Name: Mickell Als
Title: Linear Regression Assignment
Course: Basic Machine Learning

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)*
   - Demand for bikes is higher in the warmer months AND warmer seasons. Spring, Summer and Autumn have more ridership than winter AND the same from June through to October
   - Clear weather results in more demand.
   - Demand was much higher in 2019 vs 2018.
   - Median demand is much higher on non holidays.

2. *Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)*
   - For a categorical variable with n levels, we create n-1 dummy variables. The reason we do not make n dummy variables is because if all other dummy variables have a value of 0 then it is inferred that the final variable would have a value of 1. So by using drop_first=True in the pandas.get_dummies() we create n-1 dummy variables with the first level or n1 being dropped and assumed to be the default if all other dummy variables = 0 in that row.

3. *Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)*
   - Temp and atemp appear to have the highest correlations. Since atemp is derived from temp, then it can be said temp has the highest correlation with the target variable cnt

4. *How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)*
   - The assumptions of linear regression are a linear relationship exists between predictors and target variables, error terms are normally distributed, variance is constant i.e they are homoscedastic and error terms are independent of each other. These were verified by the following:
     i) In plotting the y_test vs y_pred for the test set the resultant scatterplot showed a positive linear relationship. Additionally, the high F-Statistic and high r-squared value indicated a linear relation explained much of the variation.
     ii) The residuals (y_train – y_train_pred) had a normal distribution with the mean centered at 0.
     iii) There was no pattern observed in the scatterplot of the residuals vs the predicted values on the training set therefore there is homoscedasticity.
     iv) The Durbin-Watson statistic for the final model was 2.035 which is more than 2. Values between 1.5 and 2.5 indicate there is no autocorrelation and therefore each result is independent of the previous result(s) (Kenton, 2019).

Name: Mickell Als
Title: Linear Regression Assignment
Course: Basic Machine Learning

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 based on the absolute value of their coefficients were

a) temp with a coef of 0.3577
b) Light_Precip with a coef of -0.3004
c) yr with a coef of 0.2369

General Subjective Questions

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

Linear regression is a machine linear algorithm where a linear relationship, a straight line, is sought to be established between one or more predictor variables and a continuous target/response variable. This straight line, also called the best fit line, is fit to the data using the Residual Sum of Squares (RSS) Method. This method involves fitting multiple straight lines to a scatterplot of response variable vs predictor variables. Then for each point, the distance from it to the straight line is calculated – this is the residual of that point. Each residual is then squared and all then summed. The straight line which yields the lowest value for the RSS is selected.

Note that linear regression can only be conducted if the target/response variable is a continuous value such that there are no "holes" that exist on the interval on which the response variable exists. Heights in cm for example is continuous but Height Group (tall, short average) is not.

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

This is a typical example of why data visualisations are necessary before modelling a linear relationship. This thought experiment focuses on 4 datasets with two variables (X,Y) within each. All the datasets have the same descriptive statistics – number of samples (N=11), mean, variance and correlation between X and Y. Because of these statistics, one would think the same linear model would fit each dataset but this is untrue. Each dataset would yield a different r-squared value. This is because the distributions of each is different. One dataset may be distributed linearly, while another is curvi-linear (not a straight line relationship), another may have a large outlier that is missed by the model and the final, multiple values of X yield the same Y but a single X value yields a significantly higher Y value. Just looking at the descriptive statistics may lead one to believe that these are 4 of the same datasets but in reality their distributions are all different and therefore knowing the distribution of the data helps in identifying whether a linear model is a good fit for a dataset even before modelling

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

Name: Mickell Als
Title: Linear Regression Assignment
Course: Basic Machine Learning

Pearson correlation coefficient is a measure of how well a linear relationship fits two variables. It varies between -1 and 1 with values of 0 having no correlation. The higher the absolute value of the correlation, the more linear the relationship. An absolute value of 1 means the relationship is purely linearly, i.e all points lie on the best fit line while the sign of the number indicates whether this linear relationship is positive or negative. That is to a a correlation of -0.8 indicates that as X increases, Y decreases with most of the points lying on the negatively sloped best fit line. The pandas function corr() uses this method by default to determine the correlation values between variables within a dataset.

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

Scaling is bringing the range of all variables in a dataset to the same range, that is some variables may range from 100-1000, others 1-10, in scaling we would ensure that the range of all variables is the same while retaining their relative distributions. Scaling of variables is done in order to ensure that one variable does not have a significantly higher impact on a model than another. Features with large ranges and values tend to have very small coefficients, near 0, which makes it appear as though their influence on the model is less significant than another feature with a smaller range. With all features on the same scale, we can mitigate this effect on coefficients and example each equally. There are two main types of scaling used:
Normalization – places all values between 0 and 1 with 0 being the minimum and one being the maximum value of a specified feature's range. It is calculated using the formula (x – min(x))/(max(x) – min(x) . This type of normalisation is commonly called MinMaxScaling
Standardization – in this type scaling, each point for a feature is compared to its distance from the mean of that feature, its z-score. The mean is given a value of 0 and the values other than the mean are given a value based on how many standard deviations from the mean a value is. This type of scaling is calculated using the formula (x – mean(x))/sd(x). A value of 1.5 would mean that x is 1.5 standard deviations to the right of the mean. In standardisations, values can be positive or negative with no lower or upper bound and with the sign indicating whether a value is to the right or left of the mean. It also allows outliers to be more readily seen as values with z-scores of more than +/- 3 is usually noted to be an outlier as it means it lies outside of 99.7% of the other data points

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

A VIF value of infinity can only occur if the denominator of the VIF formula is equal to 0. Such a cause can only occur if the r-squared value is 1. An r-squared of 1 means that the feature is perfectly captured by a linear combination of other features within your dataset. That is, there is perfect multicollinearity between that specific feature i and one or more other features in the dataset. Since multicollinearity is not ideal in a regression, we seek to remove any features whose variance is explained by other features within the dataset.

Name: Mickell Als
Title: Linear Regression Assignment
Course: Basic Machine Learning

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 plot, also called a quantile-quantile plot, is a technique that graphs two datasets to check the normality assumption of a linear model. To construct this graph we have the take the theoretical quantiles of the data points on a normal distribution for the x value and compare them to the actual quantile of the dataset as the y value. This will yield a scatterplot, we then fit a straight line to the data. If a distribution is truly normal then it will very closely match the best fit line, if the distribution is not, there will be significant variation above and below the straight line. In linear regression, one of the assumptions is that the residuals are normally distributed with a mean of 0. If they are truly normally distributed, then a Q-Q plot of the residuals should essentially be a straight line of points which match very closely, the best fit line of those points. We can therefore use a Q-Q plot to confirm whether the normality assumption of linear regression holds true for our model.

References

Kenton, Will. (2019). Understanding the Durbin Watson Statistic. Investopedia. https://www.investopedia.com/terms/d/durbin-watson-statistic.asp