Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)

Answer: -

The categorical variables from the dataset can have following inferences:

- Season:
  - Bike Rentals are happening more during the fall (Monsoon) season.
- Weather:
  - Bikes seem to be rented more in clear or partly cloudy or Mist weather.
- Working day:
  - In a week (excluding holidays) there are usually 5 weekdays/working days and 2 weekends (sat and sun). Though the count of bike sharing is more on working day, Holiday counts are almost nearer to working day. This implies that people tend to use bikes more on Holidays.
- Year:
  - Bike Rental popularity has increased in 2019 when compared to 2018.
- Month:
  - Bike Rental has increased approximately during the months between June to October
- Week day:
  - Bike Rentals are maximum on Sundays and Saturdays
- Temperature:
  - Bike Rentals are observed at higher temperatures.
- Humidity:
  - Temperature being directly proportional to Humidity, Bike Rentals are making duringhigh humidity.
- Wind speed:
  - Wind speeds increase with a greater temperature difference. Wind speed near thesurface is most highly correlated with the temperature.

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

Answer: -

When creating dummy variables from categorical variables, the drop_first=True parameter is used to drop the first level or category of the variable. This is important for several reasons:

1) To avoid Multicollinearity: Including dummy variables for all categories of a categorical variable can lead to multicollinearity in the dataset. Multicollinearity occurs when two or more predictor variables are highly correlated with each other. By dropping the first category, we create a baseline or reference level for the variable, and the remaining dummy variables capture the information about the other categories relative to the baseline. The remaining variables gives the same information without the category we dropped and hence this eliminates perfect multicollinearity among the dummy variables.

2) Interpretability of coefficients: Dropping the first category helps in interpreting the coefficients of the dummy variables. By dropping the first category, the coefficients of the remaining dummy variables represent the difference in the outcome variable between each category and the baseline category directly.

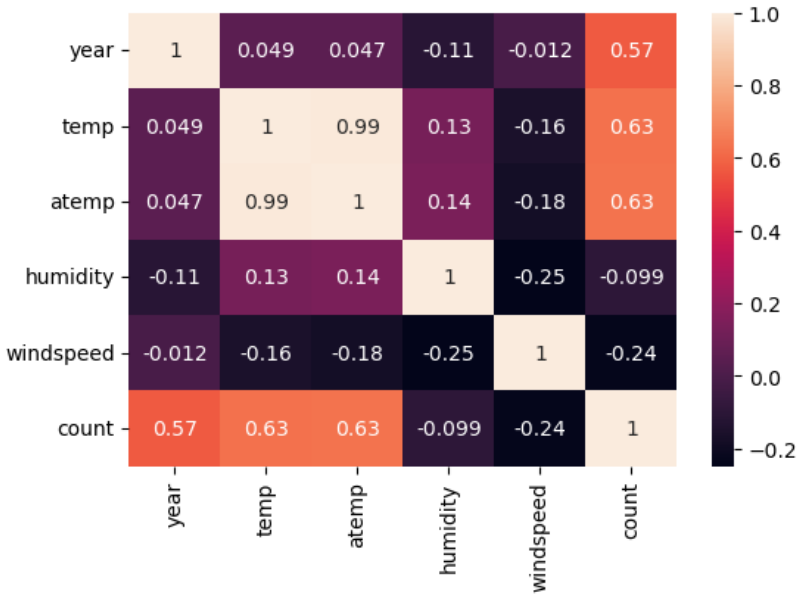
3) Reducing dimensionality: Including dummy variables for all categories increases the dimensionality of the dataset, which can be problematic for certain algorithms when dealing with large datasets. Dropping the first category reduces the number of dummy variables by one, which can help in improving computational efficiency of an algorithm.

Overall, using drop_first=True when creating dummy variables is important to address multicollinearity, enhance interpretability of coefficients, and reduce dimensionality, resulting in a more meaningful and efficient analysis.

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 which is temperature is highly correlated with the target variable (cnt which is count) when compared to other variables.This can be portrayed based on the results of either pair plot or heatmap:



---

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

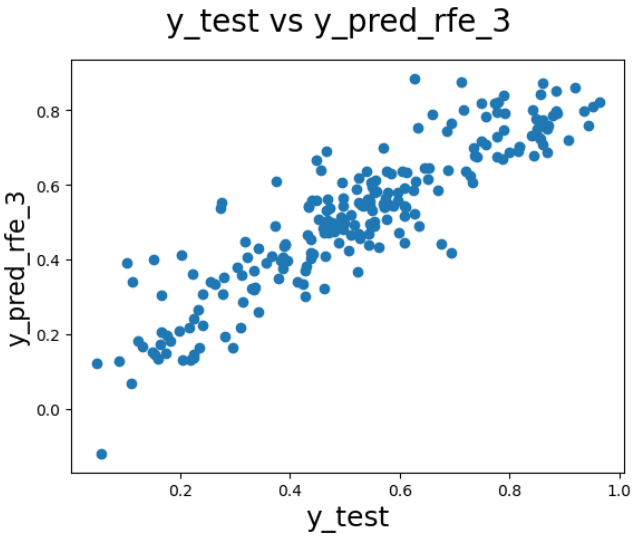
Answer: -

Let's take a look at what the assumptions of simple linear regression were:

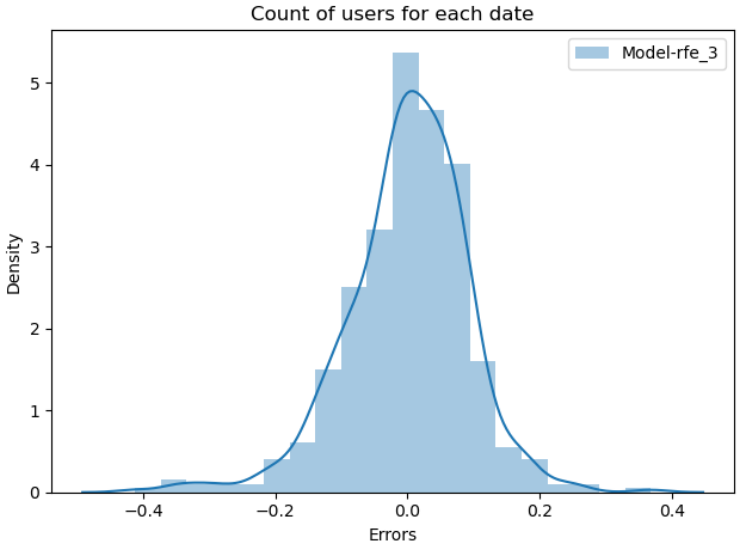
1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Validating the assumptions after building the multiple linear regression model on the training set involves examining several aspects of the model's performance and the residuals. Here are steps performed to validate the assumptions:

1) **Linearity:** Checked the linearity assumption by plotting the predicted values against the actual values. From below figure as the relationship appears to be linear, it suggests that the linearity assumption is reasonable. Nonlinear patterns may indicate that additional transformations or nonlinear models are needed.

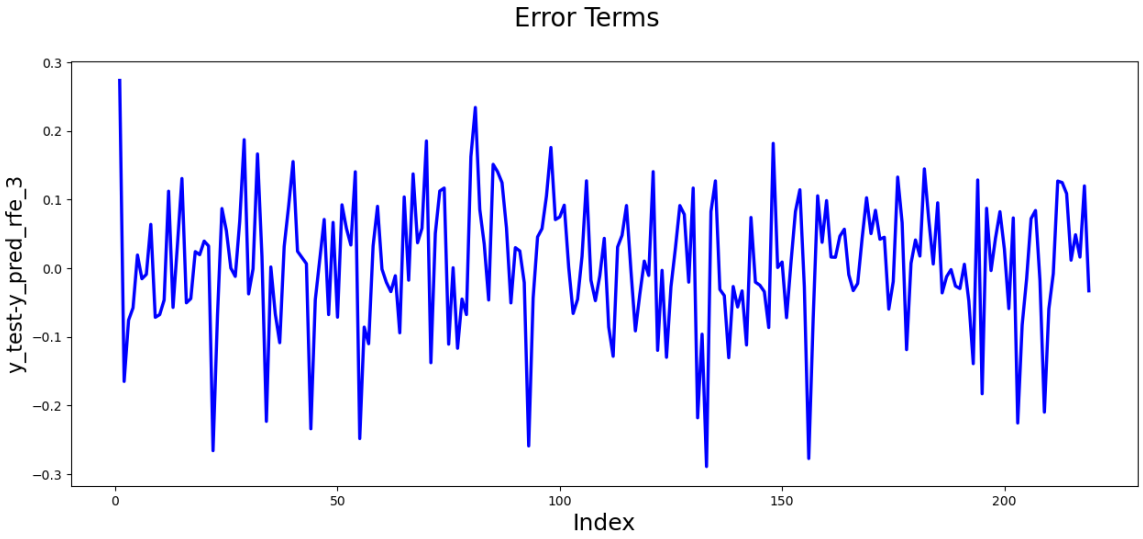


2) **Normality of residuals**: We checked if the residuals are normally distributed by ploting a distplot plot of the residuals to assess their distribution.



3) **Independence of residuals:** We need to examine the residuals to ensure they are independent of each other. We do this by Plotting the residuals against the predicted values. We can do this by plotting residuals against the independent variables as well. If there is a pattern or structure in the residuals, such as a funnel shape or distinct groups, it suggests violations of the independence assumption. For below diagram we can confirm that there is no independence of the residuals.

4) **Homoscedasticity**: Assess whether the residuals exhibit a constant variance (homoscedasticity) across the range of predicted values. We do this by Plotting the residuals against the predicted values. We can also do this by plotting residuals against the independent variables. If the spread of the residuals appears to change systematically or there is a funnel-like pattern, it indicates heteroscedasticity.



By examining these aspects and conducting the appropriate plots, validated the assumptions of linear regression model.

---

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: -

Among some variables which are contributing and benefitting the demand of the shared bikes count are as follows:

Label	coef	std err	t	P>|t|	[0.025	0.975]
temperature	0.3883	0.026	14.944	0	0.337	0.439
Sunday	0.0641	0.015	4.418	0	0.036	0.093
working_day	0.0515	0.011	4.577	0	0.029	0.074

In addition to this we can consider year also.

General Subjective Questions

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

Answer:

Linear regression is a widely used statistical algorithm for modeling the relationship between a dependent variable and one or more independent variables which are also called as predictive variables. It assumes a linear relationship between the variables and aims to find the best-fit line that minimizes the difference between the observed data points and the predicted values. This model can be used to forecast the independent variable.

Examples: Forecast sales, to analyze customer behavior.

When there is only dependent feature it is called Uni-variate Linear Regression and if there are multiple dependent features, it is called Multiple Linear Regression.

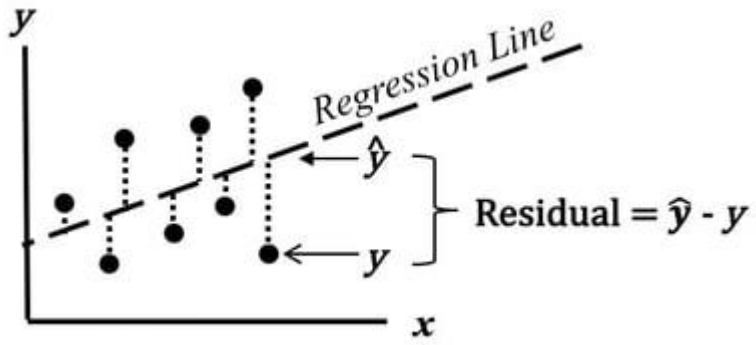
Explanation of the linear regression algorithm:

1. Data Representation:

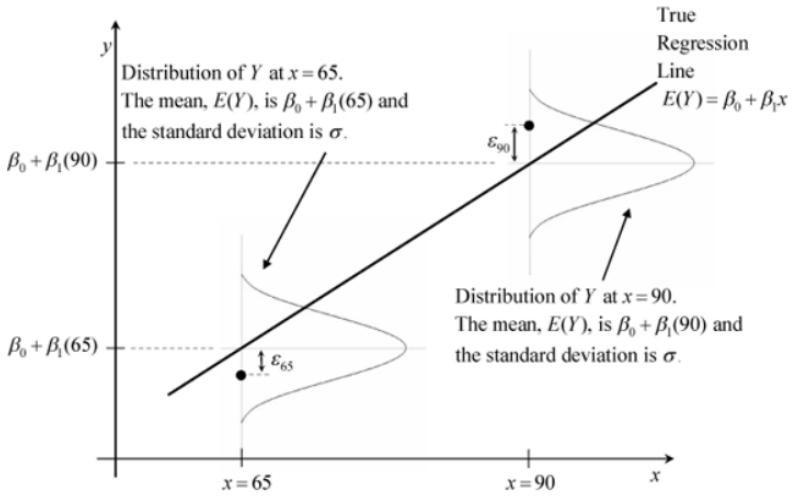
- Let's denote the dependent variable (the variable we want to predict) as "Y" and the independent variable(s) (the variables we use to predict Y) as "X".
- Suppose we have "n" data points, so we have n pairs of observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n).

2. Assumptions:

- Linearity: Linear regression assumes that there is a linear relationship between X and Y. That is, Y can be represented as a linear combination of X and a constant term.
- There are assumptions related to residuals which is difference of actual Y value and predictive Y value from the model:



- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity) : The variance of the errors (the differences between the observed and predicted values) should be constant across all levels of X.



3. Model Representation:

- Linear regression assumes a linear equation of the form:

  $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_pX_p + \varepsilon$

  where Y is the dependent variable, X_1, X_2, ..., X_p are the independent variables, β_0, β_1, β_2, ..., β_p are the coefficients to be estimated, and ε is the error term.

4. Estimating Coefficients:

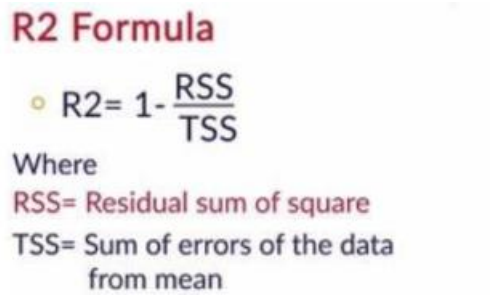
- The goal is to find the values of β_0, β_1, β_2, ..., β_p that minimize the difference between the observed Y values and the predicted Y values.
- This is typically done by minimizing the sum of squared errors (SSE), also known as the ordinary least squares (OLS) method.
- The coefficients can be estimated using various techniques such as matrix operations, gradient descent, or statistical libraries.

5. Model Evaluation:

- Once the coefficients are estimated, we evaluate the model's performance and determine its effectiveness in predicting Y.
- Common evaluation metrics for linear regression include the coefficient of

determination (R^2), root mean squared error (RMSE), mean squared error (MSE), etc.

- R^2 measures the proportion of the variance in the dependent variable that can be explained by the independent variables.
- We also consider Adjusted R-square which penalizes the model for adding more and more variables to our model while building it.



6. Predictions:

- Once the model is built and evaluated, it can be used to make predictions on new or unseen or test data by substituting the values of X into the equation.
- The predicted value of Y represents the expected value given the input values of X.

7. Assumptions Checking and Residual analysis of the data:

- It is important to check the assumptions of linear regression to ensure the validity of the model.
- Assumptions can be examined by assessing the residuals (the differences between the observed and predicted values).
- If the assumptions are violated, alternative regression techniques or transformations may be necessary.

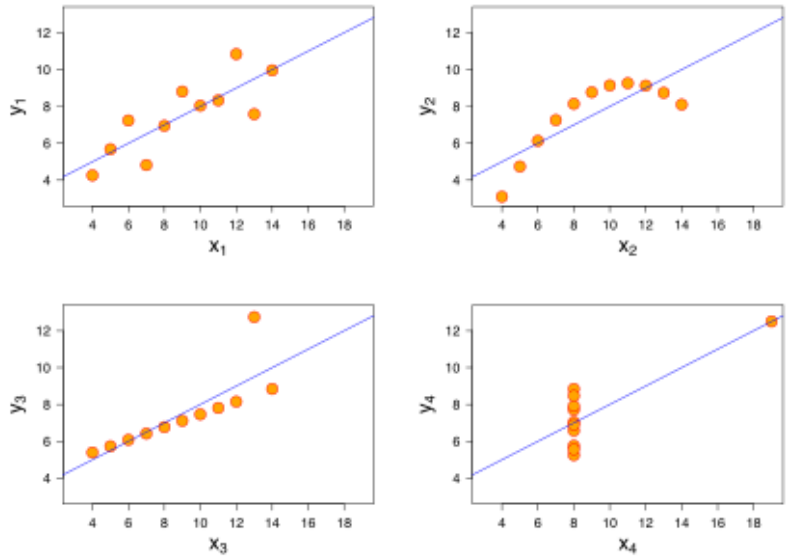
Linear regression is a simple yet powerful algorithm that can provide valuable insights into the relationship between variables. It is widely used in various fields such as economics, finance, social sciences, and machine learning.

---

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

Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.



Note: It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Explanation of this above graph:

a) In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
b) In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
c) In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
d) Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

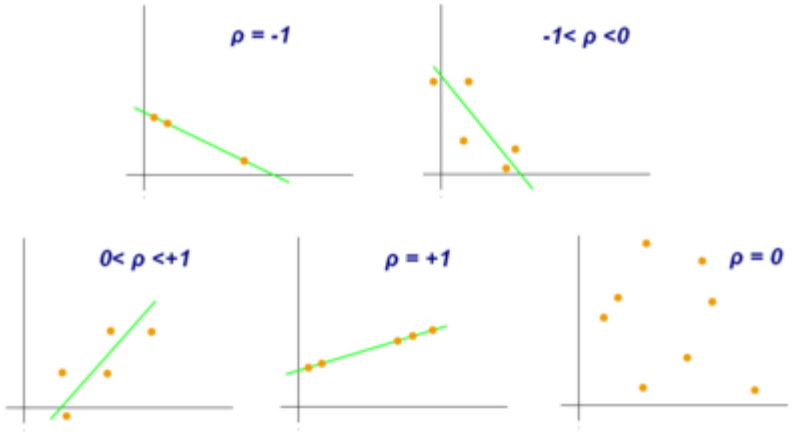
Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Answer:

Generally, Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in statistics is the Pearson Correlation also known as Pearson ProductMoment Correlation coefficient (PPMCC). 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: rho (ρ) for a population and the letter "r"for a sample.



The bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no-linear correlation, and −1 is total negative linear correlation.

For a population:

Pearson's correlation coefficient, when applied to a population, may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables (X,Y), the formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y.

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.

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.

Normalization vs. Standardization:

The terms normalization and standardization are sometimes used interchangeably, but they usually refer to different things. Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a z-score, and data points can be standardized with the following formula:

1. Normalization: It brings all of the data in the range of 0 and 1. Can be implemented through ***sklearn.preprocessing.MinMaxScaler*** in python.

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

2. 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). ***sklearn.preprocessing.scale*** helps to implement standardization in python.

$$X = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially information related to outliers.

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

Answer:

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equalto the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

If all the independent variables are orthogonal to each other, then VIF = 1.0. 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. This would mean that that standard errorof this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

VIF	Conclusion
1	No Multicollinearity
2 to 5	Moderate
> 5	not be ignored and inspected appropriately
10 or greater	Severe

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

$$VIF_i = 1/(1 - R2_i)$$

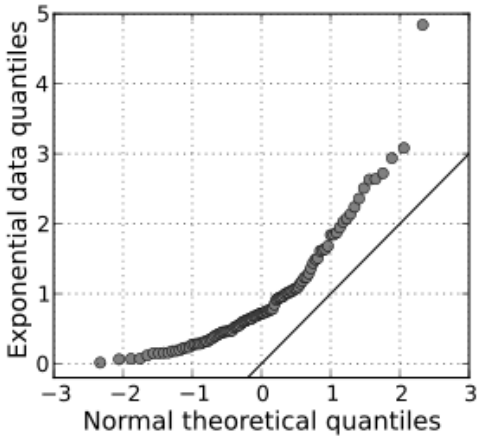
In the case of perfect correlation, we get R2 =1, which lead to 1 / (1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

A Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First,the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to oneof the quantiles of the second distribution (y-coordinate) plotted against the same quantileof the first distribution (x-coordinate). Thus, the line is a parametric curve with the parameter which is the number of the interval for the quantile.

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 canalso be used as a graphical means of estimating parameters in a location-scale family of distributions.



A normal Q–Q plot of randomly generated, independent standard exponential data, (X ~ Exp (1)). This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a strongly nonlinear pattern, suggesting that the data are not distributed as a standard normal (X ~ N (0,1)). The offset between the line and the points suggests that the mean of the data is not 0. The median of the points can be determined to be near 0.7

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based ondata, there are multiple quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions