Linear Regression 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?

Interpretation of Coefficients

1. **Intercept (const)**
    - **Coefficient: 2590.3541**
    - **Interpretation: When all predictors are zero, the expected value of the dependent variable is approximately 2590.35.**
2. **Season**
    - **Coefficient: 1234.7963**
    - **Interpretation: For each unit increase in the 'season' variable, the dependent variable increases by approximately 1234.80 units, holding all other factors constant. This effect is statistically significant ($p < 0.05$).**
3. **Month (mnth)**
    - **Coefficient: -148.8346**
    - **Interpretation: For each unit increase in the 'mnth' variable, the dependent variable decreases by approximately 148.83 units, but this effect is not statistically significant ($p > 0.05$).**
4. **Weekday**
    - **Coefficient: 391.5961**
    - **Interpretation: For each unit increase in the 'weekday' variable, the dependent variable increases by approximately 391.60 units. This effect is statistically significant ($p < 0.05$).**
5. **Weather Situation (weathersit)**
    - **Coefficient: -688.7959**
    - **Interpretation: For each unit increase in the 'weathersit' variable, the dependent variable decreases by approximately 688.80 units. This effect is statistically significant ($p < 0.05$).**
6. **Holiday**
    - **Coefficient: -731.0819**
    - **Interpretation: Being a holiday is associated with a decrease of approximately 731.08 units in the dependent variable. This effect is statistically significant ($p < 0.05$).**
7. **Year (yr)**

- **Coefficient: 2044.1048**
- **Interpretation: For each unit increase in the 'yr' variable, the dependent variable increases by approximately 2044.10 units. This effect is statistically significant ($p < 0.05$).**

8. **Temperature (temp)**
   - **Coefficient: 4141.5319**
   - **Interpretation: For each unit increase in the 'temp' variable, the dependent variable increases by approximately 4141.53 units. This effect is statistically significant ($p < 0.05$).**
9. **Humidity (hum)**
   - **Coefficient: -802.4864**
   - **Interpretation: For each unit increase in the 'hum' variable, the dependent variable decreases by approximately 802.49 units. This effect is statistically significant ($p < 0.05$).**
10. **Windspeed**
    - **Coefficient: -47.0984**
    - **Interpretation: For each unit increase in the 'windspeed' variable, the dependent variable decreases by approximately 47.10 units. This effect is statistically significant ($p < 0.05$).**

Statistical Significance

- **Significant Predictors: Season, weekday, weathersit, holiday, year, temperature, humidity, and windspeed all have p-values less than 0.05, indicating that their effects on the dependent variable are statistically significant.**
- **Non-Significant Predictor: Month (mnth) has a p-value greater than 0.05, indicating that its effect is not statistically significant.**

Confidence Intervals

- **The 95% confidence intervals for each coefficient provide a range within which we can be 95% confident that the true coefficient lies. For example, for the 'season' variable, the true coefficient is likely between 825.047 and 1644.545.**

Conclusion

From this regression analysis, we can conclude that several factors significantly influence the dependent variable. Specifically, season, weekday, weathersit, holiday, year, temperature, humidity, and windspeed all have notable impacts, while the month does not show a significant effect. This kind of analysis helps in understanding which factors are most influential and can guide decision-making and further analysis.

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

- **Avoid Multicollinearity: Dropping one dummy variable prevents perfect multicollinearity.**
- **Clear Interpretation: It provides a baseline category for clear interpretation of the coefficients.**
- **Model Stability: Improves the stability and reliability of the regression estimates.**

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

Temp and a_temp have highest positive correlation 0.65 with the target variable cnt.

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

Validating the assumptions of linear regression is crucial to ensure the reliability and validity of the model. Here are the steps to validate these assumptions:

1. Linearity

- **Assumption**: The relationship between the independent variables and the dependent variable is linear.
- **Validation**: Plot the residuals vs. fitted values.
    - **How to check**: If the plot shows a random scatter (no pattern), the linearity assumption is likely satisfied.

2. Independence

- **Assumption**: The residuals are independent.
- **Validation**: Check for any patterns or correlations in the residuals.
    - **How to check**: Plot the residuals over time or the order of observations to look for any patterns.

3. Homoscedasticity

- **Assumption**: The residuals have constant variance (homoscedasticity).
- **Validation**: Plot the residuals vs. fitted values.
    - **How to check**: If the spread of residuals is consistent across all levels of fitted values, the homoscedasticity assumption is met. A funnel shape indicates heteroscedasticity.

4. Normality of Residuals

- **Assumption**: The residuals are normally distributed.
- **Validation**: Use a Q-Q plot and a histogram of residuals.
    - **How to check**: The Q-Q plot should show the residuals falling approximately along the reference line. The histogram should show a roughly bell-shaped distribution.

5. No Multicollinearity

- **Assumption**: The independent variables are not highly correlated.
- **Validation**: Calculate the Variance Inflation Factor (VIF) for each predictor.
    - **How to check**: VIF values greater than 10 indicate significant multicollinearity that needs to be addressed.

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

To identify the top 3 features, we look at the predictors with the highest absolute t-values and the smallest p-values:

1. **Year (yr)**
    - **Coefficient**: 2044.1048
    - **t-value**: 25.935
    - **P>|t|**: 0.000
    - **Interpretation**: Year has a very large positive effect on bike demand and is highly significant. This implies that bike demand has increased significantly over time.
2. **Temperature (temp)**
    - **Coefficient**: 4141.5319
    - **t-value**: 21.725
    - **P>|t|**: 0.000
    - **Interpretation**: Temperature has a large positive effect on bike demand and is highly significant. Warmer temperatures are associated with higher bike demand.
3. **Weather Situation (weathersit)**
    - **Coefficient**: -688.7959
    - **t-value**: -7.533
    - **P>|t|**: 0.000
    - **Interpretation**: Weather situation has a significant negative effect on bike demand. Poor weather conditions (e.g., rain or snow) decrease bike demand.

The top 3 features contributing significantly towards explaining the demand for shared bikes, based on their statistical significance and effect size, are:

1. **Year (yr)**
2. **Temperature (temp)**
3. **Weather Situation (weathersit)**

These features are crucial for understanding and predicting the demand for shared bikes, and any strategies or policies aimed at improving bike usage should consider these factors.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables.

Basic Concept

Linear regression aims to find the best-fitting straight line through the data points. This line, called the regression line, is used to predict the value of the dependent variable based on the values of the independent variables.

Mathematical Representation

For a simple linear regression (one independent variable), the model is represented as:

$y=\beta 0+\beta 1x+\epsilon$

y=β0 +β1x+ϵ

where:y is the dependent variable.

x is the independent variable.

β0 is the intercept of the regression line.

β1 is the slope of the regression line.

ϵ is the error term (the difference between the actual and predicted values).

For multiple linear regression (multiple independent variables), the model is represented as:

$y=\beta 0+\beta 1x1+\beta 2x2+...+\beta nxn+\epsilon$

y=β0 +β1x1 +β2x2 +...+βnxn +ϵ

where:

y is the dependent variable.

$x1,x2,...,xn$ are the independent variables.

$\beta 0,\beta 1,\beta 2,...,\beta n$ are the coefficients.

ϵ is the error term.

Assumptions

Linear regression relies on several key assumptions:

Linearity: The relationship between the independent and dependent variables should be linear.

Independence: Observations should be independent of each other.

Homoscedasticity: The variance of error terms should be constant across all levels of the independent variables.

Normality: The error terms should be normally distributed.

Estimation of Coefficients

The coefficients ($\beta 0, \beta 1, \ldots, \beta n$) are estimated using the method of least squares. This method minimizes the sum of the squared differences between the observed values and the values predicted by the linear model.

The objective is to find the coefficients that minimize the cost function:

$J(\beta 0, \beta 1, \ldots, \beta n) = \sum i=1m(yi - y\hat{}i)2$

J(β0,β1 ,…,β n)=∑ i=1m (y i − y^ i) 2 where:

m is the number of observations.

yi is the actual value of the dependent variable for the i-th observation.

$y\hat{}i$ is the predicted value of the dependent variable for the i-th observation.

Matrix Formulation

In matrix form, the linear regression model can be expressed as:

$y = X\beta + \epsilon$

y=Xβ+ϵ

where:y is an m×1 vector of observed values.

X is an $m \times (n+1)$ matrix of input features (including a column of ones for the intercept).

β is a (n+1)×1 vector of coefficients.

ϵ is an m×1 vector of errors.

The coefficients are estimated using the normal equation:

$\beta = (XTX) - 1XTy$

Model Evaluation

The performance of the linear regression model can be evaluated using several metrics:

R-squared (Coefficient of Determination): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. Ranges from 0 to 1.

$R2 = 1 - (\sum i=1m(yi - y\hat{}i)2 / \sum i=1m(yi - y^{-})2)$

Mean Squared Error (MSE): The average of the squared differences between the observed and predicted values.

MSE=1/m∑i=1m(yi−y^i)2

Root Mean Squared Error (RMSE): The square root of the MSE.

$RMSE$=square root of MSE

Advantages and Disadvantages

Advantages:

Simple to implement and understand.

Computationally efficient.

Provides interpretable results (coefficients).

Disadvantages:

Assumes a linear relationship, which may not always be true.

Sensitive to outliers.

Can be affected by multicollinearity in the case of multiple linear regression.

Conclusion

Linear regression is a powerful tool for modeling relationships between variables. By understanding its assumptions, estimation methods, and evaluation metrics, one can effectively apply this technique to various predictive modeling tasks.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics yet appear very different when graphed. These datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analysing it and to show how statistical properties can be misleading without visual inspection.

Descriptive Statistics

Each of the four datasets in Anscombe's quartet shares the following characteristics:

- The mean of xxx values is 9.
- The mean of yyy values is approximately 7.50.
- The variance of xxx values is 11.
- The variance of yyy values is approximately 4.12.

- The correlation coefficient between xxx and yyy is approximately 0.816.
- The linear regression line is y=3.00+0.500xy = 3.00 + 0.500xy=3.00+0.500x.

Importance of Anscombe's Quartet

Anscombe's quartet emphasizes several key points in data analysis:

1. **Visual Inspection**: Always visualize data before performing any analysis. Graphical representation can reveal patterns, trends, and anomalies that simple statistics might miss.
2. **Outliers**: Identifying outliers is crucial, as they can significantly influence statistical results and lead to misleading interpretations.
3. **Model Selection**: The apparent relationship between variables might require different models (linear, polynomial, etc.) for accurate representation.
4. **Descriptive Statistics Limitations**: Relying solely on descriptive statistics without considering the data's distribution and characteristics can be misleading.

Conclusion

Anscombe's quartet serves as a powerful reminder that statistical properties are not always sufficient to understand the data. Visual exploration and thorough analysis are essential to accurately interpret and model data.

3. What is Pearson's R?

Pearson's r, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two continuous variables. It quantifies the degree to which the variables are linearly related, indicating both the strength and direction of the relationship.

Key Features of Pearson's r:

1. **Range**: Pearson's r ranges from -1 to 1.
    - r=1indicates a perfect positive linear relationship.
    - r=−1 indicates a perfect negative linear relationship.
    - r=0 indicates no linear relationship.
2. **Direction**:
    - A positive r value indicates that as one variable increases, the other variable tends to increase.
    - A negative r value indicates that as one variable increases, the other variable tends to decrease.
3. **Strength**:
    - The closer r is to 1 or -1, the stronger the linear relationship.
    - The closer r is to 0, the weaker the linear relationship.

Formula:

The formula for Pearson's r is:

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

where:

- x_i and y_i are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

Interpretation:

- **Strong positive correlation**: r is close to 1.
- **Strong negative correlation**: r is close to -1.
- **Weak or no correlation**: r is close to 0.

Example:

Suppose you have the following data:

X Y

1 2

2 4

3 6

4 8

5 10

For this data set, Pearson's r would be 1, indicating a perfect positive linear relationship.

Use Cases:

- **Social Sciences**: To determine the relationship between variables like income and education level.
- **Medicine**: To study the correlation between dosage of a drug and its effectiveness.
- **Finance**: To explore the relationship between stock prices and economic indicators.

Pearson's r is a widely used and powerful tool for summarizing the linear relationship between two continuous variables in various fields.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of adjusting the range and distribution of data values so that they fall within a specified range or have certain statistical properties. This process is crucial in data preprocessing, particularly in machine learning and statistical modelling, to ensure that different features contribute appropriately to the model and to improve the convergence of optimization algorithms.

Different types of scaling are:
a. Min Max Scaling (Normalization).
b. Standardisation
c. Robust scaling
d. Log Scaling
e. Max Abs scaling

Why is the scaling performed?

- **Equal Contribution**: Ensures that all features contribute equally to the model, particularly important for algorithms that rely on distance measurements, such as k-nearest neighbours (KNN) and support vector machines (SVM).
- **Improved Convergence**: Helps in accelerating the convergence of gradient descent optimization algorithms, as features with similar scales result in more stable and faster training.
- **Interpretability**: Makes the model parameters more interpretable, particularly in linear models where coefficients are influenced by the scale of the features

What is the difference between Normalisation and Standardisation?

Normalization and standardization are two techniques used to transform data into a common scale. Normalization is a technique used to scale numerical data in the range of 0 to 1. This technique is useful when the distribution of the data is not known or when the data is not normally distributed.

On the other hand, standardization is a technique used to transform data into a standard normal distribution. This technique is useful when the distribution of the data is known and when the data is normally distributed. Both techniques have different applications, and choosing the right technique based on the data and the problem you're trying to solve is important.

Example

Suppose you have the following data:

Height (cm)	**Weight (kg)**
150	55
160	60
170	65
180	70
190	75

- **Min-Max Scaling**:
    - Height: (150−150)/(190−150)=0, (160−150)/(190−150)=0.25, ...
    - Weight: (55−55)/(75−55)=0, (60−55)/(75−55)=0.25, ...

- **Standardization**:
    - Compute the mean and standard deviation for height and weight.
    - Height: (150−170)/15.81=−1.27, (160−170)/15.81=−0.63, ...
    - Weight: (55−65)/7.91=−1.27\, (60−65)/7.91=−0.63, ...

5. You might have observed sometimes the VIF is infinite. Why does this happen?

Variance Inflation Factor (VIF) measures the extent of multicollinearity in a set of multiple regression variables. Specifically, VIF quantifies how much the variance of a regression coefficient is inflated due to the correlation with other predictors in the model. VIF values are calculated for each predictor variable and a high VIF indicates a high level of multicollinearity.

Why VIF Can Be Infinite

A VIF value becomes infinite when there is perfect multicollinearity between one predictor variable and the others in the regression model. Perfect multicollinearity means that one predictor variable can be exactly predicted by a linear combination of the other predictor variables.

Understanding the Calculation

The VIF for a predictor variable Xi is calculated as:

VIF(Xi)=1/(1−Ri^2)

where Ri^2 is the coefficient of determination from a regression of Xi on all the other predictor variables.

- **Ri^2 close to 1**: This indicates a high degree of multicollinearity, as most of the variability in XiX_iXi can be explained by the other predictors.
- **Ri^2=1**: This implies perfect multicollinearity, where Xi is a perfect linear combination of the other predictors.

In such a case, the denominator 1−Ri^2 becomes zero, leading to a VIF value that is mathematically infinite.

Consequences of Infinite VIF

1. **Unstable Estimates**: Coefficients of the regression model become unstable and highly sensitive to small changes in the model.
2. **Inflated Standard Errors**: The standard errors of the coefficients become inflated, leading to wider confidence intervals and less reliable hypothesis tests.
3. **Model Identification Issues**: Perfect multicollinearity makes it impossible to uniquely estimate the regression coefficients. The model cannot distinguish the individual effect of correlated predictors.

Example

Consider a regression model with predictors X1 and X2, where X2=2X1. Here, X2 is perfectly collinear with X1:

- Running a regression of X2 on X1 will yield R2^2=1..

- Consequently, VIF(X2)=1/(1−1)=∞.

Addressing Infinite VIF

1. **Remove Redundant Predictors**: If predictors are perfectly collinear, one of the collinear variables should be removed from the model.
2. **Combine Predictors**: Create composite variables or use techniques like Principal Component Analysis (PCA) to combine correlated predictors into a single predictor.
3. **Regularization Techniques**: Methods like Ridge Regression can help mitigate the effects of multicollinearity by adding a penalty term to the regression.

Understanding and addressing infinite VIF values is crucial for building reliable and interpretable regression models.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a data set follows a specific theoretical distribution, typically the normal distribution. It compares the quantiles of the sample data with the quantiles of a theoretical distribution. In the context of linear regression, Q-Q plots are primarily used to check the normality assumption of the residuals.

Key Components of a Q-Q Plot

1. **Theoretical Quantiles**: These are quantiles from the theoretical distribution you are comparing your data against (e.g., the normal distribution).
2. **Sample Quantiles**: These are the quantiles from your sample data.

How to Interpret a Q-Q Plot

- **Straight Line**: If the points in the Q-Q plot lie approximately along a straight line, it indicates that the sample data follows the theoretical distribution closely.
- **Deviations from Line**:
    - **Heavy Tails**: If the points form an S-shape or deviate significantly at the ends, it suggests that the data has heavier tails than the theoretical distribution.
    - **Light Tails**: If the points form an inverted S-shape, it suggests that the data has lighter tails.
    - **Skewness**: Systematic deviations from the line in one direction (upwards or downwards) can indicate skewness in the data.

Importance of Q-Q Plots in Linear Regression

1. **Assessing Normality of Residuals**:
    - One of the key assumptions in linear regression is that the residuals (errors) are normally distributed.

- A Q-Q plot helps in visualizing whether the residuals deviate from normality. If the residuals fall approximately along the 45-degree reference line in the Q-Q plot, it suggests they are normally distributed.

2. **Identifying Outliers**:
    - Q-Q plots can help identify outliers that may not be apparent in other residual diagnostics. Points that deviate significantly from the reference line might indicate outliers.
3. **Model Validation**:
    - Checking the normality of residuals is crucial for valid hypothesis testing and confidence interval estimation. If residuals are not normally distributed, the standard errors, confidence intervals, and hypothesis tests may not be reliable.

Example

Suppose you fit a linear regression model and obtain the residuals. You create a Q-Q plot of the residuals to check for normality. Here's what you might observe:

- **Good Fit**: Residuals lie close to the straight line, indicating they are normally distributed.
- **Deviations**: Residuals deviate from the line, suggesting potential non-normality, which might require transformations or a different model.

Q-Q plots are a powerful diagnostic tool in linear regression for assessing the normality of residuals. Ensuring residuals are normally distributed helps validate the regression model's assumptions, leading to more reliable and interpretable results.