subjective questions related to linear regression

Assignment-based Subjective Questions

Q.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 :

Affect of Categorical data on dependent variable `cnt`:

1. season : summer, fall, winter these three season has some positive trend with dependen variable CNT.
2. yr : cnt is progressively increasing by yr.
3. mnth : month like `may` to `October` we can see positive trend.
4. holiday : Despite the working day the number of bike count is maximum when day is not holiday.
5. weekday : weekday show no impact on dependent variable.
6. weatersit : As w move to Clear weather to Heavy Rain count decreases heavily.

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

Answer : Using `drop_first=True` during dummy variable creation in pandas is important for the following reasons:

1. Avoiding Multicollinearity

Multicollinearity occurs when one predictor variable in a model can be linearly predicted from the others with a substantial degree of accuracy. This can cause problems in the estimation of the regression coefficients. When creating dummy variables, if you have (k) categories, you'll create (k) binary variables (one for each category). However, these variables are perfectly collinear (i.e., their sum equals 1 for each observation), which leads to multicollinearity. By using `drop_first=True`, you drop one of the (k) dummy variables, which removes this perfect multicollinearity. This allows the regression model to function correctly.

Example

Consider a categorical variable `Color` with three categories: `Red`, `Green`, and `Blue`. Without `drop_first=True`, you'd get three dummy variables:

- `Color_Red`
- `Color_Green`

- `Color_Blue` If you include all three in your model, you'll have the following relationship for each observation: [\text{Color_Red} + \text{Color_Green} + \text{Color_Blue} = 1] This perfect multicollinearity makes the matrix inversion step in regression problematic. With `drop_first=True`, you'd get:
- `Color_Green`
- `Color_Blue` Here, `Color_Red` is dropped, and the dropped category acts as the reference category.

Summary

Using `drop_first=True`:

- **Prevents Multicollinearity**: Ensures that the model doesn't suffer from perfect multicollinearity, making the regression coefficients estimable.
- **Simplifies Interpretation**: The coefficients of the dummy variables are interpreted relative to the dropped category, providing a clear reference point. Here's a sample code snippet to illustrate:

```
import pandas as pd
# Sample DataFrame
data = {'Color': ['Red', 'Green', 'Blue', 'Green', 'Red', 'Blue']}
df = pd.DataFrame(data)
# Create dummy variables with drop_first=True
df_dummies = pd.get_dummies(df['Color'], drop_first=True)
print(df_dummies)
```

Output:

```
  Blue  Green
0     0      0
1     0      1
2     1      0
3     0      1
4     0      0
5     1      0
```

In this output, `Red` is the reference category (dropped), and the model uses `Blue` and `Green` to capture the effect of these categories relative to `Red`.

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

Answer : To determine which numerical variable has the highest correlation with the target variable, you can look at the pair-plot and calculate the correlation coefficients. A pair-plot visually shows the relationships between variables, and correlation coefficients provide a numerical measure of the strength and direction of these relationships.

- Here Feature `registered` having the highest correlation with the target variable `cnt`

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

Answer : Validating the assumptions of linear regression is crucial to ensure that the model's inferences and predictions are reliable. Here are the main assumptions of linear regression and how you can validate them after building the model on the training set:

Assumptions and Validation Methods

1. **Linearity**:
- **Assumption**: The relationship between the independent variables and the dependent variable is linear.
- **Validation**:
  - **Residual Plots**: Plot the residuals (errors) against the predicted values. The residuals should be randomly scattered around zero.
  - **Partial Regression Plots**: These plots show the relationship between the dependent variable and each independent variable, holding other variables constant.

1. **Homoscedasticity**:
- **Assumption**: The residuals have constant variance at every level of the independent variables.
- **Validation**:
  - **Residual Plots**: Plot the residuals against the predicted values. The spread of residuals should be roughly constant across all levels of the predicted values.
  - **Breusch-Pagan Test**: A statistical test to detect heteroscedasticity.

1. **Independence**:
- **Assumption**: The residuals are independent.
- **Validation**:
  - **Durbin-Watson Test**: This test checks for autocorrelation in the residuals, particularly in time series data.

1. **Normality of Residuals**:
- **Assumption**: The residuals are normally distributed.
- **Validation**:
  - **Q-Q Plot (Quantile-Quantile Plot)**: Plot the quantiles of the residuals against the quantiles of a normal distribution. The points should fall approximately along a straight line.
  - **Shapiro-Wilk Test**: A statistical test to check the normality of the residuals.

1. **No Multicollinearity**:
- **Assumption**: The independent variables are not highly correlated with each other.
- **Validation**:
  - **Variance Inflation Factor (VIF)**: Calculate the VIF for each independent variable. VIF values greater than 10 indicate high multicollinearity.

Q.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 :

1. `Very High` bin from `Registered` Feature
2. `Temp`
3. `Year`
4. `September`

- These are the features contributing significantly towards the demand of the shared bikes.

General Subjective Questions

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

Answer : Linear regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting linear relationship that explains how changes in the independent variables affect the dependent variable. Here's a detailed explanation of the linear regression algorithm:

1. Basic Concept

In simple linear regression, we model the relationship between a dependent variable (Y) and a single independent variable (X) using a linear equation: [Y = \beta_0 + \beta_1 X + \epsilon]

- (Y): Dependent variable
- (X): Independent variable
- (\beta_0): Intercept (constant term)
- (\beta_1): Slope (coefficient of the independent variable)
- (\epsilon): Error term (residual) In multiple linear regression, the model includes multiple independent variables: [Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon]
- (X_1, X_2, \ldots, X_p): Independent variables
- (\beta_1, \beta_2, \ldots, \beta_p): Coefficients of the independent variables

2. Objective

The objective of linear regression is to find the values of (\beta_0, \beta_1, \ldots, \beta_p) that minimize the sum of squared residuals (errors). This is known as the method of least squares.

3. Method of Least Squares

The residual (or error) for each observation is the difference between the observed value and the predicted value

4. Finding the Coefficients

To find the coefficients that minimize SSR, we take partial derivatives of SSR with respect to each (\beta) and set them to zero. This results in a system of normal equations.

5. Assumptions of Linear Regression

Linear regression relies on several assumptions:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The residuals have constant variance at all levels of the independent variables.
4. **Normality**: The residuals of the model are normally distributed.
5. **No Multicollinearity**: Independent variables are not highly correlated with each other.

6. Model Evaluation

After fitting the model, evaluate its performance using:

- **R-squared ((R^2))**: Proportion of variance in the dependent variable that is explained by the independent variables.
- **Adjusted R-squared**: Adjusted for the number of predictors in the model.
- **F-statistic**: Tests whether at least one predictor is significantly related to the dependent variable.
- **p-values**: Tests the null hypothesis that the coefficient of a predictor is zero (no effect).

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

Answer : Anscombe's quartet is a set of four distinct datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Created by the statistician Francis Anscombe in 1973, the quartet demonstrates the importance of graphing data before analyzing it and illustrates how different statistical methods can reveal different aspects of data.

Datasets

Each dataset in Anscombe's quartet consists of eleven (x, y) points. Despite their visual differences, they share the following nearly identical statistical properties:

1. **Mean of x**: 9 (for all datasets)

2. **Mean of y**: 7.50 (for all datasets)
3. **Variance of x**: 11 (for all datasets)
4. **Variance of y**: 4.125 (for all datasets)
5. **Correlation between x and y**: 0.816 (for all datasets)
6. **Linear regression line**: (y = 3 + 0.5x) (for all datasets)

The Four Datasets

7. **Dataset I**: A typical linear relationship with some random noise.
8. **Dataset II**: A perfect linear relationship with an outlier.
9. **Dataset III**: A dataset where the linear relationship is not appropriate (quadratic).
10. **Dataset IV**: A dataset where most points are constant except for one extreme outlier.

Visualization and Analysis

Graphing the four datasets reveals their differences and emphasizes the point that relying solely on summary statistics can be misleading. Below are the visualizations and interpretations of each dataset.

Dataset I

- **Scatter plot**: Shows a roughly linear trend with some noise.
- **Interpretation**: The linear regression model is appropriate.

Dataset II

- **Scatter plot**: Shows a linear relationship with one outlier that significantly affects the regression line.
- **Interpretation**: The outlier has a large influence on the linear model, highlighting the importance of identifying and handling outliers.

Dataset III

- **Scatter plot**: Shows a clear nonlinear (quadratic) relationship.
- **Interpretation**: A linear regression model is not appropriate; a quadratic model would be more suitable.

Dataset IV

- **Scatter plot**: Most points are the same with one extreme outlier.
- **Interpretation**: The outlier skews the results, demonstrating how a single data point can disproportionately affect summary statistics and regression models.

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

Answer : Pearson's r, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of this relationship, ranging from -1 to 1.

Interpretation

- **(r = 1)**: Perfect positive linear relationship.
- **(r = -1)**: Perfect negative linear relationship.
- **(r = 0)**: No linear relationship.
- **(0 < r < 1)**: Positive linear relationship.
- **(-1 < r < 0)**: Negative linear relationship. The closer (r) is to 1 or -1, the stronger the linear relationship. The closer (r) is to 0, the weaker the linear relationship.

Assumptions

1. **Linearity**: The relationship between the variables is linear.
2. **Homoscedasticity**: The variability of one variable is similar across the range of the other variable.
3. **Normality**: The variables are approximately normally distributed (especially important for small sample sizes).

Interpretation of Example

- **Pearson's r**: This will be close to 1, indicating a strong positive linear relationship.
- **P-value**: This tests the null hypothesis that the true correlation is zero. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, suggesting that the observed correlation is statistically significant.

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

Answer : Scaling is a preprocessing technique used in data analysis and machine learning to adjust the range of independent variables or features. The primary purpose of scaling is to ensure that the features contribute equally to the model and improve the performance of many machine learning algorithms.

Why Scaling is Performed

1. **Algorithm Requirements**: Some machine learning algorithms, such as Support Vector Machines (SVM) and K-Nearest Neighbors (KNN), assume that the features are on a similar scale. Without scaling, these algorithms might produce biased results.

2. **Improved Convergence**: Gradient-based algorithms, such as Gradient Descent, can converge more quickly when features are scaled, as it ensures that all features are treated equally in terms of their influence on the cost function.

3. **Interpretability**: In some cases, scaling can make the results more interpretable by normalizing the magnitude of different features.

Types of Scaling

4. **Normalization (Min-Max Scaling)**:

- **Definition**: Rescales the feature values to a fixed range, typically [0, 1] or [-1, 1].
- **Usage**: Useful when the data has a known minimum and maximum, or when the distribution of the data is not Gaussian and you want to retain the original distribution's shape.
- **Example**: If you have feature values ranging from 50 to 200, normalization will rescale these values to the range [0, 1].

1. **Standardization (Z-Score Scaling)**:

- **Definition**: Transforms the feature values so that they have a mean of 0 and a standard deviation of 1.
- **Usage**: Useful when the data follows a Gaussian distribution (or approximately so) and when you want to standardize the data based on its statistical properties.
- **Example**: If you have feature values with a mean of 100 and a standard deviation of 20, standardization will rescale these values to have a mean of 0 and a standard deviation of 1.

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

Answer : The Variance Inflation Factor (VIF) is a measure used to detect the presence of multicollinearity in a regression analysis. Specifically, it quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.

Understanding VIF

The VIF for a predictor variable 1 / (1 - R-Squared)

Why VIF Can Be Infinite

VIF becomes infinite when R-Squared = 1. This occurs when there is perfect multicollinearity, meaning that the predictor X_i is a perfect linear combination of one or more other predictors in the model. In other words, X_i can be expressed exactly as a sum of other predictors, which makes it impossible to uniquely estimate the coefficients of the predictors.

Causes of Perfect Multicollinearity

1. **Duplicate Variables**: Including the same variable more than once in the model.
2. **Linear Dependencies**: One predictor is a perfect linear combination of other predictors (e.g., if (X_3 = 2X_1 + 3X_2)).

3. **Interaction Terms**: Including interaction terms without including the main effects can sometimes lead to perfect multicollinearity.
4. **Dummy Variable Trap**: Including all dummy variables for a categorical predictor without dropping one category (reference category).

Handling Perfect Multicollinearity

To address perfect multicollinearity:

1. **Remove Redundant Predictors**: Identify and remove one of the predictors causing the perfect multicollinearity.
2. **Combine Predictors**: If predictors are highly correlated, consider combining them into a single predictor (e.g., using Principal Component Analysis).
3. **Use Regularization**: Techniques like Ridge Regression can mitigate the effects of multicollinearity by adding a penalty term to the regression.

Q.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 (Quantile-Quantile) plot is a graphical tool used to assess whether a set of data follows a given distribution, typically the normal distribution. It compares the quantiles of the data with the quantiles of a specified theoretical distribution.

Construction of a Q-Q Plot

1. **Sort the Data**: Arrange the sample data in ascending order.
2. **Calculate Theoretical Quantiles**: Compute the quantiles of the specified theoretical distribution (e.g., the normal distribution).
3. **Plot the Points**: Plot the quantiles of the sample data against the theoretical quantiles.
4. **Interpretation**:

- If the data follow the specified distribution, the points will approximately lie on a straight line.
- Deviations from the straight line indicate departures from the specified distribution.

Use and Importance in Linear Regression

In linear regression, several assumptions must be satisfied to ensure the validity of the model. One of these assumptions is that the residuals (errors) are normally distributed. The Q-Q plot is a crucial diagnostic tool for checking this assumption.

Steps to Use a Q-Q Plot in Linear Regression

1. **Fit the Linear Regression Model**: Estimate the model parameters using the training data.
2. **Obtain Residuals**: Calculate the residuals, which are the differences between the observed values and the predicted values from the regression model.
3. **Create the Q-Q Plot**: Plot the quantiles of the residuals against the quantiles of a normal distribution.

Interpretation

- **Straight Line**: If the residuals follow a normal distribution, the points in the Q-Q plot will lie on or near the 45-degree reference line.
- **S-shaped Curve**: This indicates heavy tails in the distribution of residuals, suggesting that the residuals have more extreme values than expected under normality.
- **Inverted S-shaped Curve**: This indicates light tails in the distribution of residuals, suggesting that the residuals have fewer extreme values than expected under normality.
- **Other Patterns**: Any systematic deviation from the reference line suggests that the residuals do not follow a normal distribution, which may violate the assumptions of linear regression.

Importance

1. **Assumption Checking**: Ensuring that the residuals are normally distributed is crucial for the validity of hypothesis tests (e.g., t-tests for regression coefficients) and confidence intervals.
2. **Model Diagnostics**: Identifying non-normality in the residuals can prompt further investigation into potential issues, such as outliers, missing variables, or incorrect model specification.
3. **Transformations**: If residuals are not normally distributed, transformations of the dependent variable (e.g., log, square root) or the use of robust regression methods might be necessary.