Multiple linear regression, a statistical method used to predict the value of a variable based on the values of two or more other variables, rests on several key assumptions:

Assumption 1: Linear Relationship:

The relationship between the dependent variable and each independent variable is assumed to be linear.

This means that changes in the independent variables should result in proportional changes in the dependent variable. This assumption can be checked by using scatter plot of residual vs predicted Y values or observed vs predicted Y values.

Assumption 2: Normality of Errors:

The errors (residuals) are assumed to be normally distributed. This is crucial for valid hypothesis testing. This assumption can be checked using histograms, Q-Q plots, or statistical tests.

Assumption 3: Independence of Errors:

The errors (residuals) are assumed to be independent of each other. This means that the error in one observation should not be related to the error in another observation. This is particularly important for time-series data.

This assumption can be checked by using graph of residual vs observation order or time. We can use confirmatory test called the Durbin-Watson test which tells us whether there is a violation of this assumption or not.

Assumption 4: Homoscedasticity (Equal variance assumption)

This assumption states that the variance of the errors (residuals) is constant across all levels of the independent variables. Heteroscedasticity (non-constant variance) can distort estimates. This assumption can be checked by using graph of residual vs predicted Y values

Assumption 5: No Multicollinearity

Multicollinearity occurs when independent variables are highly correlated with each other. This makes it difficult to determine the unique contribution of each variable to the model. Presence of multicollinearity is confirmed by using statistics call Variance Inflationary Factor (VIF).

Assumption 6: No Measurement Errors in Predictors

The independent variables should be measured accurately. Measurement errors can bias estimates.

Residual Analysis

Residual analysis is a crucial step in evaluating the quality of a linear regression model. It involves examining the residuals, which are the differences between the observed (actual) values of the dependent variable and the predicted values generated by the regression model. Residuals provide insight into how well the model fits the data and whether the assumptions of linear regression are met.

Purpose of Residual Analysis:

Residual analysis helps to:

- Assess the goodness-of-fit of the regression model.

- Check whether the assumptions of linear regression are satisfied.
- Identify potential issues such as non-linearity, heteroscedasticity, or outliers.

Checking the assumption of linearity

A Multiple linear regression can only accurately estimate the relationship between dependent and independent variables if the relationships are linear in nature.

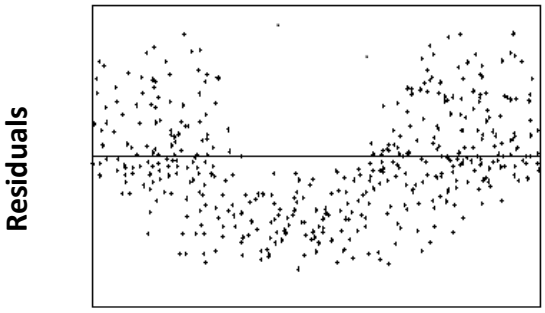
To detect linearity, we can use one of the following graphs:

- plot of the observed values of Y (taking along X-axis) versus predicted values of Y (taken along Y-axis)
- plot of residuals (taken along Y-axis) versus predicted values (taken along X-axis)
  (or standardized residuals versus standardized predicted values)

If the residuals show a systematic pattern (e.g., curvature), it suggests that the relationship between the independent and dependent variables is not linear.

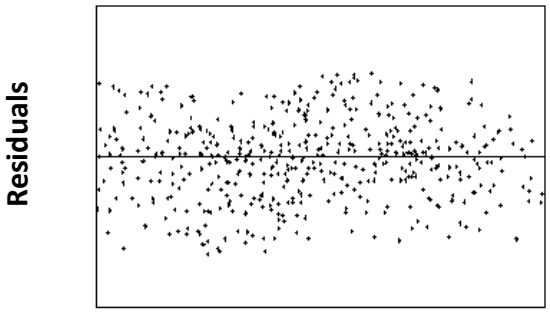
A plot showing Curvilinear Relationship

A curved pattern suggests non-linearity; consider adding polynomial terms or transforming variables.



A plot showing Linear Relationship

Random scatter around e = 0 line (reference line) indicates a good fit with no major violations of assumptions.



To fix this problem we may apply a *nonlinear transformation* to the dependent and/or independent variables. Some of the transformation rules are: log transformation, square transformation etc.

Checking the assumption of Normality of Errors

It's important that the residuals are normally distributed. If residuals are not normally distributed, confidence intervals and hypothesis tests may not be valid.

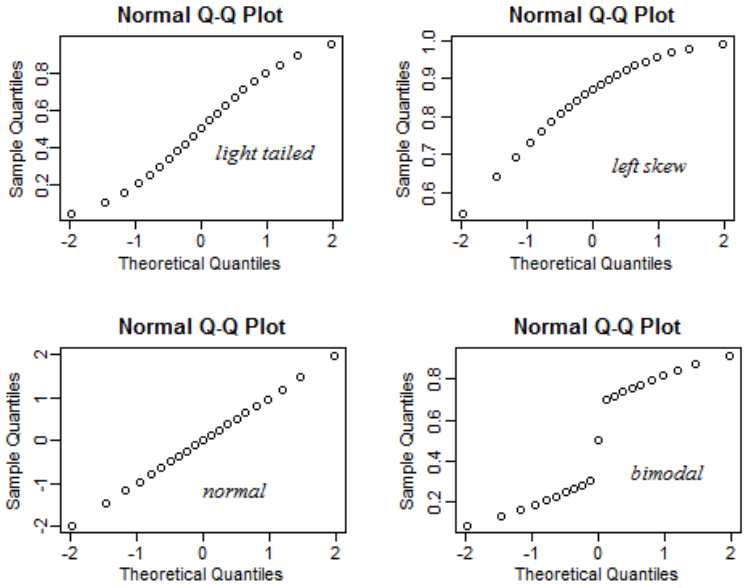
We can use following graphical tools to check the normality of error distribution:

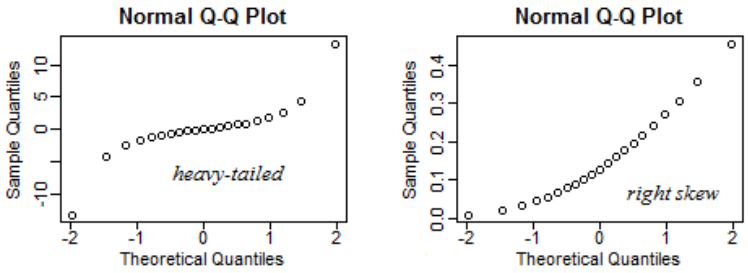
1. Histogram of residuals
2. QQ plot of residuals
3. PP plot of residuals

Commonly used tool is QQ plot. A **Q-Q plot** helps visually assess whether a dataset conforms to a specific distribution, particularly normality, by comparing observed and theoretical quantiles. Theoretical quantiles are taken along X-axis and empirical quantiles are taken along y-axis.If the distribution is normal, the points on this plot should fall close to the diagonal line.

- A *bow-shaped* pattern of deviations from the diagonal indicates that the residuals have excessive *skewness* (i.e., they are not symmetrically distributed, with too many large errors in the same direction).
- An S-shaped pattern of deviations indicates that the residuals have excessive *kurtosis*--i.e., there are either too many or too few large errors in both directions.

Sample Q-Q plots are given below.





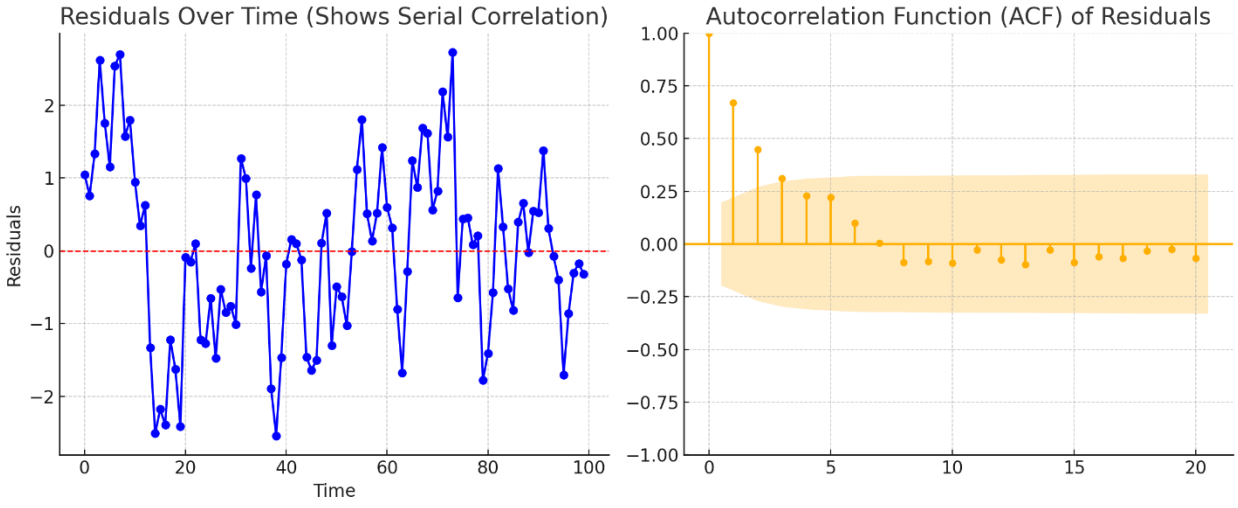
Checking the assumption of Independence of errors

This assumption demands that error terms for different observations are uncorrelated or there is no serial correlation. Issue of autocorrelation of residual mainly occurs in **time-series data**

One of the basic assumptions of the regression model we have been considering is the independence of the errors. This assumption is often violated when data are collected over sequential periods of time because a residual at any point in time may tend to be similar to residuals at adjacent points in time. Thus, positive residuals would be more likely followed by positive residuals and negative residuals would be more likely followed by negative residuals. Such a pattern in the residuals is called autocorrelation. When substantial autocorrelation is present in a set of data, the validity of fitted regression model may be in serious doubt. Serial correlation is also sometimes a byproduct of a violation of the linearity assumption.

Detecting autocorrelation:

1. The easiest way to detect autocorrelation in a set of data is to plot the residuals in time order or observation order. If an autocorrelation effect is present, clusters of residuals with the same sign will be present and an apparent Patten will be readily detected.
2. **The Durbin-Watson test**



Residuals Over Time (Left Plot):

- The residuals show a noticeable pattern rather than being randomly scattered.

- This suggests that errors are correlated over time, violating the independence assumption.

Autocorrelation Function (ACF) Plot (Right Plot):

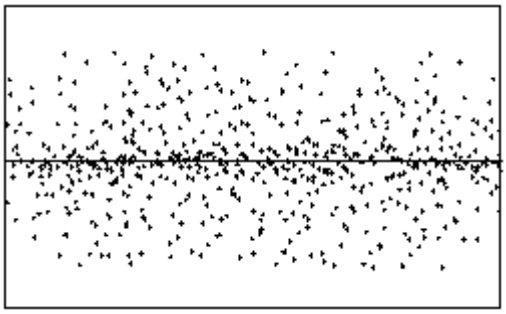
- The bars at different lags indicate that past residuals influence current residuals.
- Significant autocorrelation at lag 1 confirms serial correlation in the residuals.

Checking the assumption of Homoscedasticity of the errors (Equal variance assumption):

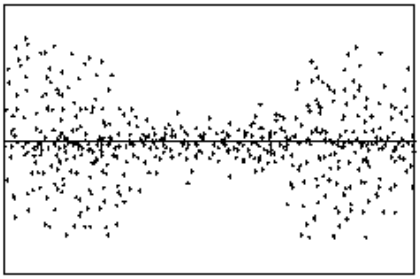
Homoscedasticity means that the variance of errors is the same across all levels of the independent variables. When the variance of errors differs at different values of the independent variables, heteroscedasticity is indicated.

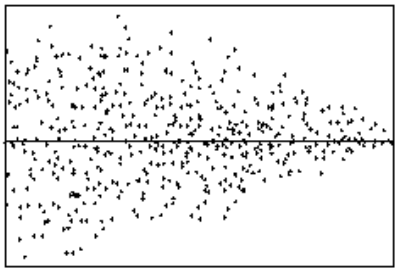
This assumption can be checked by visual examination of a plot of residuals (taken along Y-axis) versus predicted Y value (taken along X-axis)

Homoscedastic data



Heteroscedastic data





Ideally, residuals are randomly scattered around 0 (the horizontal line) providing a relatively even distribution. Heteroscedasticity is indicated when the residuals are not evenly scattered around the line. There are many forms heteroscedasticity can take, such as a bow-tie or fan shape.

Summary of Patterns and Their Implications:

PATTERN	INTERPRETATION
Random scatter	Homoscedasticity: Variance is constant; assumption is satisfied.
Funnel shape	Heteroscedasticity: Variance increases with fitted values.
Reverse funnel shape	Heteroscedasticity: Variance decreases with fitted values.
Curved or nonlinear pattern	Non-linearity in the relationship; may co-occur with heteroscedasticity.
Outliers	Potential influential points; may distort variance estimates.
Systematic clustering	Omitted variables or grouping effects; may indicate heteroscedasticity.

Remedies for Heteroscedasticity:

If heteroscedasticity is detected, consider the following remedies:

1. **Transformations:** Apply transformations to the dependent variable (e.g., log, square root) to stabilize variance.
2. **Weighted Least Squares (WLS)** : Use WLS to give less weight to observations with higher variance.
3. **Robust Standard Errors** : Use robust standard errors to adjust inference for heteroscedasticity.
4. **Add Missing Variables** : Include additional predictors that explain the changing variance.
5. **Nonlinear Models** : Consider using nonlinear models if the relationship is inherently nonlinear.