptions of linear regression that were validated after building the model on the training set.

1. Linearity

- **Assumption**: The relationship between the independent and dependent variables should be linear.
- **Validation**: Using the scatter plots to visualize the relationship between each independent variable and the dependent variable. The plots showed a straight-line relationship between the dependent variable(temp and the independent variables(cnt).

2. Independence

- **Assumption**: Observations should be independent of each other.
- **Validation**: This is often ensured through study design. For time series data, check for autocorrelation using the Durbin-Watson test. A value close to 2 indicates no autocorrelation.

3. Homoscedasticity

- **Assumption**: The variance of residuals (errors) should be constant across all levels of the independent variables.
- **Validation**: Create a scatter plot of residuals versus predicted values. We saw a consistent spread across the range, indicating homoscedasticity. We did not see any see patterns (e.g., a funnel shape) confirming homoscedasticity.

4. Normality of Residuals

- **Assumption**: The residuals should be approximately normally distributed.
- **Validation**: I displot plot to check the distribution of residuals if they form a normal distribution and I could see there was a normal distribution in the error distribution and was centred around zero.

5. No Multicollinearity

- **Assumption**: Independent variables should not be too highly correlated with each other.
- **Validation**: Variance Inflation Factor (VIF) for final predictor were well below 5 which indicates multicollinearity issues were removed from the model.

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: Looking at the magnitude of the coefficient of the features, temp, year, winter, September are the top 4 features. I have included year as well but not much can we done based on the year as the situation may not repeat itself again in another year. If the year was politically motivated year due to elections or some major government policy then it can be of importance to the demand of the shared bikes.

General Subjective Questions

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

Answer: Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is a supervised learning algorithm commonly used in data science and machine learning for predictive analysis.

Key Concepts of Linear Regression:

1. **Objective**: The primary goal of linear regression is to find the best-fit line (or hyperplane in higher dimensions) that describes the relationship between the dependent variable (response) and independent variables (predictors). This line minimizes the discrepancies between the observed values and the values predicted by the model.
2. **Equation**: In its simplest form, linear regression can be represented by the equation:

$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta pxp+\varepsilon y=\beta0+\beta1x1+\beta2x2+\cdots+\beta pxp+\varepsilon$

- y is the dependent variable.
- x1,x2,…,xpx1,x2,…,xp are independent variables.
- β0 is the intercept.
- β1,β2,…,βpβ1,β2,…,βp are the coefficients that represent the change in yy for a one-unit change in xx.
- ε is the error term accounting for variability not explained by the model.

3. **Least Squares Method**: Linear regression uses the least squares method to estimate the coefficients. This method minimizes the sum of squared differences between observed and predicted values, known as residuals.
4. **Evaluation Metrics**:
    - **R-squared**: Indicates how well data points fit a statistical model – an R-squared of 0.8 means 80% of the variance in the dependent variable is predictable from the independent variables.
    - **P-value**: Tests if an independent variable significantly predicts the dependent variable. A low p-value (< 0.05) indicates that you can reject the null hypothesis, meaning there is a significant relationship.
5. **Assumptions**:
    - Linearity: The relationship between dependent and independent variables should be linear.
    - Independence: Observations should be independent of each other.
    - Homoscedasticity: Constant variance of errors.
    - Normality: Errors should be normally distributed.
6. **Applications**: Linear regression is widely used in various fields such as economics for forecasting, biology for predicting growth rates, and finance for risk management.

By understanding these concepts, one can effectively use linear regression to analyze relationships between variables and make informed predictions based on data.

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

Answer: Anscombe's Quartet is a group of four datasets that are nearly identical in their simple descriptive statistics, yet they reveal very different distributions when graphed. Created by statistician Francis Anscombe in 1973, the quartet demonstrates the importance of data visualization in statistical analysis and the limitations of relying solely on summary statistics.

Key Features of Anscombe's Quartet:

1. **Identical Summary Statistics**: All four datasets have the same mean, variance, correlation, and linear regression line:
    - Mean of xx: 9
    - Mean of yy: 7.50
    - Variance of xx: 11
    - Variance of yy: 4.125
    - Correlation between xx and yy: 0.816
    - Linear regression line: y=0.5x+3y=0.5x+3
2. **Different Distributions**: Despite having similar statistical properties, the datasets differ significantly when plotted:
    - **Dataset 1**: Appears to follow a linear relationship closely, fitting well with the regression line.
    - **Dataset 2**: Displays a clear non-linear relationship, suggesting that a simple linear regression is inappropriate.
    - **Dataset 3**: Contains an outlier that heavily influences the regression line, masking the underlying linear relationship.

- **Dataset 4**: Shows a vertical pattern with one influential outlier that distorts the regression analysis.

3. **Implications for Data Analysis**:

    - **Importance of Visualization**: Anscombe's Quartet highlights how visualizing data can reveal patterns, relationships, and anomalies that summary statistics alone cannot capture.
    - **Caution with Regression Models**: The datasets illustrate how outliers and non-linear relationships can mislead regression models if not properly addressed.

Anscombe's Quartet serves as a powerful reminder for analysts to always visualize their data before drawing conclusions or building models, ensuring that they capture the true nature of the data's structure and relationships.

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

Answer: Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure used to quantify the strength and direction of a linear relationship between two continuous variables. Here are the key points about Pearson's R:

Definition and Range

- **Definition**: Pearson's R is a measure of linear correlation between two variables, denoted as r*r*. It is calculated as the covariance of the two variables divided by the product of their standard deviations.
- **Range**: The value of Pearson's R ranges from -1 to +1. A value of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Interpretation

- **Positive Correlation**: A positive value (e.g., 0.8) suggests that as one variable increases, the other variable tends to increase as well.
- **Negative Correlation**: A negative value (e.g., -0.5) indicates that as one variable increases, the other tends to decrease.
- **No Correlation**: A value close to 0 suggests no linear relationship between the variables.

Applications and Limitations

- **Applications**: Pearson's R is widely used in statistics and data analysis to assess relationships between variables. It is particularly useful in fields like finance, psychology, and natural sciences.
- **Limitations**: Pearson's R only measures linear relationships and may not capture non-linear associations. It is also sensitive to outliers, which can distort the correlation coefficient.

Pearson's correlation coefficient is a fundamental tool for understanding relationships between variables, but it should be used alongside other analyses to ensure comprehensive insights into data relationships.

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

Answer: **Scaling** is a data preprocessing technique used in machine learning to adjust the range of independent variables (features) so that they are on a similar scale. This is important because many machine learning algorithms, especially those that calculate distances or rely on gradient descent, perform better when features are on a comparable scale. Scaling helps improve the performance, stability, and convergence speed of these algorithms.

Why Scaling is Performed

1. **Improves Algorithm Performance**: Algorithms like k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and neural networks are sensitive to the scale of input features. Scaling ensures that no particular feature dominates others due to differences in units or scales.
2. **Accelerates Convergence**: For optimization algorithms like gradient descent, scaling can help achieve faster convergence by ensuring that the cost function contours are more circular rather than elongated ellipses.
3. **Enhances Interpretability**: In some cases, scaling can make models easier to interpret by bringing features to a common scale.

Difference Between Normalized Scaling and Standardized Scaling

Feature	Normalization (Min-Max Scaling)	Standardization (Z-score Normalization)
Purpose	Rescales features to a specific range, typically [0, 1]	Transforms features to have zero mean and unit variance
Formula	$\text{normalized value} = \frac{\text{value} - \min}{\max - \min}$	$\text{standardized value} = \frac{\text{value} - \mu}{\sigma}$
Range	[0, 1] or [-1, 1]	No specific range; values can be negative or positive
When to Use	When the data does not follow a Gaussian distribution; suitable for algorithms sensitive to feature scales like k-NN and neural networks [1] [3]	When data follows a Gaussian distribution or when assumptions about data distribution are made; suitable for algorithms like linear regression and SVM [1] [4]
Effect on Outliers	Can be affected by outliers as it relies on minimum and maximum values [3]	Less sensitive to outliers as it centers data around the mean [1] [3]

Normalization is ideal when you want to bound your feature values within a specific range, which is useful for certain machine learning models. Standardization is preferred when you need your features to have properties of a standard normal distribution, particularly when the algorithm assumes a Gaussian distribution of input data.

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

Answer: The occurrence of infinite values for the Variance Inflation Factor (VIF) is primarily due to the presence of perfect multicollinearity among the independent variables in a regression model. Here's a detailed explanation:

Perfect Multicollinearity

1. **Definition**: Perfect multicollinearity occurs when one or more independent variables in a regression model are perfectly linearly dependent on others. This means that one variable can be expressed as an exact linear combination of others.
2. **Impact on VIF**: The VIF measures how much the variance of an estimated regression coefficient increases due to multicollinearity. When multicollinearity is perfect, the VIF tends to infinity because the variance of the coefficient becomes undefined. This is because the denominator in the VIF calculation, which involves the coefficient of determination (R2R2) from regressing one variable on others, becomes 1, leading to

division by zero.

Causes of Infinite VIF

- **Duplicate Variables**: Including duplicate columns or variables that are linearly dependent (e.g., one column is a multiple of another) can lead to perfect multicollinearity.
- **High Correlation**: While not perfect, very high correlations between variables can also result in large VIF values that approach infinity, indicating severe multicollinearity issues.

Solution

To resolve issues with infinite VIF values, you can:

- **Remove or Combine Variables**: Identify and remove or combine variables that are causing perfect multicollinearity.
- **Check for Data Errors**: Ensure there are no data entry errors or duplicated columns in your dataset.

By addressing these issues, you can reduce multicollinearity and obtain more reliable regression coefficients.

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

Answer: A Q-Q plot, or quantile-quantile plot, is a graphical tool used in statistics to compare the quantiles of two probability distributions by plotting them against each other. Here's a detailed explanation of its use and importance, particularly in the context of linear regression:

What is a Q-Q Plot?

- **Definition**: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If the distributions being compared are similar, the points in the Q-Q plot will approximately lie on a 45-degree line (the identity line) .
- **Construction**: To create a Q-Q plot, you calculate or estimate the quantiles for both datasets. If one of the datasets is theoretical, such as a normal distribution, its quantiles are derived from its cumulative distribution function (CDF) .

Use and Importance in Linear Regression

1. **Assessing Normality**: In linear regression, one key assumption is that the residuals (errors) are normally distributed. A Q-Q plot can be used to visually assess whether this assumption holds. If the residuals are normally distributed, the points will fall approximately along the straight line .
2. **Identifying Deviations**: The Q-Q plot helps identify deviations from normality, such as skewness or kurtosis. Deviations from the straight line at the tails indicate potential issues with normality, which could affect model validity and inference .
3. **Model Diagnostics**: Beyond checking normality, Q-Q plots can diagnose other issues in regression models by comparing residuals to theoretical distributions. This helps in understanding how well the model fits and whether assumptions are violated .

In summary, Q-Q plots are crucial for validating assumptions in linear regression models and ensuring that statistical inferences drawn from these models are reliable. They provide a visual method to check if data meet the normality assumption and help identify any anomalies in distribution that could impact model performance.