Linear Regression Assignments

Siddharth Shankar

2024-07-04

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? (3 marks)**
   - The categorical variables in the dataset include season, weather situation, month, and weekday. From the analysis:
     - **Season**: Different seasons have varying impacts on bike rentals, with some seasons showing higher demand due to favorable weather conditions for biking.
     - **Weather Situation**: Clear weather is associated with higher bike rentals, while adverse weather conditions (like mist and light rain) are associated with lower rentals.
     - **Month**: Certain months (like summer months) show higher bike rentals due to more favorable weather conditions.
     - **Weekday**: Weekdays and weekends show different patterns in bike rentals, likely due to commuting patterns during weekdays and recreational use during weekends.
2. **Why is it important to use drop_first=True during dummy variable creation? (2 marks)**
   - Using `drop_first=True` during dummy variable creation helps avoid the dummy variable trap, which occurs when there is multicollinearity among the dummy variables. This means that one of the dummy variables can be perfectly predicted from the others, leading to issues in regression analysis. By dropping the first category, we prevent this problem and ensure that the model can run without multicollinearity issues.
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   - From the pair-plot analysis, the variable `registered` has the highest correlation with the target variable `cnt` (total bike rentals). This is expected as registered users make up a significant portion of the total rentals.
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - After building the model, the following steps were taken to validate the assumptions of Linear Regression:
     - **Linearity**: Checked the scatter plot of residuals vs. predicted values to ensure no patterns were present, indicating linear relationships.
     - **Homoscedasticity**: Examined the residual plot to ensure that residuals are evenly distributed around zero, indicating constant variance.
     - **Normality of Residuals**: Plotted a histogram of residuals and a Q-Q plot to check if residuals followed a normal distribution.
     - **No Multicollinearity**: Calculated Variance Inflation Factor (VIF) for predictor variables to ensure no high multicollinearity was present.
     - **Independence of Errors**: Checked the Durbin-Watson statistic to ensure that residuals were not autocorrelated.
5. **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 3 features contributing significantly towards explaining the demand for shared bikes are:
     - **yr**: Indicates the year and captures the increase in demand over time.

- **temp**: Represents the temperature and shows a positive relationship with bike rentals.
- **workingday**: Indicates whether the day is a working day and shows higher bike rentals on working days.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   - Linear regression is a 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 straight line (regression line) that minimizes the sum of the squared differences between observed and predicted values. The algorithm involves:
     1. **Hypothesis Representation**: The relationship is modeled as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$, where y is the dependent variable, x_i are independent variables, β_i are coefficients, and ϵ is the error term.
     2. **Cost Function**: The cost function (mean squared error) is $J(\beta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\beta(x_i) - y_i)^2$, where m is the number of observations.
     3. **Gradient Descent**: An optimization algorithm used to minimize the cost function by iteratively updating the coefficients β_i based on the gradient of the cost function.
     4. **Assumptions**: The algorithm assumes linearity, independence, homoscedasticity, normality of residuals, and no multicollinearity among predictors.
2. **Explain the Anscombe's quartet in detail. (3 marks)**
   - Anscombe's quartet comprises four datasets with nearly identical statistical properties (mean, variance, correlation, and regression line) but vastly different distributions and scatter plots. It demonstrates the importance of visualizing data before analysis. Each dataset has:
     - The same mean for x and y.
     - The same variance for x and y.
     - The same correlation coefficient between x and y.
     - The same linear regression line.
     - Visual inspection reveals distinct patterns in each dataset, highlighting that relying solely on summary statistics can be misleading.
3. **What is Pearson's R? (3 marks)**
   - Pearson's R, or the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:
     - $R = 1$: Perfect positive linear correlation.
     - $R = -1$: Perfect negative linear correlation.
     - $R = 0$: No linear correlation.
   - It is calculated as $R = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$, where $\text{cov}(X, Y)$ is the covariance of X and Y, and σ_X and σ_Y are the standard deviations of X and Y.
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   - **Scaling**: The process of transforming the features to a common scale without distorting differences in the ranges of values.
     - **Why Scaling is Performed**: Scaling improves the performance of many machine learning algorithms, especially those that use distance-based metrics (e.g., KNN, SVM) and gradient descent optimization (e.g., Linear Regression, Neural Networks).
     - **Normalized Scaling**: Rescales the data to a range of [0, 1]. It is calculated as $X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$.
     - **Standardized Scaling**: Transforms the data to have a mean of 0 and a standard deviation of 1. It is calculated as $X' = \frac{X - \mu}{\sigma}$, where μ is the mean and σ is the standard deviation.
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   - A VIF value becomes infinite when there is perfect multicollinearity among the predictor variables. This occurs when one predictor variable is a perfect linear combination of other predictors. In such cases, the determinant of the correlation matrix becomes zero, making it impossible to invert,

leading to an undefined (infinite) VIF value.

6. **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 to assess if a dataset follows a particular distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the specified distribution.
     - **Use and Importance in Linear Regression**: In linear regression, a Q-Q plot is used to check the normality assumption of the residuals. If the residuals follow a straight line in the Q-Q plot, it indicates that they are normally distributed, satisfying one of the key assumptions for valid inference in linear regression models. Deviations from the line suggest departures from normality, indicating potential issues with the model's assumptions.