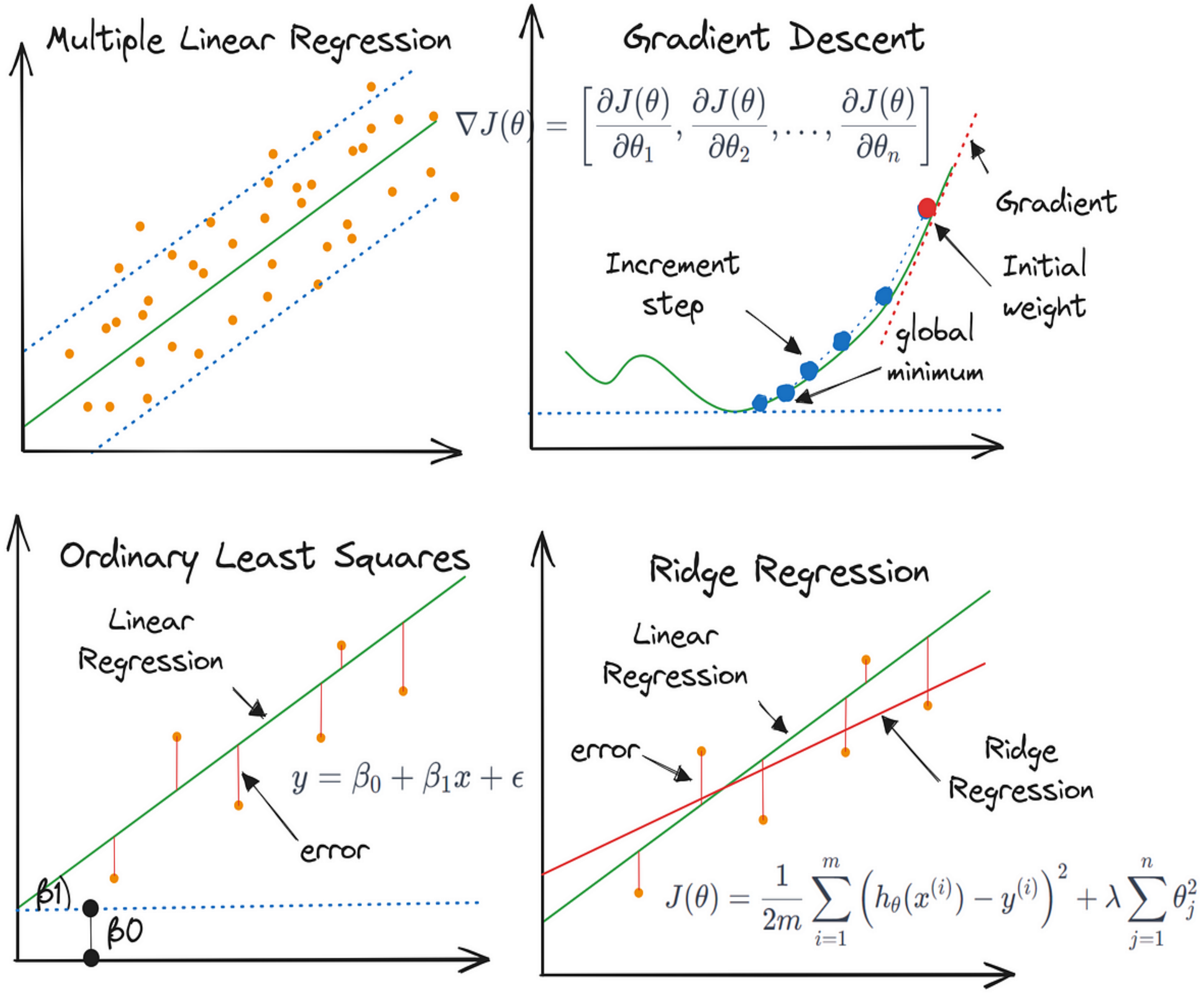

Multiple Linear Regression
Gradient Descent
$\nabla J(\theta) = \left[\frac{\partial J(\theta)}{\partial \theta_1}, \frac{\partial J(\theta)}{\partial \theta_2}, \ldots, \frac{\partial J(\theta)}{\partial \theta_n}\right]$
Gradient
Initial weight
Increment step
global minimum
Ordinary Least Squares
Linear Regression
$y = \beta_0 + \beta_1 x + \epsilon$
error
β1
β0
Ridge Regression
Linear Regression
error
Ridge Regression
$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2$

Boom Bike Assignment

Linear Regression Subjective Questions

Subhrajit Pyne 8/24/24 ML C-65

Assignment-based Subjective Questions:-

Q. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: The categorical variables in the dataset were “season”, “year”, “month”, “holiday”, “weekday”, “workingday” and “weathersit”.

These were visualized using a boxplot. These variables had the following effect on our dependent variable.

Season:

- Fall: Bike demand is highest in Fall.
- Summer and Winter: Summer and winter have intermediate value of count with summer having greater count among the two.

Spring: The demand for bikes is lowest in spring, possibly due to less favorable weather.

Year:

- 2019 vs. 2018: There is a clear increase in bike demand from the year 2018 to 2019. This trend suggests that the bike sharing program gained popularity over this period.

Month:

- High-Demand Months: June, July, August and September are the months with the highest bike demand. Out of all these months, September has seen the highest no of rentals. This could be due to the warm summer weather, which is ideal for biking.
- Low-Demand Months: December has seen the lowest no of rentals. January, February, and December see the less bike demand, likely due to colder winter weather, which discourages biking.

Holiday:

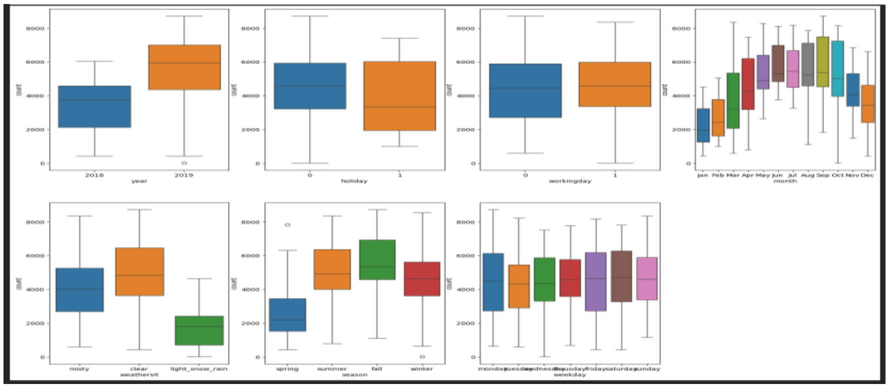
- Holidays vs Non-holidays: Bike demand is higher on holidays compared to non-holidays. This increase can be attributed to people having more leisure time and choosing to bike for recreation or errands on holidays.

Weekday:

- Even Distribution: Bike demand is relatively evenly distributed across all weekdays, indicating consistent usage throughout the week.

Weathersit:

- Clear Weather: The highest bike demand occurs during clear weather conditions, due to favorable weather.
- Adverse Weather: Bike demand decreases significantly during misty conditions, light snow/rain, and heavy snow/rain. There are no users when there is heavy snow/rain. The least demand is observed during light snow/rain, as adverse weather conditions make biking less appealing and potentially hazardous.



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

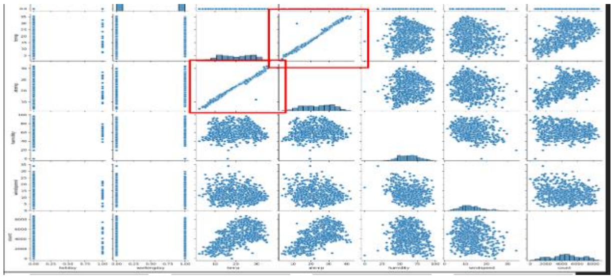
Ans: Using ‘drop_first=True’ during dummy variable creation is important for the following reasons:

- Preventing Multicollinearity: Including all dummy variables for a categorical feature can lead to multicollinearity, which occurs when predictor variables are highly correlated. This can make it difficult to determine the individual effect of each variable on the target variable. By dropping the first dummy variable, we avoid this issue and ensure that the remaining dummies can provide the necessary information without redundancy.

- Reducing Redundancy: By dropping the first category, the total number of dummy variables is reduced, which simplifies the model and improves efficiency.
- Model Interpretability: This practice ensures that the model remains interpretable and free from redundant variables, making it easier to understand and analyze the effects of other predictors.

Q. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: In the pair-plot analysis, the two temperature variables, "temp" and "atemp" , show the highest correlation with the target variable "count" or "cnt" . This strong positive correlation indicates that higher temperatures are associated with an increase in bike bookings.



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

Ans:- We have validated the assumption of Linear Regression Model based on below 5 assumptions -

- Normality of Error Terms: If the residuals follow a normal distribution, the assumption is met.

  Histogram: Plotted a histogram of the residuals. If the residuals are normally distributed, the histogram should resemble a bell curve.

Q-Q Plot: Plotted a Q-Q plot of the residuals. If the residuals are normally distributed, the points should lie along the 45-degree line.

- Multicollinearity Check:

  Variance Inflation Factor (VIF): Calculated the VIF for each predictor variable. VIF values less than 10 indicate that multicollinearity is not a concern.

- Linear Relationship Validation:

  Residual Plot: Plotted residuals against the predicted values. If the residuals are randomly scattered around zero, it suggests that there is a linear relationship between the predictors and the response variable.

- Homoscedasticity:

  Residuals vs. Predicted Plot: Plotted residuals against the predicted values to check for constant variance. The absence of a clear pattern indicates homoscedasticity.

- Independence of residuals:

  Tested with residuals and curves.

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

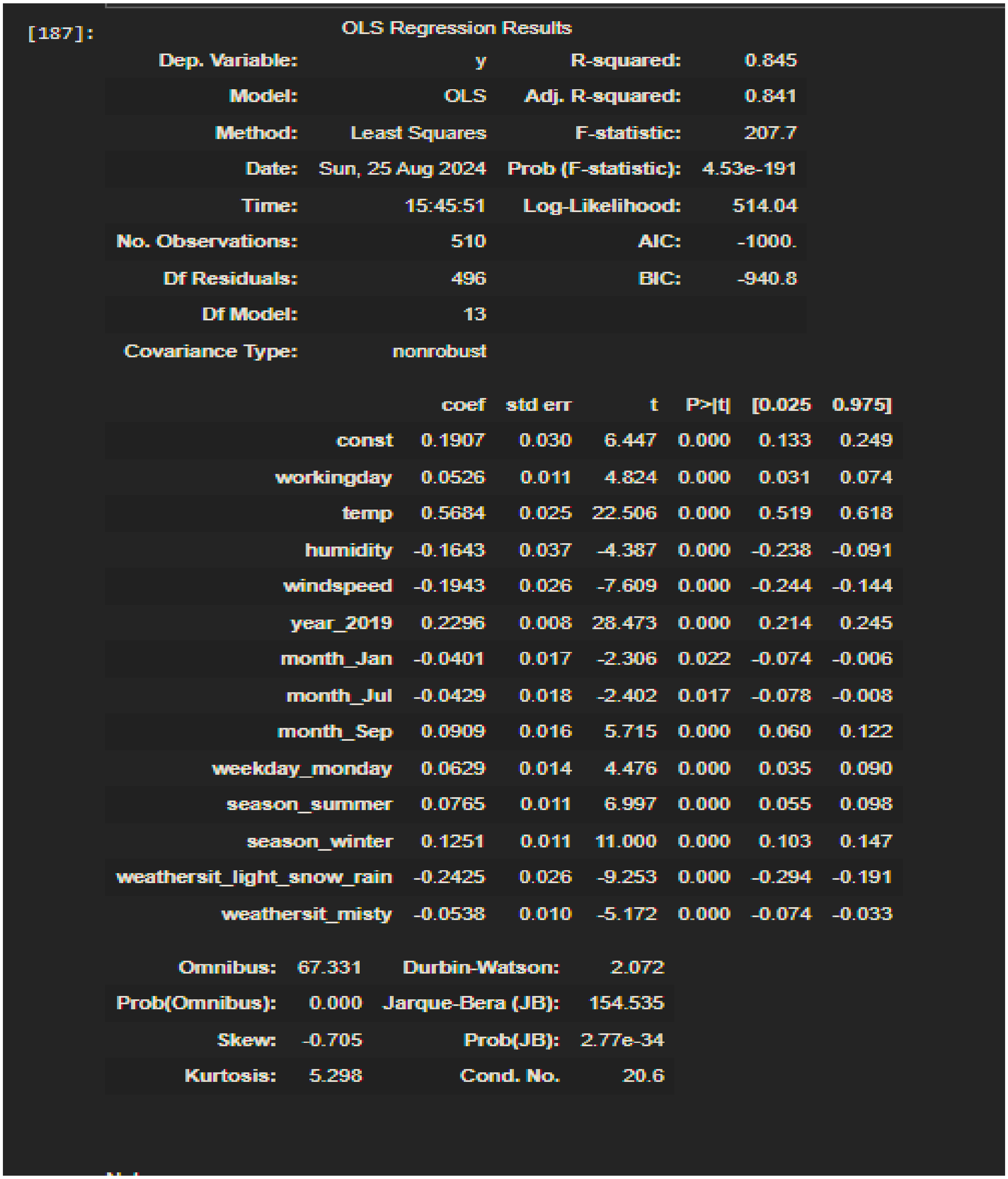
Ans: -

Top features for the model

- temp coef:0.57
- year_2019 coef:0.23
- light snow coef:-0.2425

Final Equation

count (cnt) = 0.1907 + workingday * 0.0526+ temp * 0.5684 - humidity * 0.1643 - windspeed * 0.1943 + year_2019 * 0.2296 - month_Jan * 0.0401 - month_Jul * 0.0429 + month_Sep * 0.0909 + weekday_monday * 0.0629 + season_summer * 0.0765 + season_winter * 0.1251 - weathersit_light_snow_rain * 0.2425 - weathersit_misty * -0.0538



```
[187]:
                              OLS Regression Results
        Dep. Variable:                 y          R-squared:      0.845
               Model:                OLS      Adj. R-squared:      0.841
              Method:      Least Squares          F-statistic:      207.7
                Date:   Sun, 25 Aug 2024  Prob (F-statistic):  4.53e-191
                Time:           15:45:51      Log-Likelihood:     514.04
   No. Observations:                510                  AIC:     -1000.
        Df Residuals:                496                  BIC:     -940.8
            Df Model:                 13
     Covariance Type:          nonrobust

                                 coef  std err        t   P>|t|  [0.025   0.975]
                      const    0.1907    0.030    6.447   0.000   0.133    0.249
                 workingday    0.0526    0.011    4.824   0.000   0.031    0.074
                       temp    0.5684    0.025   22.506   0.000   0.519    0.618
                   humidity   -0.1643    0.037   -4.387   0.000  -0.238   -0.091
                  windspeed   -0.1943    0.026   -7.609   0.000  -0.244   -0.144
                  year_2019    0.2296    0.008   28.473   0.000   0.214    0.245
                  month_Jan   -0.0401    0.017   -2.306   0.022  -0.074   -0.006
                  month_Jul   -0.0429    0.018   -2.402   0.017  -0.078   -0.008
                  month_Sep    0.0909    0.016    5.715   0.000   0.060    0.122
             weekday_monday    0.0629    0.014    4.476   0.000   0.035    0.090
             season_summer    0.0765    0.011    6.997   0.000   0.055    0.098
              season_winter    0.1251    0.011   11.000   0.000   0.103    0.147
 weathersit_light_snow_rain   -0.2425    0.026   -9.253   0.000  -0.294   -0.191
          weathersit_misty   -0.0538    0.010   -5.172   0.000  -0.074   -0.033

          Omnibus:  67.331     Durbin-Watson:     2.072
   Prob(Omnibus):    0.000   Jarque-Bera (JB):   154.535
             Skew:  -0.705           Prob(JB):  2.77e-34
         Kurtosis:   5.298           Cond. No.      20.6
```

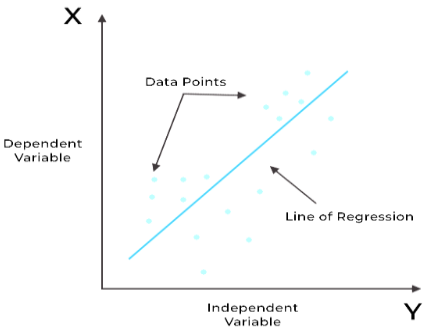
General Subjective Questions: -

Q. Explain the linear regression algorithm in detail.

Ans: Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis. The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

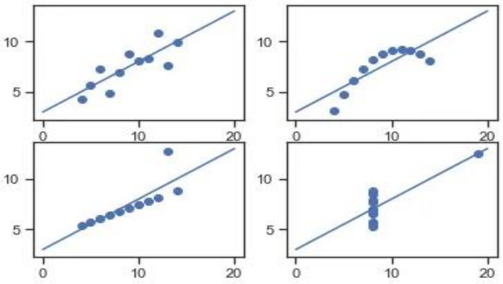
This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.



Q. Explain the Anscombe's quartet in detail.

Ans: ***Anscombe's Quartet*** is the modal example to demonstrate the importance of data visualization which was developed by the statistician ***Francis Anscombe*** in 1973 to signify both the importance of plotting data before analyzing, it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these datasets is that they all share the same descriptive statistics (mean, variance, standard deviation etc) but different graphical representations. Each graph plot shows the different behavior irrespective of statistical analysis.

However, the statistical analysis of these four datasets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represents a different behavior.



Graphical Representation of Anscombe's Quartet

- 1st data set fits linear regression model as it seems to be linear relationship between X and y.

- 2nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.

- 3rd data set shows some outliers present in the dataset which can’t be handled by a linear regression model.

- 4th data set has a high leverage point means it produces a high correlation Coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data

Q. What is Pearson’s R?

Ans: In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

Ans: Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed, then algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.
Difference between Normalizing Scaling and Standardize Scaling:

1.In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.

2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.

3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.

4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.

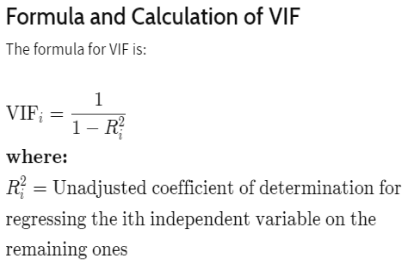
5. Normalized scaling is used when we don’t know about the distribution whereas standardized scaling is used when distribution is normal.

6. Normalized scaling is called scaling normalization whereas standardized scaling is called Z Score Normalization.

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

Ans: VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables.
The formulation of VIF is given below:

A VIF value of greater than 10 is high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.



Formula and Calculation of VIF

The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

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

Ans: Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether two distributions are similar or not.

If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots.
Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression:

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also

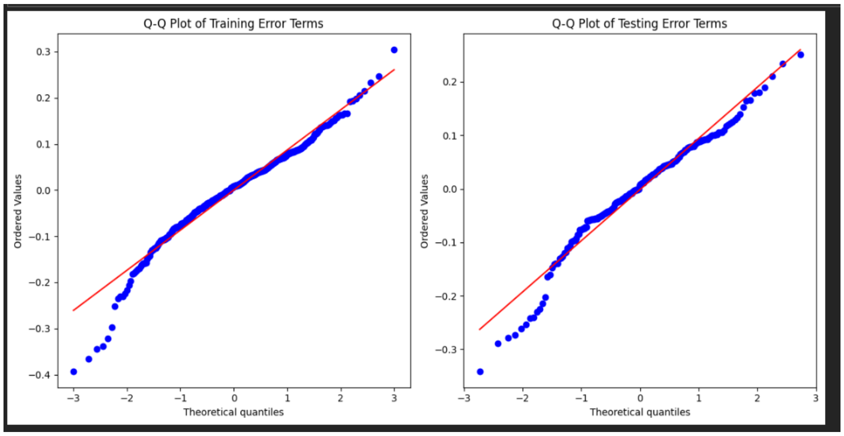
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check.

- If both datasets came from population with common distribution.

- If both datasets have common location and common scale.

- If both datasets have similar type of distribution shape.

- If both datasets have tail behavior.


Q-Q Plot of Training Error Terms
Ordered Values
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−3
−2
−1
0
1
2
3
Theoretical quantiles
Q-Q Plot of Testing Error Terms
Ordered Values
0.2
0.1
0.0
−0.1
−0.2
−0.3
−3
−2
−1
0
1
2
3
Theoretical quantiles