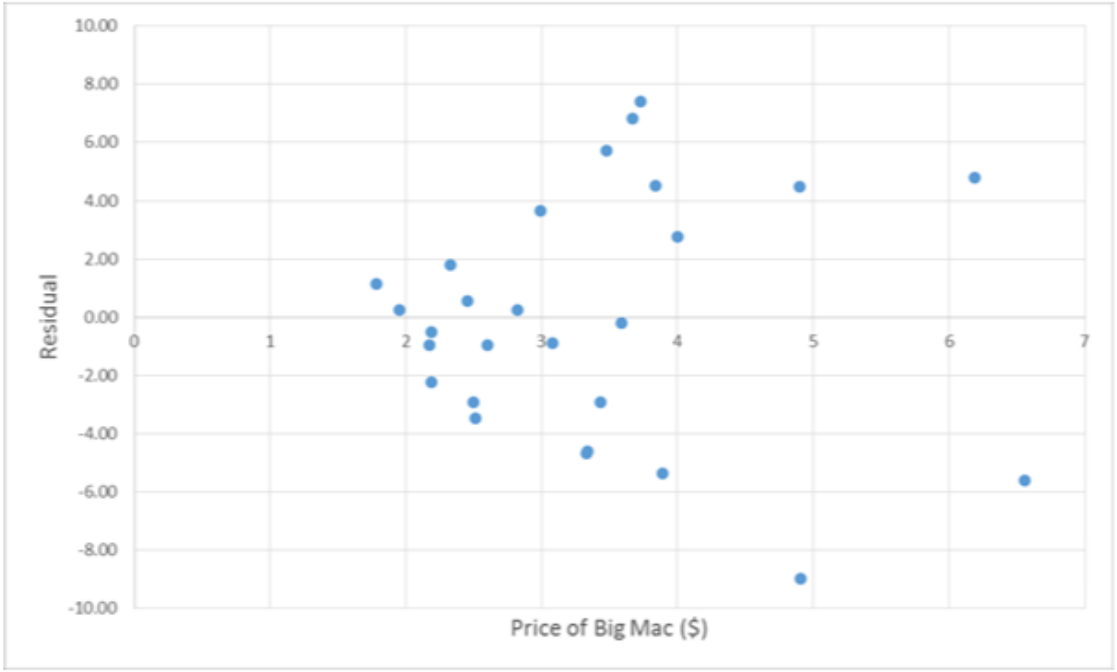
Residual Analysis

Error in the prediction

Residual Analysis



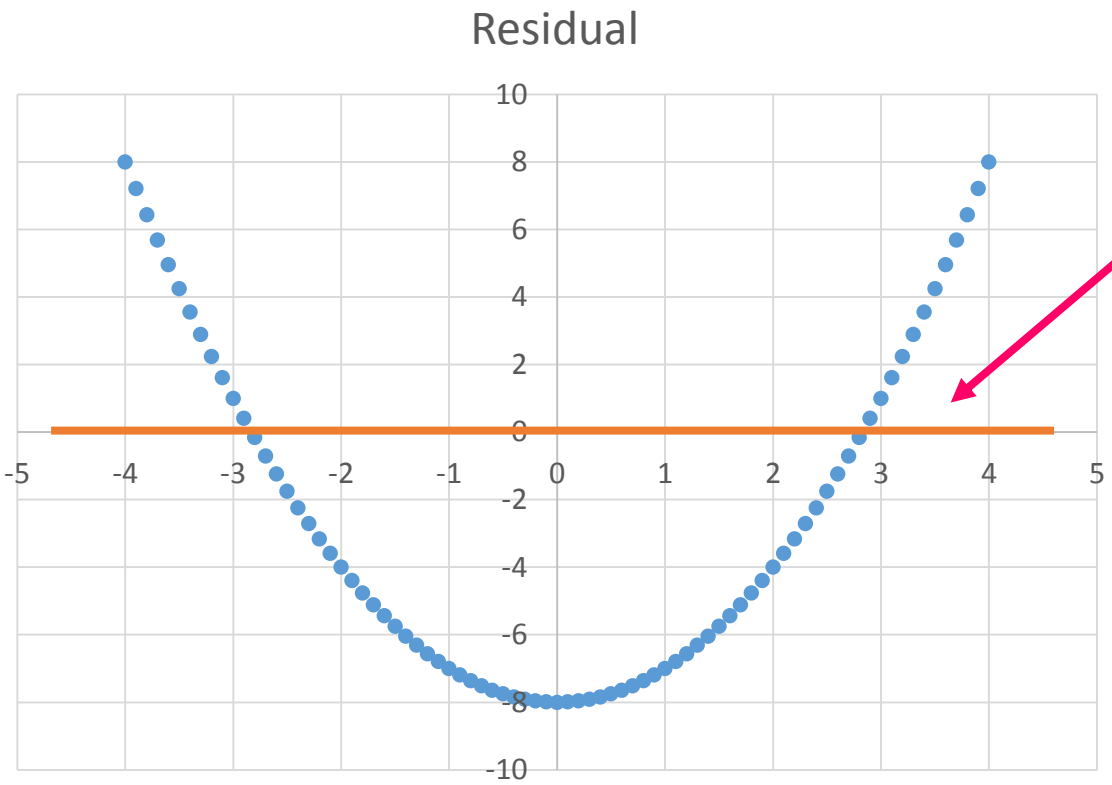
Can be used to locate Outliers.

Residual = Forecast errors

$= (y_i - \hat{y}_i)$

Assumptions of the Regression Model

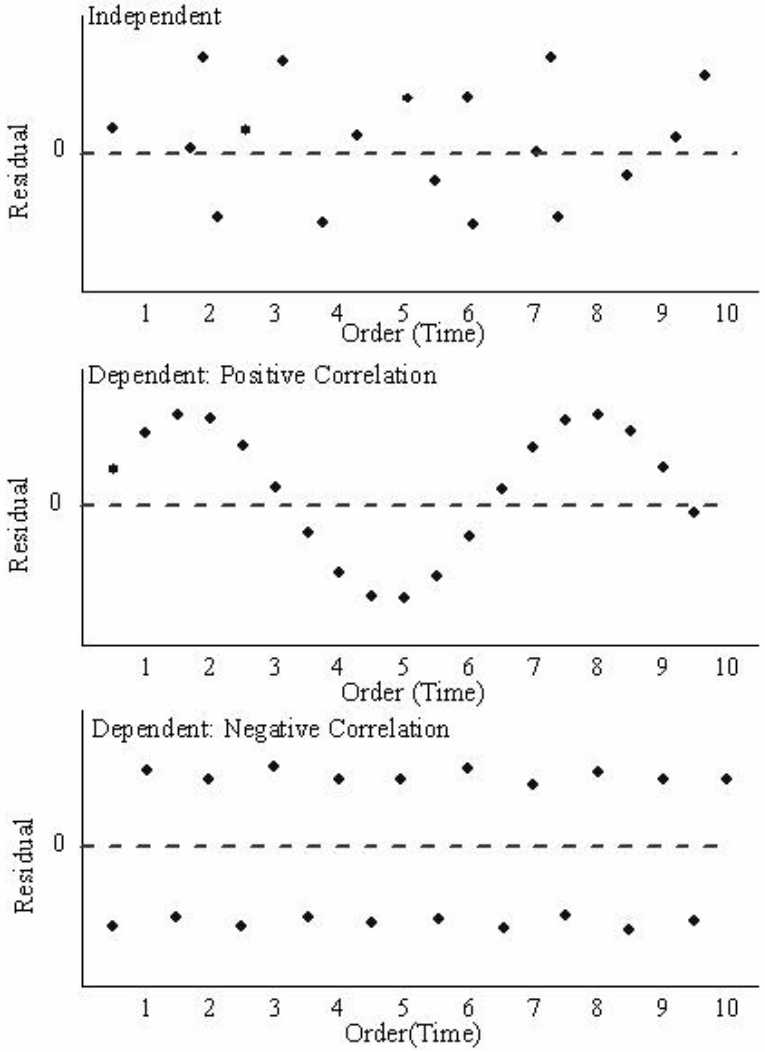
- The model is linear



Zero Residual line

Assumptions of the Regression Model

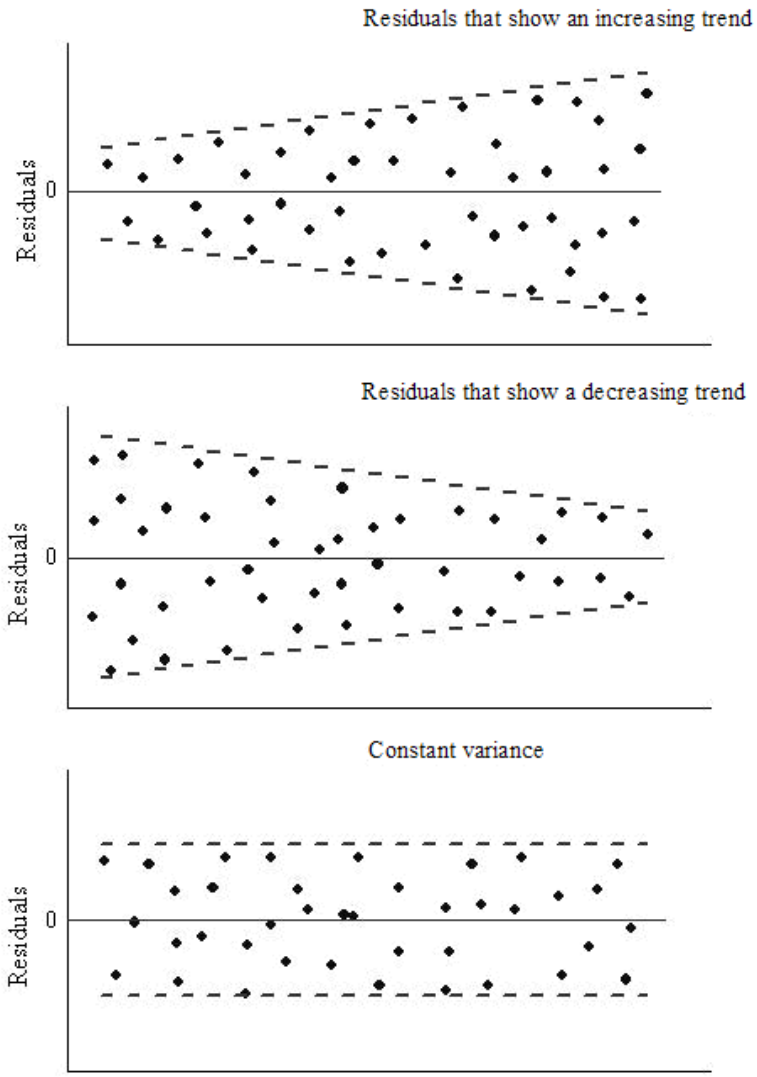
- The error term are independent
  - Plot against any time (order of observation) of spatial variables preferably. Plots against independent variables may also detect independence.



Assumption of Regression Model

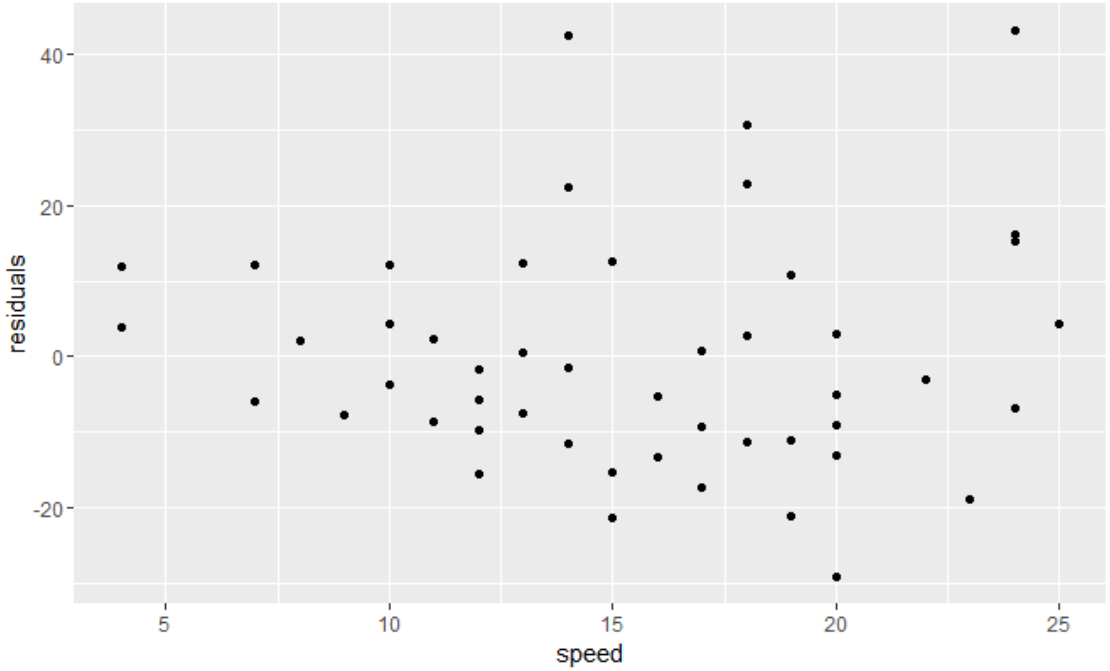
- The error terms have constant variance (homoscedasticity as opposed to hetroscedasticity)

- RMSE (Root Mean Square Error) of Regression or Standard Error of the estimate will be misleading as it will underestimate the spread for some x_i and overestimate for others.



Assumption of Regression Model

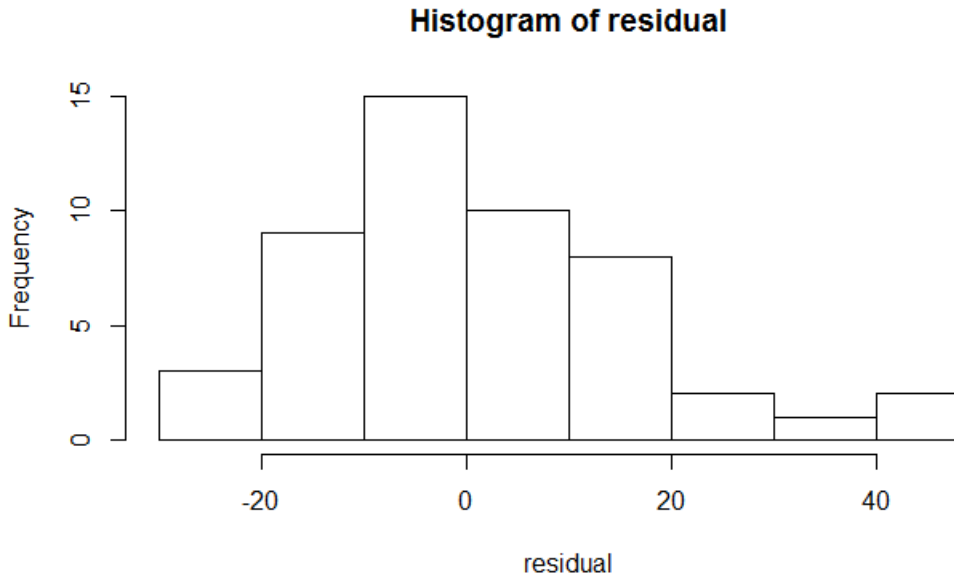
- The residual errors are normally distributed

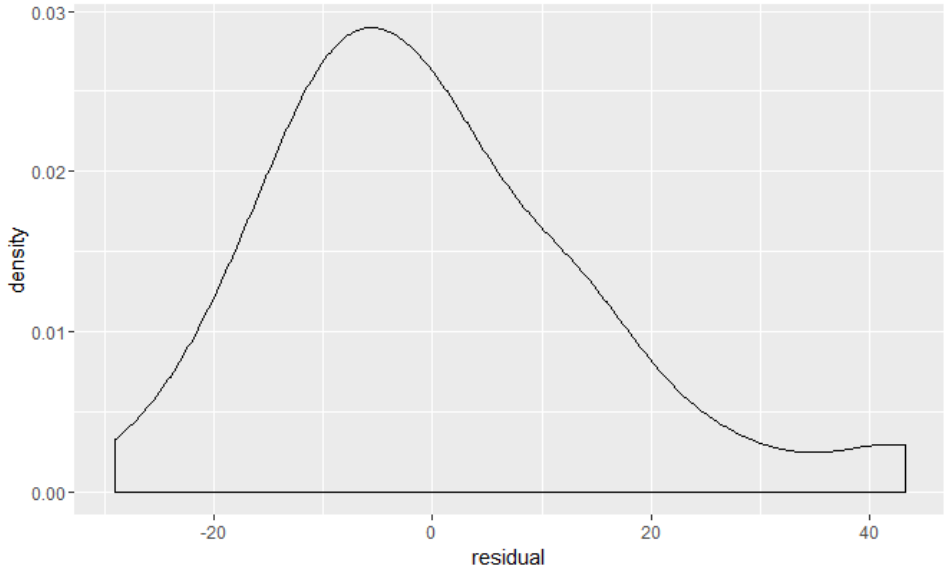


But, how do we know if something is normally distribution ?

Checking for Normality

- Start by plotting the data





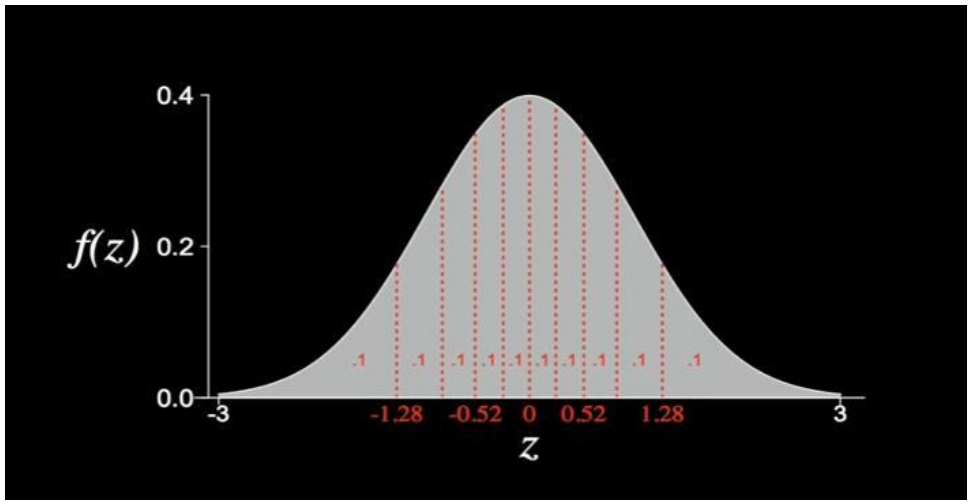
Quantile Quantile-Plot

- Its used to assess if the given data-set follows a particular distribution
- For example is the 9-point (sorted) data-set below normal?

-1.2,-1.11, -1.08, -0.28, -0.25, 0.33, 0.41, 1.37, 1.41

- Lets start with assumption that the data is from normal distribution.
- Lets divide the normal distribution into 9+1 equal areas.
- The boundary point would represent a 0.1 quantile

Quantile Quantile-Plot



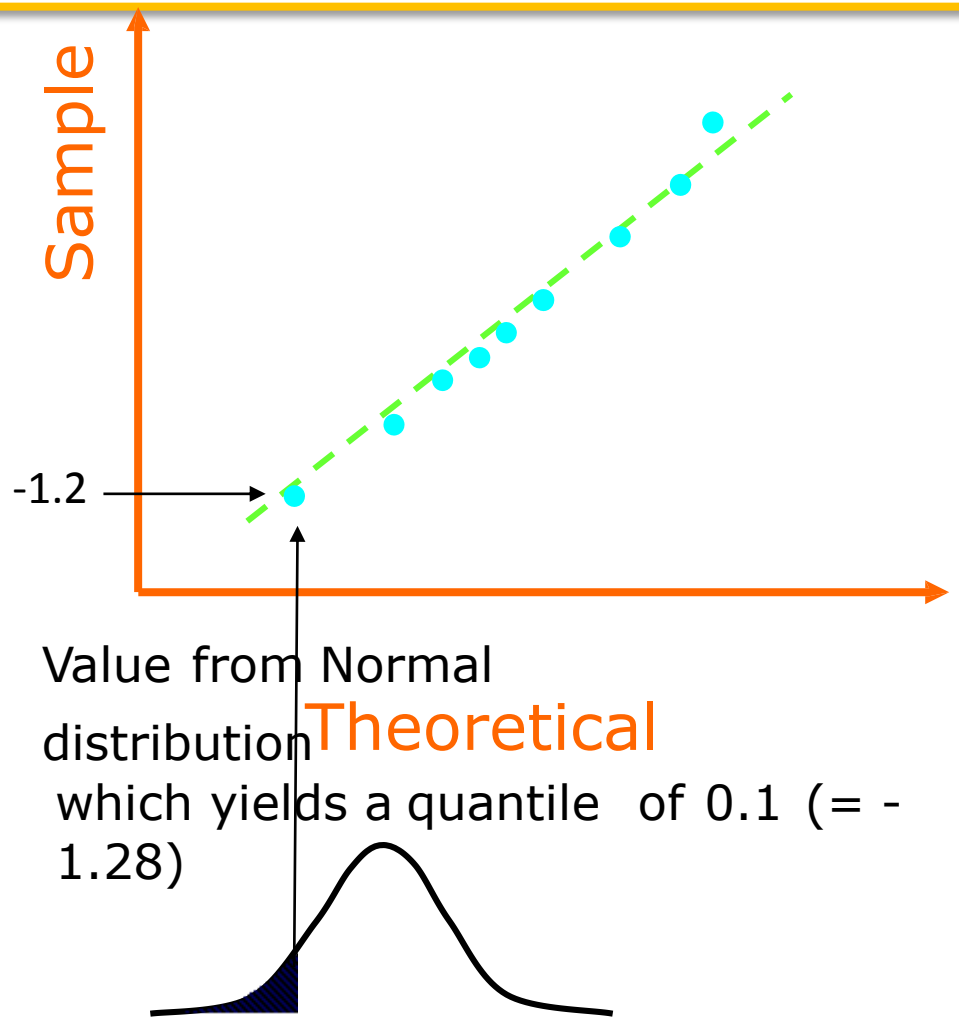
- Then one might expect the smallest of the 9 data points to be from the lowest quantile (0.1)
- Similarly, the largest value would be from the largest quantile (0.9) of the normal distribution

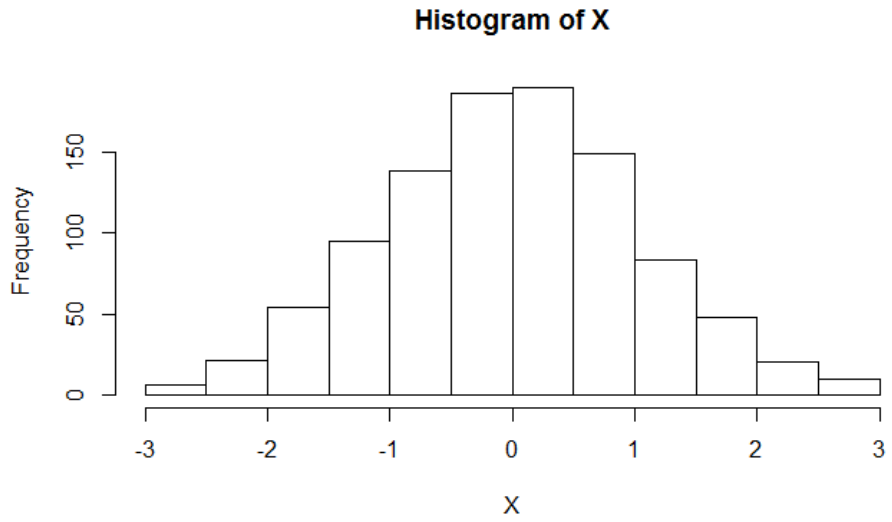
Quantile Quantile-Plot

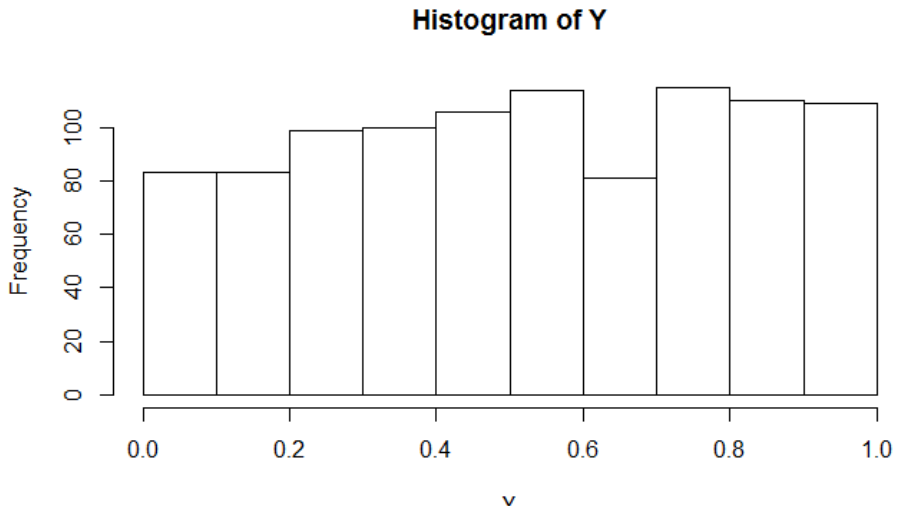
-1.2, -1.11, -1.08, -0.28, -0.25, 0.33, 0.41, 1.37, 1.41

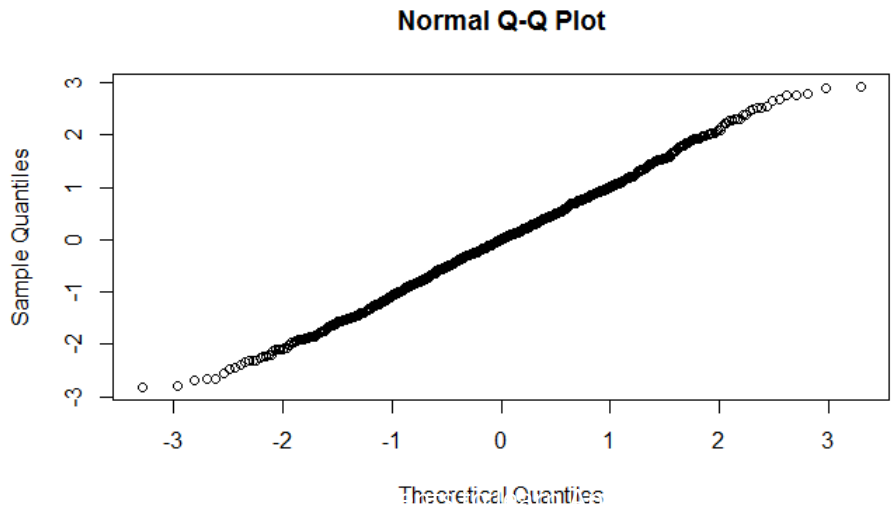
We plot the quantile values for the distribution on the x-axis and the values of the sample on the y-axis

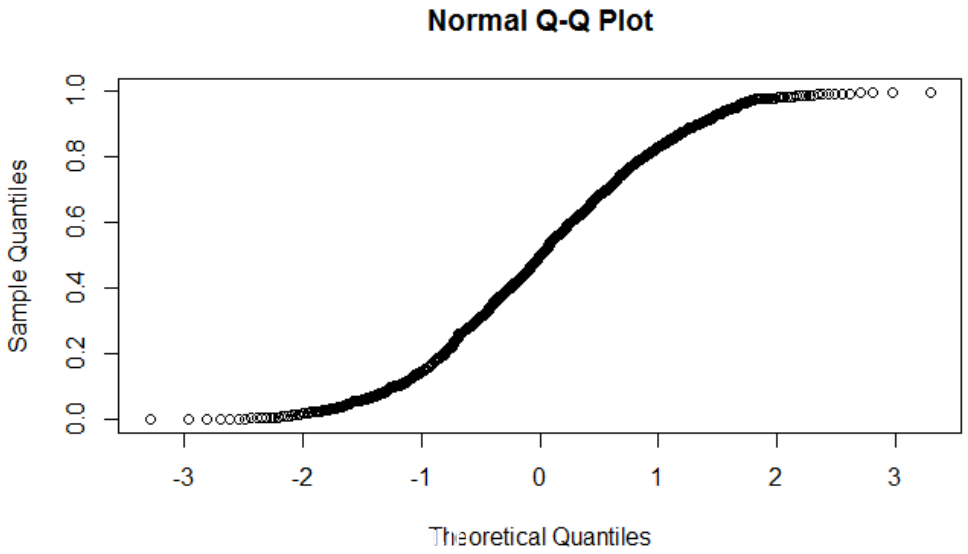
If the points lie on close to a straight line, then the sample is normal







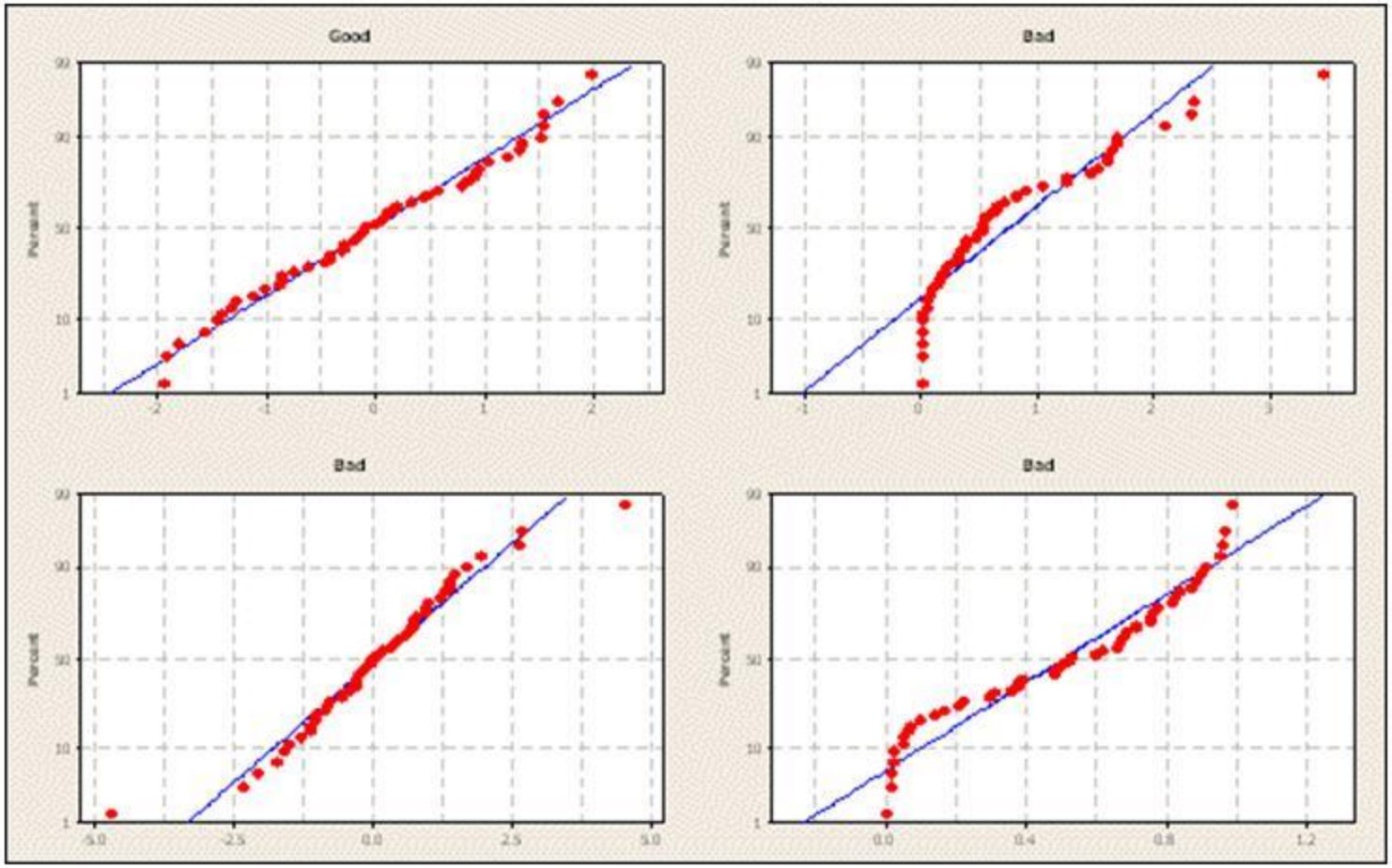




QQ Plot for Normal vs Uniform Distribution

Assumptions of the Regression Model

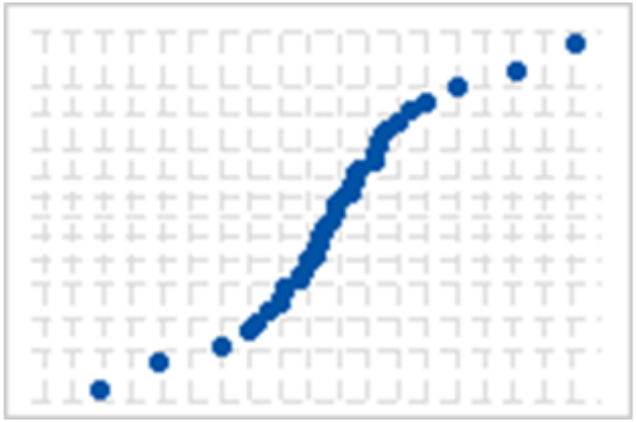
- The error terms are normally distributed



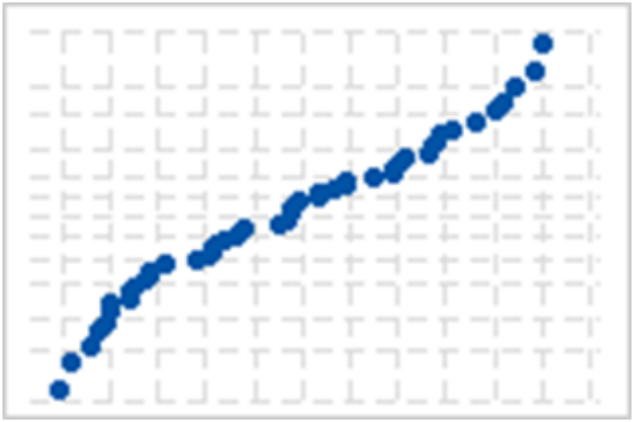
Checking for Normal Distribution

- Other objective methods of checking for normality also exist
- Shapiro-Wilk Test gives a probability value (p-value) that the given data sample is actually from a Normal distribution
- If p-value is less than 0.05, then its unlikely to be from Normal distribution

Interpreting Residuals – Non-normality



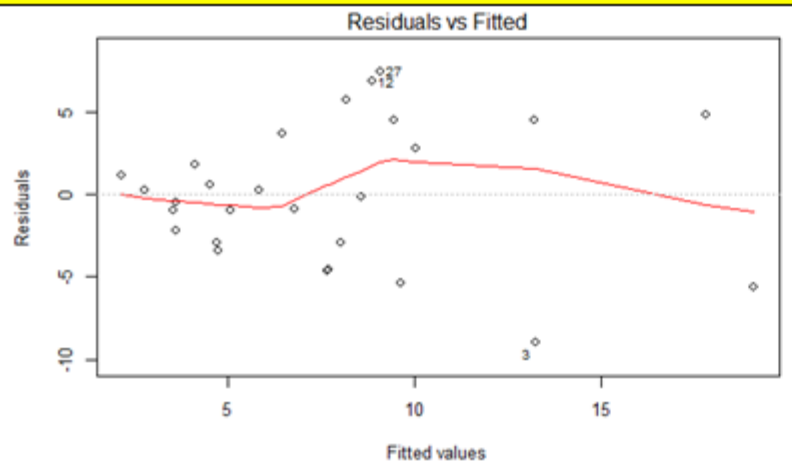
S-curve implies a distribution with long tails



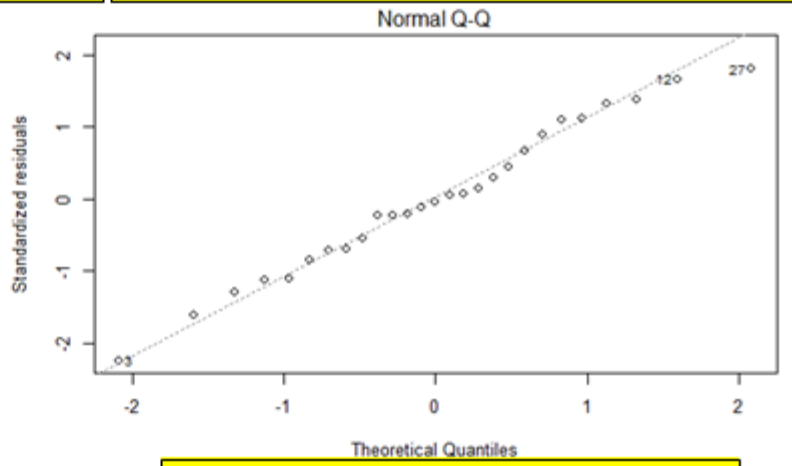
Inverted S-curve implies a distribution with short tails

Residuals – Big Mac

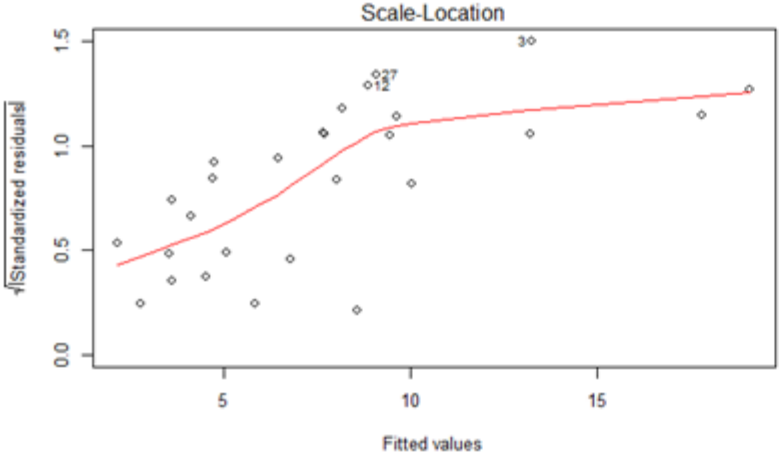
Is a wrong model fitted (linear or quadratic, etc.)?



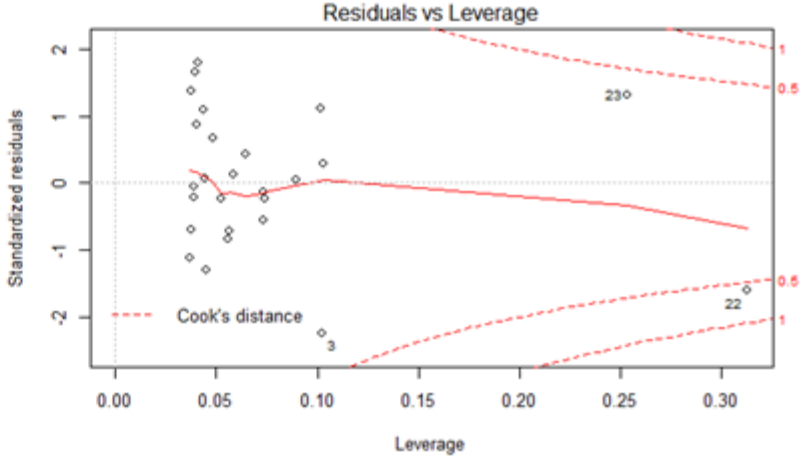
Are the residuals normally distributed?



Is the data homoscedastic?



Are there influential outliers?



INNOMATICS TECHNOLOGY HUB

Fixing Non-normality and Heteroscedasticity

Transformation of data can help correct normality and unequal variances problems.

Hypothesis test for the slope of the Regression model and testing the overall model

Reference

Head First Statistics

Business Statistics