Linear Regression: Bike-Sharing Assignment

Submission by Sangbeda Das

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)

Answer: The preliminary analysis of the categorical variables has revealed that while the data distributions across all seasons, the two years, and all days of the week do not exhibit drastic differences, the clear-weather, working-day, and no-holiday conditions are observed to be more prevalent over other conditions in the dataset.

Overall, the dependent/target variable 'cnt' conveying the count of the total rental bikes (including both casual and registered) increases in the second year (2019).

The mean of 'cnt' drastically increases with the seasonal transition from spring to summer, increasing further in fall, thereafter decreasing moderately in winter.

Furthermore, the mean of the dependent variable 'cnt' overall increases throughout the year from January until December—the increase is continual until July, after which it decreases slightly in August, nonetheless increasing again in September, and subsequently gradually decreasing until December. However, despite the slight decrease in the last three months of the year, the mean of 'cnt' in December is much higher than that in January.

Moreover, on holidays, the count of the total rental bikes is higher. It has also been observed that the mean 'cnt' does not vary much regardless of whether the day is a working or non-working day.

The mean 'cnt' varies substantially across the diverse weather conditions, demonstrating the maximum value under the "Clear, Few clouds, Partly cloudy, Partly cloudy" condition, and the minimum under the "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" condition. (There are no records corresponding to the "Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog" condition in the dataset.)

The mean 'cnt' does not vary significantly across all days of the week. Nonetheless, it is found to be slightly lower on Tuesdays and Wednesdays, relative to that on other days of the week.

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

Answer: When creating dummy variables from categorical variables, it is common practice to set drop_first=True when calling the get_dummies() method. This parameter is used to drop one of the levels (the first column) of each categorical variable, resulting in a set of n-1 dummy variables for a categorical variable with n levels. The information that is eliminated by dropping the first column is already implicit in the combination of the remaining n-1 columns.

For example, let 'Relationship_Status' be a categorical variable with three levels: Single, In a relationship, Married. If we drop the first dummy column 'Single', we can still be able to confirm whether a person is single—if both the dummy variables 'In a relationship' and 'Married' are equal to zero, it is implied that the person is single.

By dropping the first category, we can capture the information about the categories effectively using fewer variables. This reduction in the number of variables helps to reduce redundancy and collinearity in the regression model, which can improve the stability and interpretability of the model.

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

Answer: The variable 'temp' has the highest corelation with the target variable 'cnt'.

In fact both 'atemp' and 'temp' showed high collinearity with 'cnt'.

Owing to the high corelation between 'atemp' and 'temp' (0.99 correlation coefficient), 'atemp' was dropped towards the end of the exploratory data analysis.

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

Answer: The linear regression assumptions have been successfully validated after building the model.

1. Assumption of Linearity: The scatter plot of the observed values against the predicted values indicates a linear relationship.
2. Assumption of Normality of Residuals: The residual plot obtained by creating a shows a fairly symmetric bell-shaped distribution with the mean at zero.
3. Assumption of Independence of Residuals: Plotting the residuals against the predicted values did not reveal any systematic patterns, thus indicating their independence.
4. Assumption of Homoscedasticity: Plotting the residuals against the predicted values exhibited a constant variance of residuals across the range of predicted values. Furthermore, the variance did not reveal any pattern with the change in the error values.

5. Assumption of Minimum Multicollinearity: The low variance inflation factor (VIF) of the predictor variables (which were less than 5) indicated low multicollinearity. Thus, stability and interpretability of the model are not affected.

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

Answer: Year (‘yr’), temperature (‘temp’), and “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds” weather condition (‘weathersit_light’) are three most effective predictors of the demand for shared bikes. All three variables have p-value of zero, indicating high significance. These variables have high absolute value of coefficient (both positive and negative), indicating their predictive power. Furthermore, their VIFs are also low, suggesting low multicollinearity.

1. Temperature (‘temp’): It has a low VIF of 5.17, a p-value of zero, and a strong coefficient of 0.4377.
2. Year (‘yr’): It has a low VIF of 2.07, a p-value of zero, and a strong coefficient of 0.2343.
3. Weather situation (‘weathersit_light’): It has a low VIF of 1.09, a p-value of zero, and a strong coefficient of -0.2929.

These features exhibit both statistical significance and low multicollinearity as well as strong coefficients, making them the best predictors of the demand of shared bikes.

General Subjective Questions

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

Answer: Linear regression is widely used for predicting a continuous target variable based on one or more input variables. It assumes a linear relationship between the input variables and the target variable. In addition, it also assumes normality, independence, and homoscedasticity of the residuals (error terms). The goal of linear regression is to find the best-fit line that minimizes the difference between the predicted values and the actual values of the target variable.

The algorithm works by estimating the coefficients (slope and intercept) of the linear equation that represents the relationship between the input variables and the target variable. This is done using a method called Ordinary Least Squares (OLS). The OLS method minimizes the sum of the squared differences between the predicted values and the actual values.

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable

- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- B_1X_1 = the regression coefficient (B_1) of the first independent variable (X_1) (the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- B_nX_n = the regression coefficient of the last independent variable
- ϵ = model error (the variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates the regression coefficients that lead to the smallest overall model error, the F-statistic of the overall model, and the associated p-value (how likely it is that the F-statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true). It then calculates the F-statistic and *p* value for each regression coefficient in the model.

The steps involved in the linear regression algorithm are as follows:

1. Data Preparation: Firstly, the data is also cleaned and pre-processed if necessary. In case of multiple categorical variables, they must be adequately dealt with by encoding them using dummies. The input data is prepared by splitting it into Train and Test datasets.
2. Model Building and Training: The Train dataset is further segmented into independent variables (X) and the dependent/target variable (Y). Only the Train dataset is used to build the linear regression model. The training process involves finding the optimal values for the coefficients that minimize the error between the predicted values and the actual values. In case of multiple independent variables, the best approach to model building is through the recursive feature elimination approach. Furthermore, the features should be selected such that they are not only highly significant but also are less corelated with each other, i.e., they should have low variance inflation factor (VIF must be below the threshold (around 5)).
3. Model Evaluation: At each iteration of model building, the trained model is evaluated using evaluation metrics such as R-squared and F-statistic. These metrics measure the accuracy and significance of the overall model. In addition, the coefficients are analyzed and the p-values of the features are keenly monitored (p-value must be below 0.05). Furthermore, multicollinearity between the features should be minimized to the extent feasible. Before employing the model on the test dataset, the assumptions of linear regression are verified (through residual analysis).
4. Prediction: Once the model is trained and evaluated, it can be used to make predictions on new, unseen data. At this stage, the independent variables of the Test dataset are fed into the model, and it calculates the predicted values of the target variable.

The r2_score on the test set must be close to the R-squared value obtained on the train set, i.e., the difference must be within 5%. If this is satisfied, we can say that what the developed model has learnt on the train set is appropriately generalized on the test set. Thus, the developed model is a good fit and makes appropriate predictions.

Although linear regression is a simple yet powerful algorithm that provides interpretable, owing to it's assumptions, it may not capture complex non-linear patterns. In such cases, more advanced regression techniques may be required.

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

Answer: Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and the limitations of relying solely on summary statistics. Despite having very different patterns, these datasets have nearly identical summary statistics, highlighting the need for graphical analysis.

The four datasets in Anscombe's quartet consist of x and y coordinate pairs. Here is a brief description of each dataset:

1. Dataset I: This dataset forms a simple linear relationship with a slight positive slope. It resembles a typical linear regression scenario.
2. Dataset II: This dataset also exhibits a linear relationship but with an outlier. The relationship remains roughly linear but is influenced by the presence of the outlier.
3. Dataset III: This dataset has a nonlinear relationship that resembles a quadratic curve. The relationship between x and y is not linear, yet the summary statistics (mean, variance, correlation) are similar to those of the first two datasets.
4. Dataset IV: This dataset consists of several distinct groups with the same x-values but different y-values. Each group exhibits a linear relationship. When analyzed as a whole, it gives a misleading linear regression line, suggesting a linear relationship when there isn't one.

The significance of Anscombe's quartet lies in its ability to challenge the assumption that summary statistics alone provide a complete understanding of data. It highlights the importance of visually exploring and understanding the data through plots and graphs. By presenting datasets with different patterns but identical summary statistics, it emphasizes the need to consider the context and distribution of data points. Anscombe's quartet serves as a reminder to always visualize and analyze data graphically before drawing conclusions or making predictions based solely on summary statistics.

3. What is Pearson's R? (3 marks)

Answer: Pearson's correlation coefficient, commonly referred to as Pearson's R or simply the correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after the mathematician Karl Pearson, who developed the coefficient in the early 20th century.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. The formula for Pearson's correlation coefficient is as follows:

$r = (\Sigma((X - \bar{X})(Y - \bar{Y}))) / (n * \sigma X * \sigma Y)$

where:

- r represents Pearson's correlation coefficient
- X and Y are the individual values of the two variables
- $\bar{X}$ and $\bar{Y}$ are the means (averages) of X and Y, respectively
- σX and σY are the standard deviations of X and Y, respectively
- n represents the number of data points

The resulting value of Pearson's R ranges between -1 and 1. A value of -1 indicates a perfect negative linear relationship, where one variable decreases as the other increases. A value of 1 indicates a perfect positive linear relationship, where both variables increase or decrease together. A value of 0 suggests no linear relationship between the variables.

Pearson's correlation coefficient is widely used in statistical analysis and research to assess the strength and direction of the relationship between variables. It helps in determining the degree of association, identifying patterns, and making predictions based on observed data. However, it should be noted that Pearson's R only measures the linear relationship and may not capture other types of associations, such as nonlinear relationships or dependencies.

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

Answer: Scaling, in the context of data preprocessing, refers to the transformation of numerical features or variables to a specific range or distribution. It involves adjusting the values of the variables to ensure that they are comparable and do not create biases or distortions in the analysis. Scaling is typically performed as a pre-processing step before model building.

The main reasons for scaling are:

1. Comparable Magnitudes: Scaling helps to bring the features onto a similar scale or magnitude. Variables with different scales can have a disproportionate impact on the analysis. Scaling ensures that all variables are treated equally and prevents any one variable from dominating the analysis.
2. Gradient Descent Optimization: Many machine learning algorithms, such as gradient descent-based optimization algorithms, converge faster and perform better when the input features are on a similar scale. Scaling ensures that the optimization process is more efficient and effective.
3. Interpretability and Stability: Scaling makes the variables more interpretable and stable. It removes the unit dependence of variables, allowing for easier interpretation and comparison of their effects on the target variable.

Normalized scaling and standardized scaling are two commonly used scaling techniques:

1. Normalized Scaling (Min-Max Scaling):
   - Normalized scaling transforms the data to a specific range, typically between 0 and 1.
   - The formula for normalized scaling is: $X' = (X - X\_min) / (X\_max - X\_min)$, where X' is the scaled value, X is the original value, X_min is the minimum value of the variable, and X_max is the maximum value of the variable.
   - Normalized scaling preserves the relative relationships between the data points and ensures that the maximum and minimum values are mapped to 1 and 0, respectively.
2. Standardized Scaling (Z-score Scaling):
   - Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.
   - The formula for standardized scaling is: $X' = (X - \mu) / \sigma$, where X' is the scaled value, X is the original value, μ is the mean of the variable, and σ is the standard deviation of the variable.
   - Standardized scaling centers the data around the mean and adjusts the scale based on the standard deviation.
   - Standardized scaling preserves the shape of the distribution and can handle outliers effectively.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the analysis and the characteristics of the data. Normalized scaling is useful when the range of the data needs to be preserved, while standardized scaling is beneficial when the distribution and comparison of data points are important.

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

Answer: The occurrence of infinite Variance Inflation Factor (VIF) typically arises due to perfect multicollinearity within the dataset. Perfect multicollinearity refers to a situation where there is an exact linear relationship between two or more predictor variables in a regression model. In other words, one predictor variable can be expressed as a perfect linear combination of other predictor variables.

When perfect multicollinearity exists, the mathematical calculation of VIF leads to infinite values. This happens because VIF is computed as the ratio of the variance of the estimated coefficient of a predictor variable to the variance of the same predictor variable. In the presence of perfect multicollinearity, the variance of the estimated coefficient becomes extremely large, leading to division by zero and resulting in infinite VIF.

Perfect multicollinearity can be detected by examining the correlation matrix or by calculating the determinant of the matrix of predictor variables. If the determinant is zero or close to zero, it indicates the presence of perfect multicollinearity.

Perfect multicollinearity can occur due to various reasons, such as:

1. Data Errors: Data entry mistakes or measurement errors can sometimes introduce perfect multicollinearity.
2. Redundant Variables: Including redundant or highly similar variables in the model can create perfect multicollinearity.
3. Incorrect Model Specification: Incorrectly specifying the model, such as including derived variables that are already represented by other variables, can lead to perfect multicollinearity.
4. Data Transformation: Improper data transformations, such as dividing one variable by another, can introduce perfect multicollinearity.

Dealing with perfect multicollinearity is crucial because it can affect the interpretation and stability of the regression model.

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

Answer: A quantile-quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of the expected theoretical distribution. The Q-Q plot is particularly useful in linear regression to evaluate the assumption of normality for the residuals.

The importance of a Q-Q plot in linear regression lies in its ability to visually inspect if the residuals (the differences between the observed and predicted values) are normally distributed. If the residuals are not normally distributed, it indicates a violation of the assumption, which can affect the validity of the regression analysis and the interpretation of the results.

In a Q-Q plot, the observed residuals are plotted on the y-axis, and the expected quantiles from a normal distribution (assuming the residuals are normally distributed) are plotted on the x-axis. If the residuals follow a normal distribution, the points in the Q-Q plot will roughly align along a straight line. Any deviations from the straight line suggest departures from normality.

Interpreting a Q-Q plot involves examining the pattern of the plotted points. If the points closely follow the straight line, it indicates that the residuals are normally distributed. On the other hand, if the points deviate significantly from the line, it suggests non-normality. Common deviations include heavy tails, skewness, or systematic patterns.

By assessing the Q-Q plot, one can identify the need for any transformations or adjustments to meet the assumption of normality. It provides insights into potential issues with the model, such as influential outliers or the presence of heteroscedasticity. Additionally, it allows for the comparison of different distributions and helps in selecting the most appropriate distributional assumption for the data.

In summary, the Q-Q plot is a valuable tool in linear regression to evaluate the normality assumption of residuals, detect departures from normality, and guide necessary adjustments to ensure the validity of the regression analysis.