Linear Regression Assignment – Abhishek Pandey

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:

- Categorical variables significantly enhance model development, boost accuracy (as reflected by high R-squared values), and improve the precision of predicting the dependent variable.

- Some of the important categorical variables considered in the final model are:

- **Weathersit:** It denoted the weather situation,
  For 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds: just 3% sales which could mean when people were not sure of going out due to weather situation then sales is very less.

- **Yr(Year):** It denotes the year of sales with 0 for (2018) or 1 for 2019

- **Season:** Season 3(Fall) is having the maximum 32% followed by Season 2 and 4

- **Holiday:** 98% sales made in Non holidays which suggests people are using the services on non-holidays for their commute

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

Answer:

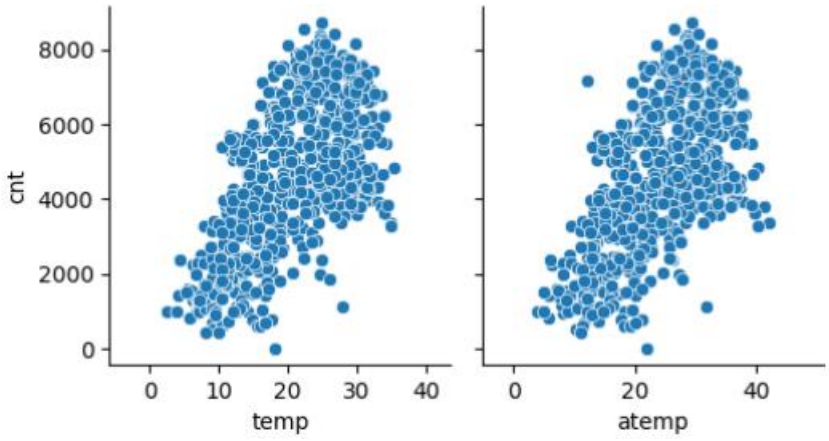
- The drop_first=True parameter in dummy variable creation helps avoid the Dummy Variable Trap by omitting one category, making interpretation easier and improving model efficiency

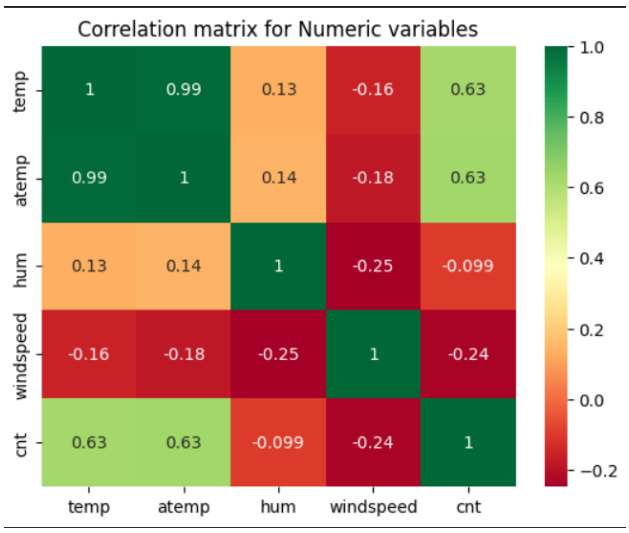
- Additionally, it helps us handle multi-collinearrity among the dummy variables

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

Answer:

- From the pair plot we can see temp and atemp is having highest correlation with the target variable cnt which is 0.63

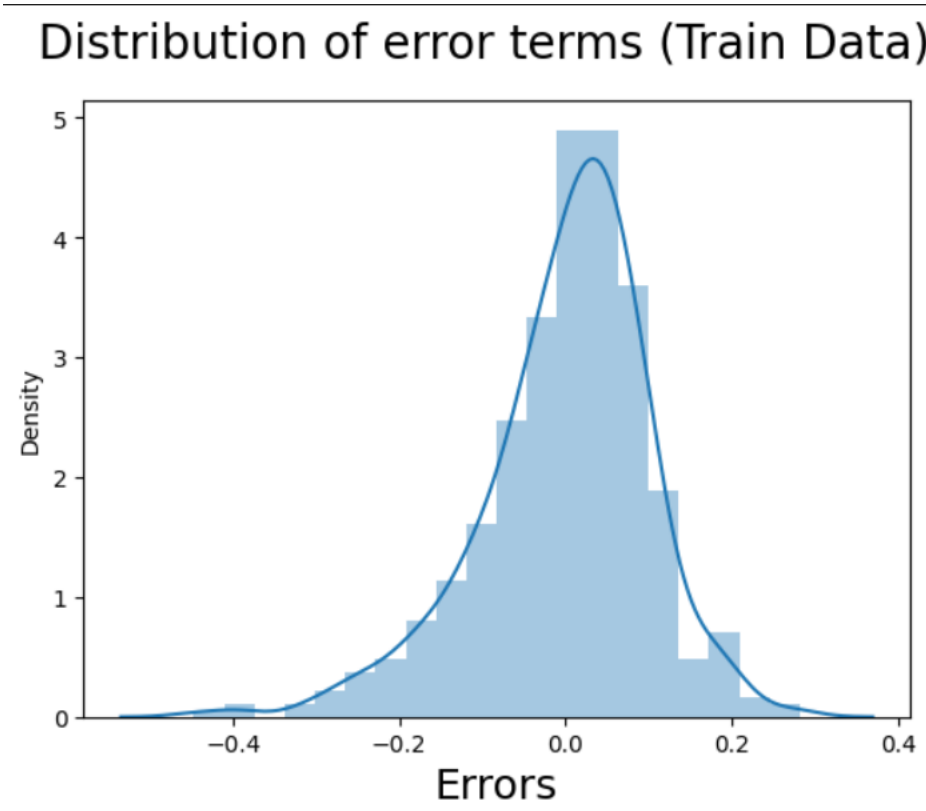


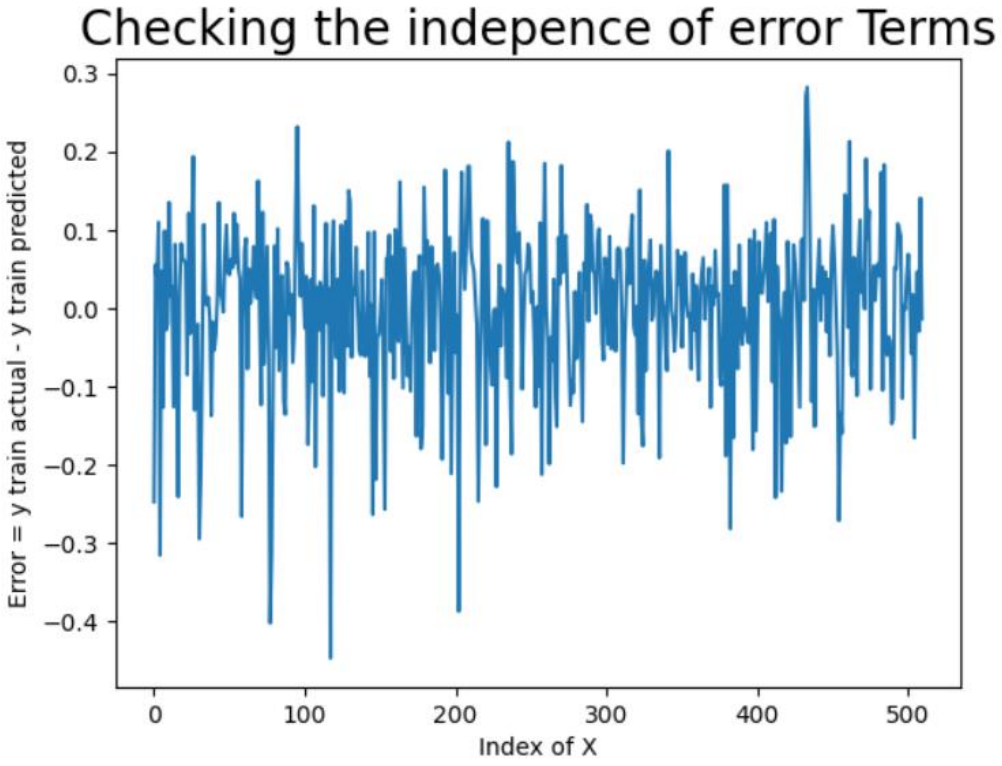


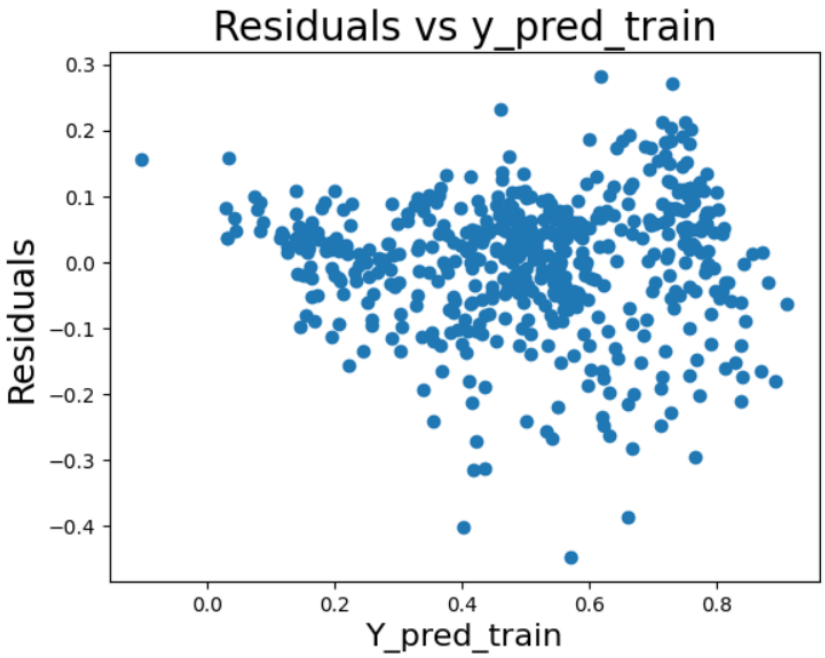
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

- We validated the following assumptions of linear regression
    - Analysis of Residuals
        - Residuals are normally distributed and centred around 0
        - Independence of error terms. It means that there should not be any meaningful distribution between independent variable and error term
        - Constant variance of Error terms. Thus, variance should not increase or decrease as the error values change. Also, variance should not follow any pattern as the error terms change


Distribution of error terms (Train Data)
Density
Errors
0
1
2
3
4
5
−0.4
−0.2
0.0
0.2
0.4


Checking the indepence of error Terms
Error = y train actual - y train predicted
Index of X
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
0
100
200
300
400
500



- Check for the linear relationship
  - The final model obtained is a linear relationship between the output variables 'cnt' and independent features.

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:

In the final model top 3 features are mentioned as below:

- **Temp:** It denotes the temperature in Celsius having the coefficient value as 0.5500, it indicates that unit increase in temp will lead to +0.5500 units increase in sales and it is seen that temperature has biggest influence of bike sales

- **Yr:** It denotes the year (0 = 2018 and 1 = 2019) having the coefficient value as 0.2347, it indicates that when year = 1 it will lead to 0.2347 increase in sales.

- **Weather_xtrm:** it denotes extreme weather and wethersit = 3 and have the coefficient value as -0.2212, hence it indicates that when the weather is poor then it will lead to decrease in sales.

General Subjective Questions

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

Answer:

Linear regression is a statistical technique used to model the relationship between a **dependent variable** (often denoted as *y*) and one or more **independent variables** (usually denoted as *x*). Let's break it down:

1. **Dependent Variable (y)**: This is the outcome or response variable that we want to predict or explain. For example, in a study examining the relationship between hours of study (*x*) and exam scores (*y*), the exam scores would be the dependent variable.
2. **Independent Variable(s) (x)**: These are the predictors or explanatory variables. In our example, the independent variable would be the hours of study.
3. **Linear Relationship**: Linear regression assumes that the relationship between the dependent variable and the independent variable(s) can be represented by a straight line. Hence the term "linear."
4. **Regression Line Equation**: The goal of linear regression is to find the best-fitting line (regression line) that represents the overall trend of the data. The equation for the regression line is typically expressed as:

   [y = mx + b]

   - (m) represents the slope of the line (also known as the regression coefficient). It indicates how much the dependent variable changes for a one-unit change in the independent variable.
   - (b) is the y-intercept, which is the value of (y) when (x) is zero (i.e., where the line intersects the y-axis).
5. **Least Squares Method**: Linear regression finds the best-fitting line by minimizing the sum of the squared differences (residuals) between the actual data points and the predicted values on the regression line. This method is called the least squares method.
6. **Predictions**: Once we have the regression line, we can use it to predict values of the dependent variable for new data points that were not part of the original dataset.

Simple Linear Regression:

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points

The standard equation of the regression line is given by the following expression: $Y = \beta_o + \beta_1.X$

Where :

Y is output variable

β_o is intercept or we can say constant

β1 is coefficient of X

Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables). The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Equation :

Y = B0 + B1.X1 + B2.X2+ …..+BnXn + e

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

Answer:

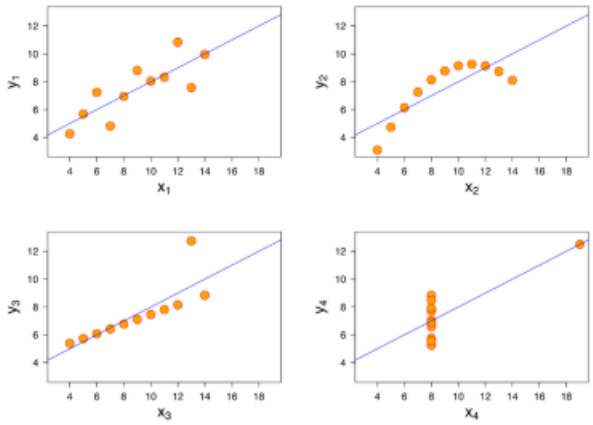
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises **four datasets, each containing eleven (x, y) pairs**. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9
- Mean of y is 7.50 for each dataset.
- Variance of x is 11 and
- Variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

- When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story



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

Answer:

Pearson's correlation coefficient, often denoted as **(r)**, is a fundamental statistical measure used to quantify the **strength and direction of the linear relationship** between **two quantitative variables**.

1. Pearson's (r) assesses how closely the data points in a scatter plot align with a straight line (the "line of best fit"). It ranges between **–1** and **1**:
    - When (r) is **positive**, it indicates a **positive correlation**:
        - As one variable increases, the other tends to increase as well.
    - When (r) is **negative**, it signifies a **negative correlation**:
        - As one variable increases, the other tends to decrease.
    - When (r) is **close to 0**, there's **no linear relationship**:
        - The variables don't move together predictably.

2. **Interpretation**:
    - (r > 0.5): **Strong positive correlation**
    - (0.3 < r \leq 0.5): **Moderate positive correlation**
    - (0 < r \leq 0.3): **Weak positive correlation**
    - (r = 0): **No linear correlation**
    - (-0.3 \leq r < 0): **Weak negative correlation**
    - (-0.5 \leq r < -0.3): **Moderate negative correlation**
    - (r < -0.5): **Strong negative correlation**

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

Answer:

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Now, let's talk about two common scaling techniques:

Normalization

- **Objective**: Normalize your data to a specific user-defined range, typically between 0 and 1 or -1 and 1.
- **How It Works**: Map the minimum feature value to 0 and the maximum to 1. Essentially, squeeze your data into that desired interval.
- **Example**: Imagine you have a bag of colorful balls, but some colors dominate. Normalization equalizes the color distribution, making sure each color gets its fair share.
- **Assumption**: Normalization doesn't assume anything about the underlying data distribution. It's more flexible in that sense.

Standardization (Z-score Normalization)

- **Objective**: Transform your data to have a mean of 0 and a standard deviation of 1.
- **How It Works**: Instead of enforcing a specific range, standardization centralizes your data. It's like shifting the data to be centered around zero.
- **Example**: Think of it as adjusting your measuring tape to fit what you're measuring. It doesn't change the shape of your data—just its size.
- **When to Use**: Standardization is often preferred because it's suitable for most cases. Plus, it's handy when you suspect your data follows a normal distribution.

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

Answer:

If there is perfect correlation, then VIF = infinity.

A large value of VIF indicates that there is a correlation between the variables.

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables.

In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot: When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests