Linear Regression Subjective Questions

Linh Pham – DS72

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?

Answer: Weather, season, year, and month significantly impact bike rentals, making them key predictors. Rentals peak in clear weather, fall, and summer, with the highest demand from July to October, especially in September. The second year (yr=1) sees significantly higher rentals than the first. In contrast, weekday and working day have minimal influence, making them less critical for prediction.

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

Answer: Using drop_first=True in linear regression helps prevent multicollinearity by removing redundancy in dummy variables. When all dummy variables are included, one can be predicted from the others, causing perfect multicollinearity and making the model unstable. Dropping one dummy variable sets it as the baseline category, allowing the remaining variables' coefficients to show their impact relative to this reference group.

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

Answer: The registered variable has the strongest correlation with the target variable cnt (target variable), as seen in the clear positive linear relationship in the scatter plot. Following that, casual, temp, and atemp also show notable correlations with cnt.

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

Answer: After building the model, I validated the assumptions of Linear Regression through the following checks:

- Checked the Variance Inflation Factor (VIF) for the independent variables in the final model to ensure multicollinearity was removed (VIF < 5).
- Plotted the histogram of residuals to verify if they followed a normal distribution.
- Examined the Residual Plot to ensure that residuals were randomly scattered around zero, with no clear pattern, indicating a good fit for linear regression.

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

Answer: The year (yr), temperature (temp), and winter windspeed are the most significant factors influencing bike-sharing demand.

- yr (Year) – Coefficient: 0.2332, p-value: 0.000. This has a positive impact on bike-sharing demand, indicating that rentals increase each year compared to the previous one.
- temp (Temperature) – Coefficient: 0.5464, p-value: 0.000. Higher temperatures are strongly correlated with an increase in bike rentals, suggesting that warmer weather leads to greater demand for bike-sharing.
- windspeed – Coefficient: -0.1427, p-value: 0.000, which has a negative impact, meaning stronger wind speeds decrease bike demand.

General Subjective Questions

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

Answer: Linear Regression is a machine learning algorithm used to predict a continuous value (y) based on one or more input features (X). It finds the best-fitting straight line that describes the relationship between the input and output.
For example, predicting bike rentals (cnt) based on temperature (temp), wind speed (windspeed), and season (season).
The formula:

$$y = w_0 + w_1x_1 + w_2x_2 + ... + w_nx_n + \varepsilon$$

The goal of Linear Regression is to find the best values for w (weights) so that the predicted values are as close as possible to the actual values.It does this by minimizing the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

For the model to work well, these assumptions should be met:

- **Linearity:** The relationship between input and output is linear.
- **No Multicollinearity:** Input variables should not be highly correlated.
- **Homoscedasticity:** The variance of errors should be constant.
- **Normality of Residuals:** Errors should follow a normal distribution.
- **Independence of Errors:** Residuals should not be correlated.

To measure how well the model fits the data, we use:

- **R² Score:** Measures how much variance in y is explained by X (closer to 1 is better).
- **Residual Plots:** Ensures errors are randomly distributed.
- **Variance Inflation Factor (VIF):** Detects multicollinearity.

General Subjective Questions

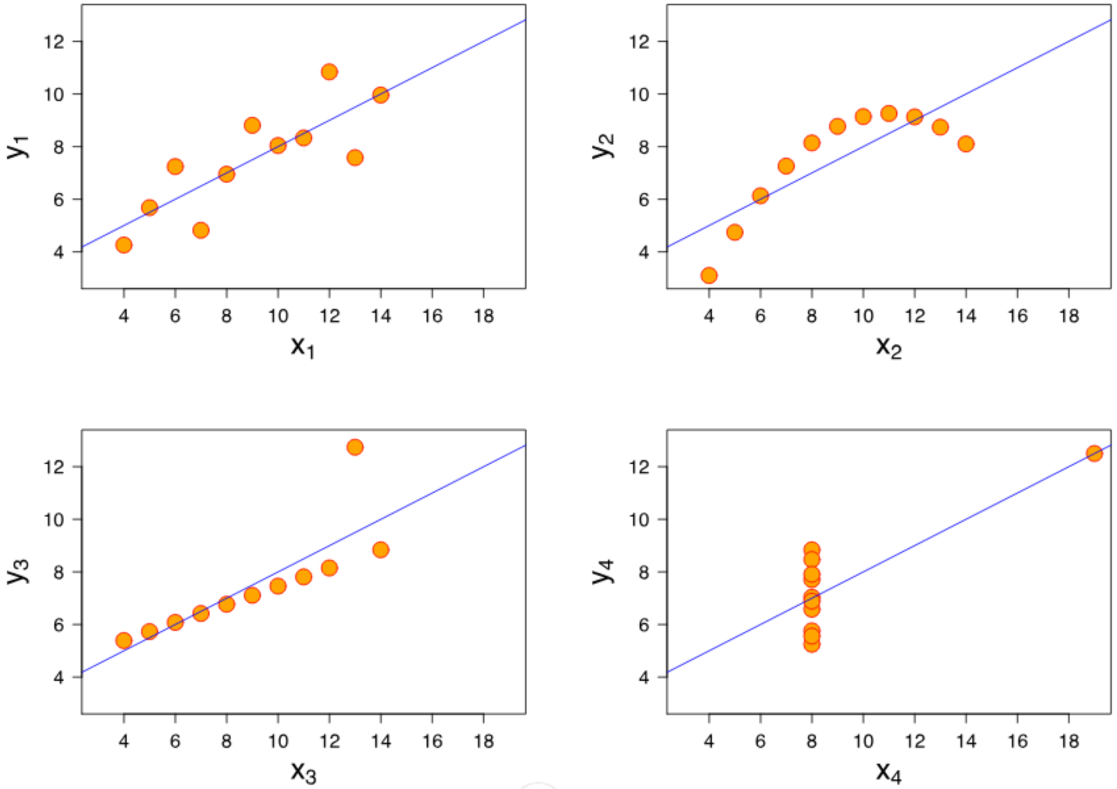
2. Explain the Anscombe's quartet in detail

Answer: Anscombe's Quartet is a set of four datasets that look very different when plotted but have almost identical statistical properties, such as:

- Mean (average) of x and y
- Variance (spread of data)
- Correlation (relationship between x and y)
- Linear regression equation

Here's what each dataset looks like when plotted:

- Dataset 1 (X1,Y1): Data follows a linear pattern, making it a good fit for linear regression.
- Dataset 2 (X2,Y2): Data is curved, so a linear model is not appropriate.
- Dataset 3 (X3,Y3): One outlier is affecting the regression line.
- Dataset 4 (X4,Y4): Most points are the same, except for one extreme outlier that skews the regression.



Using Anscombe's Quartet helps us:

- **Avoid misleading statistics:** Visualizing data reveals differences that summary statistics (mean, variance, correlation) may not show.
- **Detect non-linear relationships** : Some datasets in the quartet have curved patterns (e.g., Dataset 2), making linear regression inappropriate.
- **Identify outliers** : Certain datasets contain extreme values that significantly impact analysis and model accuracy.
- **Ensure the right model selection:** If data is non-linear or affected by outliers, linear regression may not be the best choice.

General Subjective Questions

3. What is Pearson’s R?

Answer: Pearson’s R is a statistical measure used to evaluate the strength and direction of the linear relationship between two **continuous variables**.

Range of values: Pearson’s R ranges from -1 to 1.

- R close to 1: A strong positive linear relationship (as one variable increases, the other also increases).
- R close to -1: A strong negative linear relationship (as one variable increases, the other decreases).

When to use it:

- To measure the strength of a linear relationship between two continuous variables.
- When the data has few or no outliers, as outliers can distort the result.
- To check relationships between variables before building a model (e.g., linear regression).

Formular of Pearson’s R:

$$R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

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

Answer: Scaling is the process of transforming data values to bring them to the same scale or range. This process doesn’t change the distribution or information within the data; it only changes how the data is represented.

Why is scaling performed?

- Improves model performance – Prevents large values from dominating smaller ones.
- Speeds up training – Makes optimization algorithms like gradient descent faster.
- Removes unit bias – Ensures fair treatment of features with different units (e.g., cm vs. kg).

Difference between normalized scaling and standardized scaling

- Standardized scaling: Changes the data so it has a mean (average) of 0 and a standard deviation of 1. This makes the data follow a normal distribution. Values can be negative or positive, and this method is best when the data is normally distributed or close to it.
- Normalized Scaling changes the data so all values fit within a range, usually between 0 and 1. This method is useful when the scale of data is important, like in Neural Networks, where inputs need to be small and consistent. However, it’s sensitive to outliers since extreme values can stretch the range.

General Subjective Questions

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

When the VIF value equals infinity, it happens because there is a perfect linear relationship between that variable and the other independent variables in the model, leading to severe multicollinearity. In simple terms, this means that the variable can be 100% predicted using the other variables.

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

A Q-Q Plot (Quantile-Quantile Plot) compares the distribution of actual data to a theoretical distribution (usually normal). In linear regression, it checks if the residuals (errors) follow a normal distribution.

How to Interpret a Q-Q Plot?

- If residuals are normal: Points align closely with the diagonal line (y = x).
- If residuals are not normal: Points deviate from the diagonal line, indicating potential issues like outliers or a poor model fit.

Why is it Important in Linear Regression?

- Checks a key assumption: Verifies that residuals are normally distributed, ensuring reliable results.
- Identifies issues: Detects outliers or non-linear relationships that could harm the model.
- Guides adjustments: Suggests data transformations (e.g., log) or alternative models (e.g., Polynomial Regression).