Linear Regression Assignment Questions

-Akhil Suresh

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?**

   From the analysis carried out we see a notable reduction in the usage of services (count variable/Target variable) during the spring season. This decline is significant and suggests a seasonal pattern in service utilization. On the other hand, an increase in service usage during non-holiday periods could indicate that people are primarily using the service for commuting to work. Interestingly, there appears to be no substantial variation in service usage across different days of the week, whether they are weekdays or weekends, or between working and non-working days.
   Additionally, customers seem to prefer using the service on days with clear weather, as demonstrated by the weather column. The year-over-year usage plot indicates a general upward trend in service utilization over time, suggesting that overall demand has been increasing. Furthermore, the analysis of usage across various months reinforces the idea that there are distinct seasonal patterns influencing service usage.
   The above inferences were evident once the model was built and proved to be true based on the model equation

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

   When creating dummy variables for a categorical feature with k categories, you generate k dummy variables. These variables are perfectly collinear because their values always sum to one for every observation. This perfect collinearity can cause multicollinearity problems in regression models, making it challenging to estimate the coefficients accurately.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   atemp and temp both have same correlation with target variable of 0.65 which is the highest among all numerical variables.

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

   Validated the assumptions of linear regression by checking the following:

   - By Scatter and pair plot analysis to visualize linear relationship between independent and dependent variables
   - By using VIF to check Multicollinearity between independent variables
   - Residual analysis of errors: that residual errors follow normal distribution

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

   Top three significant features would be, Temperature, Year and Holiday variable

General Subjective Questions

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

   Linear regression aims to model the relationship between a dependent variable (target) and one or more independent variables (predictors) by fitting a linear equation to observed data. The simplest form is the simple linear regression, which involves one predictor variable. If more than one independent variable is involved its known as multiple linear regression

   The linear regression model can be represented as: $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon$

   Where y is the dependent variable or target variable and x1,x2 ... xn are the independent variables and β1, β2 are their respective coefficients and β0 is the intercept,

   The simple linear regression model tries to fit a line that pass through most of the data points such that error between the data point and the line is minimal that error is known as residual error. In case of multiple linear regression, the model tries to fit a hyperplane.

   The goodness of fit is measured based on RSS- residual sum of squares and TSS total sum of squares. Total sum of squares is calculated as the difference between the observed value and the mean value, both RSS and TSS together is used to calculate the metric r-square which is given by 1-(RSS/TSS) and range lies between 0-1.

2. **Explain the Anscombe's quartet in detail?**

Anscombe's quartet is a set of four datasets designed by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before performing statistical analyses. Despite having nearly identical summary statistics (such as mean, variance, correlation, and regression line) across all four datasets, the datasets exhibit vastly different distributions and patterns. This quartet is used to illustrate that descriptive statistics alone can be misleading if data is not visualized.

Dataset 1: This dataset follows a linear trend with some random noise.

Dataset 2 This dataset also follows a linear trend but has a clear outlier in the (x, y) data.

Dataset 3: This dataset forms a perfect quadratic curve (a parabolic shape), where y increases with x but the relationship is not linear.

Dataset 4: This dataset has a vertical line of points with a single horizontal outlier.

All the above datasets have similar summary statistics but when you visualize it you'll see following outputs and trends:

Dataset I: A scatterplot shows a clear linear relationship.

Dataset II: A scatterplot reveals a linear trend with an outlier that significantly affects the regression line.

Dataset III: A scatterplot illustrates a quadratic relationship, showing that a linear model does not fit well.

Dataset IV: A scatterplot displays a vertical line with one outlier that distorts the regression line.

Anscombe's quartet serves as a powerful reminder of the importance of data visualization in statistical analysis. It emphasizes that while summary statistics provide useful information, they do not capture the full story of the data's distribution and underlying patterns.

3. **What is Pearson's R?**

Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

Positive Correlation: When r is between 0 and +1, it indicates a positive linear relationship. As one variable increases, the other tends to increase as well.

Negative Correlation: When r is between -1 and 0, it indicates a negative linear relationship. As one variable increases, the other tends to decrease.

No Correlation: When r is around 0, it suggests no linear relationship between the variables. However, this does not rule out the possibility of a non-linear relationship

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

   Scaling is a preprocessing technique used when building a machine learning model to standardize independent feature variables within a fixed range. Datasets often contain features with varying magnitudes and units, and without scaling, this discrepancy can lead to inaccurate modelling due to mismatched units among the features. The key difference between normalization and standardization is that normalization adjusts the data to fall within a range of 0 to 1, while standardization transforms the values into their Z-scores, which represent how many standard deviations each value is from the mean.

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

   The VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. Specifically, it's used to detect multicollinearity in multiple regression models.
   VIF is calculated as VIF=1/(1-R^2), so VIF will become infinite only when the term 1-R^2 is 0 for that to happen R^2 should be 1. Which indicates perfect multicollinearity, the means the predictor in test has perfect linear combination with other predictors.

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

   Q-Q plots are also known as Quantile-Quantile plots. They plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let's discuss some of the probability distributions. Probability distributions are essential in data analysis and decision-making. Some machine learning models work best under some distribution assumptions. Knowing which distribution, we are working with can help us select the best model.
   QQ plots is very useful to determine:
   - If two populations are of the same distribution
   - If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
   - Skewness of distribution