# Linear Regression Assignment

Answers to the subjective questions

## 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 are season, holiday, workingday, weather, weekday and month. To analyse their effect on the dependent variable, which is the count of bike rentals, I used the following steps:

   - I created dummy variables for each categorical variable using the pandas get_dummies function with drop_first=True option.
   - I performed a multiple linear regression using the statsmodels OLS function with the count as the response variable and the dummy variables as the predictor variables.
   - I checked the summary of the regression output and looked at the p-values, coefficients, and R-squared of the model.
   - Based on the analysis, I could infer that:
     Some categorical variables have a significant effect on the count of bike rentals, while others do not. For example, the season variable has four levels: spring, summer, fall, and winter. The dummy variables for winter and spring all have p-values less than 0.05, which means they are statistically significant predictors of the count. The coefficients are also positive and negative, which means that the count of bike rentals increases and decreases in these seasons

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

   It is important to use drop_first=True during dummy variable creation to avoid the problem of multicollinearity. Multicollinearity occurs when two or more predictor variables are highly correlated with each other, which can cause instability and bias in the regression coefficients and p-values. When creating dummy variables for a categorical variable with k levels, we only need k-1 dummy variables to represent the information. If we use k dummy variables, then one of them will be redundant and perfectly correlated with the others. For example, if we have a categorical variable with two levels: yes and no, we only need one dummy variable to indicate whether the value is yes or no. If we use two dummy variables, then they will be the exact opposite of each other and have a correlation of -1. This will cause multicollinearity and make the regression results unreliable. By using drop_first=True, we can drop one of the dummy variables and avoid this problem.

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

   The pair-plot among the numerical variables shows the scatter plots and histograms of the variables: temp, atemp, humidity, windspeed, and count. The variable that has the highest correlation with the target variable, which is the count of bike rentals, is the temp variable. The scatter plot between temp and count shows a positive linear relationship, which means that as the temperature increases, the count of bike rentals also increases. The correlation between temp and count is 0.63, which is the highest among the numerical variables.

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

After building the model on the training set, I validated the assumptions of linear regression using the following methods:

- I checked the linearity assumption by plotting the predicted values versus the residuals and looking for any patterns or curvature. The linearity assumption states that the relationship between the response and the predictor variables should be linear. The plot showed a random scatter of points around zero, which indicates that the linearity assumption is not violated.
- I checked the independence assumption by plotting the residuals versus the time index and looking for any autocorrelation or trend. The independence assumption states that the residuals should be independent of each other and not influenced by the previous or future observations. The plot showed a random scatter of points without any obvious pattern or trend, which indicates that the independence assumption is not violated.

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

Based on the final model, which includes both the numerical and the categorical variables, the top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- The temp variable, which has a positive coefficient of 0.42 and a p-value of 0.00. This means that for every unit increase in the temperature, the count of bike rentals increases by 0.42 units, holding all other variables constant. This also means that the temp variable is a highly significant predictor of the count variable.
- The light_snow variable, which has a negative coefficient of -1.37 and a p-value of 0.00. This means that for every unit increase in the windspeed, the count of bike rentals decreases by 1.37 units, holding all other variables constant. This also means that the light_snow variable is a highly significant predictor of the count variable.
- The winter variable, which has a positive coefficient of 0.31 and a p-value of 0.00. This means that the count of bike rentals in winter is 0.31 units higher than the count of bike rentals in spring, holding all other variables constant. This also means that the winter variable is a highly significant predictor of the count variable.

# General Subjective Questions

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

Linear regression is a supervised machine learning algorithm that is used to model the relationship between a response variable and one or more predictor variables. The goal of linear regression is to find the best-fitting line or plane that minimizes the sum of squared errors between the observed and the predicted values of the response variable.

The linear regression algorithm can be divided into two types: simple linear regression and multiple linear regression. Simple linear regression is when there is only one predictor variable, while multiple linear regression is when there are two or more predictor variables.

The linear regression algorithm can also be divided into two methods: ordinary least squares (OLS) and gradient descent. OLS is a mathematical method that uses matrix operations to find the optimal values of the regression coefficients that minimize the sum of squared errors. Gradient descent is an iterative method that uses a learning rate and a cost function to update the values of the regression coefficients until they converge to the minimum of the cost function.

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

The Anscombe's quartet is a set of four datasets that have the same summary statistics, such as the mean, the variance, the correlation, and the regression line, but have very different patterns and distributions when plotted. The Anscombe's quartet was created by Francis Anscombe in 1973 to demonstrate the importance of visualizing the data before performing any statistical analysis.

The Anscombe's quartet illustrates that the same summary statistics can hide very different patterns and distributions in the data, and that it is essential to visualize the data before performing any statistical analysis.

Datasets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

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

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. Pearson's R can be calculated as the ratio of the covariance between the two variables and the product of their standard deviations.

Pearson's R is widely used to measure the correlation between two variables in statistics and data analysis. It can be used to test the hypothesis of no correlation, to assess the strength of the linear relationship, and to select the predictor variables for linear regression. However, Pearson's R has some limitations and assumptions that need to be considered. For example, Pearson's R only measures the linear relationship and does not capture any nonlinear or curved relationship. Pearson's R is also sensitive to outliers and extreme values that can inflate or deflate the correlation. Pearson's R also assumes that the two variables are normally distributed and have a constant variance across the range of values.

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

Scaling is a process of transforming the values of a variable or a feature to a common range or scale. Scaling is performed for various reasons, such as:

- To make the data comparable and consistent across different units or scales. For example, if one variable is measured in meters and another variable is measured in kilometres, scaling can convert them to the same unit or scale.
- To reduce the effect of outliers and extreme values on the data. For example, if one variable has a very large range and variance compared to another variable, scaling can reduce the skewness and spread of the data.
- To improve the performance and accuracy of machine learning algorithms that are sensitive to the scale and magnitude of the features. For example, algorithms that use distance-based metrics, such as k-nearest neighbors and k-means clustering, can be affected by the scale of the features. Scaling can make the features have equal weight and importance in the algorithm.

There are different methods of scaling, such as min-max scaling, mean normalization, standardization, and unit vector scaling. Two common methods of scaling are normalized scaling and standardized scaling. The difference between normalized scaling and standardized scaling are:

- Normalized scaling transforms the values of a variable to a range between 0 and 1, where the minimum value becomes 0 and the maximum value becomes 1.
- Standardized scaling transforms the values of a variable to have a mean of zero and a standard deviation of one, where the values are centered and scaled by the standard deviation.

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

VIF, or variance inflation factor, is a measure of the multicollinearity among the predictor variables in a regression model. Multicollinearity occurs when two or more predictor variables are highly correlated with each other, which can cause instability and bias in the regression coefficients and p-values. VIF quantifies the extent of multicollinearity by measuring how much the variance of a regression coefficient is inflated due to the correlation with the other predictor variables. The formula for VIF is:

VIF = 1 / (1 - R2)

where R2 is the coefficient of determination of the regression of a predictor variable on the other predictor variables. A high value of VIF indicates a high degree of multicollinearity, while a low value of VIF indicates a low degree of multicollinearity. A rule of thumb is that a VIF value above 10 is considered problematic and indicates a serious multicollinearity issue.

Sometimes, the value of VIF is infinite, which means that the denominator of the formula is zero. This happens when the predictor variable is perfectly correlated with the other predictor variables, which means that the R2 of the regression is 1. This also means that the predictor variable is redundant and can be expressed as a linear combination of the other predictor variables. For example, if we have three predictor variables: x1, x2, and x3, and x3 = 2x1 + 3x2, then the VIF of x3 will be infinite, because x3 can be completely explained by x1 and x2. This situation is also known as perfect multicollinearity, which violates the assumption of linear regression and makes the estimation of the regression coefficients impossible. To avoid this problem, we need to remove the predictor variable that causes the infinite VIF or use a different method of regression that can handle multicollinearity.