Multiple Linear Regression

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:
   I have done analysis on categorical columns using boxplot and bar chart. Below are few points that we can infer from the visualizations
   - Almost 32% of the bike booking happened in Fall. This was followed by summer & winter with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable. We could see considerably less rentals in Spring.
   - Considerable increase in bike rentals is seen from 2018 to 2019 indicating positive progress in the market. The company can capitalize on this trend by maintaining a competitive edge in the industry.
   - Most of the bike booking happened in the months 5,6,7,8 & 9. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
   - Almost 68% of the bike booking happened when the weather was clear. This was followed by moderate Misty weather with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
   - Almost 97% of the bike booking happened when it is not a holiday which means this data is clearly biased. This indicates, holiday cannot be a good predictor for the dependent variable.
   - weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week). This variable can have some or no influence towards the predictor.
   - Almost 69% of the bike booking happened on workingday. This indicates, workingday can be a good predictor for the dependent variable

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

   Answer:
   - drop_first=True helps in reducing the extra column created during the dummy variable creation.
   - Say for ex., we have a categorical variable called furnishing status that can have 3 possible levels - furnished, semi-furnished and unfurnished. When these levels are represented as columns with value 0 and 1 indicating the presence or

absence, all the three levels can be determined with just 2 columns - semifurnished and unfurnished as shown below

semi-furnished	unfurnished	furnishing status
1	0	semi-furnished
0	1	unfurnished
0	0	furnished

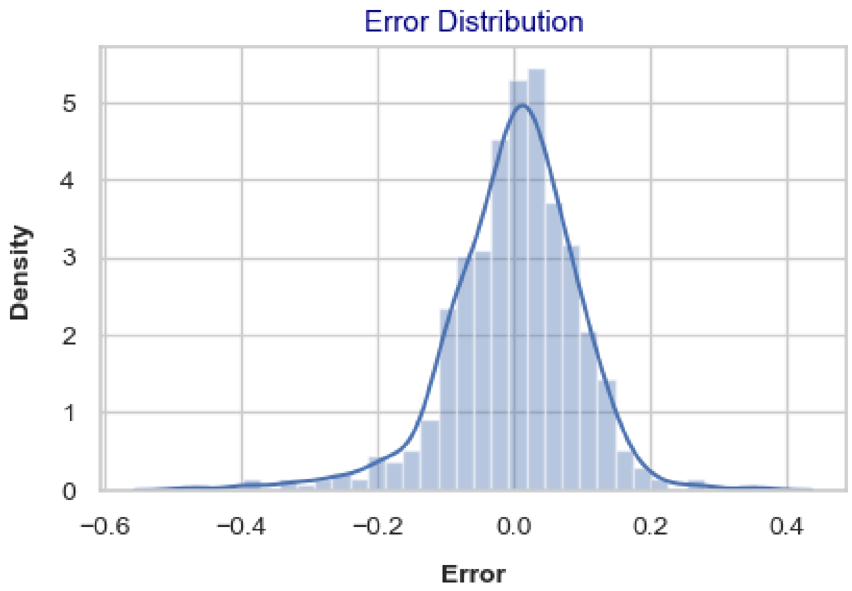
- During dummy variable creation, by default n variables will be created for n levels. In order to remove the first level, we need to use drop_first=True
- Reducing one level is with the following rationale
  - Avoid multicollinearity
  - Avoid redundancy
  - Enhance interpretation of the model

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

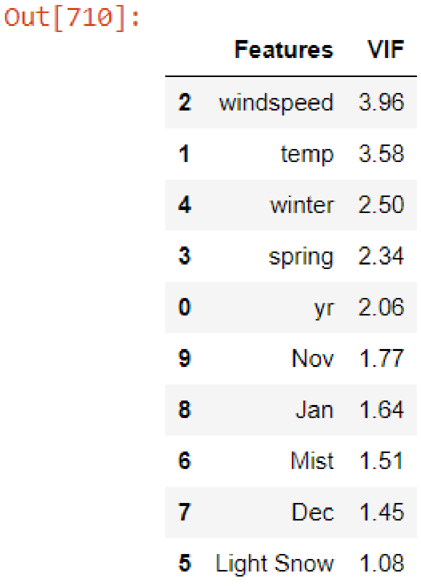
   Answer:
   'temp' variable has highest correlation with the target variable 'cnt'

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

   Answer:
   1. Normality of the error terms - Error terms should be normally distributed



2. Multicollinearity - There should be insignificant multicollinearity among variables. If the Variance Inflation Factor (VIF) of all the predictor variables are less than 5, then it confirms of no significant collinearity in the model

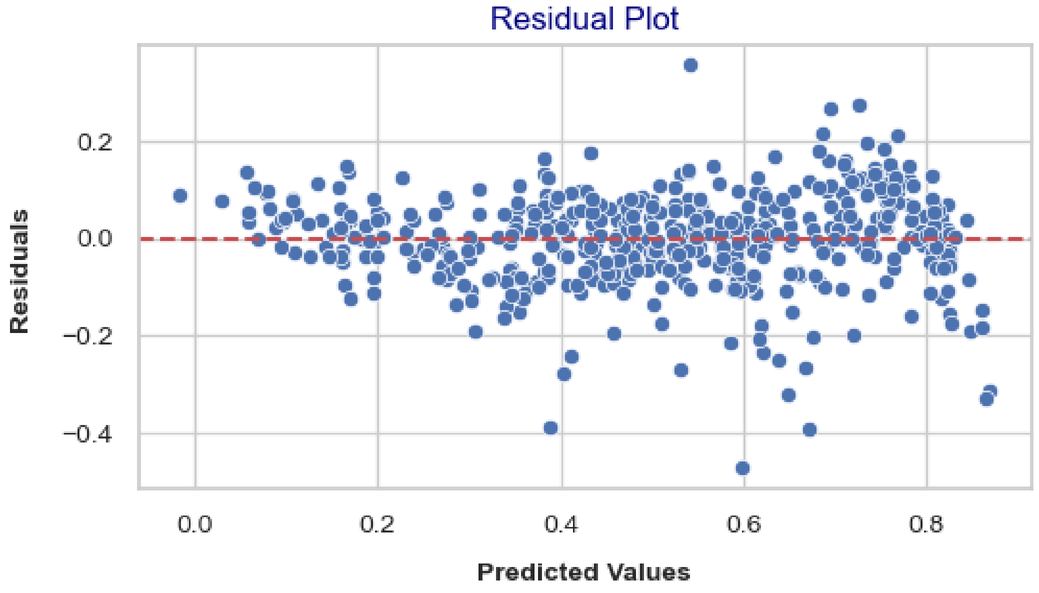


```
Out[710]:
```

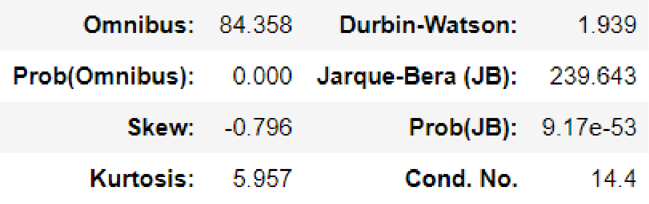
	Features	VIF
2	windspeed	3.96
1	temp	3.58
4	winter	2.50
3	spring	2.34
0	yr	2.06
9	Nov	1.77
8	Jan	1.64
6	Mist	1.51
7	Dec	1.45
5	Light Snow	1.08

3. Homoscedasticity - The spread of residuals should be roughly constant across all levels of predicted values
4. Linearity - points should fall approximately along the regression line indicating linear relationship

Above 2 assumptions are confirmed by the residual graph shown below



5. Independence of residuals - There should not be any discrete pattern seen on residuals
   As a rule of thumb, independence of residuals will be confirmed if the Durban-Watson statistic lies between 1 and 3



Omnibus:	84.358	**Durbin-Watson:**	1.939
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	239.643
Skew:	-0.796	**Prob(JB):**	9.17e-53
Kurtosis:	5.957	**Cond. No.**	14.4

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:
   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes
   - temp
   - yr
   - Light Snow/weathersit 3 (Negative correlation)

General Subjective Questions

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

   Answer:
   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Simple Linear regression, and in the case of more than one feature, it is known as Multiple linear regression.

   1) <u>Model Representation:</u>

   Simple Linear Regression
   This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.

   The equation for simple linear regression is:

y = β0 + β1X

where:

- y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable.

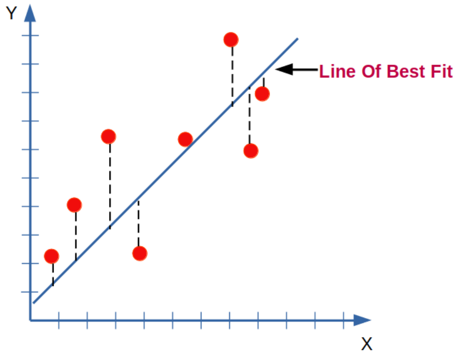
The equation for multiple linear regression is:
y = β0 + β1X1 + β2X2 + β3X3 + …… + βnXn

where:

- y is the dependent variable
- X1, X2, X3…..Xn are independent variables
- β0 is the intercept
- β1, β2, β3…. βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

2) Best Fit Line:
    - Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum.
    - The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s)



3) Hypothesis Testing:

Every time you perform a linear regression, you need to test whether the fitted line is a significant one or not or to simply put it, you need to test whether β1 is significant or not.

- Null Hypothesis (H0): β1=0
- Alternate Hypothesis (HA): β1≠0

4) Model Building:
    - Convert categorical variables into Dummy variables
    - Split data into Train and Test set (Usually 70-30 distribution)
    - Rescale numeric variables
    - Build the Model using statsmodels or sklearn or other linear regression algorithm
    - Train the model on train data set and tune the model to arrive at best fit line using top-down/ bottom-up/ regularization methods.

NOTE: As a rule of thumb, when performing manual feature selection, we consider p-values to be less than 0.05 and VIF (Variance inflation factor) to be less than 5

5) Residual Analysis:

Linear regression model is based on the below assumptions, which must be evaluated and satisfied when the model is built

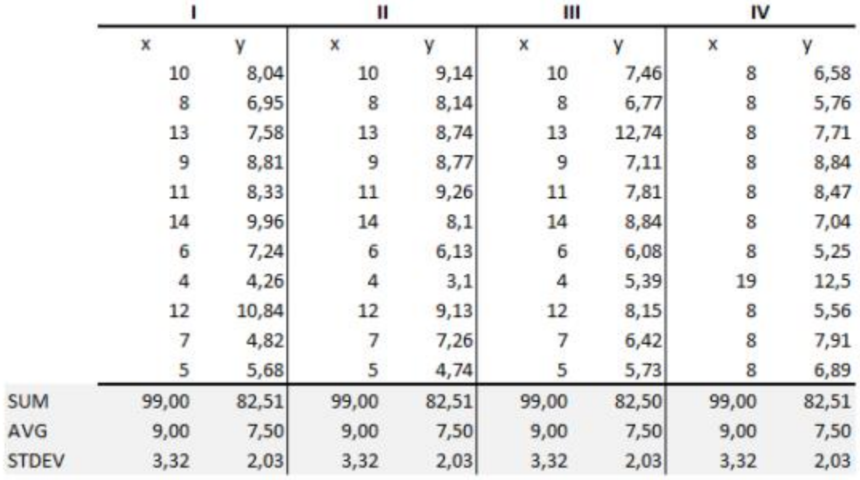
- Error terms should be normally distributed
- There should not be significant collinearity between the predictor variables.
- Homoscedasticity - The spread of residuals should be roughly constant
- Linearity - points should fall approximately along the regression line indicating linear relationship
- Independence of residuals - There should not be any discrete pattern seen on residuals

6) Making Predictions:
    - Apply scaling on Test sets
    - Predict the test set using the model built

7) Model Evaluation:
    - Calculate evaluation metrics like R-squared, Adjusted R-squared, Root mean squared error on train set and test set.
    - There should be a close alignment of R-Squared and Adjusted R-Squared values between Training and Test set indicating effective generalization
    - Evaluate model for overfitting on training data.
    - With the derived equation on best fit line, predictions will be done on the new unseen data.

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

Answer:
Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. This illustrates the importance of visualizing data and the limitations of relying solely on summary statistics.
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.



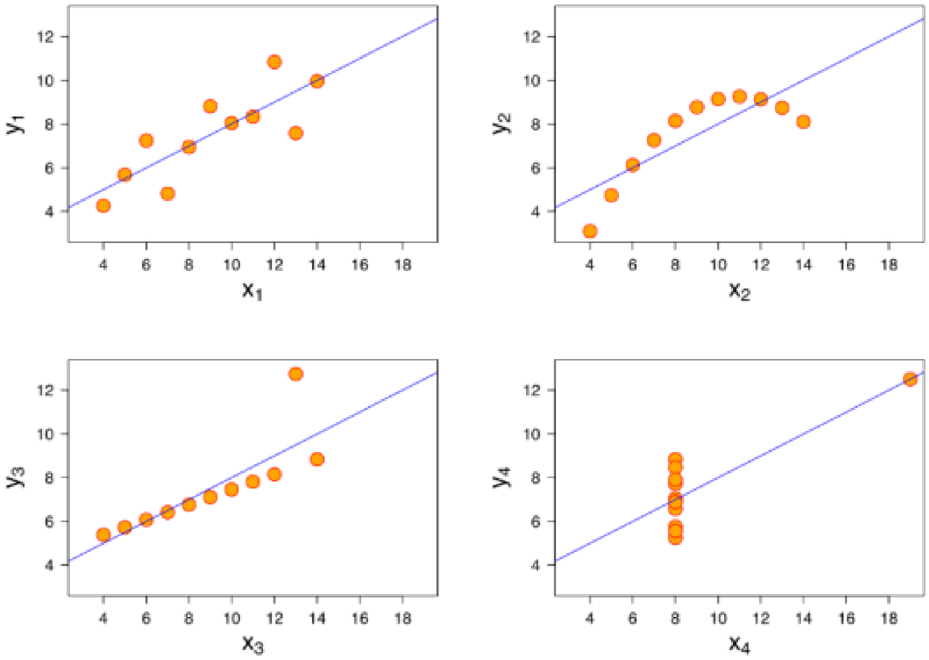
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 and mean of y is 7.50 for each dataset.
- Similarly, the 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:

- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.



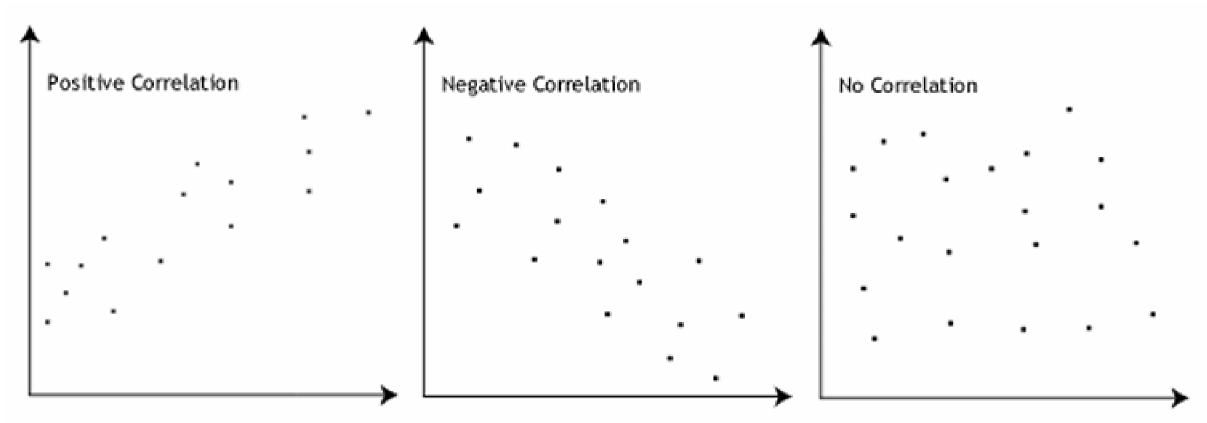
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson’s R? (3 marks)

Answer:
Pearson's correlation coefficient, often denoted as r, is a measure of the linear relationship between two variables. It quantifies the strength and direction of a linear association between two continuous variables. The coefficient takes values between -1 and 1, where:

- $r = 1$: Perfect positive linear correlation.
- $r = -1$: Perfect negative linear correlation.
- $r = 0$: No linear correlation.

A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



The formula for Pearson's correlation coefficient between two variables, X and Y, with n data points, is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Here:
X and Y are the individual data points.
$\bar{X}$ and $\bar{Y}$ are the means of X and Y.

The numerator represents the covariance between X and Y.
The denominator is the product of the standard deviations of X and Y.

Pearson's correlation coefficient is widely used in statistics to assess the strength and direction of the linear relationship between two variables. It's important to note that correlation does not imply causation, and a correlation coefficient close to zero does not necessarily mean the absence of a relationship; it only indicates the absence of a linear relationship.

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

Answer:
Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Importance of Scaling:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/ Min-Max scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

Normalization is good to use when the distribution of data does not follow a Gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors.

In Neural Networks algorithm that require data on a 0–1 scale, normalization is an essential pre-processing step. Another popular example of data normalization is image processing, where pixel intensities have to be normalized to fit within a certain range (i.e., 0 to 255 for the RGB colour range).

Standardized scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$z = \frac{x - \mu}{\sigma}$$

Standardization can be helpful in cases where the data follows a Gaussian distribution. Though this does not have to be necessarily true. Since standardization does not have a bounding range, so, even if there are outliers in the data, they will not be affected by standardization.

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

   Answer:

   The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple regression analysis. It quantifies how much the variance of the estimated regression coefficients are increased due to multicollinearity. The formula for VIF for a variable Xi is:

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

where Ri2 is the R2 value obtained by regressing Xi against all other independent variables.

When the value of VIF is infinite, it usually indicates perfect multicollinearity. Perfect multi collinearity occurs when one or more independent variables in a regression model are perfectly correlated (linearly dependent) with other variables.

In such cases:
1. There is redundant information - One variable can be expressed as a perfect linear combination of others. In case of perfect correlation, we get R-squared=1 which leads to infinite value for VIF.
2. Matrix Inversion Issues - In the computation of the VIF, there's an attempt to invert a matrix, and perfect multicollinearity leads to the matrix being singular (non invertible).

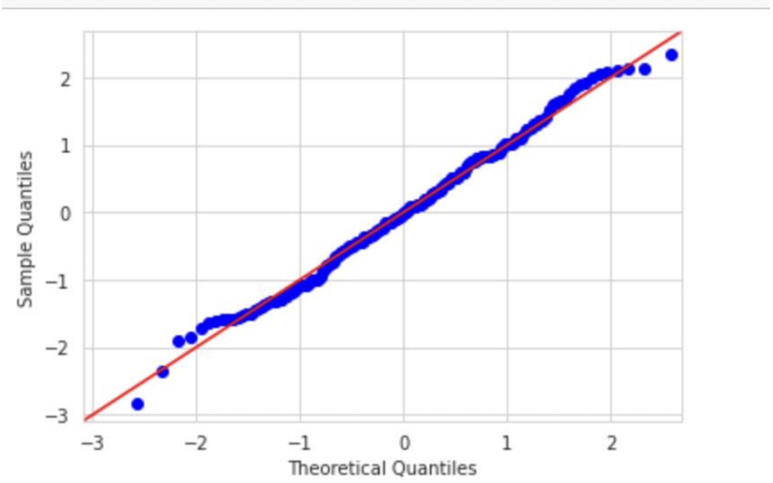
When the matrix is singular, it means that one or more variables can be predicted exactly from the others, and as a result, the computation of the VIF becomes problematic, leading to an infinite VIF value.

To address this issue, it's crucial to identify and handle multicollinearity in the dataset. This can involve removing one of the perfectly correlated variables, combining them, or using dimensionality reduction techniques. Addressing multicollinearity not only resolves the infinite VIF problem but also improves the stability and interpretability of the regression model.

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

   Answer:
   A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected distribution. If the points in the Q-Q plot approximately fall along a straight line, it suggests that the data is well-modelled by the chosen theoretical distribution.

   Sample Q-Q plot is shown below:



Importance of Q-Q Plot in linear regression:

- Normality Assessment:
  In linear regression, it is often assumed that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots are valuable for checking this assumption. If the residuals deviate significantly from normality, it can affect the reliability of statistical inferences made from the regression model.

- Identifying outliers:
  Outliers in the residuals can be detected by examining points that deviate from the expected straight line in the Q-Q plot.

- Model fit assessment:
  Q-Q plots provide a visual assessment of how well the residuals conform to a normal distribution. A good model fit is crucial for accurate predictions, and departures from normality in residuals may suggest inadequacies in the regression model.

- Validity of statistical Tests:
  When conducting hypothesis tests or constructing confidence intervals, the assumption of normality in residuals is important. Violations of this assumption can lead to inaccurate p-values and confidence intervals, affecting the validity of statistical inferences.