Unit IV- Validity of Assumptions

The major **assumptions** that we have made thus far in our study of regression analysis are as follows:

1. The relationship between the response y and the regressors is linear, at least approximately.
2. The error term ε has zero mean.
3. The error term ε has constant variance σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed.

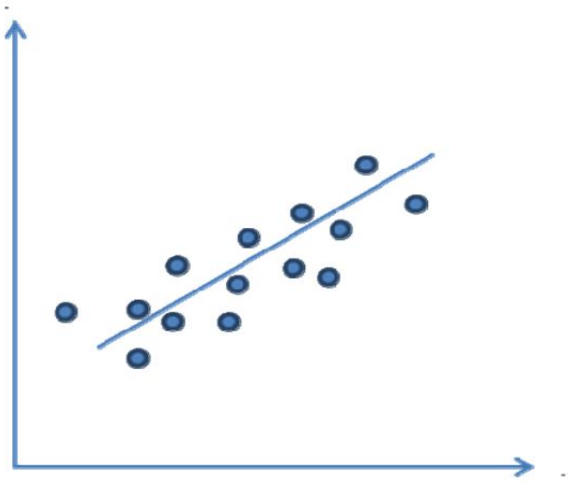
The validity of these assumptions is needed for the results to be meaningful. If these assumptions are violated, the result can be incorrect and may have serious consequences. If these departures are small, the final result may not be changed significantly. But if the deviations are large, the model obtained may become unstable in the sense that a different sample could lead to an entirely different model with opposite conclusions. So such underlying assumptions have to be verified before attempting to regression modelling. Such information is not available from the summary statistic such as t-statistic, F-statistic or coefficient of determination. We should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model.

Checking of the linear relationship between study and explanatory variables

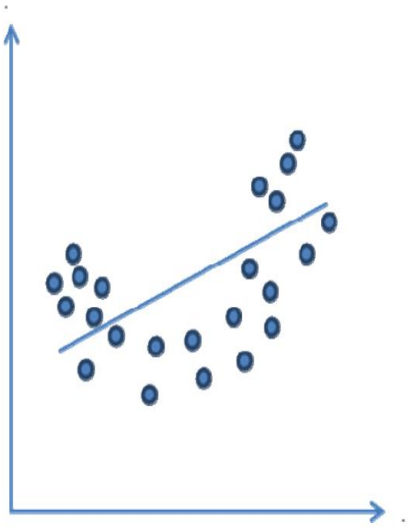
1. Case of one explanatory variable

If there is only one explanatory variable in the model, then it is easy to check the existence of the linear relationship between y and X by scatter diagram of the available data. If the scatter diagram shows a linear trend, it indicates that the relationship between y and X is linear. If the pattern is not linear, then it

suggests that the relationship between y and X is nonlinear. For example, the following figure indicates a linear trend



whereas the following graph suggests a nonlinear trend:



2. Case of more than one explanatory variables

These analyses methods are primarily based on study of the model **residuals.** The difference between the observed value y_i and the corresponding fitted value of $\widehat{y_i}$ is a residual. Mathematically the ith residual is

$e_i = y_i - \widehat{y_i} = y_i - \widehat{\beta_0} + \widehat{\beta_1} x_i$; i= 1, 2, 3, …….n

e_i is also a measure of the variability in the response variable not explained by the regression model.

Thus, any departures from the assumptions on the errors should show up in the residuals.

So, Analysis of the residuals is an effective way to discover several types of model inadequacies.

Residual Plots

Graphical analysis of residuals is a very effective way to investigate the adequacy of the fit of a regression model and to check the underlying assumptions. **Residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.

Plot of Residuals against the Fitted Values

Let's examine what this assumption means.

Every regression model inherently has some degree of error since you can never predict something 100% accurately. More importantly, randomness and unpredictability are always a part of the regression model. Hence, a regression model can be explained as:

Response= Deterministic + Stochastic

The deterministic part of the model is what we try to capture using the regression model. Ideally, our linear equation model should accurately capture the predictive information. Essentially, what this means is that if we capture all of the predictive information, all that is left behind (residuals) should be completely random & unpredictable i.e stochastic.

Characteristics of Good Residual Plots

A few characteristics of a good residual plot are as follows:

1. It has a high density of points close to the origin and a low density of points away from the origin

2. It is symmetric about the origin

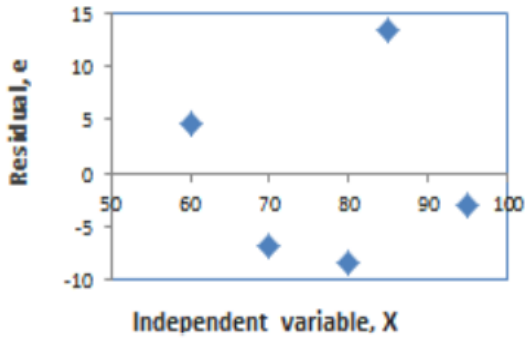
If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.

The table below shows inputs and outputs from a simple linear regression analysis.

The table below shows inputs and outputs from a simple linear regression analysis.

x	y	ŷ	e
60	70	65.411	4.589
70	65	71.849	-6.849
80	70	78.288	-8.288
85	95	81.507	13.493
95	85	87.945	-2.945

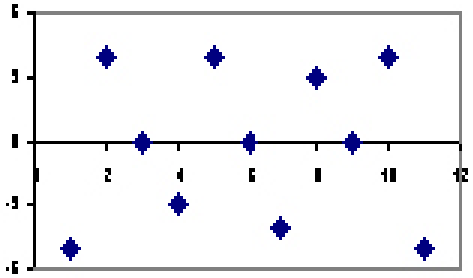
And the chart below displays the residual (e) and independent variable (X) as a residual plot.



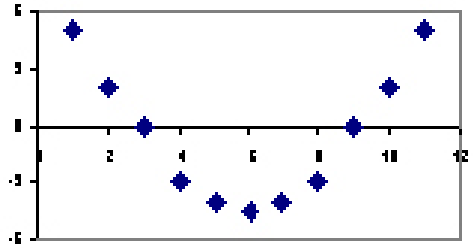
The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is

negative. This **random** pattern indicates that a linear model provides a decent fit to the data.

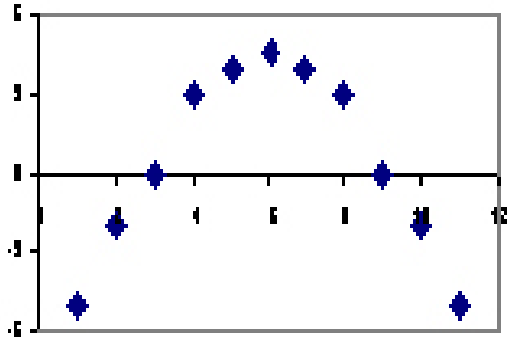
Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.



Random pattern



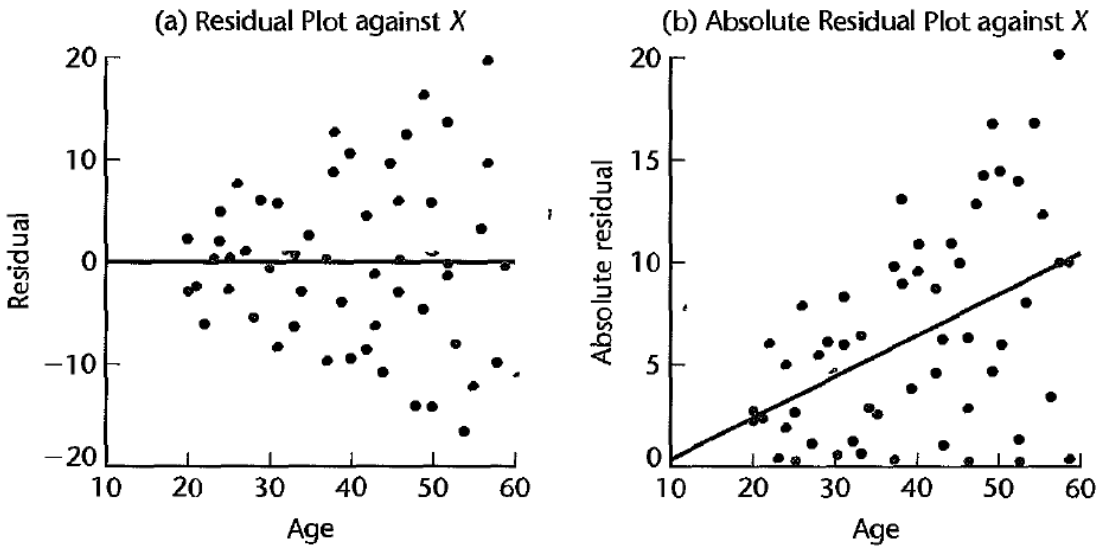
Non-random: U-shaped



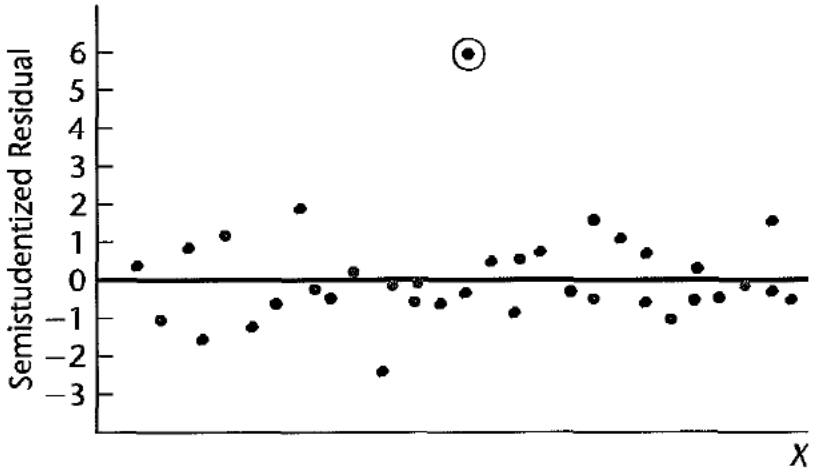
Non-random: Inverted U

The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a nonlinear model.

Residual Plots Illustrating Nonconstant Error Variance.



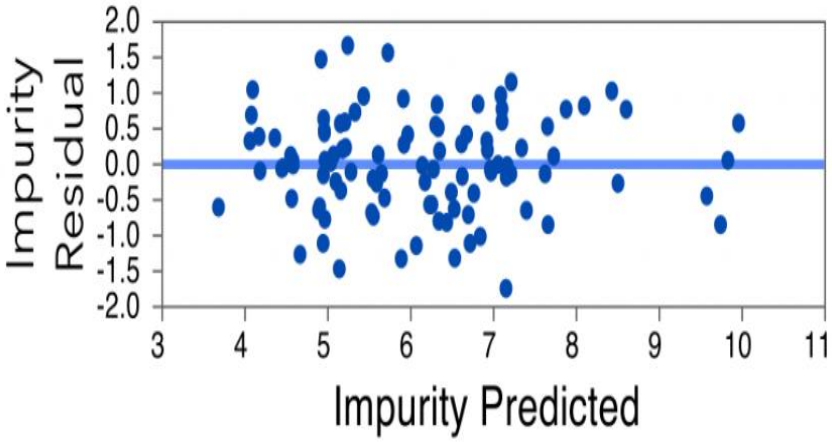
Residual Plot with Outlier.



Methods for Scaling Residuals

Sometimes it is useful to work with scaled residuals. There are four popular methods for scaling residuals. These scaled residuals are helpful in finding observations that are **outliers,** or **extreme values.**

In the residual by predicted plot, we see that the residuals are randomly scattered around the center line of zero, with no obvious non-random pattern.



One limitation of these residual plots is that the residuals reflect the scale of measurement.

As you know, the major problem with ordinary residuals is that their magnitude depends on the units of measurement, thereby making it difficult to use the residuals as a way of detecting unusual *y* values. We can eliminate the units of measurement by dividing the residuals by an estimate of their standard deviation, thereby obtaining what are known as standardized residuals.

Standardized Residuals Since the approximate average variance of a residual is estimated by *MRS,* a logical scaling for the residuals would be the **standardized residuals**

$$d_i = \frac{e_i}{\sqrt{MRS}} \qquad i = 1,2, \ldots \ldots n$$

The standardized residuals have mean zero and approximately unit variance. Consequently, a large standardized residual (*di* > 3, say) potentially indicates an outlier.

Studentized Residuals An alternative is to use studentized residuals. A studentized residual is calculated by dividing the residual by an estimate of its standard deviation. Using MRS as the variance of the *i*th residual,*ei* is only an

approximation. We can improve the residual scaling by dividing *ei* by the exact standard deviation of the *i*th residual.

e = Y- $\widehat{Y}$ = Y - $X\hat{\beta}$

e= Y – HY = (I – H)Y

where the n × n matrix H = X $(X'X)^{-1}X'$ is called the hat matrix.

e = (I-H) $(X\beta + \epsilon)$

e = $X\beta - HX\beta + (I - H)\epsilon$

$\Rightarrow e = X\beta -$ X $(X'X)^{-1}X'X\beta + (I - H)\epsilon$

$\Rightarrow e =$ (I-H) ϵ

The covariance matrix of the residuals is $Var(e) = \sigma^2\ (I - H)$

The variance of the *i*th residual is

Var$(e_i) = \sigma^2(1 - h_{ii})$

where *hii* is the *i*th diagonal element of the hat matrix **H.**

*The **studentized residuals is***

$$\boldsymbol{t_i = \frac{e_i}{\sqrt{MRS(1 - h_{ii})}}}$$

Note that the only difference between the standardized residuals considered in the previous section and the studentized residuals considered here is that standardized residuals use the mean square error for the model based on all observations, *MRS*, while studentized residuals use the mean square error based on the estimated model with the i^{th} observation deleted.

In general, studentized residuals are going to be more effective for detecting outlying *Y* observations than standardized residuals. If an observation has a studentized residual that is larger than 3 (in absolute value) we can call it an **outlier**. [Recall from the previous section that some use the term "outlier" for an observation with a standardized residual that is larger than 3 in absolute value.

To avoid any confusion, you should always clarify whether you're talking about standardized or studentized residuals when designating an observation to be an outlier.]

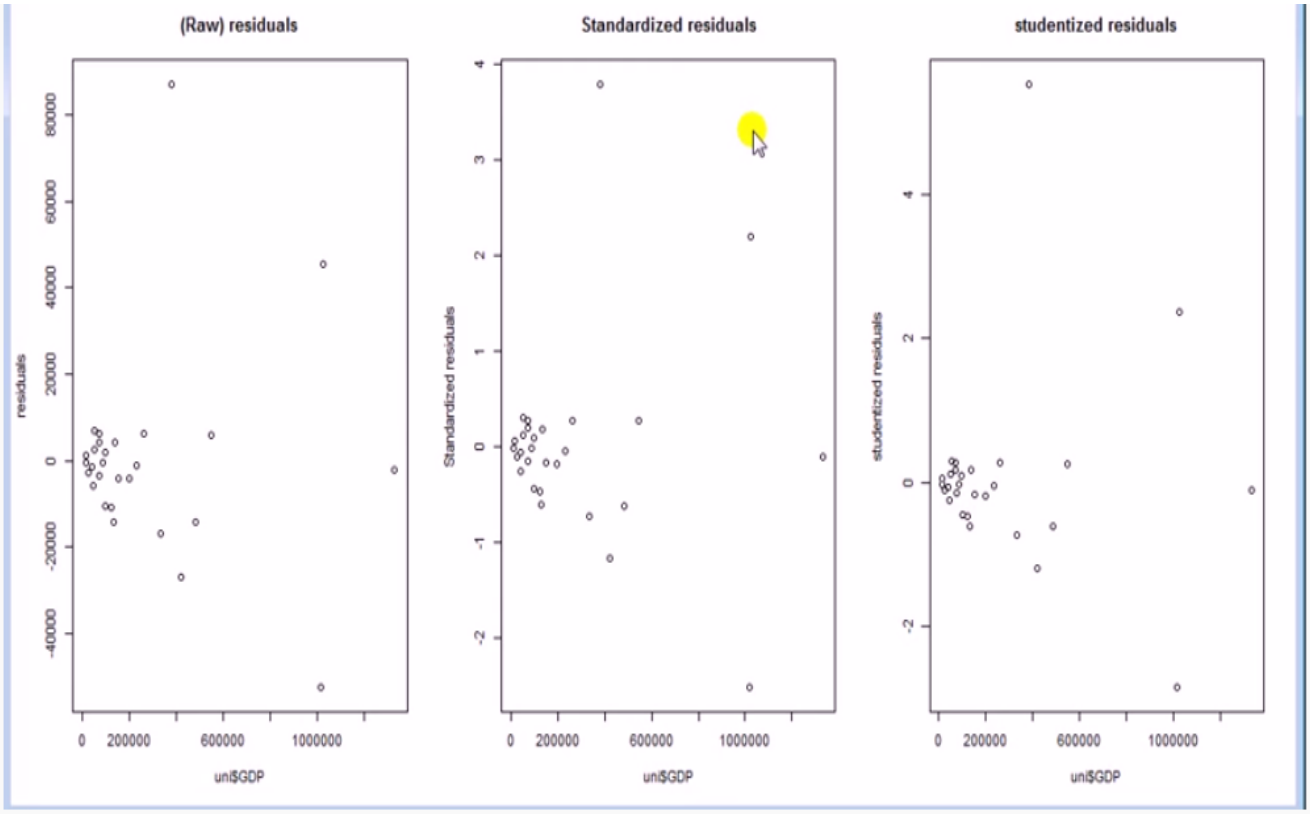
An observation with a standardized residual that is larger than 3 (in absolute value) is deemed by some to be an **outlier**. [It is technically more correct to reserve the term "outlier" for an observation with a *studentized* residual that is larger than 3 in absolute value—we consider studentized residuals in the next section.]

For example, consider again the (contrived) data set containing $n = 4$ data points (x, y):

x	y	FITS1	RESI1	HI1	SRES1
1	2	2.2	-0.2	0.7	-0.57735
2	5	4.4	0.6	0.3	1.13389
3	6	6.6	-0.6	0.3	-1.13389
4	9	8.8	0.2	0.7	0.57735

The column labeled "FITS1" contains the predicted responses, the column labeled "RESI1" contains the ordinary residuals, the column labeled "HI1" contains the h_{ii}, and the column labeled "SRES1" contains the studendized residuals. The value of *MRS* is 0.40.

The good thing about standardized residuals is that they quantify how large the residuals are in standard deviation units, and therefore can be easily used to identify outliers:

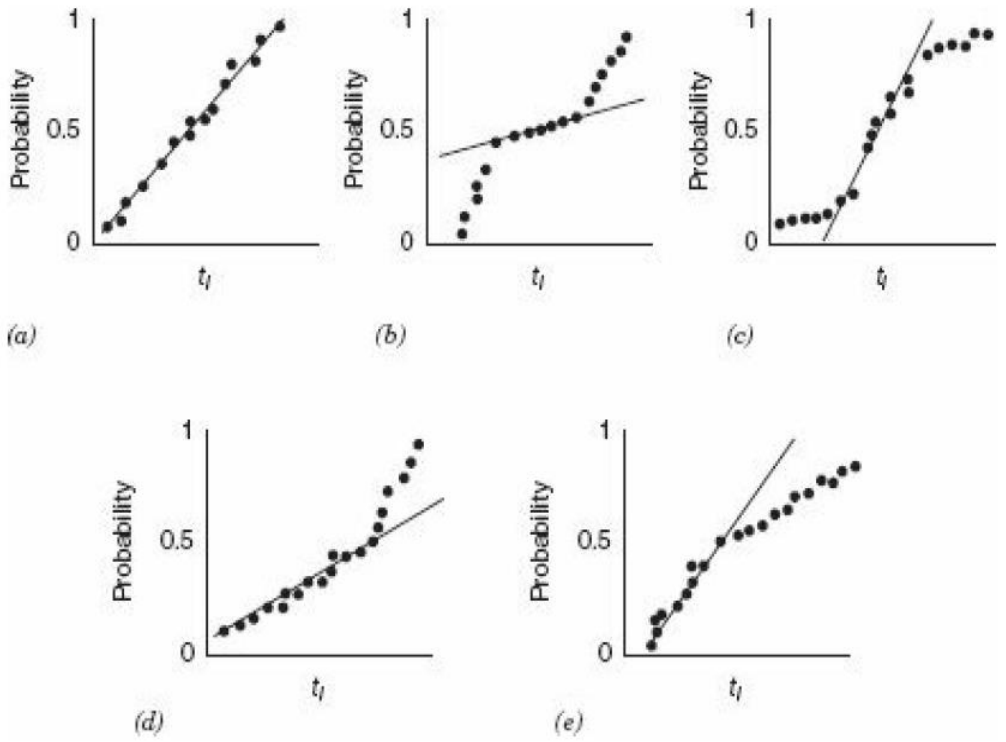


Normal Probability Plot Small departures from the normality assumption do not affect the model greatly, but gross nonnormality is potentially more serious as the t or F statistics and confidence and prediction intervals depend on the normality assumption.

Heavy-tailed error distributions often generate outlier that "pull" the least-squares fit too much in their direction. In these cases other estimation techniques (such as the **robust regression** methods should be considered.

A very simple method of checking the normality assumption is to construct a **normal probability** plot of the residuals. This is a graph designed so that the cumulative normal distribution will plot as a straight line. Let t[1] < t[2] < … < [n] be the studentized residuals ranked in increasing order. If we plot t[i] against the cumulative probability Pi i = 1, 2, …, n, on the normal probability plot, the resulting points should lie approximately on a straight line.

Figure 4.3 *Normal probability plots: (a) ideal; (b) light-tailed distribution; (c) heavy-tailed distribution; (d) positive skew; (e) negative skew.*



Residual Plot Analysis

Corrective Measures

The assumption of **constant variance** is a basic requirement of regression analysis..

Variance stabilizing transformations are often useful in these cases.

Y transformed to new variables logy or $sin^{-1}y$ or e^{y} etc.

Autocorrelation

Many applications of regression involve both predictor and response variables that are **time series**, that is, the variables are time-oriented. Regression models using time series data occur relatively often in economics, business, and many fields of engineering. The assumption of uncorrelated or independent errors that is typically made for regression is usually not appropriate for time series data. Usually the errors in time series data exhibit some type of **autocorrelated** structure. By autocorrelation we mean that the errors **are correlated** with themselves at different time periods.

Source of autocorrelation

1. Carryover effect is an important source of autocorrelation. For example, the monthly data on expenditure on the household is influenced by the expenditure of the preceding month.
2. Another source of autocorrelation is the effect of deleting some variables. For example, suppose that we wish to regress the annual sales of a product in a particular region of the country against the annual advertising expenditures for that product. Now the growth in the population in that region over the period of time used in the study will also influence the product sales. Failure to include the population size may cause the errors in the model to be positively autocorrelated.
3. The misspecification of the form of relationship can also introduce autocorrelation in the data.

The presence of autocorrelation in the errors has several effects on the ordinary least-squares regression procedure.

1. The ordinary least squares (OLS) regression coefficients are still

unbiased, but they are no longer minimum-variance estimates.

2. When the errors are positively autocorrelated, the residual mean square may seriously underestimate the error variance σ2.

3. The confidence intervals, prediction intervals, and tests of hypotheses based on the t and F distributions are, strictly speaking, no longer exact procedures.

The autocovariance at lag s is defined as

$\gamma_s = E(\epsilon_i, \epsilon_{i-s})$; s=0,1,2, ,,,,

At zero lag, we have constant variance, i.e $V(\epsilon_i) = \sigma^2$

The autocorrelation coefficient at lag s is defined as

$$\rho_s = \frac{E(\epsilon_i, \epsilon_{i-s})}{\sqrt{V(\epsilon_i)}\sqrt{V(\epsilon_{i-s})}} = \frac{\gamma_s}{\sigma^2}$$

,

DETECTING OF AUTOCORRELATION:

THE DURBIN–WATSON TEST

Residual plots can be useful for the detection of autocorrelation.

If there is positive autocorrelation, residuals of identical sign occur in clusters. That is, there are not enough changes of sign in the pattern of residuals. On the other hand, if there is negative autocorrelation, the residuals will alternate signs too rapidly.

Various **statistical tests** can be used to detect the presence of autocorrelation. The test developed by Durbin and Watson (1950, 1951, 1971) a very widely used procedure. This test is based on the assumption that the errors in the regression model are generated by a **first-order autoregressive process** observed at equally spaced time periods, that is,

$$\epsilon_i = \phi\, \epsilon_{i-1} + a_i$$

where ϵ_i is the error term in the model at time period i, a_i is an NID random variable, ϕ is a parameter that defines the relationship between successive values of the model errors εi and $\varepsilon i-1$

$i = 1, 2, \ldots, n$

Thus, a simple linear regression model with **first-order autoregressive errors** would be

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i = \phi\, \epsilon_{i-1} + a_i$

Similarly we can define for multiple linear regression.

The difference between the observed value y_i and the corresponding fitted value of $\widehat{y_i}$ is a residual. Mathematically the ith residual is

$e_i = y_i - \widehat{y_i}$

The null hypothesis is $H_0: \phi = 0$

The *D-W* test statistic is

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

$$= \frac{\sum_{t=2}^{n} e_t^2}{\sum_{t=1}^{n} e_t^2} + \frac{\sum_{t=2}^{n} e_{t-1}}{\sum_{t=1}^{n} e_t^2} - 2\frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}.$$

For large n,

$$d \approx 1+1-2r$$

$$d \simeq 2(1-r)$$

where r is the sample autocorrelation coefficient from residuals based on OLSE and can be regarded as the regression coefficient of e_t on e_{t-1}. Here

positive autocorrelation of e_t's $\Rightarrow d < 2$

negative autocorrelation of e_t's $\Rightarrow d > 2$

zero autocorrelation of e_t's $\Rightarrow d \approx 2$

As $-1 < r < 1$, so

if $-1 < r < 0$, then $2 < d < 4$ and

if $0 < r < 1$, then $0 < d < 2$.

So d lies between 0 and 4.

Heteroscedasticity

Heteroscedasticity as the condition in which the variance of error term or the residual term in a regression model varies.

$$\mathrm{Var}(\epsilon) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & . & . & . & 0 \\ 0 & \sigma_2^2 & 0 & . & . & . & 0 \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & \sigma_n^2 \end{bmatrix}$$

Possible reasons of arising Heteroscedasticity:

1. Often occurs in those data sets which have a large range between the largest and the smallest observed values i.e. when there are outliers.
2. When model is not correctly specified.
3. If observations are mixed with different measures of scale.
4. When incorrect transformation of data is used to perform the regression.
5. Skewness in the distribution of a regressor, and may be some other sources.

Effects of Heteroscedasticity:

- As mentioned above that one of the assumption (assumption number 2) of linear regression is that there is no heteroscedasticity. Breaking this assumption means that OLS (Ordinary Least Square) estimators are not the Best Linear Unbiased Estimator(BLUE) and their variance is not the lowest of all other unbiased estimators.
- Estimators are no longer best/efficient.
- The tests of hypothesis (like t-test, F-test) are no longer valid due to the inconsistency in the co-variance matrix of the estimated regression coefficients.

DETECTING OF HETEROSCEDASTICITY:

THE BREUSH – PEGAN TEST:

Residual plots can be useful for the detection of **Heteroscedasticity**

Breusch–Pagan test, developed in 1979 by Trevor **Breusch** and Adrian **Pagan**, is used to **test** for heteroskedasticity in a linear regression model.

- The null hypothesis for this test is that the error variances are all equal.
- The alternate hypothesis is that the error variances are not equal. More specifically, as Y increases, the variances increase (or decrease).

H_0 : Var $(\epsilon_i) = \sigma^2 \; \forall$ i $= 1, 2, \ldots$ n against

H_1 : Var $(\epsilon_i) = \sigma^2$ f(x) = σ^2 ($\delta 0 + \delta 1 x1 + \delta 2 x2 + \ldots + \delta p x p$)

If homoskedasticity holds, then we would have

H0:δ1=δ2=...=δk =0 against H1: At least one δi ≠ 0

1. Using OLSE find $\hat{\beta}$ *and residuals* e_i.
2. Find $\hat{\sigma}^2 = MSR$
3. Regress $e_i^2/\hat{\sigma}^2$ on x1, x2,…xp and find ESS (Explained sum of squares)
4. $\frac{ESS}{2} \sim \chi^2_{p-1}$
5. If $\frac{ESS}{2} > critical\ value\ we\ reject\ H_0$.

Multicollinearity

Multicollinearity occurs when independent variables in a **regression** model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

There are four primary sources of Multicollinearity:

1. The data collection method employed is wrong
2. Constraints on the model or in the population
3. Model misspecification

4. An overdefined model

Multicollinearity causes the following two basic types of problems:

- The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.
- In Multicollinearity the variance of OLSEs becomes large. This indicates highly unreliable estimates. You might not be able to trust the p-values to identify independent variables that are statistically significant.

Detection

1. Testing for Multicollinearity with Variance Inflation Factors (VIF)

If you can identify which variables are affected by multicollinearity and the strength of the correlation, you're well on your way to determining whether you need to fix it. Fortunately, there is a very simple test to assess multicollinearity in your regression model. **The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation**.

Since the variance of the jth regression coefficients is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of is increased due to near-linear dependences among the regressors.

So VIF = $c_{jj} = \frac{1}{1-R_j^2}$

If R_j^2 denotes the coefficient of determination obtained when X_j is regressed on the remaining (k-1) variables excluding X_j

One or more large VIFs indicate multicollinearity. Practical experience indicates that if any of the VIFs exceeds 5 it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity.

2. One popular detection method is based on the pairwise correlation between predictor variables. If it’s above .8 (or .7 or .9 or some other high number), the rule of thumb says you have multicollinearity.