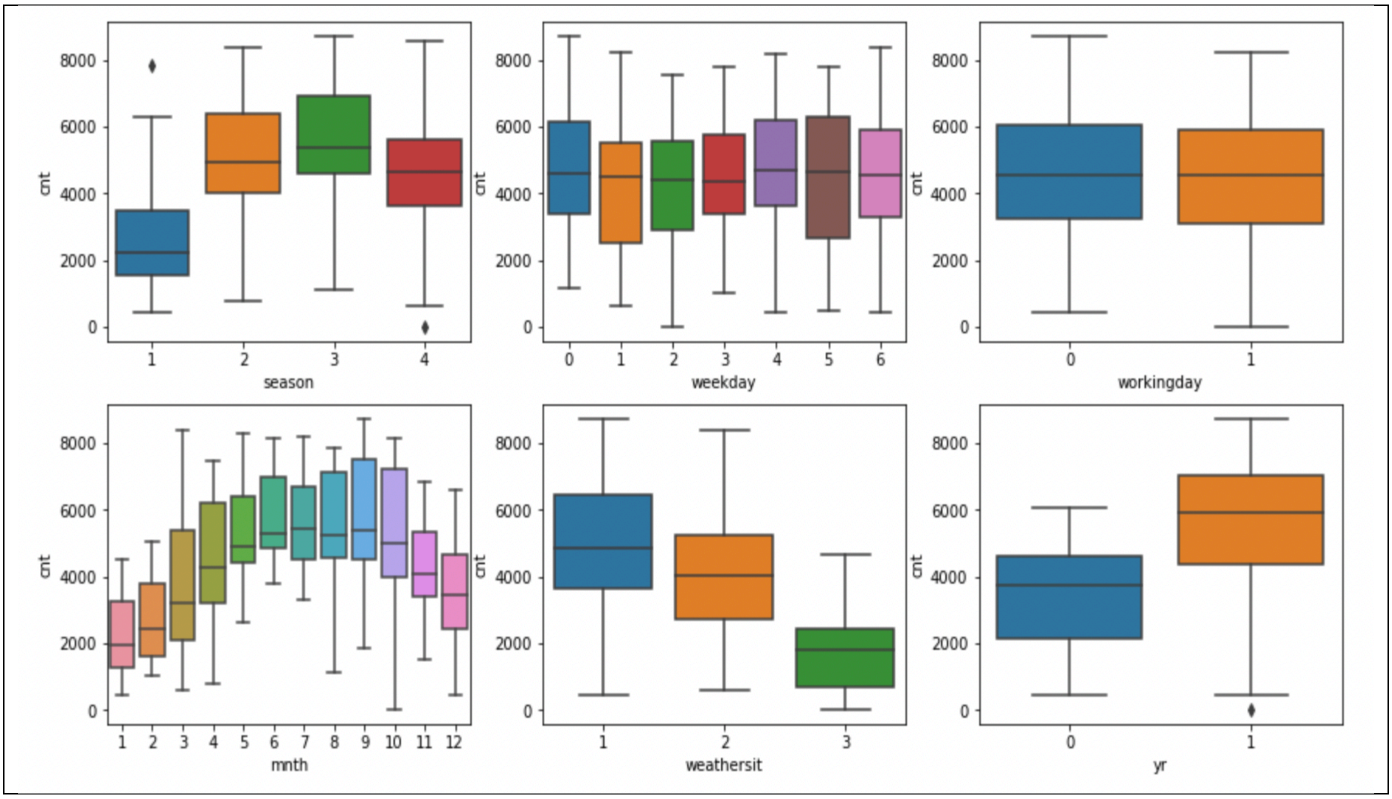
Shreeganesh S Thantri shreegs67@gmail.com

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)

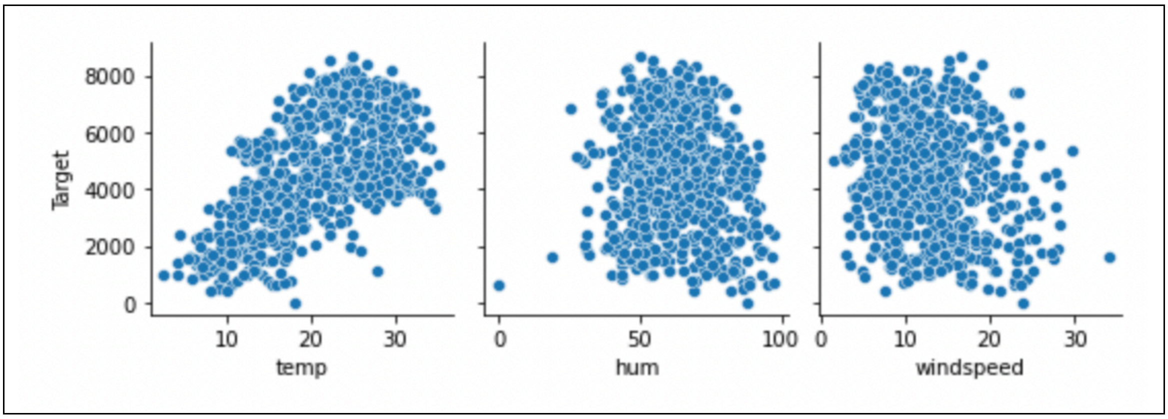


- From the above plot bike count can be seen highest in season 2 and 3 i.e, Summer and fall
- And, Sunday can be seen with highest count of bike sales Bike sale is highest in Weathersit 1, i.e., bike sale is high in Clear, Few clouds, Partly cloudy, Partly cloudy
- Bike sale is higher in the month of September and October.
- And, 2019 has more number of bike sales than that of 2018.
- Over all from the model year 2019 and season 4 from the categorical column has highest influence over bike sales.

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

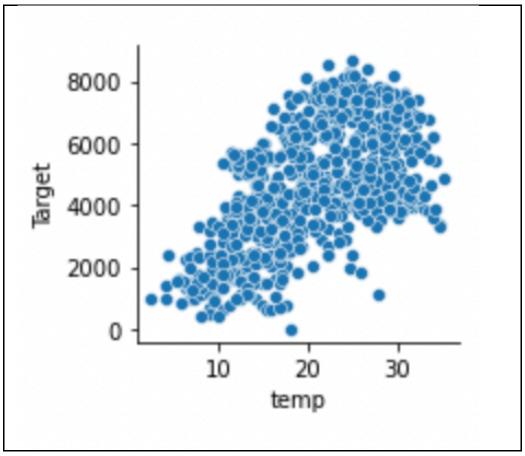
- While creation of dummy variable the algorithm will create dummy variable for each if the value counts in particular column so, drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

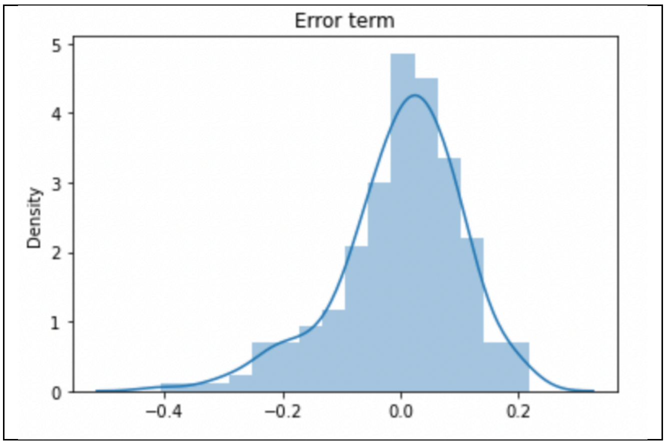


- **Temperature** variable has the highest correlation among the numerical variable

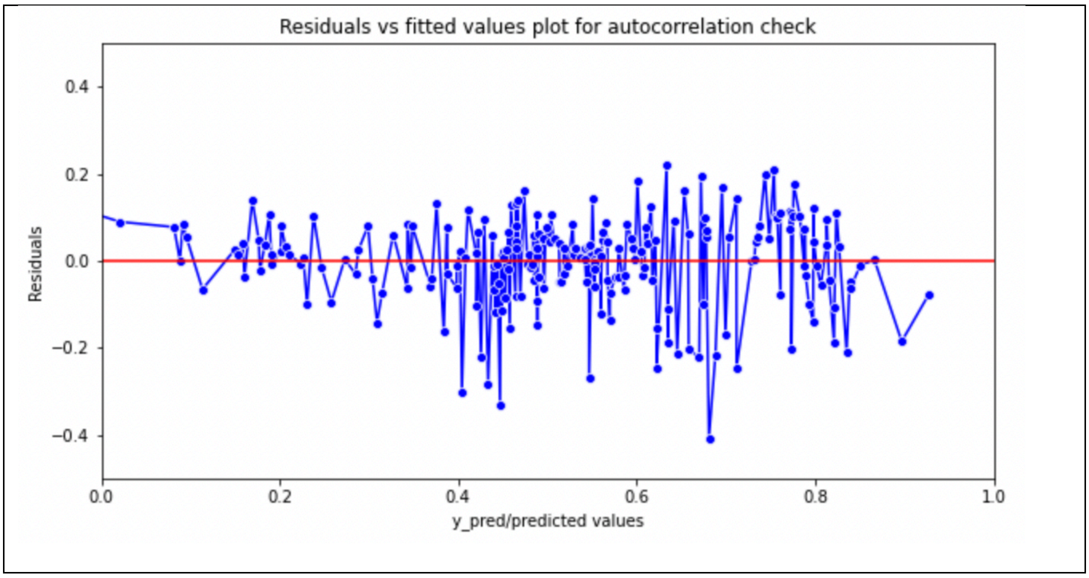
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)



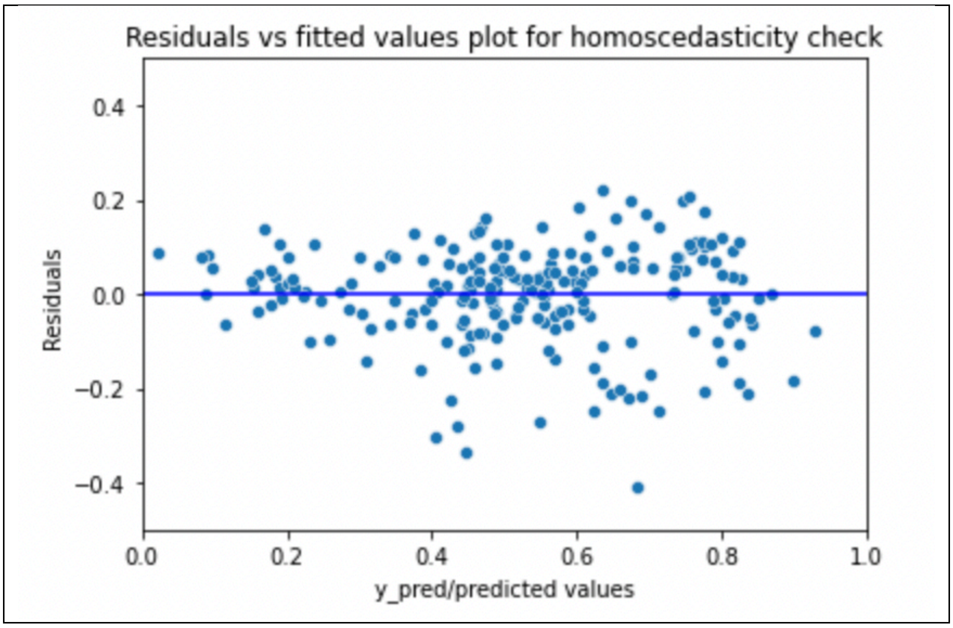
- Since the chosen numerical variable is temperature and from above figure we can see that it has linear relationship with target variable, hence 1st assumption has met



- And above figure shows that residues are normally distributed, that is 2nd assumption



- Residuals are independent of each other, i.e, no auto correlation between residuals, hence the 3rd assumption



- Errors terms should have constant variance, i.e, Homoscedasticity, hence the 4th assumption

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 model top 3 contributing features are:

- Temp (Temperature)
- yr_1 (year)
- season_4 (season)

General Subjective Questions

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

   Linear regression (LR) models are used to analyze the relationship between an independent variable (IV) or variables and a dependent variable (DV), a.k.a the predicted variable. If only one predictor variable (IV) is used in the model, then that is called a single linear regression model. However, the model will not be robust in design and will have little to no explanation power because in the real world there is no 1 variable that can fully explain, or predict, an outcome. Most commonly, the model will have multiple IVs which will make it a multiple linear regression model.

   Assumptions of LR:

   - There should be linear relationship between the dependent and independent varibales
   - Error terms should normally distributed
   - Error terms are independent of each other
   - Error terms have constant variance

Method:

Formula of a straight line and linear regression

$$\mathbf{y = mx+b} \quad \Longrightarrow \quad \mathbf{y = a_0+a_1x}$$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the mapping function that maps the input variable to the output variable. This mapping function is also known as the Hypothesis function.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation y=mx+b we can calculate MSE as:

Let’s y = actual values, yi = predicted values

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Model parameters xi, b (a0,a1) can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.

The main function to calculate values of coefficients

- Initialize the parameters.
- Predict the value of a dependent variable by given an independent variable.
- Calculate the error in prediction for all data points.
- Calculate partial derivative w.r.t a0 and a1.
- Calculate the cost for each number and add them.
- Update the values of a0 and a1.

So in Summary:

In Regression, we plot a graph between the variables which best fit the given data points. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).To calculate best-fit line linear regression uses a traditional slope-intercept form. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

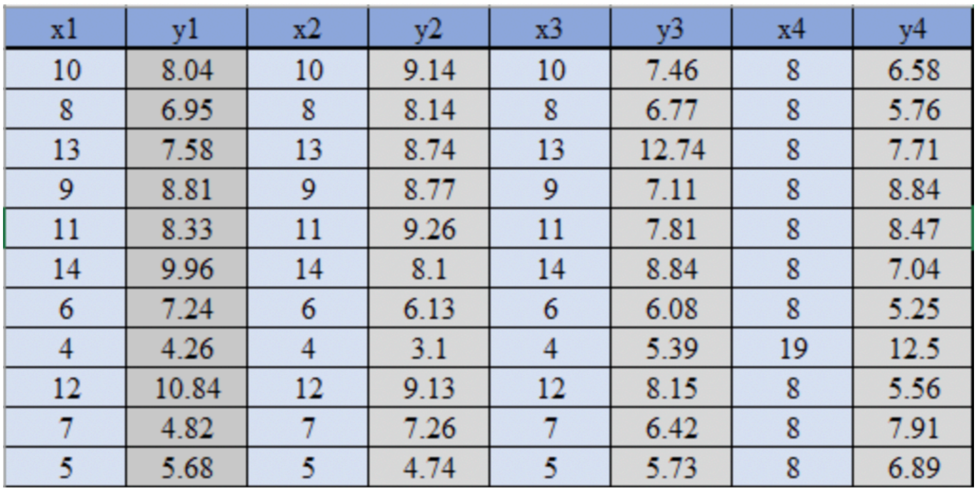
The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, Mean Squared Error (MSE) cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE)

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

Let's say we're looking at a spreadsheet of our customers. We have data about how many times they've logged in, how much revenue we've earned from them, etc. We can immediately calculate several compelling summary statistics: what's the average number of logins per customer? What's the average revenue?What's the correlation between number of logins and revenue?Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business.But there's a danger in relying only on summary statistics and ignoring the overall distribution. We took a look at this earlier as it relates to average revenue per user.

Perhaps the most elegant demonstration of the dangers of summary statistics is Anscombe's Quartet. It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x,y) pairs as follows:



x1	y1	x2	y2	x3	y3	x4	y4
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

All the summary statistics you'd think to compute are close to identical:

The average x value is 9 for each dataset

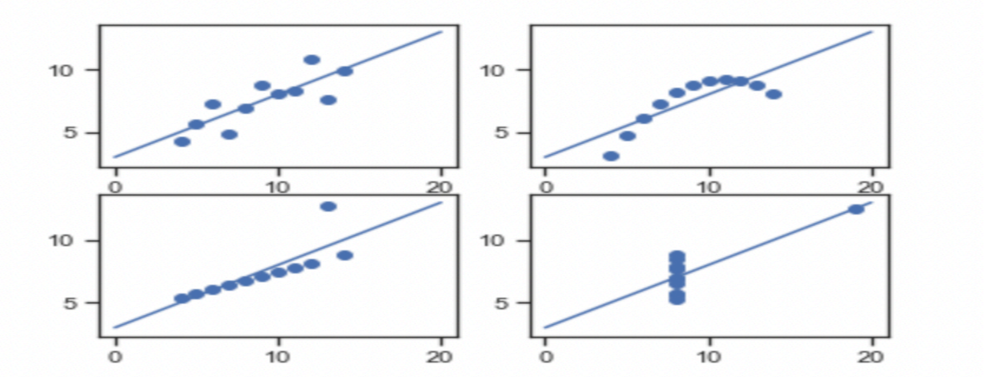
The average y value is 7.50 for each dataset

The variance for x is 11 and the variance for y is 4.12

The correlation between x and y is 0.816 for each dataset

A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:



Now we see the real relationships in the datasets start to emerge.
Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance.
Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?).
Dataset III looks like a tight linear relationship between x and y, except for one large outlier.
Dataset IV looks like x remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

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

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. If you're starting out in statistics, you'll probably learn about Pearson's R first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson's.

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ϱ) for a population and the letter "r" for a sample.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

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

Real-world datasets often contain features that are varying in degrees of magnitude, range and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

In the world of science, we all know the importance of comparing apples to apples and yet many people, especially beginners, have a tendency to overlook feature scaling as part of their data preprocessing for machine learning.

Why Scaling is important:

Gradient descent is an iterative optimisation algorithm that takes us to the minimum of a function.

Machine learning algorithms like linear regression and logistic regression rely on gradient descent to minimise their loss functions or in other words, to reduce the error between the predicted values and the actual values.

Having features with varying degrees of magnitude and range will cause different step sizes for each feature. Therefore, to ensure that gradient descent converges more smoothly and quickly, we need to scale our features so that they share a similar scale.

Normalisation

Normalisation, also known as min-max scaling, is a scaling technique whereby the values in a column are shifted so that they are bounded between a fixed range of 0 and 1.

MinMaxScaler is the Scikit-learn function for normalisation.

Standardisation

On the other hand, standardisation or Z-score normalisation is another scaling technique whereby the values in a column are rescaled so that they demonstrate the properties of a standard Gaussian distribution, that is mean = 0 and variance = 1.

StandardScaler is the Scikit-learn function for standardisation.

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

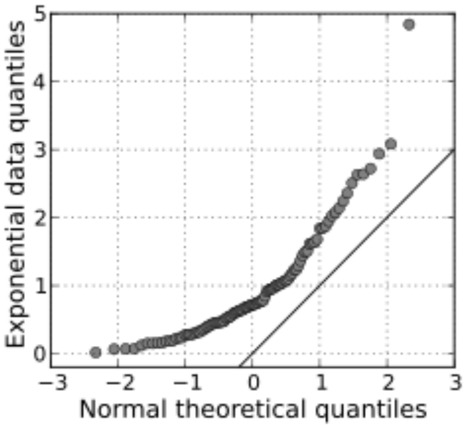
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.