Time-series Econometrics

Dr. Subrata Roy
Associate Professor
Department of Commerce
MGCUB

M.Phil/P.hd

Research Methodology (Unit-III)

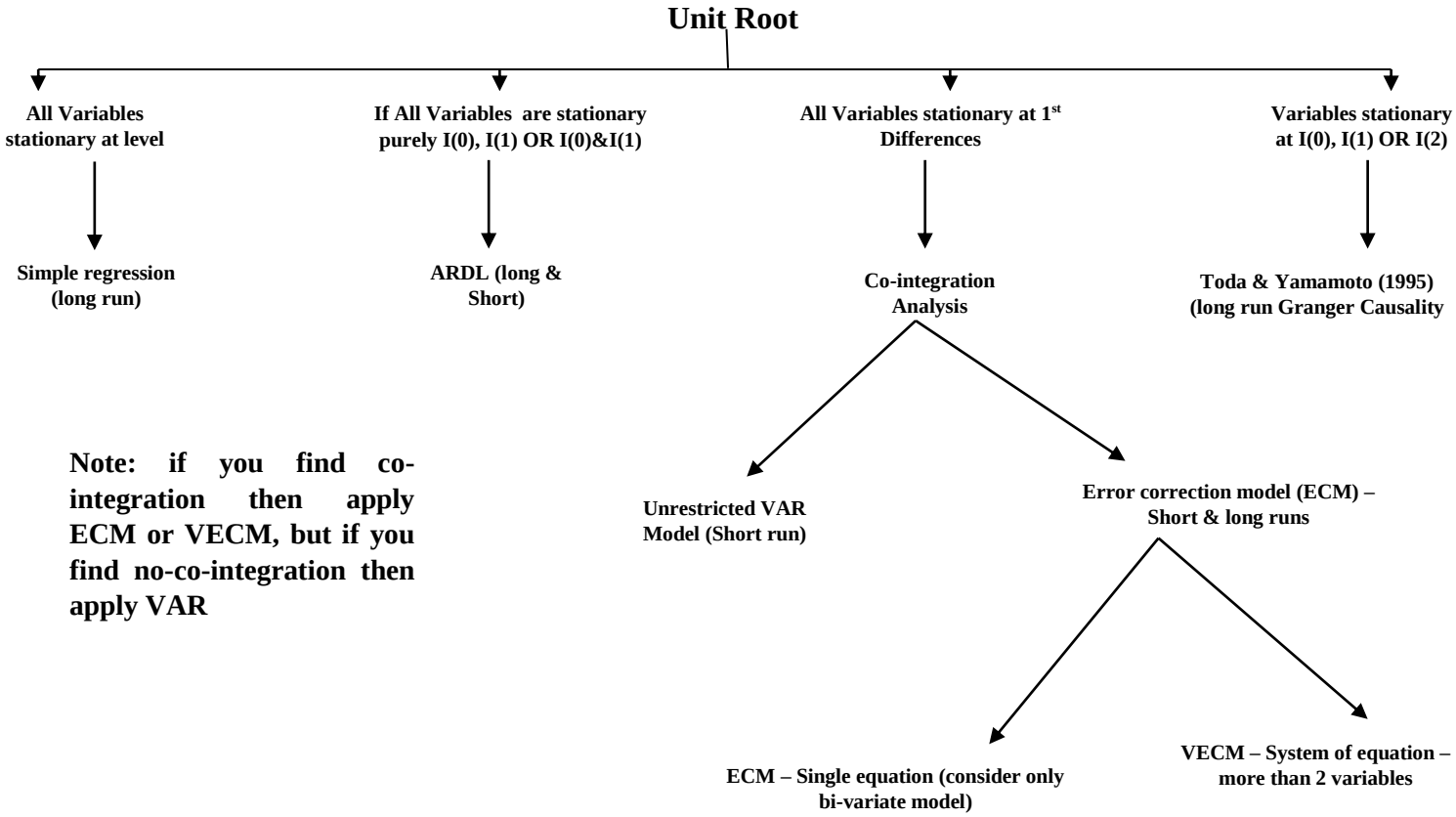
CMRC5101

I have already sent you the panel data Regression Econometrics. Now I am sending you the details regarding time series econometrics.

Research Process flow chart

Define Research Problem – Review Related Theories –Review Previous Research Findings – Determine Research Objectives and questions – Research Design – Data Collection – Analysing data by applying various statistical and econometrical tools and techniques – interpretation – Conclusion & recommendation.

Statistical Model Selection (time series) on the base of data Stationarity



Note: if you find co-integration then apply ECM or VECM, but if you find no-co-integration then apply VAR

So, time series is a collection of observations of variables collected through repeated measurements over time. A time series allows the researcher to identify presumed changes within a population over time. It can also show the impact of cyclical, seasonal and irregular events on the data item being measured. Time series can be classified into two different types: stock and flow.

A stock series is a measure of certain attributes at a point in time. A flow series is a series which is a measure of activity over a given period. An original time series shows the actual movements in the data over time and includes any movements due to trend, cyclical, seasonal and irregular events.

One approach to the analysis of time series data involves an attempt to identify the component factors that influence each of the values in a series. This identification procedure is ***called decomposition***. To understand decomposition, we start with the four components of a time series:

1. Trend: The trend is the component that represents the underlying growth (or decline) in a time series. For example, share price changes, technological changes, inflation, productivity increases etc. The trend is denoted by **capital T**.

2. Cyclical: The cyclical component is a series of wave like fluctuations or cycles of more than one year's duration. For example, changing economic conditions generally produce cycles and it is denoted by **Capital C**.

In practice, cycles are often difficult to identify and are frequently regarded as part of the trend. In this case, the underlying general growth (or decline) component is called the ***trend-cycle*** and denoted by **T**. We use the notation for the trend because the cyclical component often cannot be separated from the trend.

3. Seasonal: Seasonal fluctuations are typically found in quarterly, monthly, or weekly data. Seasonal variation refers to a more or less stable pattern of change that appears annually and repeats itself year after year. Seasonal patterns occur because of the influence of the weather or because of calendar-related events such as school vacations and national holidays and denoted by ***capital S***.

4. Irregular: The irregular component consists of unpredictable or random fluctuations. These fluctuations are the result of a myriad/numerous of events that individually may not be particularly important but whose combined effect could be large and denoted by ***capital I***.

Now the question is how the components of a time series relate to the original series?

So, this task is accomplished by specifying a model (mathematical) or equation that expresses the time series variable **Y** in terms of the components **T, C, S** and **I**.

Here, Y is the time series variable. There are two approaches/models that treats the observed value (Y_t) of a time series to the trend (T_t), seasonal (St) and irregular (I_t) components.

Additive Components Model: $\mathbf{Y_t = T_t + S_t + I_t}$ (This model works best when the time series being analyzed has roughly the same variability throughout the length of the series or in other words, all the values of the series fall essentially within a band of constant width centered on the trend)

And ***Multiplicative Components Model***: $\mathbf{Y_t = T_t * S_t * I_t}$ (This model works best when the variability of the time series increases with the level. That is the values of the series spread out as the trend increases, and the set of observations have the appearance of a megaphone or funnel. It is possible to convert a ***Multiplicative*** decomposition to an **Additive** decomposition by working with the logarithms of the data.

Take log on both sides of the ***Multiplicative Components Model*** and we get:

$$\mathbf{\log Y = \log(T * S * I) = \log T + \log S + \log I}$$

For business and economic time series, it is best to view the trend (or trend cycle) as smoothly changing overtime. Rarely can we realistically assume that the trend can be represented by some simple function such as a straight line over the whole period for which the time series is observed. However, it is often convenient to fit a trend curve to a time series for two reasons: (1) it provides some indication of the general direction of the observed series, and (2) it can be removed from the original series to get a clearer picture of the seasonality.

If the trend appears to be roughly **linear (linearity is important here)**, that is, it increases or decreases like a straight line, then it may be represented by the following equation:

$$\hat{T}_t = \alpha_0 + \beta_1 t$$

Here, $\hat{T}_t$ is the predicted value for the trend at time ***t***. the symbol t used for the independent variable represents time and ordinarily assumes integer values 1, 2, 3,…… corresponding to consecutive time periods. The slope coefficient ***beta (β_1)*** is the average increase or decrease in T for each one-period increase in time.

Time trend equations, including the straight-line trend, can be fit to the data using the method of ***least squares***.

Now consider additional Trend Curves: The life cycle of a new product has generally three stages – introduction, growth and maturity & saturation. Here, a straight-line trend would not work here. In this situation a curve other than a straight line, is needed to model the trend over a new-product life cycle. So, here such type of situation can be captured by ***modeling the quadratic trend or exponential trend curve*** as below:

$$\hat{T}_t = \alpha_0 + \beta_1 t + \beta_2 t^2$$

Actually, the trends curve models are based on the following **assumptions**:

1. Selection of the correct trend curve 2. The curve that fits the past is indicative of the future.

These assumptions suggest that judgment and expertise play a substantial role in the selection and use of a trend curve. To use a trend curve for forecasting, we must be able to argue that the correct trend has been selected and that, in all likelihood/possibility/probability/chance, the future will be like the past.

Seasonally Adjusted Data:

After the seasonal component has been isolated, it can be used to calculate ***seasonally adjusted data***. For an additive decomposition, the seasonally adjusted data are computed by substracting the seasonal component as below:

$$Y_t - S_t = T_t + I_t$$

For a multiplicative decomposition, the seasonality adjusted data are computed by dividing the original observation by the seasonal component as under:

$$\frac{Y_t}{S_t} = T_t \times I_t$$

Most economic time series published by Govt. agencies are seasonally adjusted because seasonal variation is not of primary interest. Rather, it is general pattern of economic activity, independent of the normal seasonal fluctuations, that is of interest.

Cyclical and Irregular Variations

Cycles are long-run, wave like fluctuations that occur most frequently in macro indicators of economic activity. It is assumed that cycles don't have a consistent pattern. However, some insight into the cyclical behavior of a time series can be obtained by eliminating the trend and seasonal components to give, using a multiplicative decomposition as under:

$$\frac{Y_t}{T_t \times S_t} = \frac{T_t \times C_t \times S_t \times I_t}{T_t \times S_t} = C_t \times I_t$$

A moving average can be used to smooth out the irregularities, I_t, leaving the cyclical component, C_t. to eliminate the centering problem encountered when a moving average with an even number of time periods is used, the irregularities are smoothed using a moving average with an odd number of time periods. Finally, the irregular component is estimated by

$$I_t = \frac{C_t \times I_t}{C_t}$$

The irregular component represents the variability in the time series after the other components have been removed. It is sometimes called the residual or error. With a multiplicative decomposition, both the cyclical and irregular components are expressed as indices. One reason for decomposing a time series is to isolate and examine the components of the series.

Key Formulas:

1. Time series additive decomposition: $Y_t = T_t + S_t + I_t$

2. Time series *Multiplicative Components decomposition*: $Y_t = T_t * S_t * I_t$

3. Linear trend: $\hat{T}_t = \alpha_0 + \beta_1 t$

4. Quadratic trend: $\hat{T}_t = \alpha_0 + \beta_1 t + \beta_2 t^2$

5. Exponential Trend: $T_t = \alpha_0 \beta_1^t$

6. Seasonally adjusted data (multiplicative decomposition): $\frac{Y_t}{S_t} = T_t \times I_t$

7. Cyclical-irregular component (multiplicative decomposition): $\frac{Y_t}{T_t \times S_t} = C_t \times I_t$

8. Irregular component (multiplicative decomposition): $I_t = \frac{C_t \times I_t}{C_t}$

9. Current purchasing power of $1: $\frac{100}{Consumer\ \Pr ice\ Index\ (CPI)}$

10. Deflated dollar value: (Dollar value) x (Purchasing power of $1)

Data Transformation

1. Log Transformation: The log transformation yields appealing interpretation of coefficients and model. The interpretation is good for small changes only. The log transformation makes coefficients invariant to rescaling. For example, **lnY** looks more normal than **Y**. But, **lnY** has a narrower range than **Y**.

When or not to log?

a. Don't take log zeroes or negative values

b. Don't take log dummies

c. Potentially large monetary variables are often logged (revenue, income, wages etc)

d. Large integer values are often logged (population, number of employees, number of students)

e. Small integer values are usually not (age, education, number of children)

g. Percentages can be logged or not

How to choose?

Sometimes it is unclear which form to choose, and thus we have to:

i. Rely on economic theory and previous studies

ii. Think about what is implied by particular functional forms for the relevant range of the variables

iii. Don't compare R^2 or $AdjR^2$ if the dependent variable is different.

iv. Even in the case of the same dependent variable, e.g. linear and log-linear, beware of selecting the functional form on the sole basis of R^2 and $AdjR^2$. Selecting the functional form on the basis of fit only gives you an equation that works well for your particular sample.

2. Differencing: Stationarity is an issue for time series data and is a pre-condition to perform regression analysis. Non-stationary time series data should be converted into stationary data by taking differences. The time series data which is stationary after first differencing is called stationary in order one ***I(1)***.

3. Percentage Change: It is calculated by subtracting previous one from the current one and dividing the difference by the previous one. Here, a base year is used for comparison of the two or more time series data levels. The arbitrary level of 100 is selected so that percentage changes (either increase or decrease) over year can be easily depicted. Any year can be chosen as base year, but generally recent years are chosen and the other observations are adjusted based on the base year.

Regression Analysis

Regression is an econometric technique for estimating the relationships among the variables. It helps to analyze how the typical value of the dependent variable changes when any one of the independent variables changes, while the other things are remained constant (**ceteris paribus**). Linear regression estimates how much Y changes when X changes one unit or one percent. The main purpose of linear regression analysis is to assess associations between dependent and independent variables.

In my previous classes I have discussed simple or bi-variate linear regression model and their properties. How to estimate reg. model. Significance of constant term and slope coefficient and how we will compute them and interpret. How we will forecast population regression line by using sample regression line. How we will analyze ANOVA and interpretation etc. So I am not going to that part. I will discuss here from multiple regression equation.

Multiple or Multivariate Regression Model

In simple linear regression the relationship between a single independent variable and a dependent variable in investigated. The relationship between two variables frequently allows one to accurately predict the dependent variable from knowledge of the independent variable. Unfortunately, many real-life forecasting situations are not so simple. More than one independent variable is usually necessary in order to predict a dependent variable accurately. Regression models with more than one independent variable are called ***multiple regression models***. So, multiple regressions involve the use of more than one independent variable to predict the changes of the dependent variable. Here each slope coefficient measures the rate of change in the mean value of dependent variable for a unit change in the value of an independent variable, holding the values of the other independent variables constant. Here, one important thing is that how many independent variables should be included in the regression equation depends on the theory, research objective and assumptions.

Now consider the Statistical model for Multiple Regression as under:

$$Y_i = \alpha_0 + \beta_1 X_i + \beta_2 M_2 + \beta_3 N_3 + \ldots\ldots\ldots + \beta_k \theta_i + e_i$$

Now, consider the practical example of multiple linear regression model. Suppose it is assumed that the impact of COVID-19 on economic growth. This is the problem you want to examine and thus we have considered relevant information regarding independent variable and you develop the model as under:

$$GDP = \alpha + \beta_1 attack + \beta_2 death + \beta_3 infected + \beta_4 cure + e$$

Where, alpha value is the constant and you have to estimate both alpha and slope coefficients by estimating the multiple regression equation by applying standard OLS technique.

Here, error or disturbance components denoted by e represents the deviations of the response from the true relation. They are unobservable random variables accounting for the effects of other factors on the response. Here, the errors are assumed to be independently and each is normally distributed with 0 mean and unknown standard deviation. Here, the regression coefficient together locate the regression function are unknown.

The calculation procedure of coefficients and testing are same I have discussed in developing simple regression equation in your previous classes.

Now come to, Inference for multiple regression models:

Actually, the inference for multiple regression models is analogous / similar to that for simple linear regression. The least squares estimates of the model parameters, their estimated standard errors, t statistics used to examine the significance of individual terms in the regression model, and an F statistic to check the overall significance are same like bivariate regression model.

Standard error of the estimate:

The standard error of the estimate measures the amount the actual values (Y) differ from the estimated values ($\hat{Y}$). For relatively large samples, we would expect about 67% of the differences Y - $\hat{Y}$ to be within $s_{y.x's}$ of 0 and about 95% of these differences to be within $2s_{y.x's}$ of 0.

How you will calculate standard error of the estimate below:

$$s_{y.x's} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

Where, n represents number of observations, k represents number of independent variables in the regression model.

SSE = $\sum(Y-\hat{Y})^2$ represents residual sum of squares

MSE = SSE / (n - k - 1) represents residual mean square errors

Now come to, Significance of the Multiple regression line:

The analysis of variance (ANOVA) table based on the decomposition of the total variation in Y (SST) into its (SSR) and unexplained (SSE) parts is given in a table.

ANOVA for multiple regression				
Source	**Sum of squares**	**df (degrees of freedom)**	**Mean square**	**F ratio**
Due to regression	SSR	k	MSR = SSR/k	$F = \frac{MSR}{MSE}$
Due to error	SSE	n – k - 1	MSE = SSE/(n – k – 1)	
Total	SST	n - 1		

Then you have to apply F test for the significance of the regression as under:

What is the null hypothesis here you have to formulate first as below:

$H_0: \beta_1 = \beta_2 = \ldots.= \beta_k = 0$, Ha: at least one $\beta_j \neq 0$

So, $F = \frac{MSR}{MSE}$ **with df = k, n – k – 1.** At significance level α (1% or 5% or 10%), the rejection region is $F > F_\alpha$

Where, F_α is the upper α percentage point of an F distribution with $\delta_1 = k$, $\delta_2 = n - k - 1$ degrees of freedom.

The coefficient of determination R is given as under:

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

It represents the proportion of variation in the response Y explained by the relationship of Y with the X variables.

A value of $R^2 = 1$ indicates that all the observed Y's fall exactly on the fitted regression function. All of the variation in the response is explained by the regression. A value of $R^2 = 0$ says that $\hat{Y} = \bar{Y}$, that is, SSR = 0, and none of the variation in Y is explained by the regression. In practice, $0 \le R^2 \le 1$, and the value of R^2 must be interpreted relative to the extremes, 0 and 1.

The quantity $R = \sqrt{R^2}$

Is called the multiple correlation coefficient and is the correlation between the responses Y and the fitted values $\hat{Y}$. Because the fitted values predict the responses, R is always positive so that $0 \le R \le 1$.

For multiple regression $F = \frac{R^2}{1 - R^2}\left(\frac{n - k - 1}{k}\right)$

So, everything else equal, significant regressions (large F ratios) are associated with relatively large values for R^2.

The coefficient of determination can always be increased by adding an additional independent variable X to the regression function, even if this additional variable is not important. For this reason, some analysts prefer to interpret R^2 adjusted for the number of terms in the regression function. The ***adjusted coefficient of determination***, $\bar{R}^2$, is given by

$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{n - 1}{n - k - 1}\right)$$

Like R^2, $\bar{R}^2$ is a measure of the proportion of variability in the response Y explained by the regression. It can be shown that $0 \le \bar{R}^2 \le R^2$. When the number of observations n is large relative to the number of independent variables k, $R^2 \approx \bar{R}^2$. If k = 0, $\hat{Y} = \bar{Y}$ and $\bar{R}^2 = R^2$. In many practical situations, there is not much difference between the magnitudes of $\bar{R}^2$ and R^2.

Misuse of R^2 statistic: In spite of many usefulness of the R^2 statistic, one must be cautious about its possible misuses. In particular, it is to be remembered that it is dangerous to play the game of maximizing the value of R^2. Some researchers do this by gradually increasing the number of explanatory variables in the model. However, in empirical research, quite often we come across a situation where the value of R^2 is high but very of the estimated coefficients are statistically significant and / or they have expected signs. Therefore, the researchers should be more concerned about the logical/theoretical relevance of the explanatory variables

to the dependent variable and also their statistical significance. If in this process, a high value of R^2 is obtained, well and good. On the other hand, if R^2 is low, it does not mean that the model is necessarily bad, particularly when a good number of the estimated coefficients have expected signs and are statistically significant.

Assumptions of Classical Linear Regression Model (CLRM)

i. The regression model is linear in the parameters (can be nonlinear or linear in the dependent and independent variables). The regression model should correct variables without omitted variable bias and have correct functional form.

ii. The regressors are assumed to be fixed or non-stochastic in repeated sampling. This assumption may not be appropriate for all economic data. If independent variable and error term are uncorrelated, the classical results hold true asymptotically (i.e. in large samples).

iii. Given the values of the x (independent) variables, the expected, or mean, value of the error term is 0. $E(e_i / x) = 0$. The conditional expectation of the error term, given the values of the independent variables, is zero. Since the error term represents the influence of other factors that are not included in the regression equation and may be essentially random, it is very logical to assume that their mean or average value is 0.

iv. The variance of each e_i, given the values of x, is constant, or ***homoscedastic.*** ***$Var(e_i / x) = \sigma^2$***

v. There is no correlation between two error terms meaning that absence of ***autocorrelation*** problem. ***$Cov(e_i, e_j / x) = 0$.*** **If** there is autocorrelation, an increase in the error term in one period affect the error term in the next.

vi. There are no perfect linear relationships among the independent variables. This is the assumption of no perfect ***multicollinearity***.

vii. Error term is not correlated with the independent variables. ***$E(e_i / x_{1i}, x2i, \ldots\ldots x_{ki}) = 0$.*** **They** are independently identically distributed **(i.i.d.).**

vii. The regression model should be correctly specified. Alternatively, there is no specification bias or specification error in the model used in empirical analysis. It is implicitly assumed that the number of observations, N, is greater than the number of parameters (K) estimated.

viii. All independent variables should be exogenous which are defined outside of the model.

ix. Although it is not a part of the CLRM, it is assumed that the error term follows the normal distribution with 0 mean and constant variance. It is necessary for the hypothesis testing. **Symbolically, *$e_i \sim N(0, \sigma^2)$***.

Statistical Properties of OLS

i. ***Linearity:*** The estimators are linear, that is, they are linear functions of the dependent variable, y_i. Linearity assumption can be expressed as follows:

$$E(y_i / x_i) = \beta_1 + \beta_{2i} x_{2i} + \beta_{3i} x_{3i} + \ldots\ldots + \beta_{ki} x_{ki}$$

Linear models can be expressed in a form that is linear in the parameters by a transformation of the variables. Nonlinear models on the other hand, cannot be transformed to the linear form. The non-linearity of interest here is the one which cannot be accommodated into linear conditional mean after transformation.

Now the question how you will test linearity assumption?

Reset-type test (Ramsey, 1969) is the most common test for testing the linearity assumption and this testing procedure involves the estimation of the following (auxiliary) regression by taking into consideration the errors:

$$\hat{e}_i = \omega_1 + \omega_{2i}x_{2i} + \ldots\ldots\ldots + \omega_{ki}x_{ki} + \vartheta\hat{y}_i^2 + \varepsilon_i$$

Hypothesis testing:

H_0: $\vartheta = 0$

H_a: $\vartheta \neq 0$

We can now test the statistical significance of ϑ by using ***t-test, F-test or LM*** test statistics.

When linearity does not hold, the OLS estimators are biased and inconsistent or in other words, estimation and testing results are invalid or incorrect and you have to re-structured the estimated model again.

ii. Un-biasedness: The estimators ($\hat{\beta}_i$) are unbiased, that is, in repeated applications of the method, on average, the estimators approach their true values.

$$E(\hat{\beta}_i) = \beta_i$$

iii. Efficiency: In the class of linear estimators, OLS estimators have minimum variance. As a result, the true parameter values can be estimated with least possible uncertainty; an unbiased estimator with the least variance is called an efficient estimator.

iv. Normality: The assumption of normality can be expressed as follows:

$$e_i \sim N(0, \sigma^2), \text{ or } (y_i / x_i) \sim N(\beta x_1, \sigma^2)$$

If the assumption of normality does not hold, then the OLS estimator ($\hat{\beta}$) remains the ***Best Linear Unbiased Estimator*** (BLUE), i.e. it has the minimum variance among all linear unbiased estimators. However, without normality one cannot use the standard for the ***t and F*** distributions to perform statistical tests.

The following null hypothesis should be specified before normality test.

The null hypothesis is that the **skewness** (a measure for the degree of symmetry in the variable distribution) and **kurtosis** (a measure for the degree of peakedness / flatness in the variable distribution) coefficients of the conditional distribution of y_i (or, equivalently, of the distribution of e_i) are ***0 and 3***, respectively.

H_0: ***Skewness = 0, (if skewness < 0 then $f(y_i/x_i)$ is skewed to the left side***

H_0: *kurtosis = 0, (if kurtosis > 3 then $f(y_i/x_i$ is leptokurtic)*

Now, the above assumptions can be tested jointly by using the **Jarque-Bera test (JB, 1981)** which follows asymptotically a ***chi-square*** distribution and it is a popular test for ***normality*** as under:

What is the hypothesis for this test?

H_0: Distribution is normal

H_a: Distribution is not normal

$$J - B = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right]$$

Where, n is the number of observation, S is the skewness and K is the kurtosis. In general, when the values of S and K are 0 and 3 respectively then it may be said that the distribution is normal.

Decision rule: If the probability value of the computed J-B statistics is less than the chosen significance level (1% or 5% or 10%) then reject H_0 **(not normal)** and vice-versa.

For example, J-B statistic = 232.3196 and **Probability = 0.00000** it indicates that the probability value is less than 1% or 5% or 10% level so reject null hypothesis meaning that not normal.

Another case, J-B statistic = 2.3196 and **Probability = 0.123546** (0.123546 * 100 = **12.35%** approx) it indicates that the probability value (**12.35%**) is higher than 1% or 5% or 10% level so accept null hypothesis meaning that normally distributed.

Heteroscedasticity Problem

The variance of the error term is constant (homoscedasticity) over the sample period. If the error term doesn't have constant variance, they are said to be heteroscedasticy. So, errors may increase as the value of an independent variable increases.

For example, annual family expenditures for education differ among rich and poor. A research may include two income group of families may face heteroscedasticity problem. Measurement error can be also occurred if some of respondent gives more accurate responses.

Why this problem occurs?

i. There are sub-population differences or other interaction effects. For instance, the effect of education in employment differs for villagers and urban parts.

ii. There are model mis-specifications. For instance, instead of using y, using log of y or instead of using x, using x^2.

iii. There are omitted variables. Omitting the important variables from the model may cause bias. In the correctly specified model, the patterns of heteroscedasticity are expected disappear.

How we will detect heteroscedasticity?

There are many ways to detect such problem.

1. Graphical representation: Plotting the residuals against fitted values or plot the independent variables suspected to be correlated with the variance of the error term.

2. White Test (1980): It is a special case of Breusch-Pagan test. It involves an auxiliary regression of squared residuals, but excludes any higher order terms. It is very general. If the number of the observation is small, power of the White test becomes weak. It can be performed by obtaining least squares residuals and modeling the square residuals as a multiple regression which includes independent variables and their squares and second degree products (interaction term) as under:

$$e_i^2 = \lambda_0 + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \lambda_3 (x_{i1})^2 + \lambda_4 (x_{i2})^2 + \lambda_5 (x_{i1} x_{i2}) + \varepsilon_i$$

White test for heteroscedasticity is performed through the null hypothesis of no heteroscedasticity as under:

$$\mathbf{H_0: \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = 0}$$

$$\mathbf{H_a: \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 \neq 0}$$

Decision rule: If the computed value of probability of the chi-square distribution for observed R^2 is higher than the chosen significance level say 1% or 5% then accept null hypothesis meaning that absence of heteroskedasticity or vice-versa.

3. Breusch-Pagan Test (1980): It is a Lagrange multiplier test for heteroscedasticity and designed to detect any linear form of heteroscedasticity and the test hypothesis is similar to White test.

4. Ramsey's Test (1969): RESET stands for regression specification error test. It is a general test for the following types of specification errors:

a. Omitted variables bias (if set of independent variables do not include all relevant variables)

b. Selection of wrong functional form

c. Correlation between independent variable and error term

d. The existence of lagged dependent variable values

e. Serially correlated error terms

Ramsey says that any or all of these specification errors above produce a non-zero mean vector.

Hypothesis:

H_0: $e \sim N(0, \sigma^2 I)$

Ha: $e \sim N(0, \sigma^2 I)$ $\mu \neq 0$

The test is based on following augmented regression equation:

$$y = \beta_x + \lambda_Z + e$$

The main question in constructing the RESET test is to identify the variables which should be in the **Z matrix**. The Z matrix can include the variables which are not in the original specification, so that the test of $\lambda = 0$ is called as the omitted variables test.

5. ARCH Test (Engel, 1982): It is common for financial variables. He detects that large and small forecast errors tend to occur in clusters so that the conditional variance of error term is the autoregressive function of the past errors. Ignoring ARCH effect can result inefficiency of the estimation. **ARCH(q)** effect can be written as under:

$$\sigma_t^2 = \lambda_0 + \lambda_1 e_{t-1}^2 + \lambda_2 e_{t-2}^2 + ... + \lambda_q e_{t-q}^2 + \varepsilon_t$$

This is a test for ARCH(q) vs ARCH(0).

6. Goldfeld-Quant (GQ) Test (1965): The GQ test can be used when it is assumed that the variance of the error term increases consistently or decreases consistently as x increases. This test is commonly used because it is easy to apply when one of the regressors is considered to possess proportionality factor of heteroscedasticity. **The test has two limits**: its difficult to reject the null hypothesis of homoscedasticity and the fact that it do not allow to verify other forms of heteroscedasticity.

Decision rule: If the computed value of the **F** statistics is **greater** than the critical or table value at the chosen significance level, the null hypothesis of homoscedasticity is rejected or vice-versa.

7. Part Test: This test procedure involves three steps:

a. Modeling OLS estimation to derive the OLS residuals.

b. The derivation of the $ln(e^2_i)$ which are considered as dependent variable in the regression and looks like $ln(e^2_i) = \alpha_0 + \alpha_1 lnz_i + \mu_i$

c. The estimation results of the model are used to verify the presence of heteroscedasticity.

8. Glesjer LM test (1969): This test is similar to the Breusch-Pagan-Godfrey test. This test tests against an alternative hypothesis of heteroscedasticity. The auxiliary regression that Glejser proposes regresses the absolute value of the residuals from the original equation. As with the previous tests, this statistic is distributed from a **chi-square distribution** with degrees of freedom equal to the number of variables.

9. Harvey-Godfrey LM Tets (1976): This test is also similar to the Breusch-Pagan-Godfrey test. He tests a null hypothesis of no heteroscedasticity against heteroscedasticity of the form of $\sigma_t^2 = \exp(z_t^{'}\alpha)$ where, again z_t is a vector of independent variables.

Now the question is which test you should apply among the various tests?

Really, it is a very difficult task for the researchers. So, to choose the wrong test may not detect presence of heteroscedasticity. The **most popular** test to check heteroscedasticity is the **White Test** and this test is largely used by the researchers. Although, it has limited power against large number of alternatives. Alternatively, **visual inspections/ graphical presentation** of residuals I have already stated t the beginning can be the best method to detect heteroscedasticity correctly.

What are the ways to correct Heteroscedasticity problem?

i. A common way to alleviate this problem is to use logarithms of all variables rather than their level forms. So, our first step in handling the heteroscedasticity problem is to consider a **log linear model**.

2. Sometimes heteroscedasticity results from improper model specification. There may be sub group differences. Effects of variables may not be linear. Perhaps some important variables have been left out of the model. If these represent a problem then deal with them first.

3. Using Weighted Least squares **(WLS)** is more difficult option but superior when you can make it work right. Generalized least squares **(GLS)** is a technique that will always yield estimators that are BLUE **(Best Linear Unbiased Estimators)** when either heteroscedasticity or serial correlation is present.

Autocorrelation

Autocorrelation can only occur in the models that include time series data and it means that either the model is specified with an insufficient number of lagged variables or not all the relevant explanatory variables are specified in the model.

Here, the error term catch the influence of the not included variables affecting dependent variable. Persistence effect of the excluded variables causes positive autocorrelation. If those excluded variables are observable and includable in the model, autocorrelation test result is an indication of a mis-specification model. So it is also called mis-specification test. Incorrect functional forms, omitted variables and an inadequate dynamic specification of the model can cause autocorrelation.

What are the consequences of Autocorrelation in the residuals?

i. The standard errors are underestimated, so t-values are overestimated.

ii. High values for the **t-statistics and R^2** are observed in the estimation output. It means that the result is false if the output is not correctly interpreted.

iii. OLS estimates remain unbiased, but it becomes inefficient.

How you will detect autocorrelation?

There are various tests of autocorrelation for detection. All the tests have same null hypothesis of absence of autocorrelation in the disturbance term.

H_0: Absence of autocorrelation in the disturbance term

H_a: Presence of autocorrelation

The existence of autocorrelation may be an indication of mis-specification. A possible way to eliminate autocorrelation problem is to change model specification.

How autocorrelation problem can be detected?

There are so many ways to detect autocorrelation problem.

1. Graphical method: Just plotting the error term in graph to detect autocorrelation.

2. Breusch-Godfrey LM Test (1978): The idea behind this test is as follows:

First you have to estimate the linear model with OLS.

$$Y_t = \alpha_1 + \beta_1 X_t + \beta_2 Z_t + \mu_t$$

Next, residuals are computed and following regression equation is estimated again with OLS as under:

$$\mu_t = \varphi_1 \mu_{t-1} + ... + \varphi_p \mu_{t-p} + \alpha_1^* + \beta_1^* x_t + \beta_2^* z_t + \varepsilon_t$$

What is the null hypothesis here, H_0: $\varphi_1 = \varphi_2 = ... = \varphi_p = 0$ (No autocorrelation)

Actually, the BG test is called serial correlation LM test.

Decision rule: If the computed probability value of the LM statistic is higher than 1% or 5% level then accept null hypothesis meaning that there is no autocorrelation problem in the disturbance or vice-versa.

3. Box-Pierce & Ljung-Box Tets (Q test, 1970): These two tests have asymptotic χ^2 (chi-square) distribution, with **p degrees of freedom** under the null hypothesis of no autocorrelation. This test uses autocorrelation of the residuals. The estimated autocorrelation coefficients are defines as $\hat{\rho}_i$.

It can be computed as under:

$$\hat{\rho}_i = \frac{\text{cov}(\mu_t, \mu_{t-i})}{\sqrt{\text{var}(\mu_t)}\sqrt{\text{var}(\mu_{t-i})}}$$

The theoretical autocorrelation coefficients $\hat{\rho}_i$ **are** 0 under the null hypothesis. **The Q-test** does not look at the individual autocorrelation coefficients. It considers the sum of a number of squared autocorrelation coefficients as follows:

$$Q = n \sum_{i=1}^{p} \hat{\rho}_i^2$$

But this test has low power of detecting autocorrelation. The main difference between Q-test and BG test is that one specific order of the autoregressive process specified should be chosen under the alternative hypothesis.

4. Durbin Whatson test (1950): This is the **widely used** autocorrelation test. This test is used for first order autocorrelation. **DW** test assumes that error term is stationary and normally distributed with zero mean. It tests the null hypothesis H_0 that the errors are uncorrelated.

H_0: Absence of autocorrelation
H_a: Presence of autocorrelation

This test can be used if the explanatory variables are exogenous and a constant term has been included in the model. ***DW-statistics*** is not used, if lagged dependent variables are present as explanatory variables in the model. It can be employed if the explanatory variables are

exogenous and the model includes intercept. DW-statistics should be used if all the conditions are satisfied. Otherwise it is more informative to use **Breusch-Godfrey** test in your research paper.

$$DW = \frac{\sum(\mu_t - \mu_{t-1})^2}{\sum \mu_t^2} = 2(1-\varphi_1)$$

What are the Properties of the DW-statistic?

i. $\varphi_1 = 0$, $DW \approx 2$ (No residual autocorrelation)

ii. $\varphi_1 > 0$, $DW < 2$ (Positive autocorrelation)

iii. $\varphi_1 < 0$, $DW > 2$ (Negative residual autocorrelation)

Here, one important thing is that DW-statistic cannot be tabulated. But it is possible to derive distributions of a ***lower (d_L) and an upper (d_U) bound.*** These two distributions depend on the number of the observations ***(n)*** and the number of the explanatory variables ***(K)***. Therefore, DW-statistic is tabulated as follows:

If $DW \geq d_U$: don't reject H_0.

If $d_L < DW < d_U$: the test is inconclusive

If $DW \leq d_L$: reject H_0 for the favor of first order residual autocorrelation

If there is autocorrelation, then OLS is no longer BLUE, and then EGLS (Estimated Generalized Least Squares)or FGLS (Feasible Generalized Least Squares) proposed by Cochrane-Orcutt can be used.

But DW test for autocorrelation has some limitations such as (i) The form of model (explanatory variables) should be known and (ii) The test result is sometimes inconclusive.

Now what are the ways to correct autocorrelation problem in your time series research?

Actually, the autocorrelation is an indicator of model mis-specification. If the model suffers from mis-specification then change the model not the estimation method **(from OLS to EGLS or FGLS**). There are three types of model mis-specification such as:

1. Omitted variable bias. So excluding relevant independent variables can cause autocorrelation. Find out the relevant omitted variable and include in the model.

2. Functional form; to use log transformation can eliminate autocorrelation problem.

3. Dynamic mis-specification; whether the model is static or dynamic should be decided. Inclusion of the lagged endogenous and exogenous variables eliminates the autocorrelation problem.

MULTICOLINEARITY

It is a data problem. Collinearity between variable is always present. It becomes a problem and violation of the classical assumptions if the correlations among the independent variables are very strong. It can affect accuracy of the parameter estimates. Multicollinearity misleadingly increases the standard errors. Thus, it makes some variables statistically insignificant while they should be otherwise significant. It is like two or more people singing loudly at the same time. ***One cannot discern/separate/discriminate who are singing how***. The influence of the independencies offset/compensate each other. In multicollinearity, the parameter estimates become inaccurate.

If someone uses too many ***dummy variables*** in the estimated model that cause's multicollinearity problem.

What are the causes of multicollineairty?

1. Improper use of dummy variables in the model.
2. Including a variable computed from other variables in the equation (e.g. family income = husband's income + wife's income, and the regression includes all three income measures)
3. Including the same or almost the same variable twice (height in feet and height in inches; or, more commonly, two different operationalization of the same identical concept)

How the researcher detect multicollinearity problem?

1. Variance Inflation Factor (VIF) is used to measure how much the variance of the estimated coefficients is increased over the case of no correlation among the independent variables.

- If VIF = 0, No multicollinearity
- If VIF ≥ 0, there is multicollineairity

2. Tolerance (TOL): It is the reciprocal of VIF. If the TOL value is closer to 0, the greater is the degree of collinearity of one independent variable with the others independent variables. If the value of TOL is closer to 1, the greater the evidence of no collineairity among the independent variable. But the ***rule of thumb*** is that ***TOL value of 0.10 or less*** indicates presence of severe multicollinearity.

3. Correlation Matrix: Here, you just calculate simple correlation coefficient between the independent variables and then determine the significance of the computed correlation coefficients so as to conclude about the presence of multicollineairity. But ***rule of thumb*** is that if the correlation coefficients between the independent variables are ***less than 0.90*** then absence of multicollenearity.

4. Condition Number (CN): The condition number provides an overall measure of multicollinearity. It conveys the status or condition of the data matrix X where data on explanatory variables are put in.

The rules of thumb:

i. if CN = 1, no multicollineairity. So closer the value of CN to unity, the better is the condition of the data matrix X.

ii. When $1 < CN < 10$, multicollinearity is negligible

iii. $10 \leq CN \leq 30$ represents the situation of moderate to strong multicollinearity

iv. if $CN > 30$, there is severe multicollinearity.

5. The researchers sometimes apply other techniques such as Principal component method, ridge regression etc., to solve the problem of multicollinearity. Under, PCM method, the colliner variables are grouped to form a composite index capable of representing this group of variables by itself. The ridge regression method follows an entirely different approach to the problem of multicollinearity. In ridge regression, a constant is added to the variance of each explanatory variable.

Which measure should we use to check multicollineairy?

It is very difficult to say that which measure is superior as compare to others. But, most of the empirical studies based on time series data have used correlation matrix or VIF as a standard measure. So the researchers are advised to use ***VIF or Correlation matrix or both*** to check this problem for their research.

What are the remedial measures to eliminate such problem?

1. Increasing sample size helps to reduce some but not all problems associated with mulicollinearity. Larger sample size will always improve the precision or reliability of the OLS estimates. Hence, attempts should be taken to work with larger sample.

2. It has been found that the intensity of multicollinearity gets reduced when transformed variables (ratio, first difference etc) are used instead of variables in level form.

3. If an extraneous estimate of the coefficient of one of the variables responsible for creating multicollinearity is available, it can be used and a mixed estimation method followed to correct the high-variance problem created by the multicollinearity. However, while applying this method, extra care must be taken to ensure that the extraneous estimate is relevant.

4. A popular method of avoiding this problem is by simply dropping one or more of the collinear variables. Usually, this method becomes effective when a large number of explanatory variables have been included in the model of which all are not important.

STATIONARITY/UNIT ROOT PROBLEM

The time series variables included in regression models need to be stationary because if their means and variances are changing, the computed t-statistics under the OLS regression fail to converge to their true value as sample size increases. Although the variables have strong association between them although in reality there might not be any such association between the variables. This is known as the problem of ***Spurious regression***.

Now start with the stationary stochastic process, a stochastic process (time series) $\boldsymbol{Y_t}$ is said to be stationary if its mean and variance are constant and independent of time and the covariances depend only upon the disturbance between two time periods, but not on time periods per se. So, $\mathbf{Y_t}$ is stationary when the following conditions hold:

i. $E(Y_t) = \mu$ = constant for all t

ii. $Var(Y_t) = \sigma^2$ = Constant for all t

iii. $Cov(Y_t, Y_{t-s}) = \lambda_s$ = Constant for all $t \neq s$

These conditions imply that the mean and variance of the stationary series remain constant over time. For example, if we consider monthly observations from 2010 to 2020 then the above conditions remain same during this period.

How to check Stationarity problem in your data set?

There are many ways:

1. Graphical Approach: In general, stationarity of a series can be understood simply by plotting the series over time. If the series shows no tendency to drift upwards over time, it is stationary in mean. Otherwise, it is non-stationary.

2. Autocorrelation function (ACF) and Correlogram: The stationarity of a given time series may be assessed by computing the value of its autocorrelation function (**ρ**). For the series Y_t, the value of **ρ at lag k,** denoted by $\boldsymbol{\rho_k}$**, is computed as under:**

$$\rho_k = \frac{\lambda_k}{\lambda_0} = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)}$$

Here, we may compute the value of $\boldsymbol{\rho_k}$ for different lag lengths (k). A graphical plot of $\boldsymbol{\rho_k}$ **against k is called correlogram.** The correlogram helps to understand whether the series is stationary or not. If the value of $\boldsymbol{\rho_k}$ at various hovers/stay close around 0, we say that the series is stationary. For a non-stationary series, the value of $\boldsymbol{\rho_k}$ **at various lags** are non-zero, although they may slowly decline towards 0 as the lag length increases.

3. Bartlett Test: Here, it is shown that if the series is purely random, $\boldsymbol{\rho_k}$ follows a normal distribution with mean 0 and variance **1/T where T** indicates number of observations in the series. Then 95% confidence interval for $\boldsymbol{\rho_k}$ is given by

$$\hat{\rho}_k \pm 1.96 SE(\hat{\rho}_k) = \hat{\rho}_k \pm 1.96\sqrt{\frac{1}{T}}$$

Decision rule: When ρ_k falls outside this interval, we reject the null hypothesis that $\rho_k = 0$ and conclude that Y_t series is **non-stationary**. On the other hand, if ρ_k falls within this interval, we accept the null hypothesis that means Y_t series is **stationary**.

4. Box-Pierce Test: This test is applied to examine the validity of the null hypothesis that all ρ_k's are simultaneously/at same time/concurrently 0. The Box-Pierce **Q-statistic** may be computed as under:

$$Q^*_{BP} = T\sum_{k=1}^{m} \hat{\rho}_k^2$$

Where, T means number of observations and m means maximum lag length.

Here, Q statistic follows chi-square distribution with **m** degrees of freedom. So when computed $Q^*_{BP} = \chi^{2*} > \chi^2$ at the chosen significance level and given degrees of freedom, reject null hypothesis that all ρ_k are simultaneously 0, and conclude series is non-stationary.

5. Ljung-Box Test: Actually, this test is a modified version of the Box-Pierce test. The modified Ljung-Box Q-statistic is computed as under:

$$Q^*_{LB} = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{T-k}$$

Here, Q^*_{LB} also follows Chi-square distribution with m degrees of freedom and the decision rule is same as Box-Pierce test.

6. Unit Root test: A more formal test of stationarity that has become widely popular is the unit root test. To understand this test you have to consider first an autoregressive (AR) function/model:

$$\mathbf{Y_t = \rho Y_{t-1} + \mu_t}$$

Where μ_t is the white noise error term such that

$$E(\mu_t) = 0 \text{ for all t}$$

$$E(\mu_t^2) = \sigma^2 \text{ for all t}$$

$$E(\mu_t\, \mu_s) = 0 \text{ but } t \neq s$$

We know that when $\rho = 1$, we face a non-stationary situation and conclude that Y_t has a unit root. This also implies that Y_t represents a random walk series. Therefore, one way to test if Y_t is non-stationary is to regress it on its one period lagged value, i.e., Y_{t-1}, and find out if $\hat{\rho}$ is statistically significantly equal to 1. If it is so, we conclude that Y_t is non-stationary.

Alternatively, we may rewrite the above equation as under:

$$Y_t - Y_{t-1} = (\rho - 1)Y_{t-1} + \mu_t \text{ or } \Delta Y_t = \delta Y_{t-1} + \mu_t$$

Where, $\delta = (\rho - 1)$

In this situation we cannot apply conventional t-test here for the validity of the null hypothesis (H_0): $\delta = 0$. If $\delta = 0$, then $\rho = 1$, which means that Y_t has a unit root and it is non-stationary.

To solve this problem we can apply Dickey-Fuller (DF) or Augmented Dickey-Fuller (ADF) test.

7. Dickey-Fuller test (1979): In practice D-F test is applied in three forms:

$\Delta Y_t = \delta Y_{t-1} + \mu_t$ --- (when the time series has stochastic trend but no drift)

$\Delta Y_t = \alpha + \delta Y_{t-1} + \mu_t$ ----- (when time series has both stochastic trend and drift, this model fits into the financial time series such as interest rates and exchange rates)

$\Delta Y_t = \alpha + \beta_t + \delta Y_{t-1} + \mu_t$ ----- (when the series has everything like drift, deterministic trend and stochastic trend and this model fits into trending time series like asset prices or the levels of macroeconomic aggregates like real GDP etc.)

Where, α is a constant (drift) and t is a time trend. For these entire models the null hypothesis is:

H_0: $\delta = 0$ (Presence of unit-root or non-stationary)

8. Augmented Dickey-Fuller (ADF) test: The ADF test is same as D-F test, except the D-F regression equations are augmented by including **m lags** of the dependent variable (ΔY_t) to correct serial correlation problem in the disturbance term. Here, the null hypothesis **(H_0)** is same like D-F test. Here also three models are considered as before:

$$\Delta Y_t = \delta Y_{t-1} + \sum_{i=1}^{m} \lambda_i \Delta Y_{t-i} + \mu_t$$

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \sum_{i=1}^{m} \lambda_i \Delta Y_{t-i} + \mu_t \quad \textbf{My suggestion}$$

$$\Delta Y_t = \alpha + \beta_t + \delta Y_{t-1} + \sum_{i=1}^{m} \lambda_i \Delta Y_{t-i} + \mu_t \quad \textbf{Doldado, Jenkinson and Sosvilla-Rivero}$$

Now the question here is that which model we should consider. According to **Doldado, Jenkinson and Sosvilla-Rivero** they suggest that start with the last estimation model where everything is included. But **my suggestion** is that if your time series data don't follow deterministic trend then you should start with the second model where drift and stochastic trend is included.

Now what is our decision criteria: If the absolute test statistic (ADF test statistic follows ***t*** distribution) is more than the critical value (absolute) then we reject the null hypothesis and accept alternative hypothesis.

H_0: Time series has unit root or non-stationary

Ha: Time series has no unit root or stationary

For example: Computed test statistics = 2.0671 and critical value at 5% level is 3.020686 Here, test statistic is less than the critical value (2.0671 < 3.020686), so we cannot reject null hypothesis meaning that time Y_t has unit root or non-stationary.

Or

You can check it by using the probability value. If the computed probability value corresponding to the test statistics is higher than the chosen significance level say 5% then reject H_0 and vice-versa.

9. Phillips-Perron (1988) test (P-P): The D-F/ADF tests are based on the assumptions that the error term is serially independent and has a constant variance. Thus, while using these tests, we have to ensure that these assumptions are valid. Phillips & Perron develop a generalization of the ADF test procedure that allows for less restrictive assumptions concerning the distribution of error terms. The test regression here is also the AR(1) process:

$$\Delta Y_t = \delta Y_{t-1} + \mu_t$$

While the ADF test corrects for presence of serial correlation by adding lagged differenced terms on the right hand side of the above model, the P-P test makes a correction to the computed τ-statistic of the estimated coefficient of δ to account for serial correlation in μ_t. so the P-P τ-statistic is just a modification of the ADF τ-statistic that makes into account the less restrictive nature of the error process.

Decision rule: The decision rule is same like D-F/ADF test.

Types of Models

1. Visual Models: It involves graphical representation of different economic issues. It consists of graphs with lines and curves which provides information and ideas in brief about an issue of interest. For example, **Supply and demand model**.

2. Mathematical Model: These are systems of simultaneous equations with an equal or greater number of economic variables. They consist of more complex variables and relationships. For example, **to examine the demand elasticity for luxury cars in low income countries**.

3. Empirical Model: Empirical models are one type of mathematical models designed to be used along with data. Example, **to investigate the changes in income when investment changes one percent**.

4. Static model: These models provide information about what happens over time. The model estimate generally starts with predefined equilibrium condition, and then a **shock** to the model is given. At the end, the new equilibrium is obtained without and exposition of what happened in the transition from first equilibrium to the second. ***To investigate the impact of class size on the average test score of the students***.

5. Dynamic model: Dynamic models directly incorporate time into the model. This is usually done in economic modeling by using differential equations. **To examine the role of interest rate of inflation rate movements over time**.

6. Econometric models: This model should be structured based on economic theory, experience or critical thinking. It aims to explore relationship between economic variables and interpret the results obtained through statistical techniques.

7. Structural Equation Modeling (SEM): SEMs are the equations specific for the economic theory. There are different types of structural equations such as:

i. Behavioral equation, consumption equation $Y = C + I + X - M$

ii. Technical relationships, production function $Q = f(L, K)$

iii. Identifies, Keynesian macro model $C_t = \beta_1 + \beta_2 Y_t$; $Y_t = C_t + I_t$

8. Stochastic models: Stochastic modeling is a technique of predicting outcomes and takes into account a certain degree of randomness, or unpredictability. Economic relationship is not an exact relationship; a disturbance or error term should be added to measure the impact of other variables which are not included in the model.

9. Deterministic models: Deterministic model is a mathematical model in which outcomes are determined through known relationships among variables and events without variation. In deterministic models, given input always produces the same output, such as in a known chemical reaction. In comparison, stochastic models use ranges of values for variables in the form of probability distributions.

CHOOSING A FUNCTIONAL FORM

Before running any regression equation for research first of all specify the CLRM, a specific functional form should be chosen. Any functional form that is linear in parameters can be chosen. If incorrect functional is chosen, then the model is mis-specified. If the model is mis-

specified then it may not be a reasonable approximation of the true data generation process. We make a functional form specification error when we choose the wrong functional form.

Constant term: This term should be included in the regression model unless is some strong reason for opposite such as the data is in the close neighborhood. Not including a constant term causes inflated/overstated t-ratio.

Functional forms:

i. The log-log Regression model (Double log):

$$\mathbf{\ln y_i = \alpha_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \ldots\ldots\ldots\ldots\ldots + \beta_k x_{ki} + e_i}$$

Consider the following exponential regression model: $y_i = a x_i^{\beta 1} \varepsilon^{ei}$

It can be expressed in logs: $\mathbf{\ln y_i = \alpha_0 + \beta_1 \ln x_i + e_i}$

It is called linear in logs and can be estimated by OLS on the condition that classical assumptions are satisfied.

ii. Cobb-Douglas production function:

$$Q = \alpha_0 L^{\beta 1} K^{\beta 2} \Rightarrow \log Q = \log \alpha_0 + \beta_1 \log L + \beta_2 \log K$$

When L changes 1%, Q changes by β_1%

iii. Lin-Log model (semi log): $y_i = \alpha_0 + \beta_1 \ln x_{1i} + \ldots\ldots + e_i$

When x_1 changes 1%, y changes 0.01 * β_1, holding other variables constant. The impact of a variation in x_i on y decreases as x_i gets larger.

iv. Log-in Model: $\ln y_i = \alpha_0 + \beta_1 \ln x_{1i} + \ldots\ldots + e_i$

When x_1 changes one unit Y_i changes 100 * β_1% holding the other regressors constant. The impact of a variation in x_i on y_i increases with y_i.

v. Quadratic form: $y_i = \alpha_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + e_i, \qquad i = 1,\ldots,n$

Cost curve is U-shaped and cost is quadratic in output with $\beta_1 < 0$ and $\beta_2 > 0$.

$\boldsymbol{y_i}$ increase with $\boldsymbol{x_{1i}}$ but decrease with x_{1i}^2

vi. Inverse form: $y_i = \alpha_0 + \beta_1 \dfrac{1}{x_{1i}} + \beta_2 x_{2i} + e_i$

The slope approaches to zero when $\boldsymbol{x_{1i}}$ is large.

vii. Slope dummy independent variable:

$$y_i = \alpha_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + e_i$$

There are two equations as follows:

$$y_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_i + e_i, \qquad \text{if } D_i = 1$$

$$\beta_0 + \beta_1 x_i + e_{i,} \quad \text{if } D_i = 0$$

Here, each equation should be estimated separately. Interaction term should be included if there is reason to believe that the slopes are different across categories.

viii. Lags: $y_t = \alpha_0 + \beta_1 y_{t-1} + \beta_2 x_t + e_i$

The length of time between cause and effect is called a lag. y_{t-1} is lagged independent variable. β_1 measure the impact of previous observation on the current observation. If lag structure take place over more than one time period, it is called distributed lags.

ix. Mixed functional forms: It is possible to form mix functional forms as follows:

$$y_i = \alpha_0 + \beta_1 \ln x_i + \beta_2 z_i + \beta_3 \varphi_i^2 + e_i$$

In this case y_i are a semi-log function of x_i, a quadratic function of φ_i, and a linear function of z_i. The marginal effect and elasticity for each of these variables is given by the formulas above.

SUGGESTED RESEARCH TOPIC FOR THE SCHOLARS

A. Saving, Investment and Economic Growth:

i. An analysis of the interaction among savings, investments and growth in India.

ii. Are saving and investment co-integrated: Evidence from India?

iii. Does saving really matter for growth: Evidence from any country?

iv. Do foreign direct investment and gross domestic investment promote economic growth?

v. Impact of COVID-19 on economic growth in India?

B. Trade and Economic Development:

i. How trade and foreign investment affect the growth: A case of India

ii. Trade, FDI and economic growth in India

iii. An empirical investigation of the causal relationship between openness and economic growth in India.

iv. Foreign trade and economic development in India: a Granger causality analysis

v. Is the import-led hypothesis valid for India: An empirical analysis.

vi. The dynamic relationship between the GDP, imports and domestic production of crude oil: Evidence from india

C. Stock market and economic development:

i. Financial development and the FDI growth nexus: The Indian evidence

ii. Macroeconomic environment and stock market: The Indian case

iii. Modeling the linkage between the US and Indian stock market

iv. The long-run relationship between stock returns and inflation in India

v. Testing market efficiency hypothesis: The Indian stock market

D. Economics and social issue

i. linkages between Food production and Indian poverty

ii. Relationship between Economic growth and Indian competitiveness

iii. Causal Relationship between unemployment and economic development

iv. International tourism and economic development in India: Causality analysis

v. Financial recession and economic sustainability: Indian evidence

vi. Causal relationship between Terrorism attack and social impact: Evidence from India

vi. Modeling the linkage between terrorism attack and human peace: Evidence from India

Stay at home and maintain social distancing and strictly follow Government rules and regulations during this crisis period.

Thank you very much and good luck

Contact with me if you face any problem (9432653985)