Table of Contents

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)

The categorical variables such as 'season', 'yr', 'holiday', 'weekday' and 'weathersit' significantly impact the dependent variable, which is the demand for shared bikes ('cnt').

For example, 'yr' and 'season' were found to have a considerable influence, suggesting that bike demand varies with different seasons and has shown year-over-year changes. 'Holiday' and 'weekday' also impact demand, indicating variations in bike usage patterns during holidays and across different days of the week.

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

Using drop_first=True prevents multicollinearity in the model by eliminating one dummy variable from each categorical feature.

This approach omits the first level/category of the feature, reducing the number of dummy variables by one for each categorical variable. It helps in avoiding the dummy variable trap, where highly correlated or duplicate variables can distort the results and interpretations of the model.

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

Features ‘temp’ and ‘atemp’ showed high correlation (0.63 each) with the target variable ('cnt'), but ‘temp’ was removed as feature as it closely resembles ‘atemp’.

This suggests that as the temperature increases, the demand for shared bikes tends to increase due to favourable weather conditions.

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

After building the model:

Multicollinearity: Variance Inflation Factor (VIF) analysis was performed. Higher VIF values greater than 10 (also based on the business acumen) indicated significant multicollinearity which were removed.

Homoscedasticity (Equal Variance of Residuals): To validate this assumption, a scatter plot of the residuals versus the predicted values was created. Residuals randomly dispersed around zero, showing no clear pattern.

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

The top three features contributing significantly to the model were 'yr' (year), 'season' and 'atemp' (feeling temperature).

- 'yr' (Year) being significant indicates an increasing trend in bike demand over the years.
- 'atemp' (Feeling Temparature)suggests a strong positive correlation with bike demand.
- However, ‘holiday’ negatively impacts the Bike demand as very few show interest to rent a Bike during Holidays.

General Subjective Questions

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

Linear regression models the relationship between the dependent variable (Y) and one or more independent variables (X) by fitting a linear equation to observed data.

In statistics and machine learning it is used for predicting a continuous target variable based on one or more predictor variables.

Key Steps in Linear Regression:

Apart from Data Collection, Cleaning and Feature Engineering, the key aspects are,

1. Model Fitting:

The algorithm calculates the coefficients that minimize the sum of the squared differences between the observed and predicted values (OLS approach)

2. Making Predictions:

Once the model is fitted, it can be used to make predictions. For a given set of independent variables, the model predicts the corresponding value of the dependent variable.

3. Evaluating the Model:

Common metrics for evaluating a linear regression model include R-squared, Adjusted R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

R-squared measures how well the observed values of the dependent variable are predicted by the model.

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

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, but have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. This quartet was created in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

Key Characteristics:

1. Similar Statistical Properties: All four datasets in Anscombe's quartet have nearly identical mean, variance, correlation, and linear regression lines (both slope and intercept) when applied to 'x' and 'y' values. This suggests that they share the same statistical properties.

2. Different Visual Distributions: When each dataset is graphed, they each show a distinctly different relationship between 'x' and 'y'. This discrepancy highlights the limitation of relying solely on summary statistics and the importance of visualizing data.

The Four Datasets:

Dataset I: Shows a simple linear relationship between 'x' and 'y' that appears to be a good fit for a linear model.
Dataset II: Demonstrates a clear non-linear relationship (curvilinear pattern). A linear model is not a good fit for this dataset.
Dataset III: Contains an outlier that affects the slope of the regression line. Without the outlier, the data appears to be a good fit for a linear model.
Dataset IV: Shows a relationship where 'x' values are constant for all but one point. The single outlier drives the linear regression line, highlighting the influence of outliers in regression analysis.

Implications:

Graphical Analysis Is Crucial: Anscombe's quartet emphasizes the importance of graphically analyzing data before using them in statistical models. It serves as a powerful reminder that the same statistical properties can lead to misleading conclusions if the actual distribution and pattern of the data are not considered.

Limitations of Summary Statistics: Solely relying on summary statistics can be misleading. Two datasets with the same statistical properties can represent very different relationships.

Impact of Outliers: The quartet demonstrates how outliers can significantly influence the results of statistical analyses, particularly in regression.

What is Pearson's R? (3 marks)

Pearson's R, Or Pearson correlation coefficient is a measure of the linear correlation (linear dependence) between two variables X and Y. It's a widely used statistical method that quantifies the degree to which a relationship is linear and the strength of this linear relationship.

Pearson's R is defined as the covariance of the two variables divided by the product of their standard deviations.

Interpretation:

- The value of r ranges between -1 and 1.
- A value of 1 implies a perfect positive linear relationship: as one variable increases, the other variable increases at a constant rate.
- A value of -1 implies a perfect negative linear relationship: as one variable increases, the other decreases at a constant rate.
- A value of 0 implies no linear correlation between the variables.
- Values close to 1 or -1 indicate a strong linear relationship, while values close to 0 indicate a weak linear relationship.

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

Scaling is a method used in data preprocessing to standardize the range of independent variables or features of data. In machine learning, it's a crucial step because models often assume that all features are centered around zero and have a similar variance.

Scaling helps in making the measured units consistent, avoid skewness and helps models to converge faster.

Normalized Scaling vs Standardized Scaling:

- Normalized Scaling (Min-Max Scaling):

It's useful when we want needed values in a bounded interval. However, it doesn't handle outliers well. Outliers can significantly affect the min and max values and consequently the scaling.

- Standardized Scaling (Z-score Normalization):

Standardization rescales the feature to have a mean of 0 and a standard deviation of 1.

Standardization maintains useful information about outliers and makes the algorithm less sensitive to them compared to min-max scaling.

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

The Variance Inflation Factor (VIF) is a measure used to detect the presence and severity of multicollinearity in a regression analysis.

Reasons for Infinite VIF:

Perfect or Near-Perfect Multicollinearity: If a predictor variable is a perfect or near-perfect linear combination of other predictor variables, the R-squared in

the calculation of VIF approaches 1. As R-squared gets closer to 1, the denominator in the VIF formula approaches zero causing the VIF to become extremely large or approach infinity.

Redundant Variables: This can happen if you include variables that are linear combinations of each other. For example, if one variable is the sum or difference of two or more other variables, or if one variable is an exact copy of another.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. In the context of linear regression, a Q-Q plot is typically used to analyze whether the residuals (differences between observed and predicted values) of the model are normally distributed.

After fitting a linear regression model, you calculate the residuals (the difference between the observed values and the values predicted by the model).

If the points in the Q-Q plot lie on or close to a straight line, it suggests that the residuals have a distribution close to normal.