1. What is Linear regression?

1) linear regression is Supervised machine learning algorithm used to find linear relationship between dependent variable and one or more independant variable.

2)Dependent variables will be continuous.

3)Independent variable may be discrete or may be continuous.

Linear regression expression

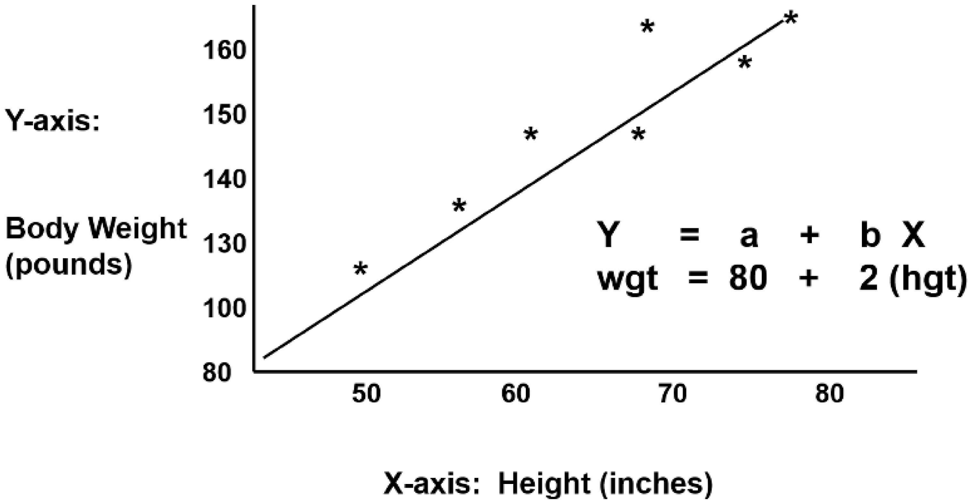
y = mx+c

m => will be the slope of linear regression

c => will be the intercept

x => will be the independent variable

y => will be the dependent variable



2. How do you represent a simple linear regression?

1)simple linear regression means we have one independant varible and one depedant variable

2) we can represent simple linear regression by below expression

y = mx+c

m => will be the slope of linear regression

c => will be the intercept

x => will be the independent variable

y => will be the dependent variable

3. What is multiple linear regression

1)multiple linear regression means we have two or more independant varible and one depedant variable

2) we can represent multiple linear regression by below expression

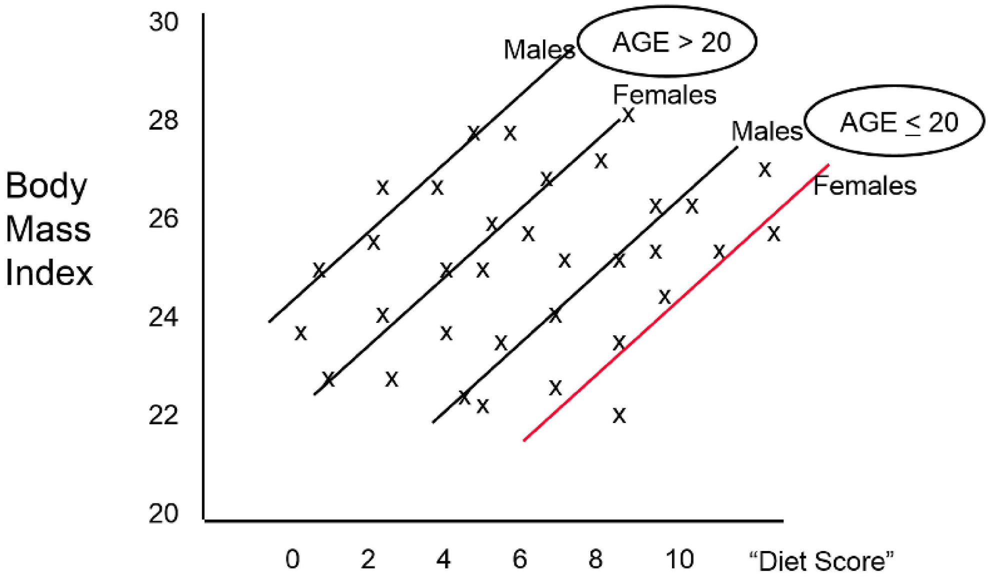
y = m1x1+m2x2 +m3x3+.....MnXn+c

m1,m2,m3 => will be the slope of dependant variable x1,x2,x3

c => will be the constant intercept

x1,x2,x3 => will be the independent variables

y => will be the dependent variable



BMI = 18.0 + 1.5 (diet score) + 1.6 (if male) + 4.2 (if adult)

$$\boxed{Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3}$$

4. What are the assumptions made in the Linear regression model?

Basically, there are four assumptions used to build model.

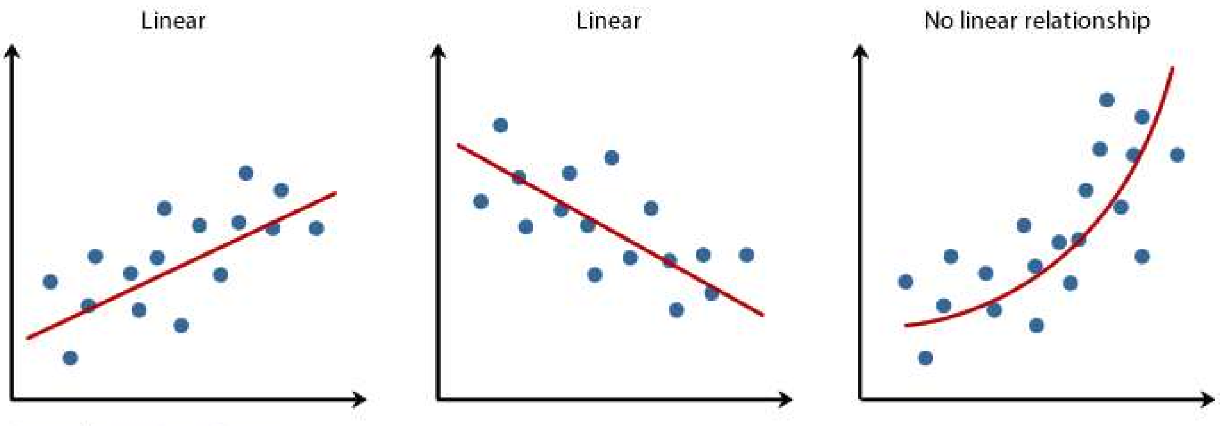
1. Linearity – There should be linear relationship between dependent and independent variable
2. No Multicollinearity- All the independent variable are independent to each other

3.Normality of Residuals - All the residuals should be normally distributed.

4.Homoscedasticity Residuals or No Heteroscedasticity Homoscedasticity of Residual or No Heteroscedasticity : Residual should follow Homoscedastic behavior (Fitted values Vs Residual)

1] Linearity:

1) Linearity means there should be linear relationship between dependent and independent variable.



Copyright 2014. Laerd Statistics.

2) For checking the linearity we have to check the coefficient of correlation which also called as R value or pearson correlation

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

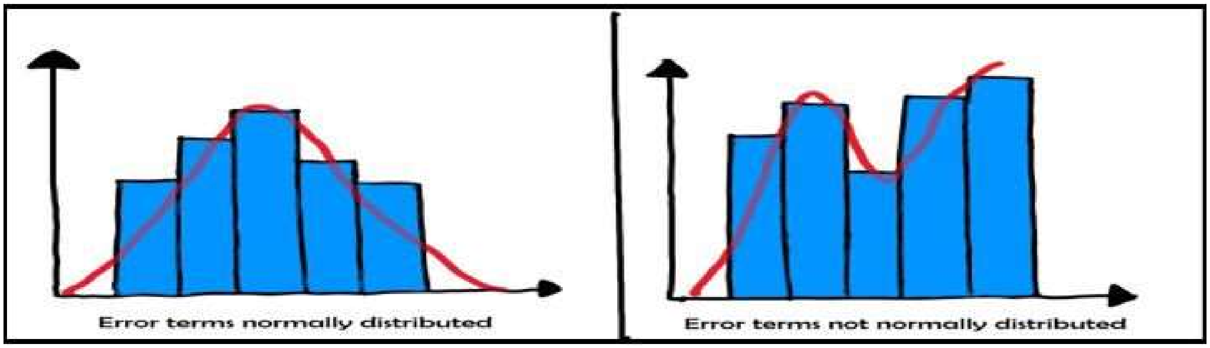
Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

2] No Multicollinearity -

All the independent variable are independent to each other

3).Normality of Residuals

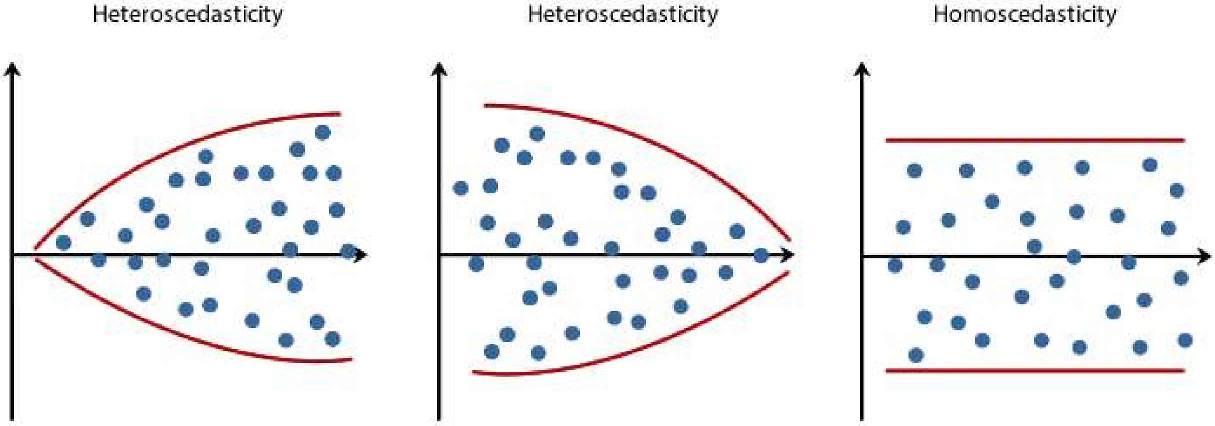
1)All the residuals(error) should be normally distributed.



4)Homoscedasticity Residuals or No Heteroscedasticity -

1) There should be Homoscedastic relation between each residual which means all the residuals should be on same range.

2)All the residuals should have constant variance.



Copyright 2014. Laerd Statistics.

5. What if these assumptions get violated?

1. Linearity assumption get violated?

If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model.

Also, this will result in erroneous predictions on an unseen data set.

2.No Multicolinearity assumption gets violated?

Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

features or independent variables.

3.Normality of Residual assumption gets violated?

If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.We can say if residuals are not normally distributed then there are some unusual datapoints in dataset.

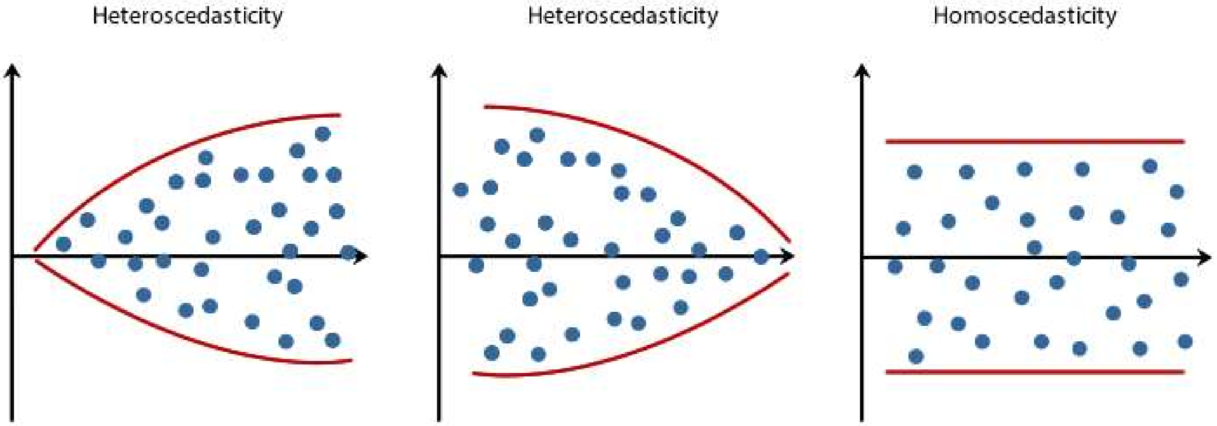
4.Homoscedasticity assumption gets violated?

we know that Homoscedasticity means constant vriance.The presence of non-constant variance in the error terms results in heteroskedasticity. Generally,non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

6. What is the assumption of homoscedasticity?

1) There should be Homoscedastic relation between each residual which means all the residuals should be on same range.

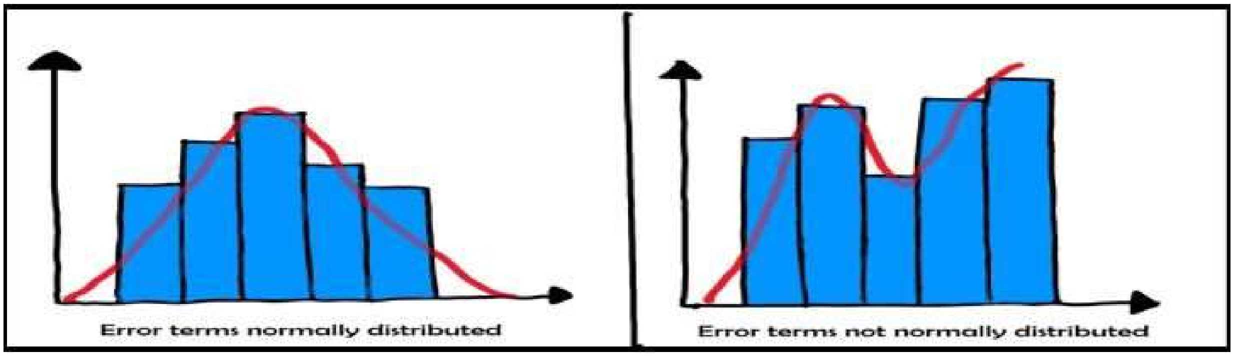
2)All the residuals should have constant variance.



Copyright 2014. Laerd Statistics.

7. What is the assumption of normality?

1)All the residuals(error) should be normally distributed. 2)If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

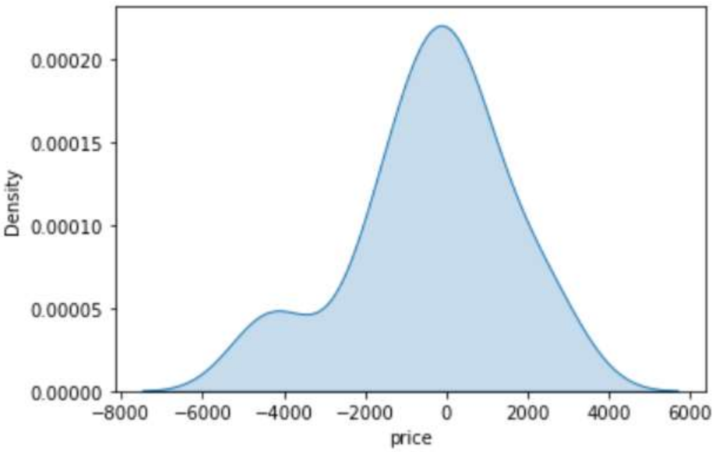

Error terms normally distributed
Error terms not normally distributed

How to check:

density plot

In [49]:
```python
sns.kdeplot(residual,fill=True)
```

Out[49]:
```
<AxesSubplot:xlabel='price', ylabel='Density'>
```



shapiro test

In [50]:
```python
_,p_val = shapiro(residual)
print("p value is :",p_val)

if p_val > 0.05:
    print('Data is normally distributed')

else:
    print('Data is not normally distributed')
```

```
p value is : 0.11151108145713806
Data is normally distributed
```

3 normaltest

In [51]:

```python
_,p_val = normaltest(residual)
print("p value is :",p_val)

if p_val > 0.05:
    print('Data is normally distributed')

else:
    print('Data is not normally distributed')
```

```
p value is : 0.2806403544131823
Data is normally distributed
```

kstest

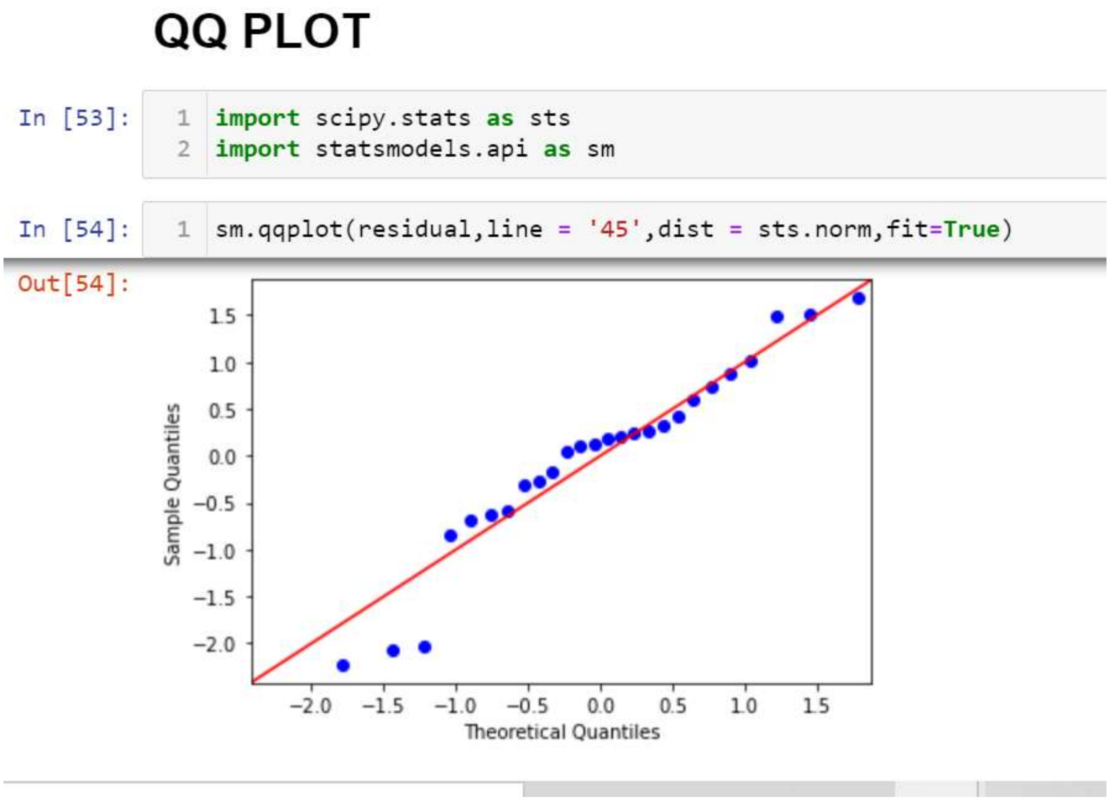
In [52]:

```python
_,p_val = kstest(residual,'norm')
print("p value is :",p_val)

if p_val > 0.05:
    print('Data is normally distributed')

else:
    print('Data is not normally distributed')
```

```
p value is : 9.851228545393985e-09
Data is not normally distributed
```



QQ PLOT

In [53]:

```
import scipy.stats as sts
import statsmodels.api as sm
```

In [54]:

```
sm.qqplot(residual,line = '45',dist = sts.norm,fit=True)
```

Out[54]:

8. How to prevent heteroscedasticity?

There are three common ways to fix heteroscedasticity:

1. Transform the dependent variable One way to fix heteroscedasticity is to transform the dependent variable in some way. One common transformation is to simply take the log of the dependent variable. For example, if we are using population size (independent variable) to predict the number of flower shops in a city (dependent variable), we may instead try to use population size to predict the log of the number of flower shops in a city.

Using the log of the dependent variable, rather than the original dependent variable, often causes heteroskedasticity to go away.

2.2. Redefine the dependent variable Another way to fix heteroscedasticity is to redefine the dependent variable. One common way to do so is to use a rate for the dependent variable, rather than the raw value.

For example, instead of using the population size to predict the number of flower shops in a city, we may instead use population size to predict the number of flower shops per capita.

In most cases, this reduces the variability that naturally occurs among larger populations since we’re measuring the number of flower shops per person, rather than the sheer amount of flower shops.

3. Use weighted regression Another way to fix heteroscedasticity is to use weighted regression. This type of regression assigns a weight to each data point based on the variance of its fitted value.

Essentially, this gives small weights to data points that have higher variances, which shrinks their squared residuals. When the proper weights are used, this can eliminate the problem of heteroscedasticity.

9. What does multicollinearity mean?

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.

This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc.

Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in your model – and that can be a problem when it comes to interpretability.

10. What are feature selection and feature scaling?

1. Feature selection:

Feature Selection is a way of selecting the subset of most relevant features from original features set by removing the redundant, irrelevant or noisy features

Need of Feature selection:

1. irrelevant
2. noisy features
3. Less important features

Benifits of Feature selection:

1. It reduced training time as well as testing time(Reducing time complexity)
2. Improve accuracy
3. It helps to Reduce overfitting
4. It helps in avoiding curse of dimensionality

Feature selection Techniques:

A. Filter Method:

1. Correlation :
2. Pearson (Cont vs Cont)
3. Spearman (Cont Vs Cont)

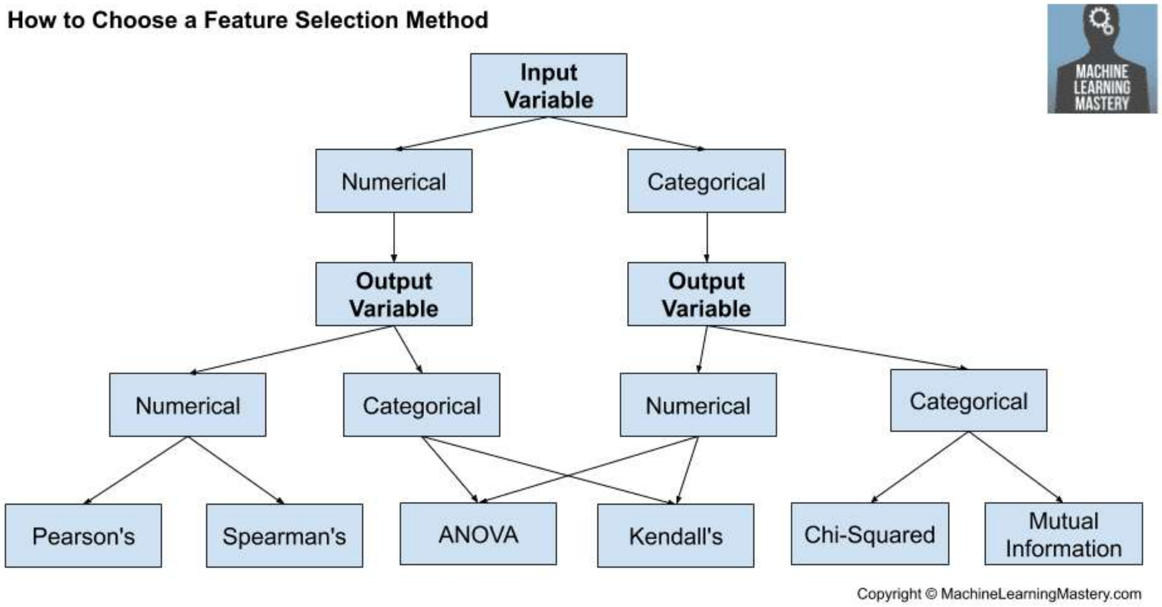
4. Kendall (Cont Vs Cat)
5. Chi-Square Test (Cat vs Cat) # hypothesis testing
6. Information Gain(Mutual Information)
7. Fishers Score
8. Mean Absolute Difference(MAD)
9. Missing Values Ratio
10. Variance Threshold
11. ANOVA Test (Numerical Vs Categorical)

B. Wrapper Method:

1. Forward Selection Method
2. Backward Elimination method
3. Exhastive Feature selection
4. Recursive Feature selection

C. Embedded Method:

1. Random Forest Imporatance (tree -based model)
2. Regularization(L1 and L2)



2. Feature scaling:

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step Just to give you an example — if you have multiple independent variables like age, salary, and height; With their range as (18–100 Years), (25,000–75,000 Euros), and (1–2 Meters) respectively, feature

scaling would help them all to be in the same range, for example- centered around 0 or in the range (0,1) depending on the scaling

technique.

a. Normalization:

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as,

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

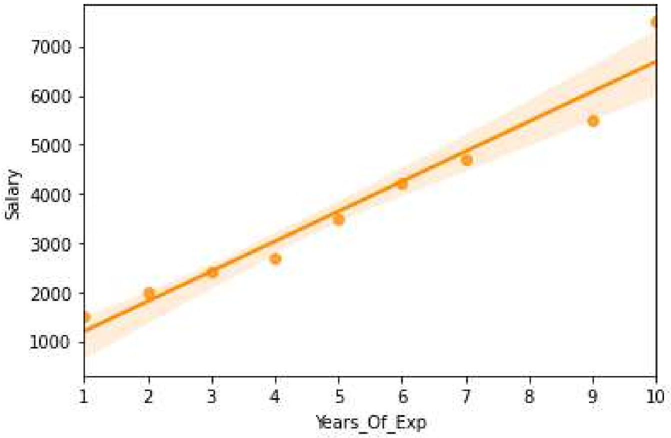
b. Standardization:

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

11. How to find the best fit line in a linear regression model?

After finding the correlation between the variables[independent variable and target variable], and if the variables are linearly correlated, we can proceed with the Linear Regression model. The Linear Regression model will find out the best fit line for the data points in the scatter cloud.



we know that equation of straight line is y = mx + c

m →slope c →intercept

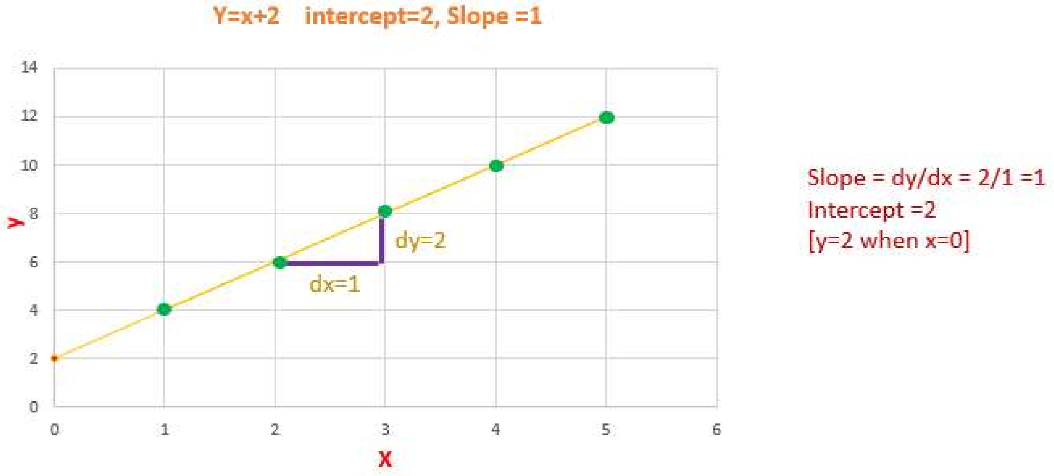
Slope m and Intercept c are model coefficient/model parameters/regression coefficients. Slope →m Slope basically says how steep the line is. The slope(m) is calculated by a change in y divided by a change in x

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Calcworkshop.com

The slope will be negative if one increases and the other one decreases. The slope will be positive if x increases and y increases.

Intercept → c The value of y when x is 0. When the straight line passes through the origin intercept is 0.

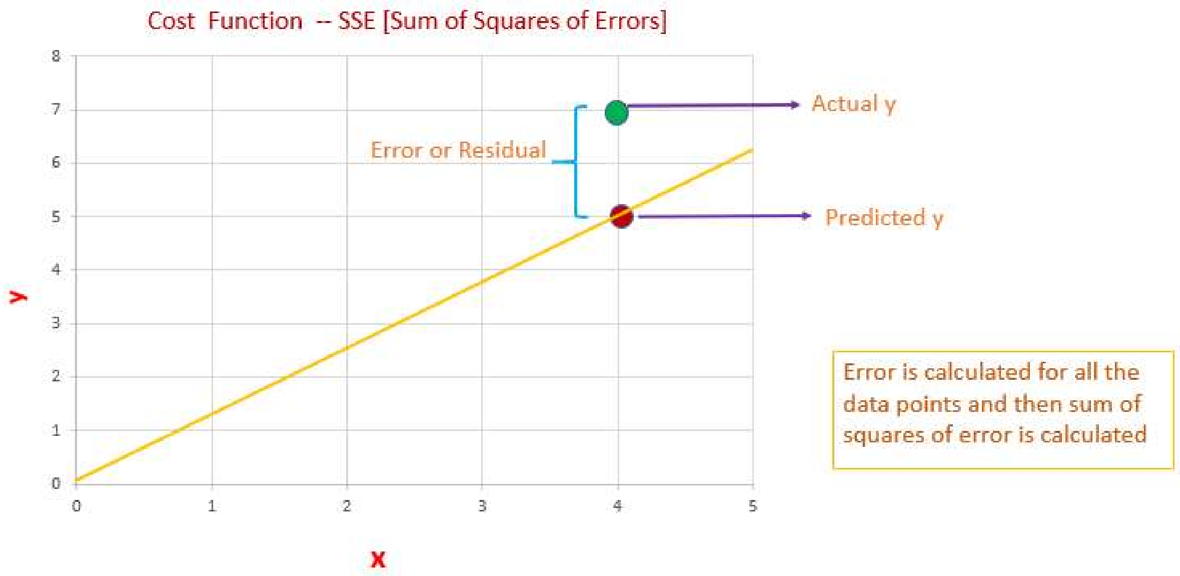


Line of Best Fit:-

We know the equation of a line is y=mx+c. There are infinite m and c possibilities, which one to chose?

The line of best fit is calculated by using the cost function — Least Sum of Squares of Errors. The line of best fit will have the least sum of squares error.

We have to calculate error/residual for all data points square the error/residuals. Then we have to calculate the sum of squares of all the errors. Out of all possible lines, the line which has the least sum of squares of errors is the line of best fit. Cost Function - The least Sum of Squares of Errors is used as the cost function for Linear Regression. For all possible lines, calculate the sum of squares of errors. The line which has the least sum of squares of errors is the best fit line.


Cost Function -- SSE [Sum of Squares of Errors]
Actual y
Error or Residual
Predicted y
Error is calculated for all the data points and then sum of squares of error is calculated
y
x
0
1
2
3
4
5
6
7
8

Error/Residuals Error is the difference between the actual value of y and the predicted value of y.

1.We have to calculate error/residual for all data points.

2.square the error/residuals.

3.Then we have to calculate the sum of squares of all the errors.

4.Out of all possible lines, the line which has the least sum of squares of errors is the line of best fit.

12. Why do we square the error instead of using modulus

1.If we are not squaring the error, the negative and positive signs will cancel. We will end up with error=0

2.So we are interested only in the magnitude of the error. How much the actual value deviates from the predicted value.

3.So, why we didn’t consider the absolute value of error. Our motive is to find the least error. If the errors are squared, it will be easy to differentiate between the errors comparing to taking the absolute value of the error.

4.Easier to differentiate the errors, it will be easier to identify the least sum of squares of error. Out of all possible lines, the linear regression model comes up with the best fit line with the least sum of squares of error. Slope and Intercept of the best fit line are the model coefficient.

13. What are techniques adopted to find the slope and the intercept of the linear regression

line which best fits the model?

1. Gradient Descent
2. Least Square Method / Normal Equation Method
3. Adams Method
4. Singular Value Decomposition (SVD)

14. What is cost Function in Linear Regression?

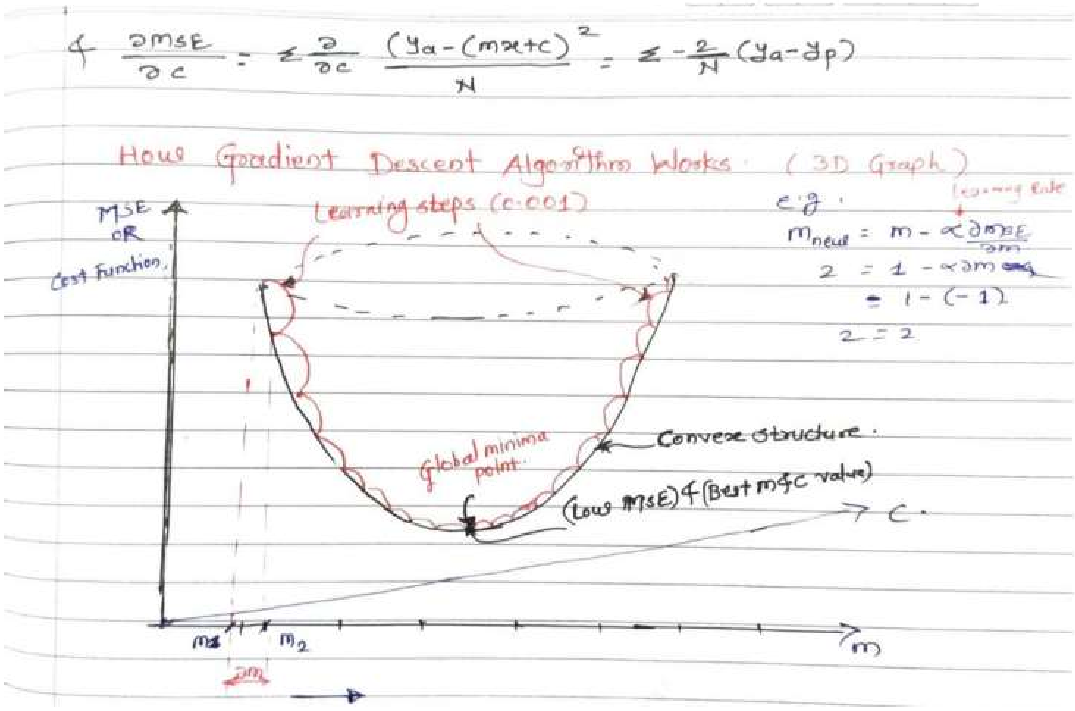
1.cost function in Linear Regression is MSE.

2.MSE is mean squared error.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y-\widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

15. briefly explain gradient descent algorithm

1. It is a iterative process that finds the global minima of a function
2. It is used to reduce the cost of function (MSE).
3. This is used to find out the best M and C values.
4. This will try the infinite m and c values till we get the best M and C values.
5. It uses partial derivative in this we have to check new ‘m’ value if our first m value is 1 and slop is 3 dimensional in that at m=1 the mse is 100 that is learning step then this algorithm will try new m values Like this will try for the multiple m and c values till the we get best M and C values or Low mse once we get the the Low MSE at some point this is called as Global Minima but when we changing the m and c values after global minima then mse will increase.



& $\frac{\partial mse}{\partial c} = \sum \frac{\partial}{\partial c} \frac{(y_a-(mx+c))^2}{N} = \sum -\frac{2}{N}(y_a-y_p)$
How Gradient Descent Algorithm Works. (3D Graph)
MSE OR Cost Function.
Learning steps (0.001)
e.g.
learning Rate
$m_{new} = m - \alpha \frac{\partial mse}{\partial m}$
$2 = 1 - \alpha \partial m$
$= 1 - (-1)$
$2 = 2$
Convex Structure.
Global minima point.
(Low MSE) & (Best m&c value)
c.
m1
m2
∂m
m

Learning Rate or Learning Rate:-

1. In first steps (i.e first ‘m’ value) we are not getting the good values of m and c, model is learning from previous values that’s why we called as learning steps.
2. If learning rate is high then there will be Overshooting problem occurs.
3. The standard value of learning rate is 0.001

16. How to evaluate regression models?

1. MSE mean squared value.
2. SSE (Sum of Squared Error) - It is squared difference between y actual and y predicted or we can say

   sum of squared difference of residuals

   SSE = sum (Ya- Yp)^2
3. SSR (Sum of Squares due to Regression) – It is difference between predicted and mean of dependent variable

   SSR = sum (Yp - Ymean)^2
4. SST (Sum of Squares of Total error) - It is squared difference between y actual and mean of dependent variable

   SST = sum (Ya - Ymean)^2

17. Which evaluation technique should you prefer to use for data having a lot of outliers in

it?

Mean Absolute Error(MAE) is preferred when we have too many outliers present in

the dataset because MAE is robust to outliers whereas MSE and RMSE are very

susceptible to outliers and these start penalizing the outliers by squaring the error

terms, commonly known as residuals.

18. What is residual? How is it computed?

Residual is also called Error. It is the difference between the predicted y value and the actual y value.

Residual = Actual y value – Predicted y value.

It can be positive or negative.

If residuals are always 0, then your model has a Perfect R square i.e. 1.

19. What are SSE, SSR, and SST? and What is the relationship between them?

1.SSE (Sum of Squared Error) - It is squared difference between y actual and y predicted or we can say sum of squared difference of residuals

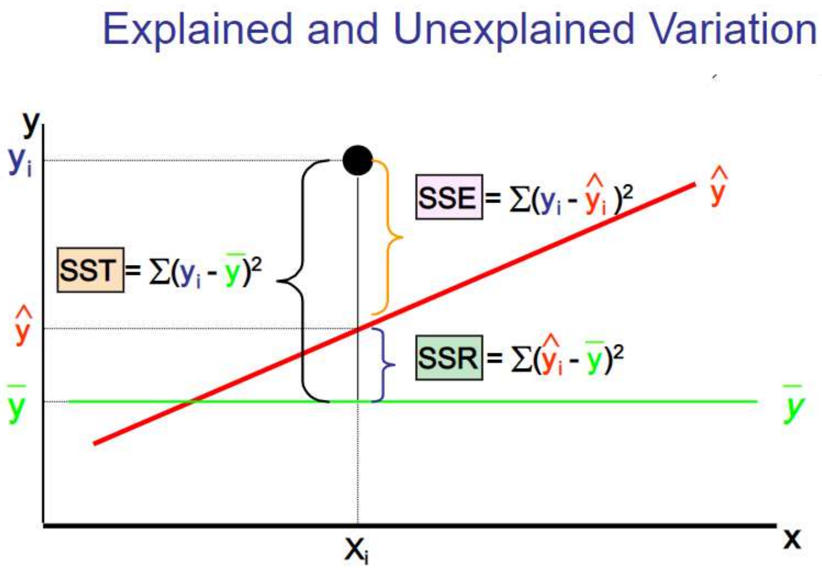
SSE = sum (Ya- Yp)^2

2. SSR (Sum of Squares due to Regression) – It is difference between predicted and mean of dependent variable

SSR = sum (Yp - Ymean)^2

3. SST (Sum of Squares of Total error) - It is squared difference between y actual and mean of dependent variable

SST = sum (Ya - Ymean)^2


Explained and Unexplained Variation
y
y_i
$\hat{y}$
$\bar{y}$
SST = $\Sigma(y_i - \bar{y})^2$
SSE = $\Sigma(y_i - \hat{y}_i)^2$
SSR = $\Sigma(\hat{y}_i - \bar{y})^2$
$\hat{y}$
$\bar{y}$
X_i
x

Relationship between SST, SSR and SSE

The total sum of squares is equal to the sum of squares due to error plus thesum of squares due to regression.

Therefore:SST = SSR + SSE

We can think of the SST as representing the total variationin the data.

Similarly SSR is the explained variation and represents the portion of the totalvariation that is explained by the regression line.

SSE is the unexplained variationand represents the portion of the total variation that is not explained by the regression line

Therefore:Total Variation = Explained Variation + Unexplained Variation When SSR is large, it explains a large portion of the variation in SST; therefore SSE is small.

This will result in the residual terms (yi- ŷi) being small.

This means that thesample observations cluster fairly close around the regression line.

In otherwords there is a strong fit between the sample observations and the regressionline. In fact, if SSE = 0, then SST = SSR,

which means all the observations lie on theregression line - i.e., a perfect fit.

As SSE gets larger, SSR gets smaller indicating a poorer fit between theobservations and the regression line.

20. What's the intuition behind R-Squared?

R2 Score (Coefficient of Determination)

R2 Score is used to evaluate the performance of linear regression model.

It is used to Check how well observed results are reproduced by the model, depending on the ratio of Explained variance to Total Variance

1. R2 score or R2 value basically it is a coefficient of Determination
2. It is used for find the goodness of best fit line R2 score = 1 - SSE/SST or (SST-SSE)/SST

R2 Score = 1 >> Best Score >> SSE = 0 >> All Data Points on Regression Line

R2 Score = 0 >> SSE = SST

R2 Score = -1(-Ve) >> Worst Score >> SSE is Greater than SST

21. What does the coefficient of determination explain?

The coefficient of determination is a measurement used to explain how much

variability of one factor can be caused by its relationship to another related

factor. This correlation, known as the "goodness of fit," is represented as a value

between 0.0 and 1.0.

how well the regression model fits the observed data. For example, a

coefficient of determination of 60% shows that 60% of the data fit the regression

model.(60% less variation around BFL than meal value)

22. Can R² be negative?

A higher coefficient is an indicator of a better goodness of fit for the

observations. The CoD can be negative, although this usually means that your

model is a poor fit for your data. It can also become negative if you didn't set an intercept.

It means regression line is making more mistakes than mean line.

23. What are the flaws in R-squared?

It will increase for good predictors as well as it will also increase for bad predictors. So we need to use Adjusted R-Squared method.

24. What is adjusted R²?

Adjusted R-Squared :-

- It will increase only for good predictors.
- Adjusted R-Squared value will always be less than or equal to R-Squared. Adjusted R-Squared <= R2 score

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent variable

25. What is the Coefficient of Correlation: Definition, Formula

Correlation coefficient formulas are used to find how strong a relationship is between data.

Theformulas return a value between -1 and 1, where: 1 indicates a strong positive relationship.

-1 indicates a strong negative relationship.

A result of zero indicates no relationship at all.

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

26. What is the relationship between R-Squared and Adjusted R-Squared?

1. R Squared is an econometric measure uses to explain the dependent and unconstrained variables where Adjusted R Squared is a value measuring that predicts the regression variables.
2. R Squared had Symbolized as R ^2 where Adjusted R Squared had written as Adjusted R ^2.

3. R squared is higher in getting the desired products, where Adjusted R Squared values are lower in measuring.
4. R Squared method had used to take the values originally where Adjusted R Squared values had been calculated mathematically.
5. Adjusted R Squared measurement requires the R Squared points for calculations.

27. What is the difference between overfitting and underfitting?

Overfitting:-

- Sometimes machine learning performs well on training data but does not perform well with the test data.
- It means the model is not able to predict the output when seals with unseen data by introducing noise in the output and hence the model is called overfitted.
- It has high training accuarcy and low testing accuarcy.
- It has low bias and high varience.

How to handle avoid Overfitting:-

1. Remove Features (15 data to 10 data)
2. Use parameter tunning. Use prunung for DT.
3. Remove Outliers
4. Increase dataset
5. Regularisation.

Underfitting :-

- Sometimes machine learning not performs well on training data and testing data.
- So this model is called Underfitted.
- It has low training accuarcy and low testing accuarcy.
- It has high bias and low varience.

How to handle avoid Underfiiting:-

1. Add Features (10 data to 13 data)
2. Handling Missing values(maen,median,etc)
3. Remove Outliers
4. Increase dataset
5. Use correlated features.

28. How to identify if the model is overfitting or

underfitting?

Overfitting is when the model's error on the training set (i.e. during training) is very low

but then, the model's error on the test set (i.e. unseen samples) is large!

that is training accuracy is high and testing accuracy is low then we can say that model is overfitting.

Underfitting is when the model's error on both the training and test sets (i.e. during training and testing) is very high.

that is both the traning and testing accuracy is low.

29. How to interpret a Q-Q plot in a Linear regression model?

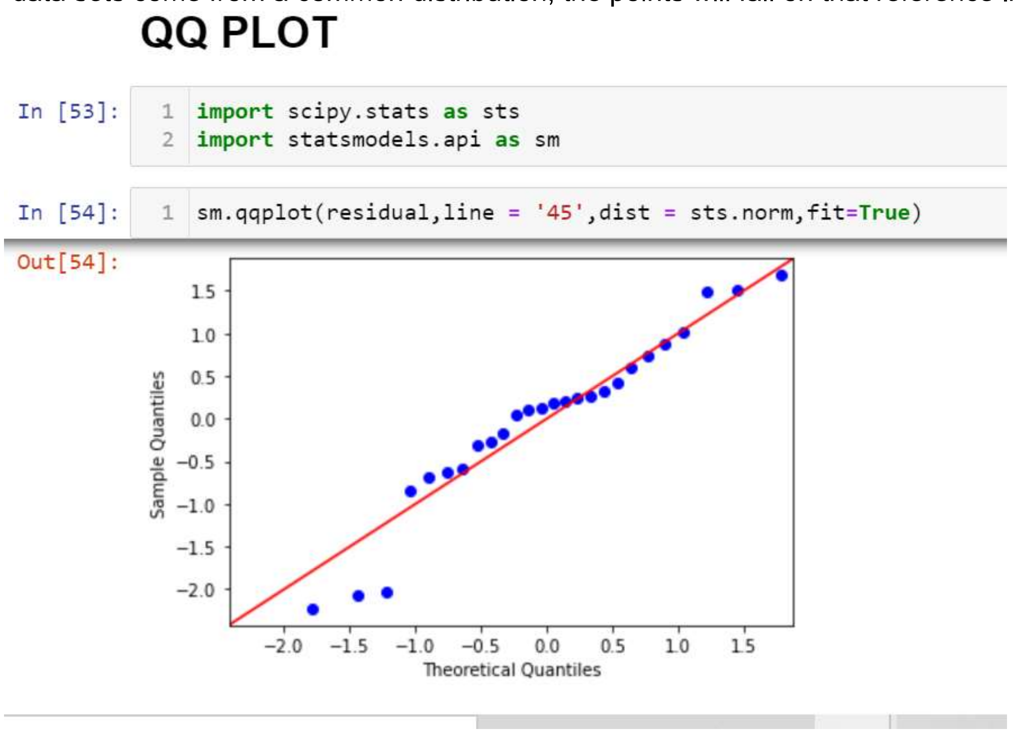
Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other.

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.

he 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.



QQ PLOT

```
In [53]: import scipy.stats as sts
         import statsmodels.api as sm
```

```
In [54]: sm.qqplot(residual,line = '45',dist = sts.norm,fit=True)
```

Out[54]:

30. What are the advantages and disadvantages

# of Linear Regression?

Advantages of Linear Regression :-

1. Linear Regression is performs exceptionally well on linearly seaprable data.
2. It is easy to implement.
3. Overfitting can be reduced by regularization L1 and L2.

Disadvantages of Linear Regression :-

1. Linearity - There should be linear realtionship between dependent variable(y) and independent variable(x).
2. Independence - All the independent variables are independent to each other.

There should not be linear relationship between independent variables(x1 and x2).

3. It is very sensitive to outliers.
4. It is sensitive to missing values.

31. What is the use of regularisation? Explain L1 and L2 regularisations

Regularisation :-

- Regularisation used for prevent model from overfitting by adding extra information to it.
- Sometimes machine learning performs well on training data but does not perform well with the test data.
- It means the model is not able to predict the output when seals with unseen data by introducing noise in the output and hence the model is called overfitted.
- So Regularisation used to reduce overfitting.
- This Technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence it maintains accuracy as well as generalization of the model.

  ## Regularisation is divided into two types:-

1. Ridge Regularisation (L2)
2. Lasso Regularisation (L1)

1. Ridge Regularisation (L1) :-

It is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It is calculated by

c.f. = (Ycatual - Ypredicted)^2 + lambda * (slope)^2 where, lambda = Hyperparameter tunning (0 to infinity range)

if we take lambda=0 this model will act as linear regression. hence try to take lambda = 0.01

2. Lasso Regularisation (L1) :-

It is used to reduce the overfitting in the model, by shrinking as well as feature selection.

It is calculated by c.f. = (Ycatual - Ypredicted)^2 + lambda * |slope|

In this method slope is "+Ve" where,

lambda = Hyperparameter tunning (0 to infinity range) if we take lambda=0 this model will act as linear regression. hence try to take lambda = 0.01

In [ ]: