Checking assumptions of Linear Regressions

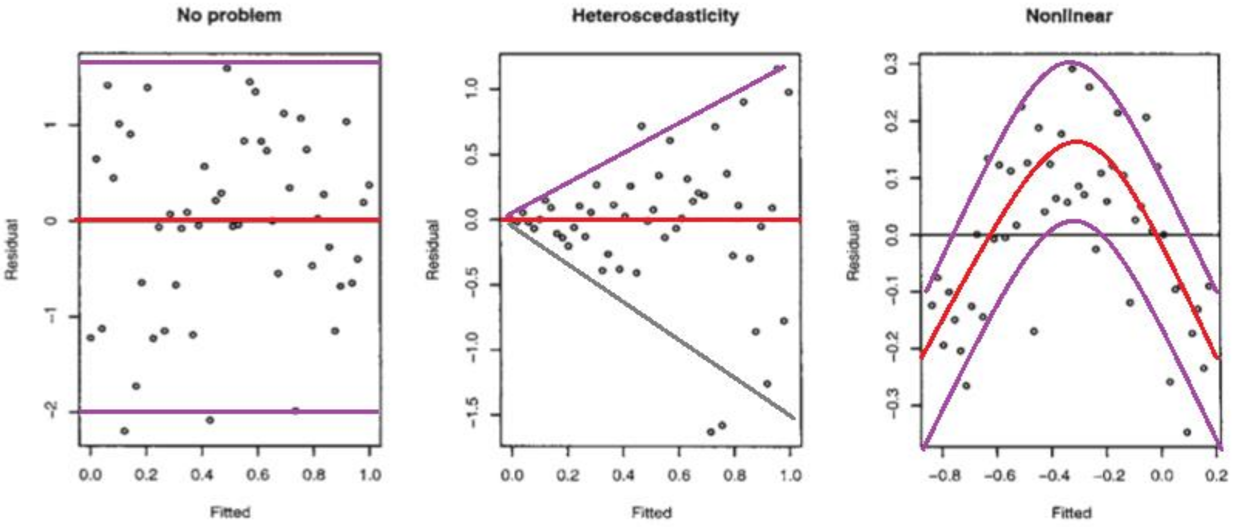
The assumptions are crucial because violating them can lead to biased or inefficient estimates, invalid statistical inferences, and unreliable predictions. Here are the key assumptions to check for Simple Linear regression & Multiple Linear regression models:

Simple Linear Regression Assumptions:

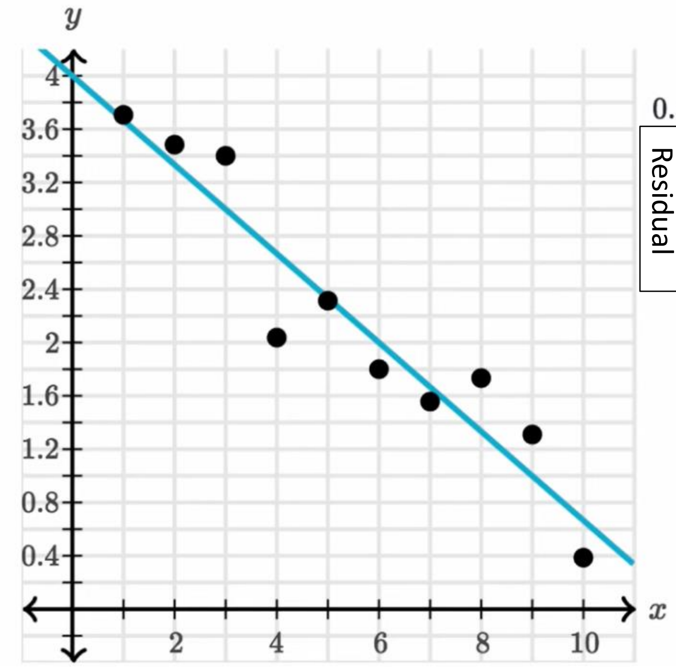
1. **Linearity:** The relationship between the independent variable (X) and the dependent variable (Y) should be linear. You can visually check this by plotting the data and examining the pattern. Following are other ways to check linearity:
   A. **Scatterplots**: Plotting the response variable (dependent variable) against each predictor variable (independent variable) can help visually inspect the linearity assumption. If the relationship appears to be linear, it supports the assumption.

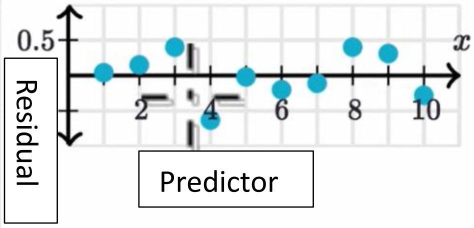
   B. **Partial Regression Plots (Component-Component plus Residual plots)**: These plots show the relationship between one predictor variable (independent variable) and the response (dependent variable) after removing the effect of all other predictors. A linear relationship in these plots suggests that the linearity assumption holds.

   C. **Residuals vs. Fitted Values Plot**: Plotting the residuals (the differences between observed and predicted values) against the fitted values (predicted values) can help identify patterns. If the points are randomly scattered around zero with no discernible pattern, it indicates that the linearity assumption is reasonable. A curvature or any systematic pattern may suggest non-linearity.

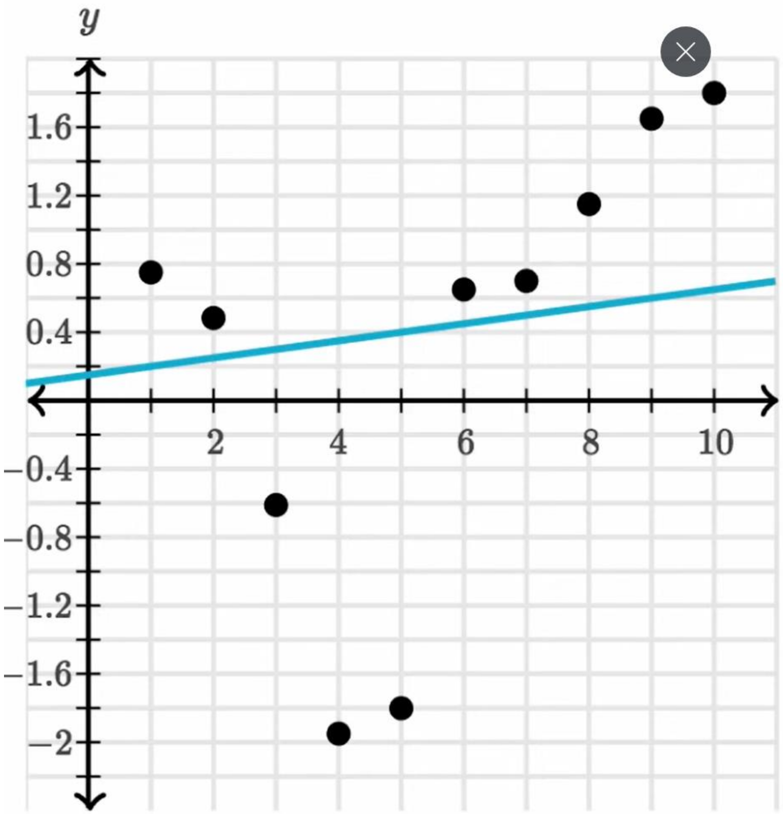


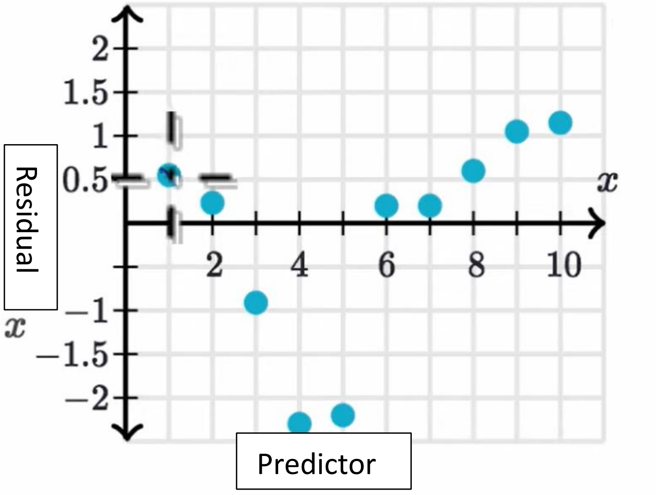
D. **Residuals vs. Predictor Plots**: Plotting residuals against each predictor variable individually can also reveal non-linear relationships. If there is no discernible pattern in these plots, it suggests linearity.





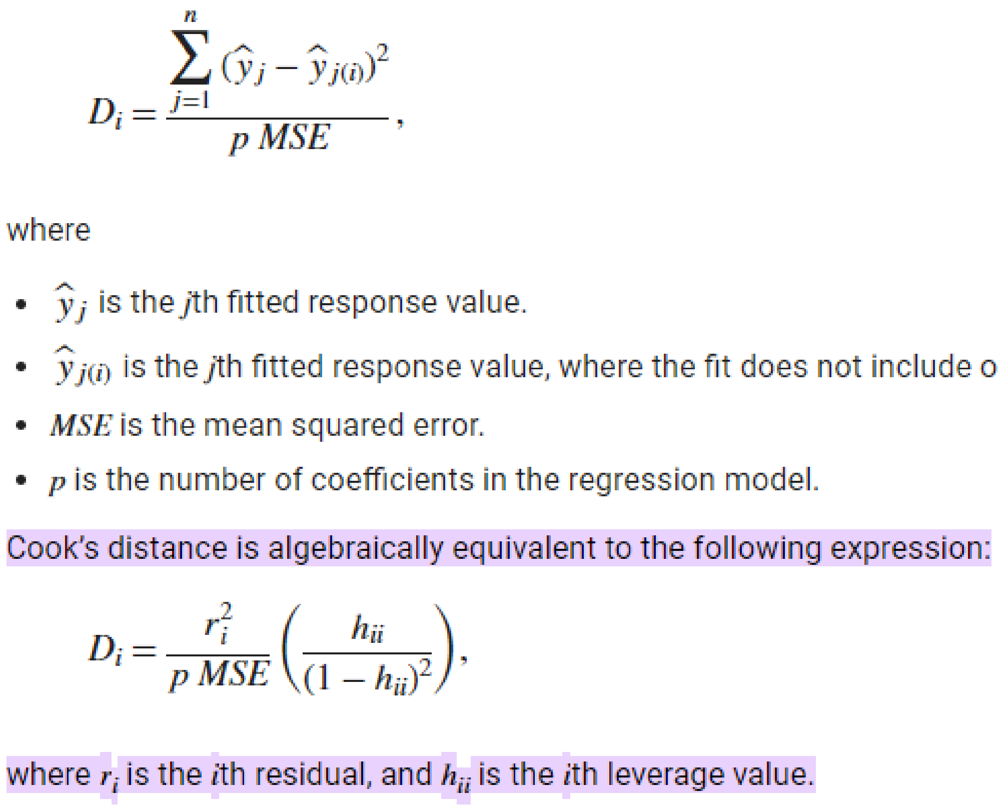
This graph shows there is no pattern in Residual vs Predictor plot. Hance, regression graph of X-values v/s Y-values are linear.





This graph shows there is up & down pattern in Residual vs Predictor plot. Hance, regression graph of X-values v/s Y-values are linear.

E. **Cook's Distance**: This measure helps identify influential data points that might be affecting the linearity assumption. Points with high Cook's distance may indicate potential non-linearity issues.
Each element in the Cook's distance D is the normalized change in the fitted response values due to the deletion of an observation. The Cook’s distance of observation i is given by,



$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p\ MSE},$$

where

- $\hat{y}_j$ is the jth fitted response value.
- $\hat{y}_{j(i)}$ is the jth fitted response value, where the fit does not include observation i.
- MSE is the mean squared error.
- p is the number of coefficients in the regression model.

Cook’s distance is algebraically equivalent to the following expression:

$$D_i = \frac{r_i^2}{p\ MSE}\left(\frac{h_{ii}}{(1-h_{ii})^2}\right),$$

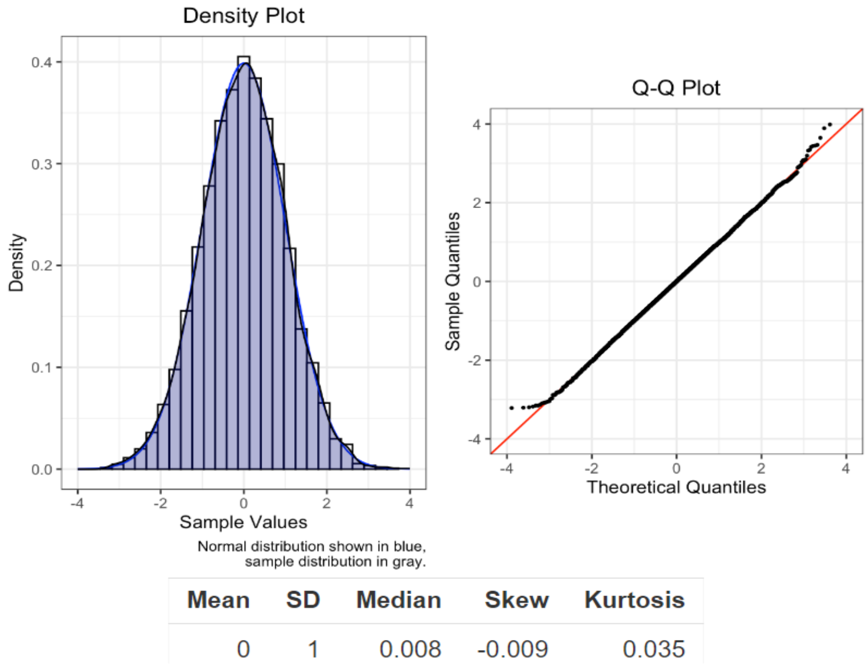
where r_i is the ith residual, and h_{ii} is the ith leverage value.

F. **Squared Term Addition**: Introducing squared terms or other transformations of predictors into the model and observing changes in model fit statistics (like R-squared) can help identify non-linear relationships. If adding squared terms significantly improves the model fit, it suggests non-linearity.

2. **Normality of residuals**: The residuals (errors) should be normally distributed. You can check this assumption by plotting a histogram or a normal probability plot of the residuals. Following are other ways to checking normality test:

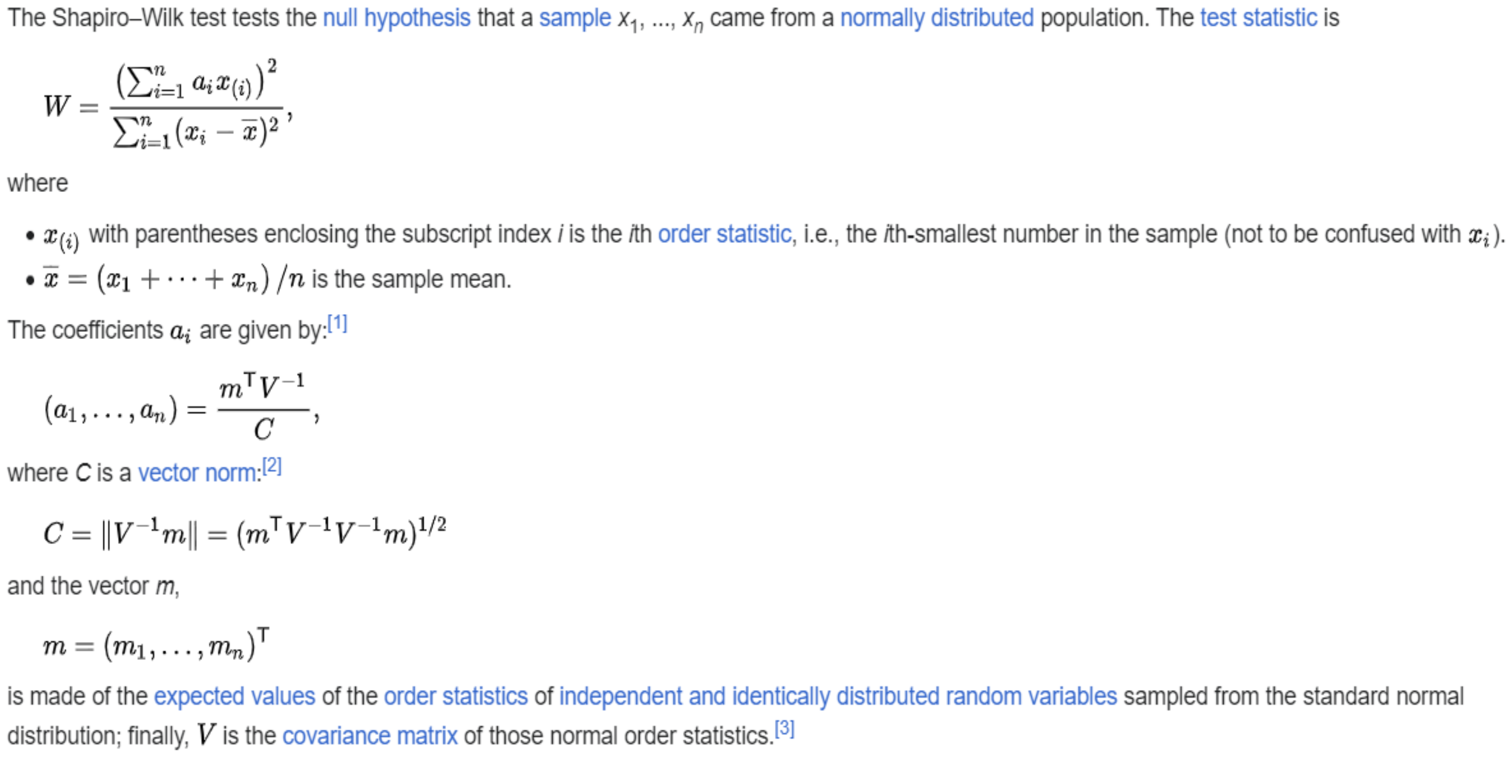
A. **Histogram and Density Plot**: Creating a histogram or density plot of the residuals can provide a visual assessment of their distribution. If the histogram or density plot resembles a bell curve, it suggests that the residuals are approximately normally distributed.

B. **Q-Q (Quantile-Quantile) Plot**: A Q-Q plot compares the quantiles of the residuals to the quantiles of a theoretical normal distribution {standard normal variate (a normal distribution with mean of zero and a standard deviation of one)}. If the residuals follow a normal distribution, the points on the Q-Q plot should fall approximately along a straight line. Deviations from a straight line indicate departures from normality.



Mean	SD	Median	Skew	Kurtosis
0	1	0.008	-0.009	0.035

C. **Shapiro-Wilk Test**: The Shapiro-Wilk test is a formal statistical test of normality. It tests the null hypothesis that the residuals are normally distributed. A low p-value (< 0.05) indicates evidence against the null hypothesis, suggesting non-normality.



The Shapiro–Wilk test tests the null hypothesis that a sample x_1, ..., x_n came from a normally distributed population. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

where

- $x_{(i)}$ with parentheses enclosing the subscript index *i* is the *i*th order statistic, i.e., the *i*th-smallest number in the sample (not to be confused with x_i).
- $\overline{x} = (x_1 + \cdots + x_n)/n$ is the sample mean.

The coefficients a_i are given by:[1]

$$(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{C},$$

where *C* is a vector norm:[2]

$$C = \|V^{-1}m\| = (m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}$$

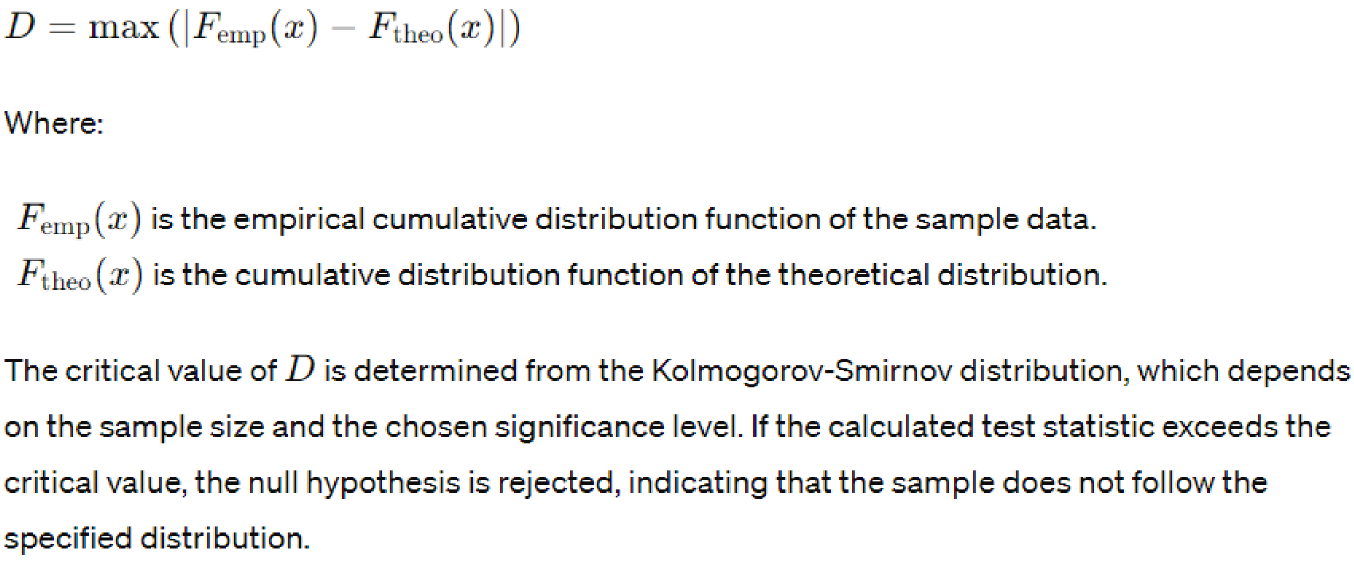
and the vector *m*,

$$m = (m_1, \ldots, m_n)^{\mathsf{T}}$$

is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics.[3]

D. **Kolmogorov-Smirnov Test**: Kolmogorov-Smirnov test is a non parametric test. It compares the empirical cumulative distribution function of the residuals to the cumulative distribution function of a normal distribution. A significant p-value suggests departure from normality.

The test statistic D is calculated as the maximum absolute difference between the two cumulative distribution functions:



$$D = \max\left(|F_{\text{emp}}(x) - F_{\text{theo}}(x)|\right)$$

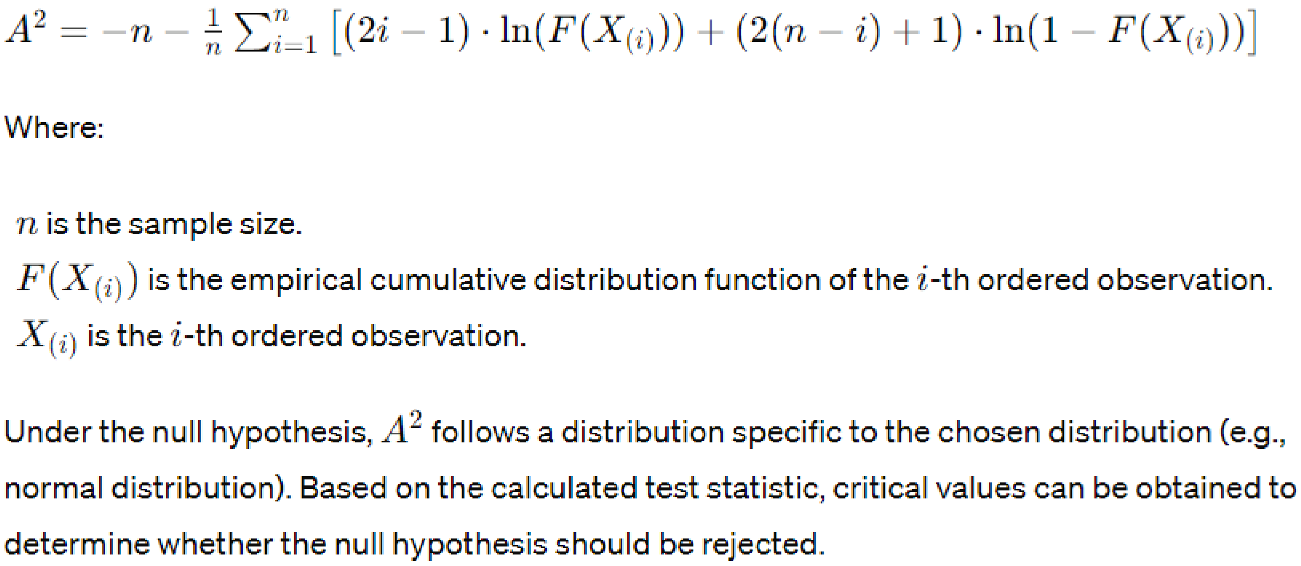
Where:

- $F_{\text{emp}}(x)$ is the empirical cumulative distribution function of the sample data.
- $F_{\text{theo}}(x)$ is the cumulative distribution function of the theoretical distribution.

The critical value of D is determined from the Kolmogorov-Smirnov distribution, which depends on the sample size and the chosen significance level. If the calculated test statistic exceeds the critical value, the null hypothesis is rejected, indicating that the sample does not follow the specified distribution.

E. **Anderson-Darling Test**: Anderson-Darling Test provides a measure of how well the data fit a normal distribution. A significant p-value suggests departure from normality.

The Anderson-Darling test statistic is calculated based on the squared differences between the observed cumulative distribution function (CDF) and the expected CDF under the null hypothesis. The test statistic takes the form:



$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\cdot\ln(F(X_{(i)})) + (2(n-i)+1)\cdot\ln(1-F(X_{(i)}))\right]$$

Where:

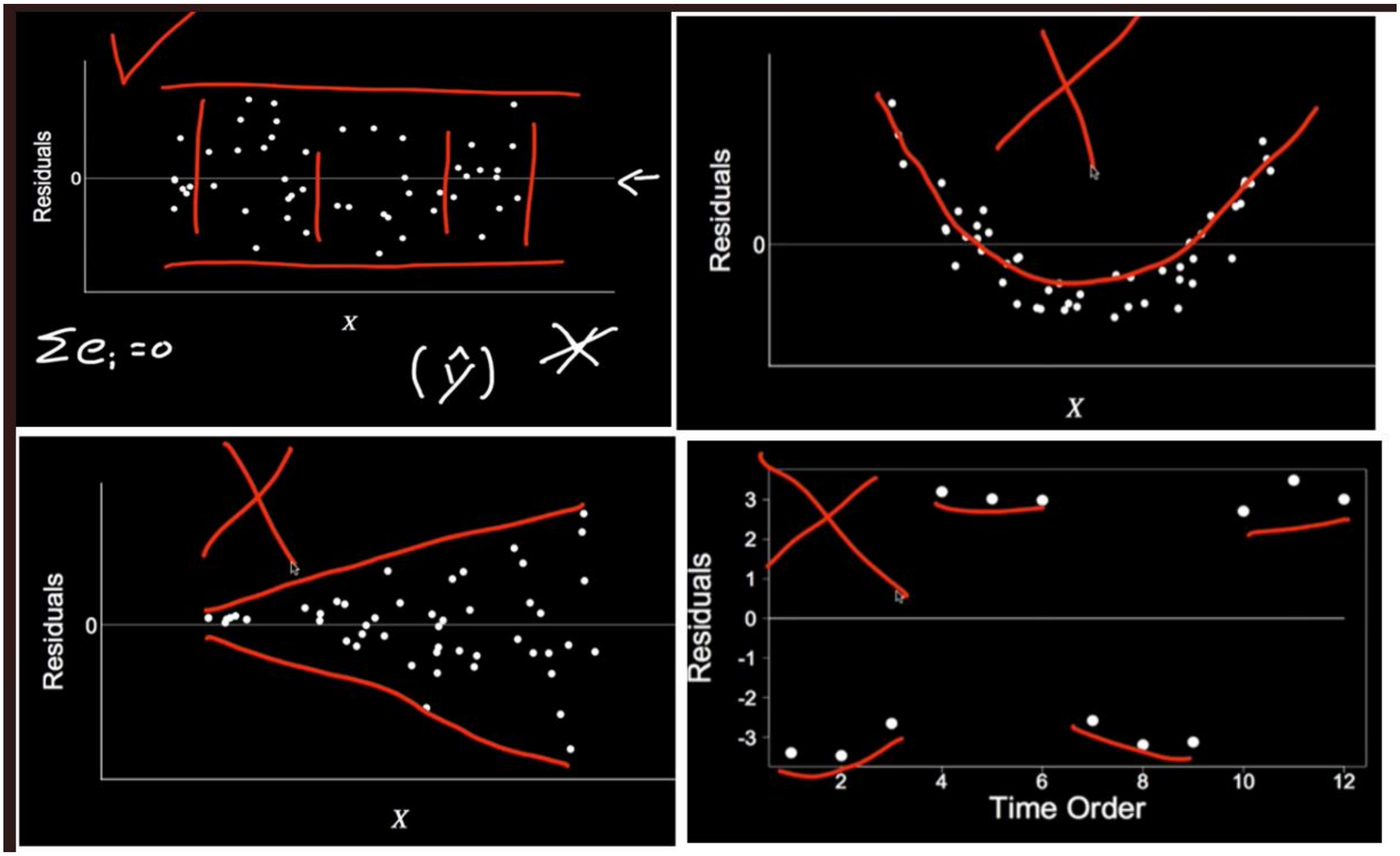
- n is the sample size.
- $F(X_{(i)})$ is the empirical cumulative distribution function of the i-th ordered observation.
- $X_{(i)}$ is the i-th ordered observation.

Under the null hypothesis, A^2 follows a distribution specific to the chosen distribution (e.g., normal distribution). Based on the calculated test statistic, critical values can be obtained to determine whether the null hypothesis should be rejected.

F. **Probability Plotting**: Plotting the ordered residuals against the corresponding quantiles from a standard normal distribution can help visualize departures from normality. If the points deviate significantly from the straight line, it indicates non-normality.

3. **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variable. Following are the ways to test Homoscedasticity:

A. **Residuals vs. Fitted Values Plot**: Plot the residuals (the differences between observed and predicted values) against the fitted values (predicted values). If there is a clear pattern, such as the spread of residuals widening or narrowing as fitted values increase, it indicates heteroscedasticity (non-constant variance). Homoscedasticity, on the other hand, would show a random scatter of residuals around zero with no discernible pattern.



First Image is shows Homoscedasticity, Rest other shows Heteroscedasticity.

B. **Residuals vs. Predictor Plot:** Plot the residuals against each predictor variable individually. If there is a systematic pattern in these plots, such as the spread of residuals varying systematically with the values of predictor variables, it suggests heteroscedasticity.

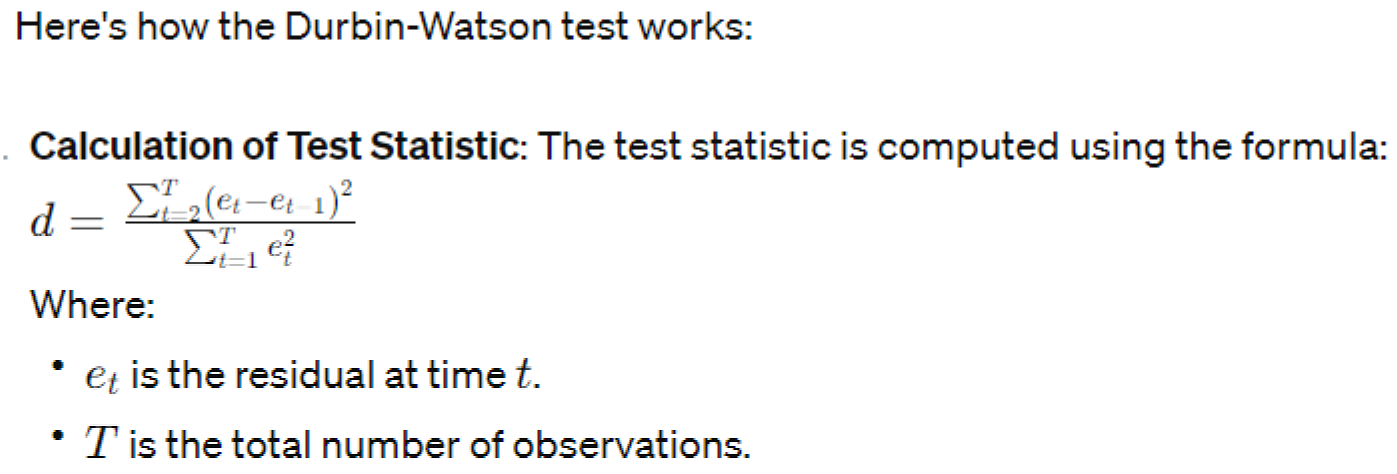
C. **White's Test**: White's test is a formal statistical test for heteroscedasticity. It involves augmenting the regression model with additional terms consisting of the squares and cross-products of the original predictors. The null hypothesis is that the variance of the errors is constant (homoscedasticity). A significant p-value suggests rejection of the null hypothesis, indicating heteroscedasticity.

4. **Independence of Errors**: The residuals should be independent of each other, meaning there should be no autocorrelation. You can check this assumption using the Durbin-Watson statistic or by plotting the residuals against their lagged values.

Following are some other ways to check independence of error:

A. **Residuals Autocorrelation**: Plot the residuals (the differences between observed and predicted values) against the order of observation or time if the data is time-series. Any discernible pattern in this plot suggests autocorrelation, indicating a lack of independence of errors.

B. **Durbin-Watson Test**: The Durbin-Watson test is a formal statistical test for autocorrelation in the residuals of a regression model. The test statistic ranges between 0 and 4, with values close to 2 indicating no autocorrelation. Values significantly lower than 2 suggest positive autocorrelation, while values significantly higher than 2 suggest negative autocorrelation.



Here's how the Durbin-Watson test works:

Calculation of Test Statistic: The test statistic is computed using the formula:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Where:

- e_t is the residual at time t.
- T is the total number of observations.

C. Plotting Residuals vs. Time or Predictor Variables: Besides plotting residuals against the order of observations, you can also plot residuals against time or predictor variables to identify patterns or trends, which may indicate lack of independence.

D. Scatterplot Matrix: Create a scatterplot matrix of the residuals against the predictor variables. If there are patterns or trends in these plots, it may suggest lack of independence.

Multiple Linear Regression Assumptions:

1. **Linearity**: The relationship between the independent variables (X1, X2, ..., Xn) and the dependent variable (Y) should be linear. You can visually inspect this by plotting partial regression plots or using residual analysis. Other ways to check Linearity assumption in Described in S.L.R.
2. **Normality**: The residuals should be normally distributed. You can check this assumption by plotting a histogram or a normal probability plot of the residuals. Other ways to check Normality assumption in Described in S.L.R.
3. **Homoscedasticity**: The variance of the residuals should be constant across all levels of the independent variables. You can visually inspect this by plotting the residuals against the fitted values or each independent variable. Other ways to check Homoscedasticity assumption in Described in S.L.R.
4. **Independence of Errors**: The residuals should be independent of each other, meaning there should be no autocorrelation. You can check this assumption using the Durbin-Watson statistic or by plotting the residuals against their lagged values. Other ways to check Independence of Error assumption in described in S.L.R.
5. **Multicollinearity**: The independent variables should not be highly correlated with each other. You can check for multicollinearity by calculating the variance inflation factor (VIF) or by inspecting the correlation matrix of the independent variables.
   Following are some other ways to check Multicollinearity:
   A. **Correlation Matrix**: Compute the correlation matrix of the predictor variables. High correlations (typically above 0.7 or 0.8) between pairs of predictor variables may indicate multicollinearity.
   B. **Variance Inflation Factor (VIF):** Calculate the VIF for each predictor variable. VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF greater than 5 or 10 is often considered indicative of multicollinearity.
   C. **Tolerance:** Tolerance is the reciprocal of the VIF. It measures the proportion of variance in a predictor variable that is not explained by the other predictor variables. Low tolerance values (typically less than 0.1) suggest multicollinearity.
   D. **Principal Component Analysis (PCA)**: Conduct PCA on the predictor variables and examine the proportion of variance explained by the principal components. High proportions of variance explained by a small number of components may suggest multicollinearity.

E. **Cross-Validation**: Split the data into training and validation sets and assess model performance. In the presence of multicollinearity, models may perform poorly on the validation set due to overfitting.

6. **No Influential Observations**: The model should not be unduly influenced by a few data points. You can identify influential observations by calculating Cook's distance, leverage values, or studentized residuals.
   Influential observations can affect the model in various ways:
   a. **Outliers:** Observations with extreme values in the predictor or response variables can unduly influence the estimated regression coefficients and reduce the precision of parameter estimates.
   b. **High Leverage Points**: Observations with extreme values in the predictor variables (far from the mean) can exert substantial leverage on the estimated regression coefficients. These points can pull the regression line closer to them, affecting the overall fit of the model.
   c. **High Influence Points**: Observations that have both high leverage and high residual values can be particularly influential. These points can substantially alter the estimated coefficients and model fit.
   Following are ways to check No Influential Observations:
   A. **Cook's Distance**: Cook's distance is a measure of the influence of each observation on the fitted values and regression coefficients. Observations with large Cook's distances are considered influential. Typically, observations with Cook's distance greater than 4/N, where N is the sample size, are flagged as influential.
   B. **Leverage**: Leverage measures the influence of each observation on the estimated regression coefficients. High leverage points are those with leverage values substantially greater than the average leverage. Leverage values are computed as the diagonal elements of the "hat" matrix.
   C. **Studentized Residuals**: Studentized residuals are residuals that have been adjusted for their standard errors. Observations with large studentized residuals are potential outliers or influential points. Typically, observations with absolute studentized residuals greater than 2 or 3 are considered influential.
   D. **Hat Matrix**: The hat matrix contains leverage values for each observation. Plotting the diagonal elements of the hat matrix against the predicted values can help identify influential observations.