

Linear Regression Assignment Questions

1. Effect of Categorical Variables on Bike Demand

Based on the analysis of categorical variables (season, weather, and month), below are the observations:

- Season Impact: Spring shows the lowest demand (-0.258 Coeff), while Summer and Winter have negligible or mild negative effects.
- Weather Impact: Rainy conditions Significantly impacts Bikes Rental/sharing (-0.302 Coeff).
- Month Impact: show seasonality. Months such as January and September have mild effects on bike demand .

Overall, categorical variables contribute significantly to explaining variations in bike rentals.

2. Importance of Using `drop_first=True` for Dummy Variables

- **`drop_first=True`**, drops the first category of the variable that is used.
- When all dummy columns are included, they cause multicollinearity, which might lead to unstable coefficients and misleading p-values.
- Dropping the first dummy ensures a well-defined and stable regression model.

3. Highest Correlated Numerical Variable with Target

- From heatmap and pair-plot, temperature (temp) has the highest positive correlation with the bike demand (cnt).

4. Validation of Linear Regression Assumptions

- **Linearity and Homoscedasticity:** Linear relationship was established between the actual and predicted values.
- **Normality of Residuals:** Normal distribution can be noticed on the Residuals
- **Multicollinearity:** Checked using VIF values and eliminated high-VIF predictors during model refinement.
- **Correlation on Error Terms:** Verified visually through residual patterns, which showed no autocorrelation.
- **R-Squared:** R Squared calculated and showed no overfit and is within 5% Deviation range.

5. Top Three Features Influencing Bike Demand

Based on the Final Model, below is the summary,

- Year (yr) -> showing major growth from 2018 to 2019.
- Seasons -> have significant impact on the bike Rentals especially Spring.
- Weather -> Has significant influence on Bike rentals, especially Rainy condition

General Objective Questions

1. Explain the Linear Regression Algorithm in Detail

Linear Regression is one of the supervised learning method used to model the linear relationship between a dependent variable (Y) and one or more independent variables (X). It's equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

The algorithm works as follows:

- It fits a straight line (or hyperplane in case of MLR) that best represents the data.
- Uses Ordinary Least Squares (OLS) to minimize the sum of squared errors between actual and predicted values.
- **Key Assumptions:**
 - Linear relationship between X and Y
 - Independence of error terms(Residuals)
 - Homoscedasticity or Constant variance of error terms(Residuals)
 - Normality of residuals,
 - No Multicollinearity.

The outputs of the model are coefficients, p-values, and R² values that help interpret the strength and influence of predictors.

2. Anscombe's Quartet – Explanation

Anscombe's Quartet is a group of four datasets that share nearly identical statistical properties (mean, variance, correlation, and regression line) but differ drastically in their graphical appearance.

Key insights from the quartet:

- Dataset 1 follows a typical linear trend.
- Dataset 2 shows a clear nonlinear pattern.
- Dataset 3 contains an outlier that heavily influences the regression line.
- Dataset 4 contains one influential point that creates a false sense of correlation.

It demonstrates the importance of visualizing data instead of relying only on summary statistics.

3. What is Pearson's R?

Pearson's R (correlation coefficient) measures the strength and direction of a linear relationship between two continuous variables.

Formula: $r = \text{Cov}(X, Y) / (\sigma_x \sigma_y)$

- $+1 \rightarrow$ Perfect positive linear relationship
- $0 \rightarrow$ No linear relationship
- $-1 \rightarrow$ Perfect negative linear relationship

Pearson's R only measures linear correlation and is sensitive to outliers.

4. What is Scaling? Why is it Performed? Difference Between Normalized and Standardized Scaling

Scaling is the transformation of numerical features, so they share a similar range or distribution.

Why scaling is needed:

- Prevents features with large ranges from dominating the model.
- Speeds up gradient-based algorithms.
- Essential for distance-based algorithms such as KNN, K-means, and SVM.

Difference between scaling types:

- Normalized Scaling (Min-Max): Scales values to the 0–1 range. Formula: $(X - \text{min}) / (\text{max} - \text{min})$.
- Standardized Scaling (Z-score): Centers data with mean 0 and standard deviation 1. Formula: $(X - \mu) / \sigma$.

5. Why Does VIF Sometimes Become Infinite?

VIF becomes infinite when $R^2 = 1$ for a predictor, meaning that the predictor is perfectly explained by other predictors. This situation is called perfect multicollinearity.

- Occurs when variables are duplicates.
- Occurs when a variable is a linear combination of others.
- Happens when all dummy variables are included without dropping the first category.

In such cases, the model cannot estimate unique coefficient values.

6. What is a Q–Q Plot? Its Use and Importance in Linear Regression

A Q–Q plot (Quantile–Quantile plot) compares the distribution of residuals to a theoretical normal distribution.

Interpretation:

- Points along the 45-degree line → residuals are normally distributed.
- Curves indicate skewness or heavy tails.
- Outliers appear as distant points.

In linear regression, Q–Q plots help verify the assumption of normality of residuals, ensuring valid p-values and confidence intervals.