Linear Regression Assignment

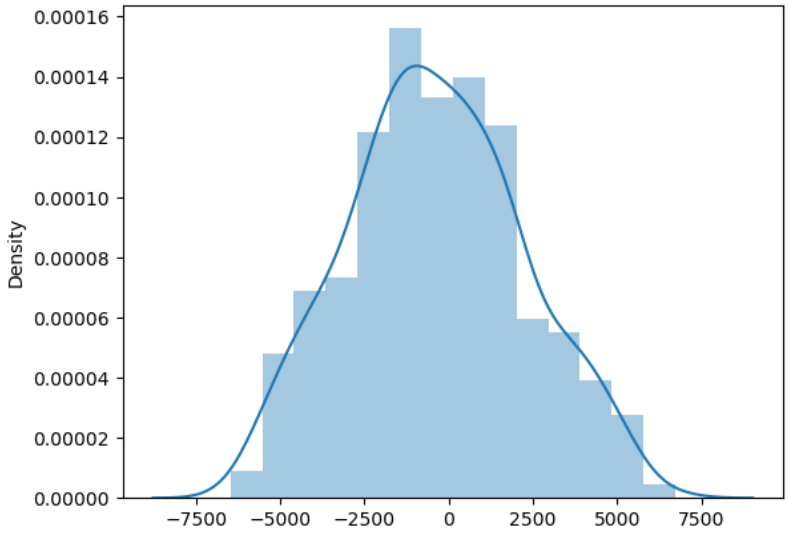
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**
   - Weather conditions have a significant impact over bike sharing. It can be observed that bike sharing is highest during fall and summer seasons. It is also observed that bike sharing increases when weather is clear.
   - Bike sharing is higher on non-holidays as compared to holidays. It can be inferred that working professionals might be using bikes for commute.
   - Bike sharing has increase significantly in 2019 as compared to 2018. It represents increasing trend of bike sharing as the years pass by.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   If a categorical variable contains n categories, we generally create (n-1) dummy variables. Dummy variables take category name and respective category is indicated as 1, while others marked as 0 for that particular record.
   We drop one category in this process. If all the dummy variables are marked as 0, it is intuitional that the record belongs to dropped category. By this way, we reduce number of columns for modelling.

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   - After building a model on training set, we predict dependant variable 'y_pred' using the linear regression model. Following the same, residual is calculated as difference in each 'y_train' and 'y pred' i.e. (y_train-y_pred).
   - By plotting histogram of the residuals, we can validate that residuals follow a normal distribution and have a mean as zero.



   - By plotting scatter plot of the residuals against independent variable, we can validate that residuals are independent of one another i.e. there is no pattern.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Three most significant features explaining the demand of shared bikes are:
   1) yr (Year)
   2) workingday (Day of the week)
   3) atemp (Temperature)

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   In linear regression, we try to define relationship between the dependant variable (also called target variable) and independent variables using linear equation.

   Assumptions in linear regression:
   - Linearity: There is a linear relationship between dependant and independent variable.
   - Normality: Residuals are normally distributed, with a mean of zero.
   - Independence: Residuals are independent of each other i.e. there is no pattern within residuals.
   - Homoscedasticity: Residuals have constant variance across all the levels.

   Linear regression modelling typically follows below steps:

   Step 1: Data collections and understanding
   Read the dataset and understand all the features and their business importance.

   Step 2: Data Cleaning
   Handling of missing values, redundant values and outliers.

   Step 3: Exploratory Data Analysis
   - Univariate Analysis: Analyse distribution of various features in the dataset
   - Bivariate Analysis: Identify key relationships between independent and dependant variable
   - Multivariate Analysis: Identify correlations withing independent variables.

   Step 4: One-hot encoding
   Convert categorical features to dummy variables

   Step 5: Splitting the data
   Split the dataframe in to train and test data

   Step 6: Feature selection
   Identify most significant features by using RFE (Recursive Feature Elimination) method

   Step 7: Model training
   Use the training set to fit the linear regression model. This includes identifying optimal values for coefficients in order to minimize RSS (Residual Sum of Squares)

   Step 8: Model evaluation
   Evaluate models performance using R-squared and RMSE (Root Mean Square Error)

   Step 9: Prediction using the model
   Use the model to predict dependant variable on test data

2. **Explain the Anscombe's quartet in detail. (3 marks)**
   Anscombe's quartet includes a set of four datasets having nearly identical descriptive summary statistics i.e. mean, standard deviation, r2 score, correlations etc. The linear regression line for the datasets is also identical, but the difference in the representation when plotted in a scatter plot.

   Each dataset consists of 11 data points (pairs of x and y). Each dataset shows same summary statistics as below:
   - Mean of X= 9
   - Mean of y=7.5
   - Variance of X=11
   - Variance of y=4.12,
   - Correlation=0.8164

   However, when plotted in a scatter plot, each dataset shows unique relationship between x and y.
   - First dataset shows linear relationship between X and y.
   - Second dataset shows non-linear relationship between X and y.
   - Third dataset shows linear relationship, but with presence of an outlier.
   - Fourth dataset shows datapoint aligned vertically.

   Importance of Anscombe's quarter: It demonstrates limitations of summary statistics that not every dataset having same summary statistics is exactly similar. It emphasizes the need for visual exploration of the datasets and diverse relationships within the features.

3. **What is Pearson's R? (3 marks)**
   Pearson's R is a measure of the linear relationship between two variables. It is the most common measure of correlation. Hence, it is also known as the Pearson correlation coefficient or simply the correlation coefficient.

   Pearson's R quantifies the extent to which two variables are linearly related to each other. The value of Pearson's R ranges between 1 to -1.
   - r=1 indicates perfect positive correlation i.e. if one variable increases, other variable increases proportionally.
   - r=0 indicates no linear relationship.
   - R=-1 indicates perfect negative correlation i.e. if one variable increases, other variable decreases proportionally.

   Assumptions of Pearson's R:

   - Linearity: relationship between two variables is linear
   - Homoscedasticity: variance of residuals is constant across all level of independent variable.
   - Normality: Both the variables are approximately normally distributed

   Pearson's R is used in various applications across the financial sector. e.g. If an investor expects bear market, he/she would prefer buying a stock that exhibits negative correlation with the benchmark index.

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

   During linear regression modelling, independent variables can vary between various ranges. Some of the features have small range, while some features can have large range.

   Scaling is preformed in data preprocessing phase in order to adjust the range of features and make sure they are on a comparable range.

   Benefits of scaling:

   - Preventing model from assigning higher weightage to the feature just because of its large range
   - Improving model performance by assigning nearly equal weightage to the features

   Normalized scaling vs Standardized scaling:

   - Normalized scaling (also known as min-max scaling) rescales the features within a fixed range, typically 0 to 1. It is useful when data is not normally distributed and has a range. However, it is sensitivity to outliers.
   - Standardized scaling rescales features to have a mean of 0 and standard deviation of 1. It can be used when data is normally distributed. It is also less sensitive to outliers.

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

   Variance Inflation Factor (VIF) is used to detect multicollinearity in linear regression models i.e. whether there exists a high correlation within any of the independent variables.

   Infinite VIF indicates perfect multicollinearity. VIF is measured as:

   $$VIF = 1/(1-R2)$$

   VIF becomes infinite only when R2 score is 1 i.e. there is a perfect correlation between one or more predictors. In this case, it can be concluded that perfectly correlated predictors convey same information and hence, one of the predictors can be dropped.

   To handle infinity VIF, we should keep only one of the highly correlated predictors for analysis. Also, we should not consider a feature which is linear combination of other features.

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

   Q-Q plot or Quantile-Quantlie plot is used to assess if the dataset follows any particular type of distribution, mostly normal distribution.

   Applications of Q-Q Plot in Linear Regression

   - Residual analysis:
     - Q-Q plots can be used to validate the assumptions in linear regression model i.e. validate normal distribution of the residuals.
     - We can product a visualization of residuals by a Q-Q plot.
   - Detection of outliers and skewness:
     - Points lying beyond reference line can be considered as outliers.
     - Skewness can be observed based on deviation from the line. Divergence from the line would suggest skewness.
   - Q-Q plot can be used to improve model performance. If we don't see residuals are normally distributed using Q-Q plot, it means we need to transform certain features in order to improve the model performance.