A BRIEF STUDY OF CRIME RATE IN UNITED STATES AND VARIOUS FACTORS AFFECTING IT DURING 1991-2019:-

NAME – MAINACK PAUL

ROLL NO – 407

SUPERVISOR – DR. DURBA BHATTACHARYA

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Mainack Paul

Signature

❖ TABLE OF CONTENTS:-

CHAPTER 1

❖ ABSTRACT

Crime Rate denotes the ratio of crimes in an area to the population of that given area. **Objective** of this study is that we will be able **to predict which factors are affecting the Crime Rate mostly**. The data being used here is a **secondary time series data and is collected for 29 years from 1991-2019.** The **response variable** is the **Crime Rate** and; the **covariates** are **Unemployment Rate, Poverty Rate, GDP Growth Rate, Annual Growth Rate, Death Rate, Depression Rate and Population Growth Rate**.

Firstly, a **Descriptive study** of the data is done followed by checking, for the **stationarity** and **autocorrelation** in the dataset, **multicollinearity** between the covariates and **heteroscedasticity** in the dataset. Next, we **fit a regression model**, test for the **normality of the residuals** obtained and measure the **goodness of fit**. Finally, we test for **significance of model parameters and interpret** them. Some **findings** are that **all the variables are non-stationary and autocorrelated, covariates face multicollinearity problem and data is homoscedastic**.

❖ INTRODUCTION:-

The word "**crime**" is derived from the Latin word "**cerno**" meant that we decide and we give judgement. One of the appropriate definitions is that **crime** is an act that is harmful not only to an individual but to the entire society and community. Those acts are forbidden and punishable by law. **Crime Rate denotes the ratio of crimes in an area to the population of that given area**. Many researchers are working day and night to study the crime rate. **It is important to study crime rate because then only we can pin point the exact causes that are causing this crime and find out the remedial measures on how to reduce it and also how to make the judicial system fairer and more effective.**

Again, due to the Global Outbreak of the **CORONAVIRUS,** there is an increase in rate of Unemployment, Poverty and Death. **No other country has suffered more than United States. More than 30 million people have been affected and more than 5 lakh people died due to this deadly virus. (Source: www.statista.com)** There is a high chance of increase in Crime Rate of the United States and so it is necessary for us to study in details which factors can cause that problem in the near future, from the previous data that we have collected from 1991 – 2019.

With a view that this study will help us to predict which factors are affecting the crime rate mostly, I have selected this topic as my project.

❖ OBJECTIVES:-

The main objectives of this project are given below:-

- **Descriptive studies of the data thoroughly and graphically, that is, by applying Time Series Analysis to thoroughly examine the data.**
- **To check for the stationarity of the data. If some time series are found non-stationary, our task is to convert them into stationary time series data.**
- **To check for the presence of autocorrelation in the dataset. If present, our task is to remove it from the time series.**
- **To check whether multicollinearity is present in the dataset. If present, our task is to remove it by thoroughly studying the explanatory variables which are causing this problem.**
- **To check for the presence of heteroscedasticiity in the dataset. If present, our task is to remove it by thoroughly studying the dataset.**
- **To fit a regression model to enhance the analysis of our project.**
- **To test for the normality of the residuals obtained from the fitted regression model.**
- **To measure the goodness of fit of the regression model.**
- **To test for the significance of the fitted regression model.**
- **To test for the significance of the model parameters and to correctly interpret them.**

CHAPTER 2

❖ DATA DESCRIPTION:-

Here, in this project, we have collected secondary data on the various factors that are affecting the CRIME RATE in UNITED STATES. It is a Time Series data of 29 years spanned from 1991 to 2019.

- **NATURE OF THE VARIABLES: Quantitative (Continuous)**
- **RESPONSE VARIABLE: Crime Rate (in %)**
- **EXPLANATORY VARIABLES:**

  1). Unemployment Rate (in %)

  2).Poverty Rate (in %)

  3).GDP Growth Rate (in %)

  4).Annual Growth Rate (in %)

  5). Death Rate (in %)

  6).Depression Rate (in %)

  7).Population Growth Rate (in %)

DESCRIPTION OF THE VARIBLES:-

1).Crime Rate denotes the ratio of crimes in an area to the population of that given area.

2).Unemployment Rate denotes the number of people who are unemployed to that of the labour force (the total number of people who are employed added to those who are unemployed).

3).Poverty Rate denotes the ratio of the number of people (all age groups taken together) whose income fall below the poverty line (taken as half the median household income of the total population) to the total population.

4).Gross Domestic Product (GDP) is the total dollar amount of all goods and services produced.**GDP Growth Rate** denotes the percentage increase or decrease from the previous measurement cycle.

5).Annual Growth Rate denotes the change in the value of an individual investment, portfolio (it is a collection of financial investment like stocks, bonds, etc), asset or cash stream over the period of a year.

6).Death Rate denotes the number of deaths in a population (all age groups taken together) to the total number of individuals of that population.

7).Depression is characterized by persistent sadness and lack of interest in previously enjoyable activities. **Depression Rate** denotes the number of people of a population in depression (all age groups taken together) to the total number of individuals of that population.

8).Population Growth Rate denotes the rate at which the number of individuals (all age groups taken together) in a given time period increases expressed as a percentage of the initial population.

Table-1: The DATASET used in the project

YEAR	CR	UR	PR	GR	AR	DEATHR	DEPR	PGR
1991	9.7	6.80	14.2	-0.1	1.90	8.871	4.66	0.96
1992	9.2	7.50	14.8	3.5	4.42	8.844	4.65	0.96
1993	9.4	6.90	15.1	2.8	3.81	8.817	4.65	0.99
1994	8.9	6.12	14.5	4.0	4.96	8.785	4.65	1.04

1995	8.1	5.65	13.8	2.7	3.60	8.754	4.65	1.11
1996	7.3	5.45	13.7	3.8	4.45	8.722	4.66	1.20
1997	6.7	5.00	13.3	4.4	4.98	8.691	4.69	1.26
1998	6.2	4.51	12.7	4.5	4.43	8.659	4.72	1.27
1999	5.6	4.22	11.9	4.8	5.05	8.631	4.75	1.23
2000	5.5	3.99	11.3	4.1	5.28	8.603	4.77	1.14
2001	6.7	4.73	11.7	1.0	2.20	8.576	4.77	1.03
2002	5.6	5.78	12.1	1.7	2.40	8.548	4.78	0.94
2003	5.7	5.99	12.5	2.9	3.87	8.520	4.78	0.88
2004	5.5	5.53	12.7	3.8	5.61	8.441	4.79	0.88
2005	5.7	5.08	12.6	3.5	5.76	8.362	4.79	0.90
2006	5.8	4.62	12.3	2.9	4.95	8.282	4.79	0.94
2007	5.7	4.62	12.5	1.9	3.62	8.203	4.79	0.96
2008	5.4	5.78	13.2	-0.1	0.85	8.124	4.78	0.96
2009	5.0	9.25	14.3	-2.5	-2.65	8.131	4.77	0.93
2010	4.8	9.63	15.1	2.6	2.90	8.138	4.76	0.88
2011	4.7	8.95	15.9	1.6	2.93	8.145	4.76	0.83
2012	4.7	8.07	15.0	2.2	3.46	8.152	4.77	0.79
2013	4.5	7.38	14.5	1.8	2.92	8.159	4.77	0.75
2014	4.5	6.17	14.8	2.5	3.63	8.264	4.78	0.72
2015	5.0	5.28	13.5	2.9	3.22	8.369	4.80	0.69
2016	5.4	4.87	12.7	1.6	1.94	8.475	4.81	0.67

2017	5.3	4.36	12.3	2.4	3.50	8.580	4.84	0.64
2018	5.3	3.90	11.8	2.9	5.07	8.685	4.87	0.62
2019	5.1	3.68	13.8	2.3	3.63	8.782	4.90	0.60

Where

- **CR: Crime Rate, UR: Unemployment Rate, PR: Poverty Rate, GR: GDP Growth Rate, AR: Annual Rate, DEATHR: Death Rate, DEPR: Depression Rate, PGR: Population Growth Rate.**
- **All the data entered are in PERCENTAGES.**

- **SCRUTINY OF THE DATA:-**

The type of data that has been used in this project is secondary data. Any data needs to be verified for its consistency and homogeneity before starting its analysis. This verification of the data is known as scrutiny.

So at first we scrutinize the data to check whether there are any misleading values or not. We observe that this data set is free from such values.

Now, at first we start our analysis using the basic Time Series Analysis and note down the important observations that can be gathered. All the calculations and the graphs are performed using MINITAB.

CHAPTER 3

❖ TIME SERIES ANALYSIS:-

- **Time Series Data :**

Data that are collected, observed or recorded at successive intervals or points of time generate an ordered set known as **Time Series Data**. A time series is said to be **‘continuous’** when observations are made continuously in time. The term **‘continuous’** is used for series of this type even when the measured variable can only take a discrete set of values. A time series is said to be **‘discrete’** when the measured variable can only take a discrete set of values. The term **‘discrete’** is used for series of this type even when the measured variable is a continuous variable.

The special feature of time-series analysis is that successive observations are usually not independent. If a time series can be predicted exactly, it is said to be **deterministic**. But most time series are **stochastic** in that the future is only partly determined by the past values.

- **COMPONENTS OF A TIME SERIES:-**

A plot of the time series gives us an overall impression of haphazard movement. A critical study of the series, however, reveals that the movements are not completely haphazard and at least a part of it can be accounted for. The part which can be accounted for is called **Systematic Part** and the other part is known as **Unsystematic or Irregular Part**. The systematic movement of a time series is broadly composed of 3 main factors:-

a).Trend (or, Secular Trend) b).Seasonal Variation c).Cyclical Fluctuation

a). TREND (or SECULAR TREND):-

It is a smooth, regular and long term movement of a time series. Some series show an upward trend or downward trend. **Increase in demand of a commodity may exhibit an upward trend while decrease in demand of a product due to unavailability of raw materials lead to a decreasing trend. Also, some series may exhibit a constant trend.**

b). SEASONAL VARIATION:-

A periodic movement is one which recurs, with some degree of regularity, within a definite period. The most frequently studied periodic movement is that which occurs within a year and which is known as **seasonal variation**.

c). CYCLICAL VARIATION:-

Cyclical movements are fluctuations which differ from seasonal movements in that they are of longer duration than a year and also in that they do not ordinarily exhibit regular periodicity. It undergoes 4 phrases – **boom, decline, depression and recovery**. The time lag between two consecutive booms or depressions is known as **period of cycle**.

IRREUGULAR COMPONENT:-

Apart from the regular variations, all the series contains factor called random or irregular which are not accounted for by trend, seasonal and cyclical variations. These fluctuations are purely random, erratic, unforeseen and unpredictable and are due to numerous non-recurring and irregular circumstances which are beyond the control of humans.

- **DECOMPOSITION OF A TIME SERIES:-**

By decomposition of a time series, we mean breaking down a time series in its various components. The two broadly used models for decomposition of a time series are:-

a) Additive Model b). Multiplicative Model

Let 'X_t' denote the value of a time series at time point 't' while 'T_t', 'S_t', 'C_t' and 'I_t' denote respectively the values of trend, seasonal component, cyclical fluctuation and irregular variation at time point 't'.

a). Additive Model:-

According to this model, decomposition of time series is done on the assumption that the effects of the 4 components are additive in nature. In other words,

$$X_t = T_t + S_t + C_t + I_t$$

This model assumes that the 4 components are independent of each other and this assumption does not hold good in general.

b). Multiplicative Model:-

According to this model, decomposition of time series is done on the assumption that the effects of the 4 components are multiplicative in nature and they are not necessarily independent of each other. In other words,

$$X_t = T_t * S_t * C_t * I_t$$

ANALYSIS OF THE COMPONENTS:-

- **The secondary data that is being used in this project is a yearly data and hence seasonal component will not be present here.**
- **As we do not know anything about the independence of the 4 components, hence, we will assume it to be a multiplicative model.**

❖ ANALYSIS OF THE RESPONSE VARIABLE: CRIME RATE (in %)

We consider the data of Crime Rate (in %) from 1991-2019. We obtain the time series plot as –

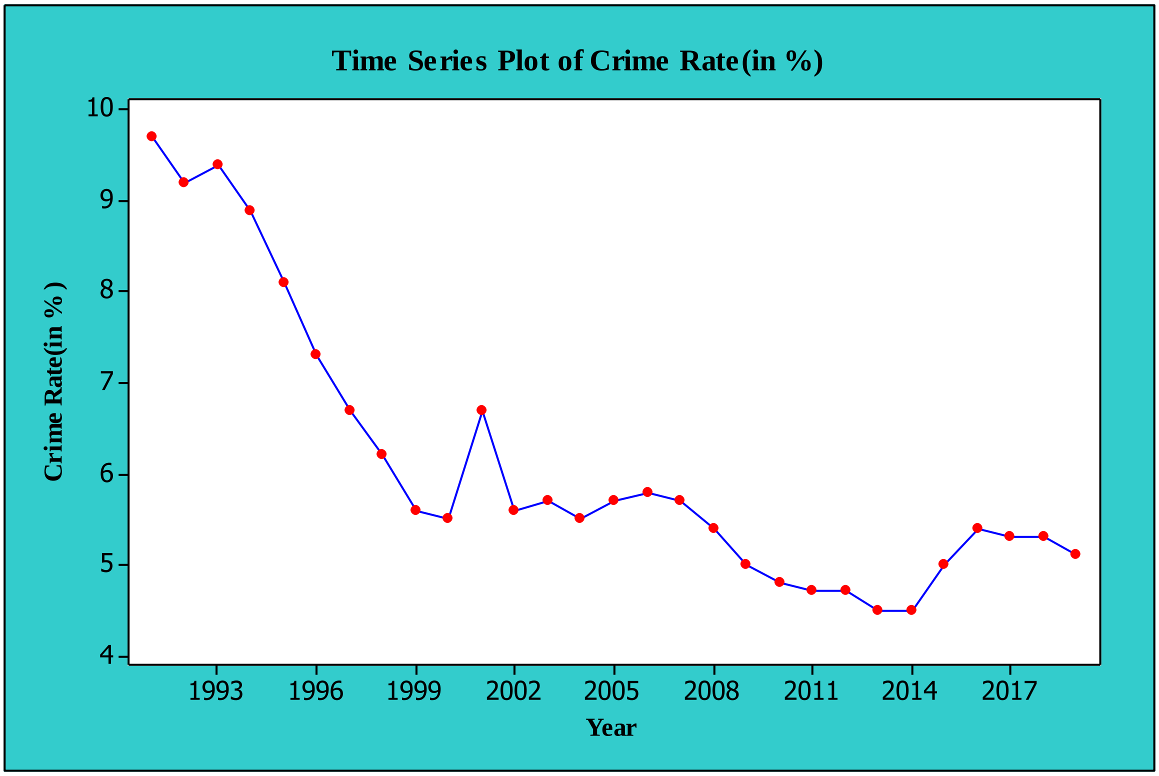


Figure-3.1: Time Series of Crime Rate (in %)

INTERPRETATION:-

Here, we can see a **decreasing trend** in general over the last 29 years.

Here, we consider the model as

$X_t = T_t * C_t * I_t$ (3.1), where t = 1(1)29.

- **TREND ANALYSIS:-**

A Quadratic Trend Model is being fitted at first and the fitted model is obtained as -

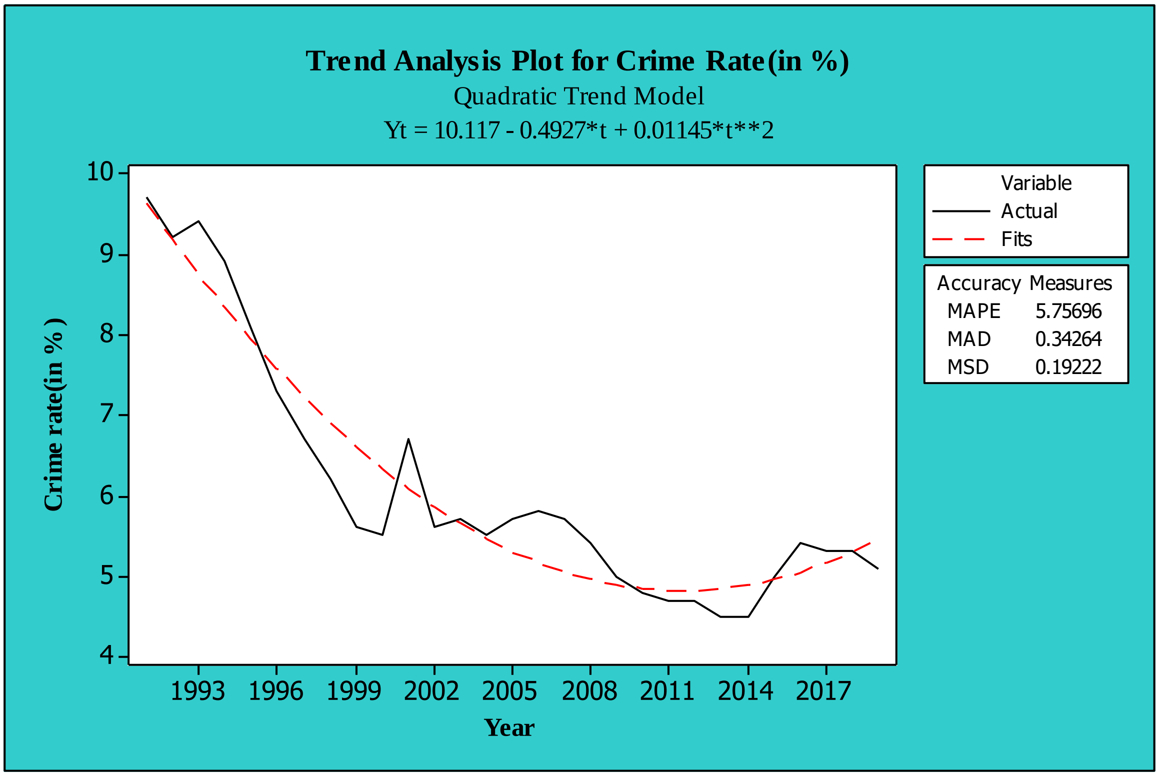


Figure-3.2: Trend Analysis Plot of Crime Rate (in %)

The fitted curve is $\hat{T}_t$, where t = 1(1)29.

Interpretations:-

From the graph, it is evident that the quadratic trend fitted well here because the Mean

Absolute Deviation (MAD) and Mean Square Deviation (MSD) values are very low here.

Now we can find a better perception about the cyclical fluctuation prevailing in this series if we do

$$X_t / \hat{T}_t = C_t * I_t \qquad (3.2), \quad \text{where } t = 1(1)29.$$

Now our task is to separate out the cyclical variation and analyzing it properly.

- **Measurement of Cyclical Variation:-**

Considering the required time series from which trend has been eliminated (given by (2)). Our main motive is to find out whether **the required time series** contains a harmonic term with period '**μ'.** A time series can be expressed as a combination of sine and cosine waves with differing periods and amplitudes and this fact is utilized to examine the periodicity (also known as **Periodogram Analysis**).

For different values of '**μ'** we do the necessary calculations and obtain the true period '**λ**'. As we find out '**λ',** we now fit a sine-cosine curve through the residual series '$\mathbf{u_t}$**'** (the process known as **Harmonic Analysis**).

After performing the necessary calculations, we finally separate out the cyclical component of the series and it is obtained as –

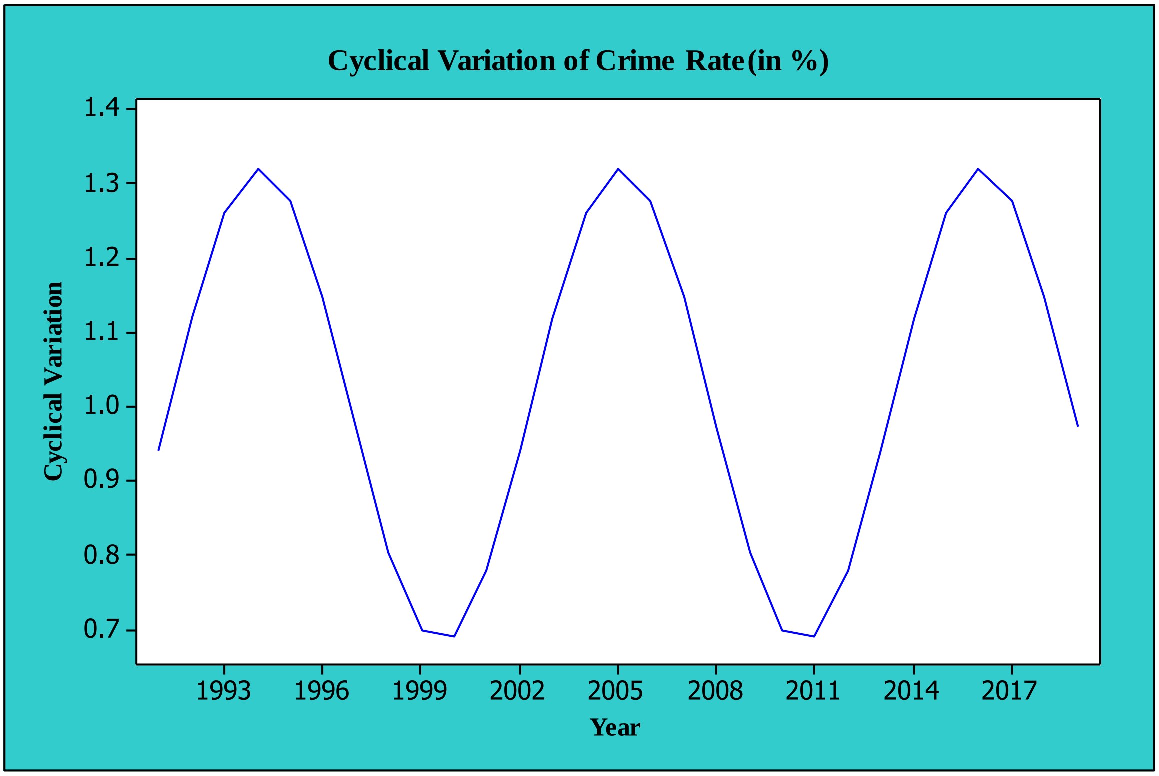


Figure-3.3: Plot of Cyclical Variation of Crime Rate (in %)

The fitted curve is $\hat{C}_t$, where t = 1(1)29.

Interpretations:-

From the above graph we can clearly see that

- The first cycle reached the boom in the middle of 1994-1995 and then started decreasing and reached the lowest value in the year 2000.
- Then, again start increasing and reach the boom in the year 2005.
- So we can roughly say that duration of a full cycle is almost 10 years.

Now we can find a better perception about the irregular variation prevailing in this series if we do

$X_t \;/\; (\hat{T}_t * \hat{C}_t) = I_t$ (3.3), where t = 1(1)29.

Now finally the irregular component of the graph is obtained as -

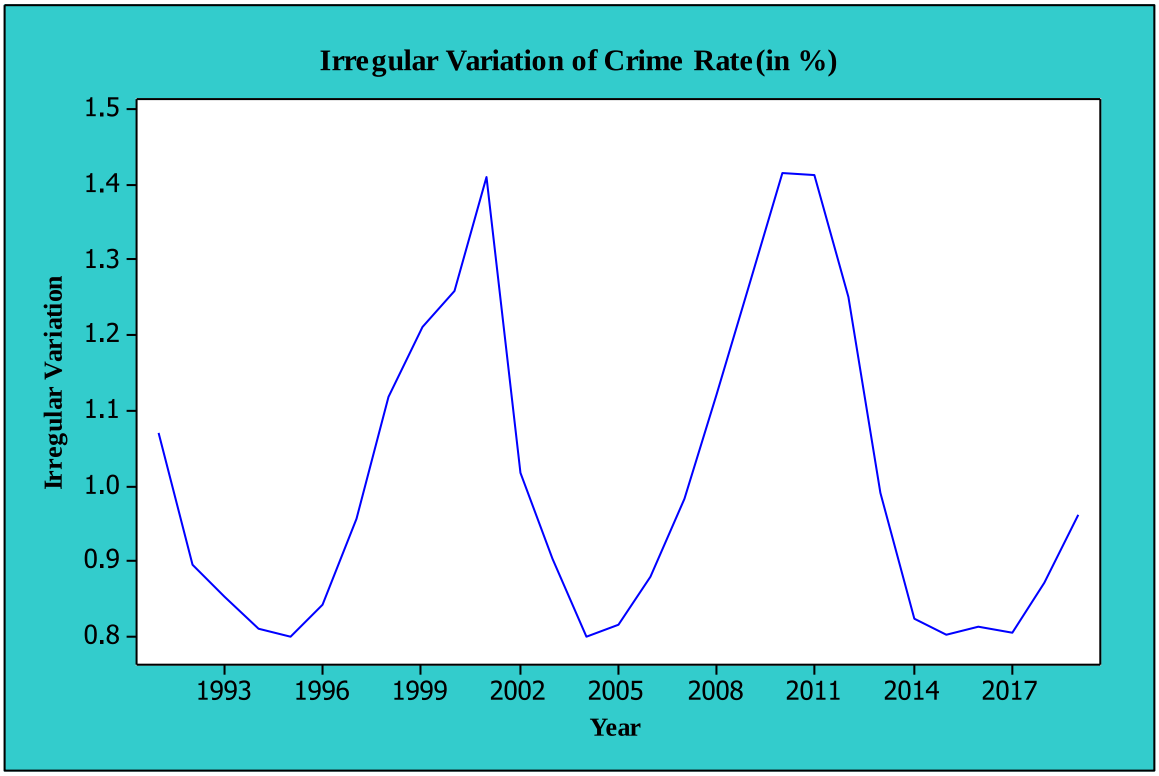


Figure-3.4: Plot of Irregular Variation of Crime Rate (in %)

The fitted curve is $\hat{I}_t$, where t = 1(1)29.

Interpretation:-

From the graph,

- We can observe a roughly cyclical pattern.
- The first cycle reached the depression at 1995 and then increases and reached its boom at 2001 and then again reached its boom at 2004.
- Roughly, we can notice that the period of the cycle is of 11 years.

❖ BRIEF ANALYSIS OF THE EXPLANATORY VARIABLES:-

All the 3 components Trend, Cyclical Fluctuations and Irregular Variations have been extracted using the similar processes and mathematical calculations as mentioned above.

The following things have been done under this brief analysis:-

i). Trend Analysis of 2 explanatory variables are performed here whose observations will be of our interest.

ii). Comparative Analysis between the explanatory variables. First comparison of trend analysis is shown, then, cyclical fluctuations and then Irregular variations.

I) TREND ANALYSIS:-

1). Trend Analysis of Unemployment Rate (in %):-

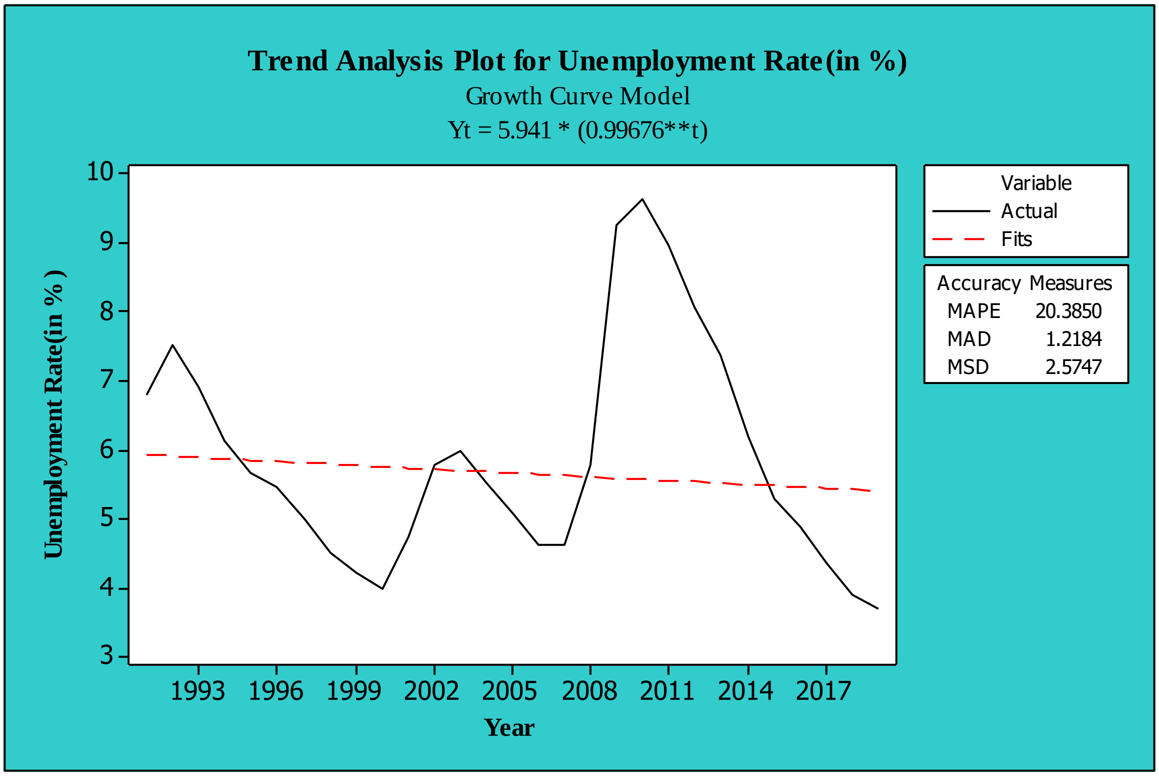


Figure-3.5: Trend Analysis Plot of Unemployment Rate (in %)

Here, a **Exponential Curve has been fitted** which is by far the best among the other fitted curves (Linear, Quadratic) because of its MAPE (Mean Absolute Percent Error), MAD (Mean Absolute Deviation) and MSD (Mean Square Deviation) values which are lower here **compared to the other curves**.

INTERPRETATION:- From the graph, it is quite observable that there is **no trend present** and hence it's MAPE, MSD and MAD values are so high and cyclical fluctuations and irregular variations may be present here.

2).Comparative Analysis between Annual Growth Rate (in %) and GDP Growth Rate (in %):-

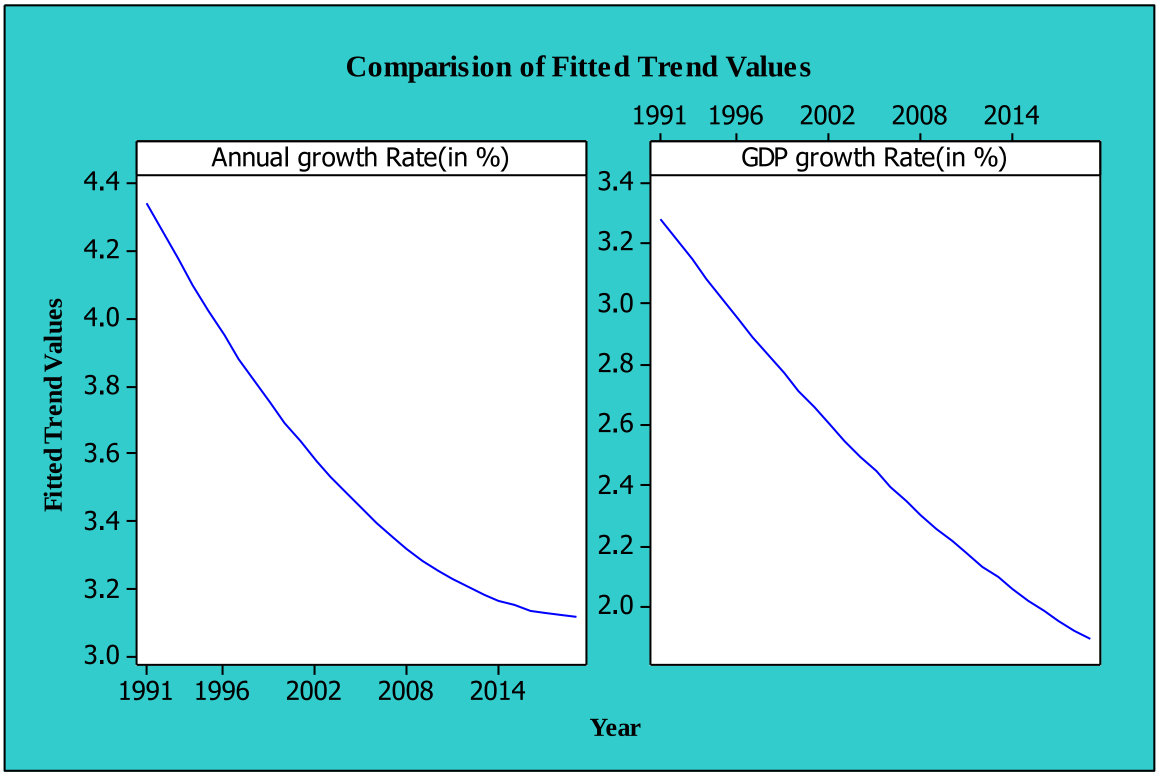


Figure-3.6: Comparison of fitted trend values between Annual Growth Rate (in %) and GDP Growth Rate (in %).

INTERPRETATION:-

From the graph, we observe that both the explanatory variables **Annual Growth Rate and GDP Growth Rate** have a **similar decreasing trend pattern** over the past 29 years, that is, from 1991-2019 and that there is a high chance that these variables may be **intercorrelated** among themselves (problem of **Multicollinearity**) or can also be due to some hidden factors .

3). Comparative Analysis between Death Rate (in %) and Population Growth Rate (in %):-

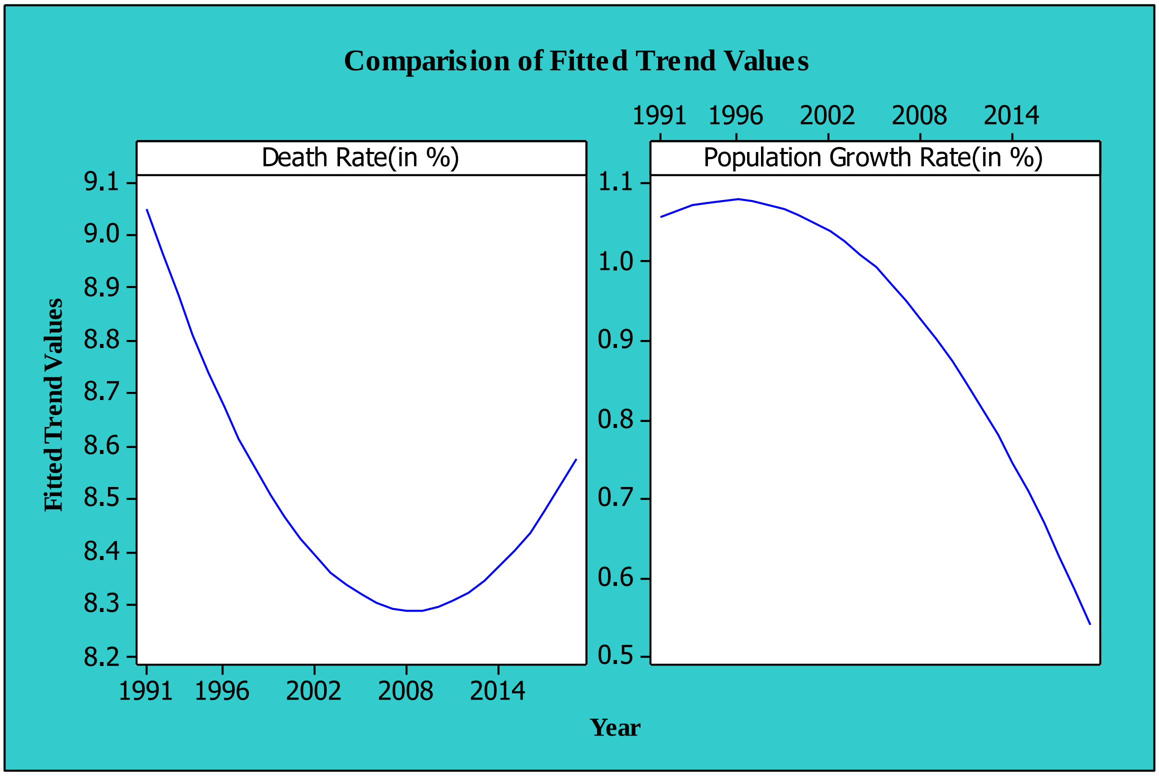


Figure-3.7: Comparison of fitted trend values between Death Rate (in %) and Population Growth Rate (in %).

INTERPRETATION:-

From the graph, we can observe that the trend of Death Rate is first decreasing and then increasing whereas the trend of Population Growth Rate is first a little bit increasing and then decreasing which is **exactly the opposite**. There is a chance that these variables may be **intercorrelated** among themselves (problem of **Multicollinearity**) or can be also due to some hidden factors.

II). CYCLICAL FLUCTUATIONS:-

1). Comparative Analysis between Annual Growth Rate (in %) and Population Growth Rate (in %):-

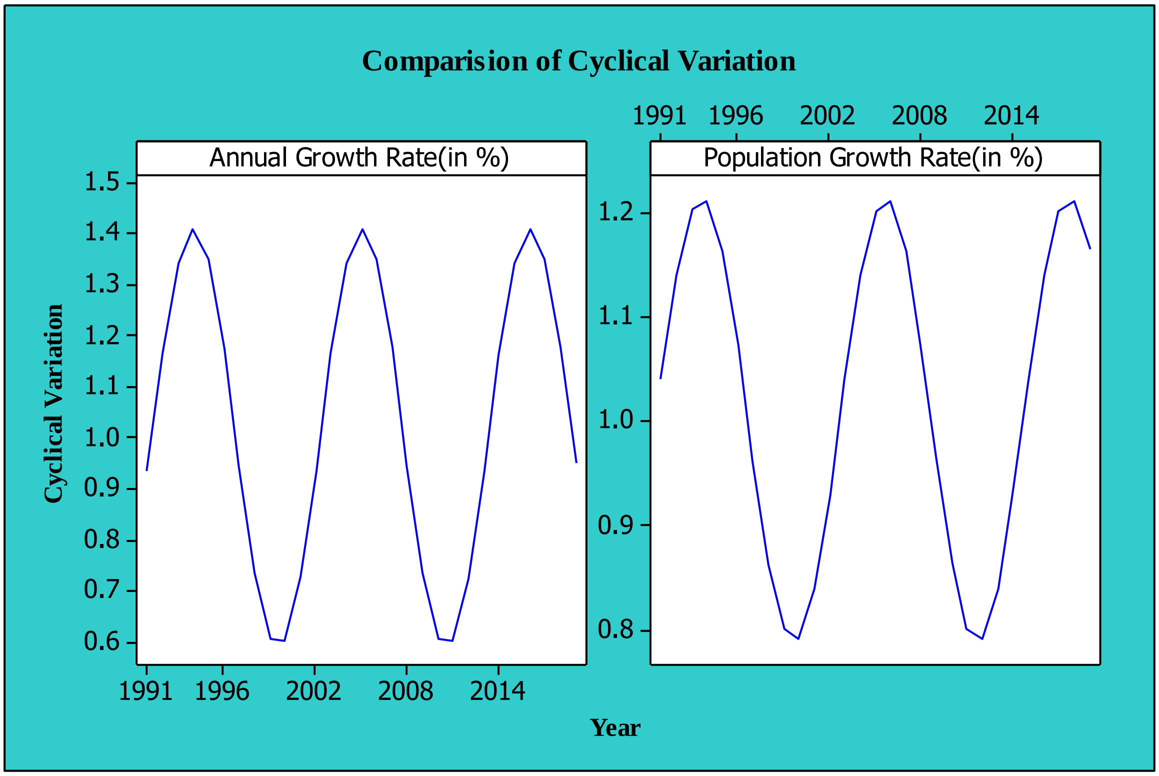


Figure-3.8: Comparison of Cyclical Fluctuation Annual Growth Rate (in %) and Population Growth Rate (in %)

INTERPRETATION:-

From the graph, we observe a **similar cyclical pattern** between these two variables and hence there is a chance that these variables may be **intercorrelated** among themselves (problem of **Multicollinearity**) or can be also due to some hidden factors.

2).Comparative Analysis between Depression Rate (in %), Death Rate (in %) and Poverty Rate (in %):-

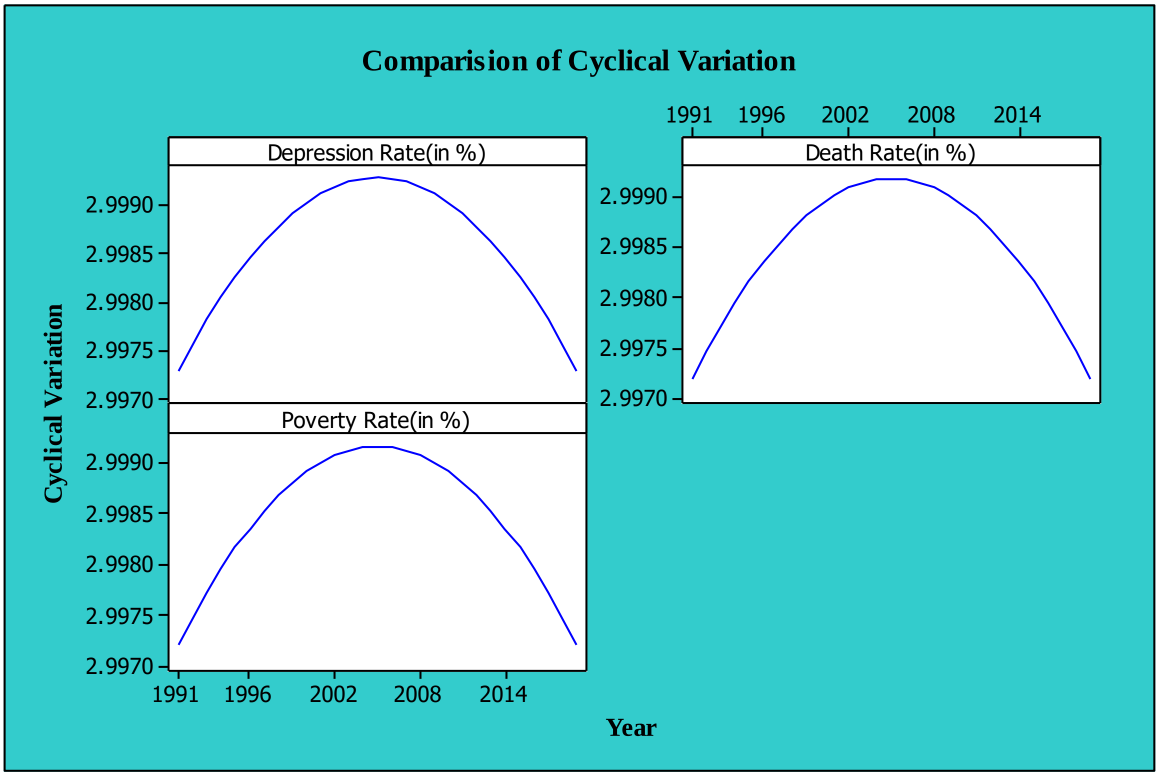


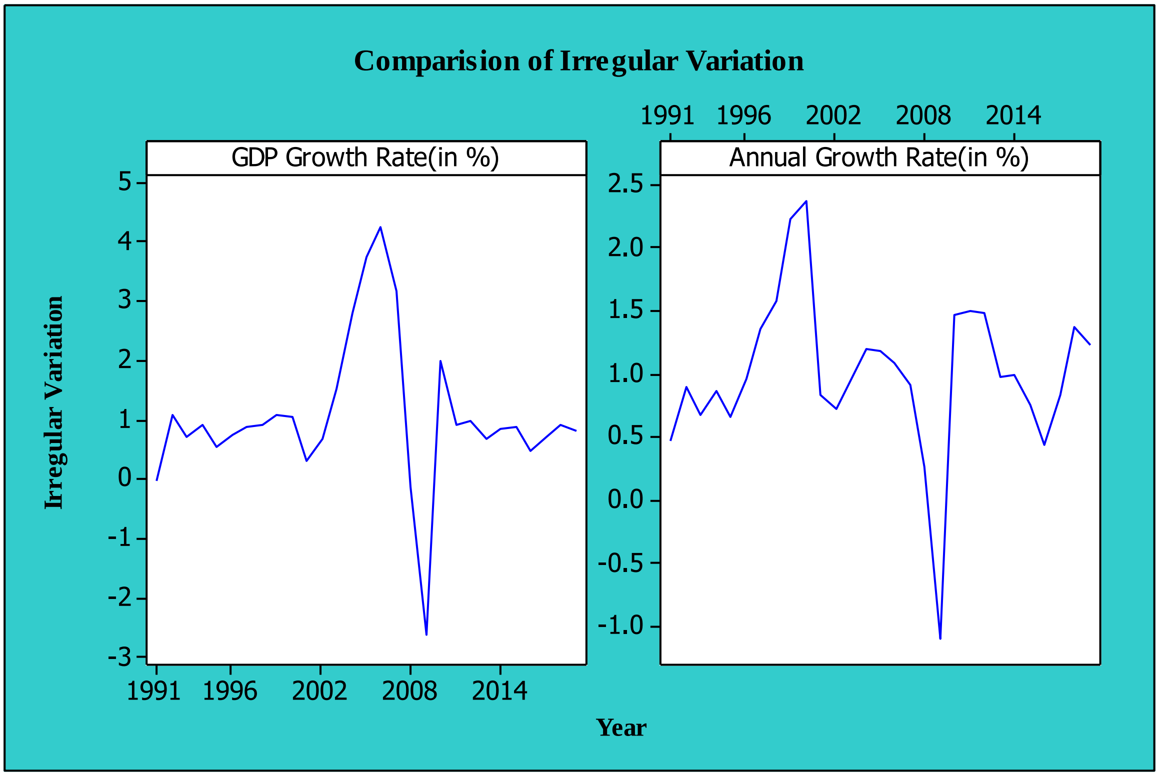
Figure-3.9: Comparison of Cyclical Fluctuation between Depression Rate (in %), Death Rate (in %) and Poverty Rate (in %).

INTERPRETATION:-

From the graph, we observe that none of the 3 above mentioned variables have completed a single cycle in the past 29 years, that is, from 1991-2019 and so **within this period, no cyclical variations are present** among these variables and if we increase the time period, there lays a slight possibility that cyclical variations could be present. Hence, there is a chance that the variables may be **intercorrelated** among themselves (problem of **Multicollinearity**) or can be also due to some hidden factors.

III). <u>IRREGULAR VARIATIONS:-</u>

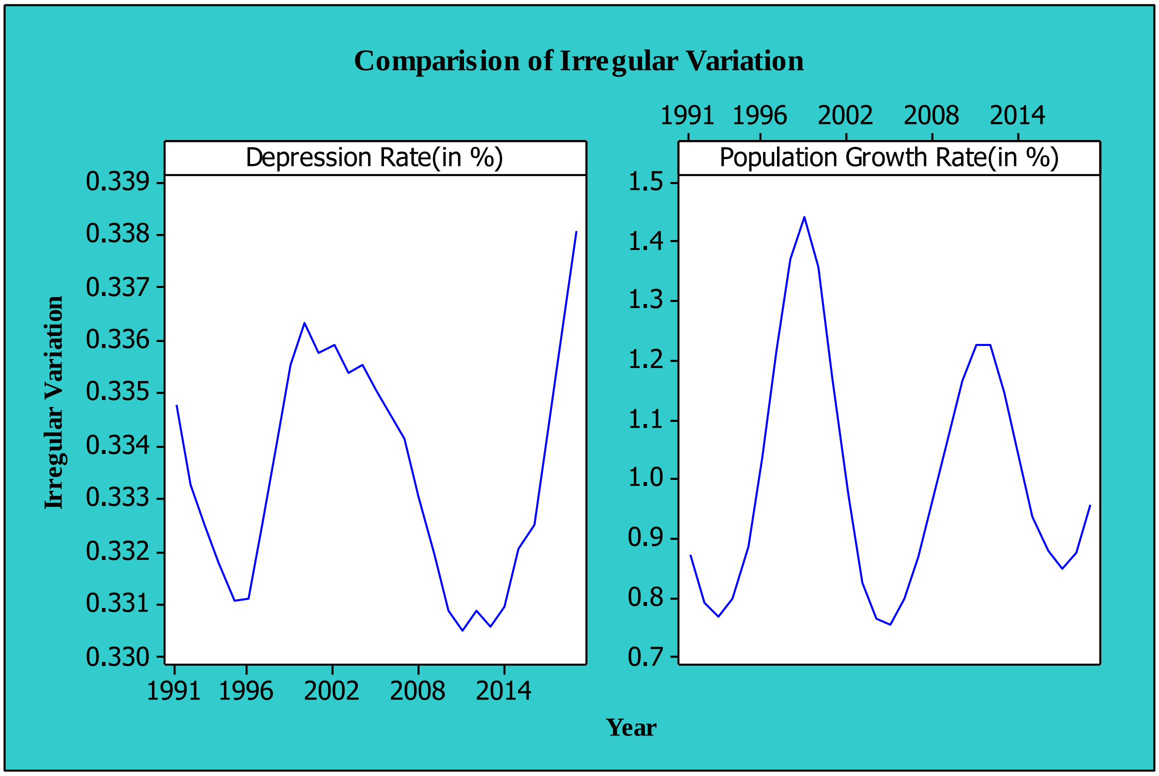
1).<u>Comparative Analysis between GDP Growth Rate (in %) and Annual Growth Rate (in %):-</u>



<u>Figure-3.10</u>: Comparison of Irregular Variation between GDP Growth Rate (in %) and Annual Growth Rate (in %).

<u>INTERPRETATION</u>:- From the graph, we can observe that both of these variables have a roughly similar pattern. A constant pattern is observed followed by a peak and then a dip has been observed in 2009 in both of these graphs which is may be **<u>due to widespread failures in financial regulation including the failure to stem the tide of toxic mortgages</u>.** Hence, there is a chance that these variables may be **intercorrelated** among themselves (problem of **Multicollinearity**) or can be also due to some hidden factors.

2). <u>Comparative Analysis between Depression Rate (in %) and Population Growth Rate (in %):-</u>



<u>Figure-3.11</u>: Comparision of Irregular Variation between Depression Rate (in %) and Population Growth Rate (in %)

<u>INTERPRETATION:-</u>

From the graph, we can observe a cyclical pattern roughly, in both of these variables but with different boom and depression points. There is a slight possibility that these variables may be

intercorrelated among themselves (problem of **Multicollinearity**) or can be also due to some hidden factors.

❖ <u>IMPORTANT CONCLUSIONS FROM TIME SERIES ANALYSIS:-</u>

From the above time series analysis, we have noticed **similar patterns** in trend, cyclical fluctuations and irregular variations between some explanatory variables. Hence, it may be possible that there exist some sorts of relationship in between those explanatory variables, out of which one of the possible cases is Multicollinearity. **<u>But before that, we have to check firstly for the Stationarity of this dataset. It is because some series if found non-stationary, then, we have to perform necessary differencing with appropriate lags to that series and in this process our present explanatory variables may get changed. Hence, first we will check for the stationarity of the time series and then, if found non-stationary, we will convert them into stationary series and hence the explanatory variables got transformed. Then, we will check for the presence of Multicollinearity in this new set of transformed explanatory variables</u>.**

<u>CHAPTER 4</u>

❖ <u>THE BASIC ASSUMPTIONS TAKEN INTO CONSIDERATION:-</u>

One of the main focuses of this project is to check for the presence of autocorrelation and heteroscedasticity in the dataset and for performing the test; **we have to assume a multiple linear regression model.** Notational wise, the response is labeled as **CR** and the predictors are labeled as **(UR, PR, AR, GR, DEATHR, DEPR and PGR).** The Linear model which we will be considering here is based on certain assumptions. The assumptions are:-

- The regression model is linear in the parameters.
- The values of the regressors are fixed and are independent of the error term 'u_i', i=1(1)29 and hence the correlation between them is zero.
- For given values of the regressors, the mean value of the error term is zero.
- For given values of the regressors, the variance of 'u_i' is constant, which means the variation of the error terms are homoscedastic.
- For given values of the regressors, there is no autocorrelation (defined below) between the disturbance terms.
- The number of the observations must be greater than the number of the parameters to be estimated.
- There must be sufficient variation in the values of the regressors.
- There is no intercorrelation among the regressor variables (that is, problem of multicollinearity is not being considered here).
- The model is correctly specified, so there is no specification bias.
- The error terms 'u_i' are normally distributed.

CHAPTER 5

❖ CHECKING THE PRESENCE OF NON-STATIONARITY AND AUTOCORRELATION IN A TIME SERIES:-

In this part, we will proceed in the following manner:-

- **First, we will check for the presence of non-stationarity in the dataset.**
- **Our second task here will be to remove the non-stationarity from the dataset, if present.**
- **And our final task here is to check for the presence of autocorrelation in the dataset.**

I). NON-STATIONARITY IN TIME SERIES DATASET:-

Let T be a subset of the set of all positive real numbers. A family of random variables $\{X_t\}_{t\,\varepsilon\,T}$, indexed by T is called a stochastic (or random) process. Intuitively, a random process $\{X_t\}_{t\,\varepsilon\,T}$ is stationary **if its statistical properties do not change by time**. For a stationary process, **X(t)** and X(**t + Δ**) have the same probability distribution. More generally for a random process, the joint distribution of $\mathbf{X(t_1)}$ and $\mathbf{X(t_2)}$ is same as the joint distribution of $\mathbf{X(t_1 + \Delta)}$ and $\mathbf{X(t_2 + \Delta)}$, that is,

$$\underline{\mathbf{Pr((X(t_1), X(t_2))\ \varepsilon\ A) = Pr((X(t_1 + \Delta)\ , X(t_2 + \Delta))\ \varepsilon\ A), \text{ for any set } A\ \varepsilon\ R^2,}}$$

where '**Pr(.)**' denotes a probability function, '**ε**' means "belong to" , '**Δ**' denotes the time difference between the two chosen time points and '$\mathbf{R^2}$' refers to the 2 dimensional space.

Time series possessing trends or with seasonality are not stationary since the trend and seasonality will affect the value of the time series at different times. Non-stationary time series data in models produces unreliable and spurious results and leads to poor understanding and forecasting. As a solution to this problem the data is made to be stationary using statistical techniques and then foretasted such that model is made less susceptible to unusual changes and fluctuating demand. Statistical techniques can be broadly classified as:

1. Detrending the series: This technique is used when we need to handle deterministic trend in the data.

2. Deseasonalizing the series: This technique is used when we need to handle deterministic seasonality in the data.

3. Differencing the series: This technique is used when we need to smoothen the random fluctuations in the data and also, in case of the presence of unit root in the data.

4. Log transforming the series: This technique is used when we need to smoothen variation in the data.

- **CHECKING THE PRESENCE OF NON-STATIONARITY USING PHILLIPS-PERRON(PP) UNIT ROOT TEST:-**

Phillips-Perron test is one of the most appreciable tests used in the detection of the presence of non-stationarity in a time series. The Phillips-Perron unit root test differs from the other tests mainly in **how it corrects for autocorrelation and heteroscedasticity in the errors**. One advantage of the PP tests over other tests is that the PP tests are robust to general forms of heteroscedasticity in the error term ‘u_t’. Another advantage is that the user does not have to specify a lag length for the test regression.

Hence , the regression equation for this test is: $\boldsymbol{\Delta y_t = \beta' * D_t + \pi * y_{t-1} + u_t}$, **where ‘$\beta'$ is a constant , ‘π’ is the parameter of interest, ‘D_t’ is a vector of deterministic terms (constant, trend etc.) and ‘u_t’ (error term) is I(0) and heteroscedastic.**

A stationary series without a trend is said to be integrated of order 0 and is referred to as I (0).

The PP tests correct for any autocorrelation and heteroscedasticity in the errors 'u_t' of the test regression with the help of test statistics 'Z_t' and 'Z_π'.

$$Z_t = \sqrt{}\,(\hat{\sigma}^2 / \hat{\lambda}^2) * t_{\pi = 0} - 0.5 * ((\hat{\lambda}^2 - \hat{\sigma}^2)/ \hat{\sigma}^2) * ((n – SE\ (\hat{\pi}))/ \hat{\sigma}^2)$$

$Z_\pi = n * \hat{\pi} - 0.5 * ((n^2 * SE\ (\hat{\pi})\ )/ \hat{\sigma}^2) * (\hat{\lambda}^2 - \hat{\sigma}^2)$, where 'n' denotes the number of data points.

The terms $\hat{\sigma}^2$ and $\hat{\lambda}^2$ are consistent estimators of the variance parameters

$$\boldsymbol{\sigma^2 = \lim_{n \,->\, \infty} \Sigma_{t=1(1)n}\ E(u_t^2) / n}$$

$$\boldsymbol{\lambda^2 = \lim_{n \,->\, \infty} \Sigma_{t=1(1)n}\ E(S_n^2 / n)}, \textbf{ where } \boldsymbol{S_n = \Sigma_{t=1(1)n}\ u_t}.$$

The sample variance of the least squares residual $\hat{u}_t$ is a consistent estimate of σ^2, and the Newey-West long-run variance estimate of u_t using $\hat{u}_t$ is a consistent estimate of $\boldsymbol{\lambda^2}$.

<u>HYPOTHESIS:-</u>

To test H_0: π=0 against H_1: Not H_0, that is, it is equivalent to test,

H_0: the series is non-stationary against H_1: Not H_0,

Where H_0 and H_1 respectively denote the null and alternative hypothesis.

<u>Sample:-</u>

The samples, collected for the 7 explanatory variables and the response variable, each are of size 29 and 29 data points correspond to 29 years from 1991-2019.

Test Statistic:-

The test statistics are given by

$$\mathbf{Z_t = \sqrt{}\ (\hat{\sigma}^2 / \hat{\lambda}^2) * t_{\pi = 0} - 0.5 * ((\hat{\lambda}^2 - \hat{\sigma}^2) / \hat{\sigma}^2) * ((n - SE\ (\hat{\pi})) / \hat{\sigma}^2)}$$

$$\mathbf{Z_\pi = n * \hat{\pi} - 0.5 * ((n^2 * SE\ (\hat{\pi})) / \hat{\sigma}^2) * (\hat{\lambda}^2 - \hat{\sigma}^2)}$$

Distribution of Test Statistic:-

Under H_0 , the distribution of the test statistic are

$\mathbf{Z_t \sim^{asy} t_{n-1}}$

$Z_\pi \sim^{asy}$ **N (δ, $\hat{\sigma}^2$), δ ε R – {0}, where R denotes the set of all real numbers.**

Here, n = 29.

Note:-

The **p-value associated with a test,** is the probability under the null hypothesis that, the given test statistic takes the observed value and more extreme values in the direction given by the alternative. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis and the chance for the rejection of the null hypothesis increases.

Testing Rule:-

If the p-value is found to be high, we accept that the time-series is non-stationary, that is, the null hypothesis is accepted else we reject the null hypothesis.

Computation:-

We have to perform the above testing procedure for the corresponding 8 variables, that is, we have to perform the test 8 times in total.

The following calculations have been computed using R software.

Table-5.1: Overall Stationary Diagnostics:-

Variable	p-value	Decision	Conclusion
CR	**0.7165**	**Acceptance of H_0**	**1**
UR	**0.6816**	**Acceptance of H_0**	**1**
PR	**0.6353**	**Acceptance of H_0**	**1**
AR	**0.05694**	**Acceptance of H_0**	**1**
GR	**0.02022**	**Acceptance of H_0**	**1**
DEATHR	**0.99**	**Acceptance of H_0**	**1**
DEPR	**0.8377**	**Acceptance of H_0**	**1**
PGR	**0.4609**	**Acceptance of H_0**	**1**

1---> indicates the given variable is a non-stationary time series.

0 ---> indicates the given variable is a stationary time series.

Decision and Conclusion:-

Since the p-values of all the 8 variables Crime Rate, Unemployment Rate, Poverty Rate, Annual Growth Rate, GDP Growth Rate, Death Rate, Depression Rate and Population

Growth Rate are very high, in the light of the given data; we can conclude that all of them are non-stationary time series. Hence, now our primary task will be to convert the non-stationary time series into stationary time series through suitable transformation of the variables.

II). REMOVING NON-STATIONARITY FROM A DATASET:-

As we have seen above that non-stationarity is present in this dataset. Among all the techniques, here we are going to remove the non-stationarity in a time series by the **difference method**. This technique is used when we need to smoothen the random fluctuations in the data and also in the presence of unit root in the data. **If the original time series is non-stationary, it is found in general that the first differences (in presence of linear trend) and the second differences (in presence of quadratic trend) often become a stationary time series. If autocorrelation is present in this dataset, the advantage of this difference method is that the differencing technique also removes the autocorrelation from the data. Hence, this difference transformation serve for dual purposes.**

Now, since most of our explanatory variables have quadratic trend in nature, hence we will go by the second difference method. The second difference method is as follows:

$$y''_t = y'_t - y'_{t-1}$$

$$= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2})$$

$$= y_t - 2 * y_{t-1} + y_{t-2}$$ **, where** y_t , y_{t-1} and y_{t-2} respectively

denotes the value of the variable at the t^{th} time point, $(t-1)^{th}$ time point and $(t-2)^{th}$ time point, where t = 3(1)29. Also, **‘y''_t’ in general denotes the values of the transformed variables and ‘y''_t’ will take 27 values, that is, from 1993 – 2019.**

Table-5.2: The Transformed Dataset

Year	CR_N	UR_N	PR_N	AR_N	GR_N	DEATHR_N	DEPR_N	PGR_N
1993	0.7	-1.3	-0.3	-3.13	-4.3	0	0.01	0.03
1994	-0.7	-0.18	-0.9	1.76	1.9	-0.005	0	0.02
1995	-0.3	0.31	-0.1	-2.51	-2.5	0.001	0	0.02
1996	0	0.27	-0.6	2.21	2.4	-0.001	0.01	0.02
1997	0.2	-0.25	-0.3	-0.32	-0.5	0.001	0.02	-0.03
1998	0.1	-0.04	-0.2	-1.08	-0.5	-0.001	0	-0.05
1999	-0.1	0.2	-0.2	1.17	0.2	0.004	0	-0.05
2000	0.5	0.06	0.2	-0.39	-1.0	0	-0.01	-0.05
2001	1.3	0.97	1	-3.31	-2.4	0.001	-0.02	-0.02
2002	-2.3	0.31	0	3.28	3.8	-0.001	0.01	0.02
2003	1.2	-0.84	0	1.27	0.5	0	-0.01	0.03
2004	-0.3	-0.67	-0.2	0.27	-0.3	-0.051	0.01	0.06
2005	0.4	0.01	-0.3	-1.59	-1.2	0	-0.01	0.02
2006	-0.1	-0.01	-0.2	-0.96	-0.3	-0.001	0	0.02
2007	-0.2	0.46	0.5	-0.52	-0.4	0.001	0	-0.02
2008	-0.2	1.16	0.5	-1.44	-1.0	0	-0.01	-0.02

2009	-0.1	2.31	0.4	-0.73	-0.4	0.086	0	-0.03
2010	0.2	-3.09	-0.3	9.05	7.5	0	0	-0.02
2011	0.1	-1.06	0	-5.52	-6.1	0	0.01	0
2012	0.1	-0.20	-1.7	0.5	1.6	0	0.01	0.01
2013	-0.2	0.19	0.4	-1.07	-1.0	0	-0.01	0
2014	0.2	-0.52	0.8	1.25	1.1	0.098	0.01	0.01
2015	0.5	0.32	-1.6	-1.12	-0.3	0	0.01	0
2016	-0.1	0.48	0.5	-0.87	-1.7	0.001	-0.01	0.01
2017	-0.5	-0.10	0.4	2.84	2.1	-0.001	0.02	-0.01
2018	0.1	0.05	-0.1	0.01	-0.3	0	0	0.01
2019	-0.2	0.24	2.5	-3.01	-1.1	-0.008	0	0

Where

- **CR_N: Transformed Crime Rate, UR_N: Transformed Unemployment Rate, PR_N: Transformed Poverty Rate, GR_N: Transformed GDP Growth Rate, AR_N: Transformed Annual Rate, DEATHR_N: Transformed Death Rate, DEPR _N: Transformed Depression Rate, PGR_N: Transformed Population Growth Rate.**
- **All the data entered are in PERCENTAGES.**

NOTE:-

We have checked the stationarity of the variables for the second difference and for the corresponding various lag values (namely, for 1, 2, 3). We have got the best result for lag 1

where the non-stationarity has been removed almost fully but the result for the other lag values are not very satisfying. So we have taken the second difference at lag 1.

Again, using the similar testing procedure as done above, PHILLIPS-PERRON TEST has been performed to check whether the non-stationarity has been removed from the dataset or not.

Testing Rule:-

If the p-value is found to be high, we accept that the time-series is non-stationary, that is, the null hypothesis is accepted else we reject the null hypothesis.

Computation:-

The following calculations have been computed using R software.

Table-5.3: Overall Stationary Diagnostics of the transformed variables

Variable	**p-value**	**Decision**	**Conclusion**
CR_N	**< 0.01**	**Rejection of H_0**	**0**
UR_N	**< 0.01**	**Rejection of H_0**	**0**
PR_N	**< 0.01**	**Rejection of H_0**	**0**
AR_N	**< 0.01**	**Rejection of H_0**	**0**
GR_N	**< 0.01**	**Rejection of H_0**	**0**
DEATHR_N	**< 0.01**	**Rejection of H_0**	**0**

DEPR_N	**< 0.01**	**Rejection of H_0**	**0**
PGR_N	**0.3742**	**Acceptance of H_0**	**1**

1---> the given variable is a non-stationary time-series.

0 ---> the given variable is a stationary time-series.

<u>Decision and Conclusion:-</u>

In the light of the given data, after applying the second difference in the dataset, a significant removal of non-stationarity has been observed. The stationarity of the 7 variables (6 explanatory + 1 response) except for the Population Growth Rate, has been achieved. Also, the p-value of transformed Population Growth Rate is significantly low than the original and hence it has also improved by a large amount. <u>Hence, our primary task of converting the non-stationary time-series into stationary time-series through double differencing method of the variables is successfully achieved.</u>

Now, our task is to check, by how much amount we are successful in removing the autocorrelation from the dataset.

III). <u>AUTOCORRELATION IN TIME SERIES DATASET:-</u>

The '**auto**' part of autocorrelation comes from the Greek word for '**self**' and autocorrelation means data that is correlated with itself as opposed to being correlated with some other data. Autocorrelation is a mathematical representation which gives us the degree of similarity between a given time-series and a lagged version of itself over successive time intervals. It is similar to that of calculating the correlation between two different time series, except autocorrelation which uses the same series two times, that is, once in its original form and once lagged one or more time periods. For a given series of numbers if there is a pattern in such a way that the values in the series can be predicted based on the preceding values in the series, the series of numbers is said to exhibit autocorrelation. It may also be termed as ‘serial correlation’ and ‘serial dependence’. An autocorrelation of ‘+1’ represents a perfect positive correlation, while an autocorrelation of ‘-1’ represents a perfect negative correlation.

Above we have defined all the assumptions of Classical Linear Model, out of which one of them is that for the given values of the regressors, there is no autocorrelation between the disturbance terms. If autocorrelation is present in the dataset, then, we cannot use ordinary least square method for estimating the model parameters and we have to modify our method. Those estimators will not be BLUE (Best Linear Unbiased Estimator) if we still use the OLS method. Standard errors are biased which leads to bias in test statistics and confidence intervals. **Hence , it is necessary to check for the presence of autocorrelation in the dataset.**

- ## CHECKING FOR THE PRESENCE OF AUTOCORRELATION USING BREUSCH-GODFREY TEST:-

It is one of the best tests for detecting the presence of autocorrelation in a time-series. The test is general in the sense that it allows for non-stochastic regressors, higher-order autoregressive schemes and simple or higher-order moving average of white noise error terms. A two-variable regression model has been used here to perform the test. Any number of lagged values of the regressand can be added to the model. Here, a special assumption of the errors has been considered. We assume that the error term ‘u_t’ follows the p^{th}-order autoregressive AR (p) scheme, that is,

$\mathbf{u_t = \rho_1 * u_{t-1} + \rho_2 * u_{t-2} + \ldots\ldots\ldots + \rho_p * u_{t-p} + \varepsilon_t}$, where **‘ε_t’** is the error term associated in the model, t =1(1)27.

<u>The model in this setup is</u>:

$CR\_N_t = f\ (UR\_N_t, PR\_N_t, AR\_N_t, GR\_N_t, DEATHR\_N_t, DEPR\_N_t, PGR\_N_t) + u_t,$ (5.1)

t=1(1)27.

Where, f (UR_N_t, PR_N_t, AR_N_t, GR_N_t, $DEATHR\_N_t$, $DEPR\_N_t$, PGR_N_t) is the functional form of the best suited linear model (linear in parameters).

<u>Basic Test Procedures:-</u>

<u>Step-1:</u> The first step is to estimate equation (5.1) using the **method of OLS** and obtaining the residuals ‘$\hat{u}_t$’. The assumptions of OLS are mentioned before.

<u>Step-2 :</u> The second step is to regress ‘u_t’ on the 7 explanatory variables along with $\hat{u}_{t-1}$, $\hat{u}_{t-2}$,….. $\hat{u}_{t-p}$, where the latter are the lagged values of the residuals that have been estimated in Step-1. Hence, the required regression is

$\hat{u}_t = \mathbf{\hat{f}\ (UR\_N_t, PR\_N_t, AR\_N_t, GR\_N_t, DEATHR\_N_t, DEPR\_N_t, PGR\_N_t) + \hat{\rho_1} * \hat{u}_{t-1} + \hat{\rho_2} * \hat{u}_{t-2} + \ldots\ldots\ldots + \hat{\rho_p} * \hat{u}_{t-p} + \varepsilon_t}$, t = 1(1)27, **where ‘$\hat{\rho_1}$’,…..,‘$\hat{\rho_p}$’ are the constants** involved in the regression and **$\hat{f}$ (UR_N$_t$, PR_N$_t$, AR_N$_t$, GR_N$_t$, DEATHR_N$_t$, DEPR_N$_t$, PGR_N$_t$) is the fitted value of the best suited linear model (linear in parameters).**

HYPOTHESIS:-

To test H_0: $\rho_1 = \rho_2 = \ldots\ldots = \rho_p = 0$ against H_1: Not H_0, that is, to test

H_0: No Autocorrelation of any order is present against H_1: Not H_0.

Sample:-

The samples, collected for the 7 explanatory variables and the response variable, each are of size 27 and 27 data points corresponds to 27 years from 1993-2019 .

Test Statistic:-

The test statistic is given by

$$(n - p) * R^2, \text{ where } R^2 = 1 – RSS/TSS, \text{ where}$$

R^2: Coefficient of Determination

RSS: Residual Sum of Squares

TSS: Total Sum of Squares

Distribution of Test Statistic:-

Under H_0, the distribution of the test statistic is χ^2_p **asymptotically,** that is Chi-square distribution with degrees of freedom ‘p’, where ‘p’ denotes the number of lags.

Testing Rule:-

If the p-value of the test is found to be **very small**, we accept the presence of autocorrelation in the model, that is, the null hypothesis is rejected else we accept the null hypothesis.

Computation:-

P-value of the test = 0.03411 (from R software).

Decision and Conclusion:-

Since the p-value is not very small, in the light of the given data, we can say that the dataset is free from the problem of autocorrelation fully. Hence, our assumed model is correct with respect to the autocorrelation assumption.

Now as we have completed our checking and both the non-stationarity and the autocorrelation, mostly has been removed from our dataset, hence, now, our task is to formally fit the linear model.

CHAPTER 6

❖ FITTING OF A MODEL:-

A linear model describes the relationship between the response variable and the explanatory variables using a linear function. Since we have fitted a linear regression model for checking the presence of autocorrelation in the dataset, here, in this part of the project, we formally define our linear model.

Hence, our model is

$CR\_N_i = b_0 + b_1 * UR\_N_i + b_2 * PR\_N_i + b_3 * AR\_N_i + b_4 * GR\_N_i + b_5 * DEATHR\_N_i + b_6 * DEPR\_N_i + b_7 * PGR\_N_i + u_i$, $i = 1(1)27$,

Where b_0, b_1, b_2, b_3, b_4, b_5, b_6 and b_7 are the constants involved in the regression determined by the method of least squares **based on the given data and u_i is the error term associated with the model. 'b_0' is the intercept of the regression equation. 'b_j' is called the partial correlation coefficient of CR_N on j^th explanatory variable for fixed values of the remaining 6 explanatory variables. It is interpreted as the rate of change in the response CR_N for a unit change in j th explanatory variable, keeping the other 6 explanatory variables fixed, j= 1(1)7.**

We get the values of the estimates of the parameters involved in our model using the R software.

Table-6.1: Diagnostics of the fitted model

Coefficients	Estimate	Standard Error	T-value	P-value
Intercept	-0.006644	0.11702	-0.059	0.9532
UR_N	-0.370121	0.166292	-2.226	0.0383
PR_N	-0.196313	0.154901	-1.267	0.2203
AR_N	-0.076411	0.176193	-0.434	0.6694
GR_N	-0.04764	0.146261	-0.270	0.7899
DEATHR_N	6.30857	4.524958	1.394	0.1794
DEP_N	-29.69916	12.370811	-2.401	0.0268
PGR_N	-4.05894	4.378870	-0.927	0.3656

Also, the value of the coefficient of determination (r^2) comes out to be **0.4491. It is a measure of efficacy of the linear regression equation of CR_N on the explanatory variables. Hence, we can interpret that around 45% of the variance in Transformed Crime Rate is explained by the 7 explanatory variables Transformed Unemployment Rate, Transformed Poverty Rate, Transformed Annual Growth Rate, Transformed GDP Growth Rate, Transformed Death Rate, Transformed Depression Rate and Transformed Population Growth Rate.**

<u>Hence , the new estimated model is</u>

$$\mathbf{CR\_N_i = -0.006644 - 0.370121 * UR_i - 0.196313 * PR_i - 0.076411 * AR_i - 0.04764 * GR_i + 6.30857 * DEATHR_i - 29.6992 * DEPR_i - 4.05894 * PGR_i + \hat{u}_i,\ i = 1(1)27} \qquad \mathbf{(6.1)}$$

<u>INTERPRETATION:</u>

From the data computed, we observe that the p-values of $\hat{b}_2$, $\hat{b}_3$, $\hat{b}_4$, $\hat{b}_5$ and $\hat{b}_7$ are large indicating these parameters are insignificant although the value of r^2 comes out to be moderate. If r^2 value is moderate, it clearly means that the regression fitted is more or less accurate and hence implies that all the regression coefficients should be of moderate value. But here, 5 some regression coefficients (mentioned above) come out to be insignificant. Hence, it implies that some error has definitely occurred in the model that we have fitted and hence the regression model we have fitted is not a good one.

<u>Now, our primary tasks are:</u>

- **To find out the causes those are making disturbances in our model.**
- **To immediately fix the causes those are making disturbances in our model.**

- **If necessary, to change our present model to a better one.**

One of the primary causes can be Multicollinearity. Hence, we check for its presence. So we have to thoroughly examine the transformed explanatory variables for finding out the defects.

❖ MULTICOLLINEARITY IN TIME SERIES DATASET:-

The term **Multicollinearity** originally meant the presence of a perfect or exact linear relationship among some or all of the independent variables in a multiple regression model.

It is necessary to check for multicollinearity while fitting a multiple regression model. Independent variables should be independent among themselves and hence this phenomenon is a problem. If the magnitude of correlation between the independent variables is high, then, it may cause problem while fitting the model. One of the main aims in a regression analysis is to isolate the relationship between the explanatory variable and the dependent variable. The interpretation of a regression coefficient is that it represents the amount of change of the dependent variable for a unit change in an independent variable, **keeping all the other variables constant.** But if **multicollinearity is present among the independent variables**, a change in one independent variable will be associated with shifts in other independent variables also. The higher the magnitude of correlation between the independent variables, the more difficult it is to change one independent variable without changing the other. Also, the estimators will have large variances and covariances making precision difficult and hence the confidence intervals tend to be wider. The other problem is that the t-ratios will be statistically insignificant. Even if r^2 value is high, there is a high probability that some of the regression coefficients will be statistically insignificant. **<u>Hence, it is necessary to detect this problem and to fix it.</u>**

- **<u>DETECTION OF MULTICOLLINEARITY:-</u>**

1). <u>USING SCATTER PLOT:</u> By drawing a scatterplot, we can see a rough picture about the nature of the relationship existing between the independent variables.

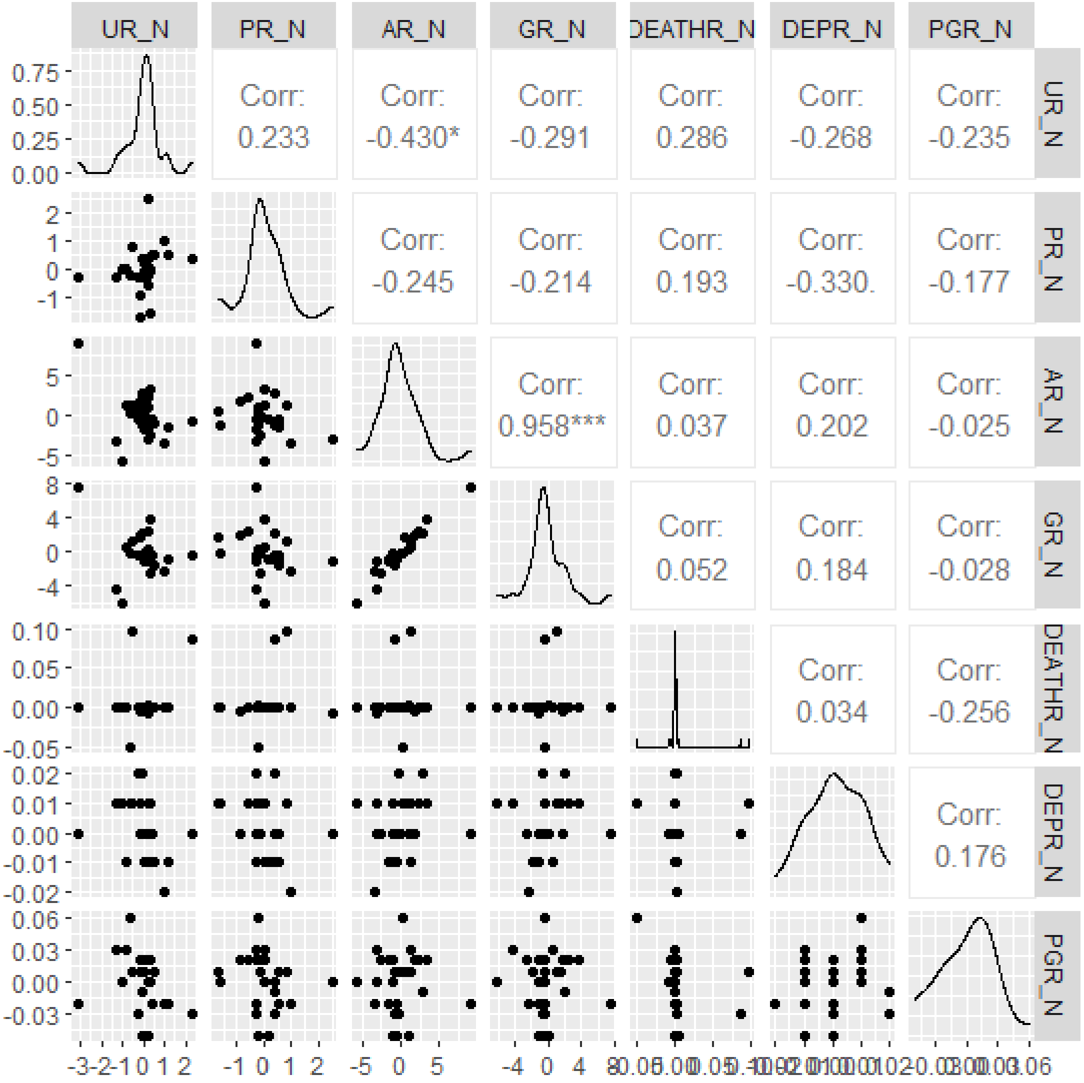


Figure -6.1: Scatterplot among the independent variables.

Where,

CR_N: Transformed Crime Rate, UR_N: Transformed Unemployment Rate, PR_N: Transformed Poverty Rate, GR_N: Transformed GDP Growth Rate, AR_N: Transformed Annual Rate, DEATHR_N: Transformed Death Rate, DEPR _N: Transformed Depression Rate, PGR_N: Transformed Population Growth Rate.

INTERPRETATION:-

From the above scatterplots, we can visualize the nature of relationship between the explanatory variables. The most observable simple correlation coefficients are in between the Transformed GDP Growth Rate and Transformed Annual Growth Rate which is **0.958**. Hence, the above mentioned pair is highly correlated between themselves although this fact is unknown whether there is any lurking effect present or not.

2). CHECKING FOR THE PRESENCE OF MULTICOLLINEARITY USING FARRAR-GLAUBER TEST:-

This statistical test is basically a set of 3 tests and hence 3 statistics has been used for testing the presence of multicollinearity. The notion of this test is that in a sample, multicollinearity is a **departure of the observed values of the explanatory variables from orthogonality.** They believed that if multicollinearity is perfect, then, the coefficients become indeterminate, and that the inter-correlations among the various explanatory variables can be measured by multiple correlation coefficients and partial correlation coefficients.

- **STEPS TO CARRY OUT THE FARRAR-GLAUBER TEST:-**

i. **The first step is to conduct a Chi - Square test to detect the existence of multicollinearity.**

ii. **The second step is to conduct a F-test to locate the variable(s) are inter-correlated.**

iii. **The final step is a t-test to detect the pair of variables that are responsible for multicollinearity.**

Here, X denotes a 27 x 7 matrix whose first column is UR_N, second column is PR_N, third column is AR_N, fourth column is GR_N, fifth column is DEATHR_N, sixth column is DEPR_N and the last column is PGR_N.

Step-1: THE CHI-SQUARE TEST:-

The multicollinearity problem may be considered as a departure from orthogonality. The stronger the departure from orthogonality, that is, the closer the value of the determinant of the matrix $\mathbf{X^TX}$ **to zero**, the stronger the degree of multicollinearity, and vice-versa. From this fact, Farrar-Glauber developed the following $\boldsymbol{\chi^2}$ **test** for detecting the strength of multicollinearity over the whole set of explanatory variables.

Hypothesis:-

To test H_0: The matrix X is orthogonal against H_1: Not H_0.

Sample:

The samples, collected for the 7 explanatory variables, each are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Test Statistic:-

The Chi-Square test statistic that has been used here is given by

$$\chi^2 = - [n – 1 – (2*k + 5)/6] * \log_e|D|$$

Where ‘n’ denotes the sample size, ‘k’ denotes the total number of independent variables and D = Determinant of X^TX.

Distribution of Test Statistic:-

Under H_0, the distribution of the Chi-Square test statistic is $\chi^2_{(k*(k-1))/2}$, that is Chi-square distribution with degrees of freedom (k*(k-1))/2.

Let d = (k*(k-1))/2.

Here, n=27, k=7, implies, d= 21.

Hence, the distribution of the Chi-Square test statistic is χ^2_{21}.

Testing Rule:-

If the p-value is found to be high, we accept that the dataset is free from multicollinearity, that is, the null hypothesis is accepted else we reject the null hypothesis.

Computation:-

Table-6.2: Overall Multicollinearity (MC) Diagnostics (From R software)

Diagnostic Measures	**MC Results**	**Detection**
Farrar Chi-Square:	**81.3767**	**1**
Red Indicator:	**0.3018**	**0**
Sum of Lambda Inverse:	**42.0078**	**1**
Theil's Method:	**0.4158**	**0**
Condition Number:	**9.3252**	**0**
Determinant |X'X|:	**0.0283**	**0**

Where

1 → indicates MULTICOLLINEARITY is detected by the test.

0 → indicates MULTICOLLINEARITY is not detected by the test.

Decision and Conclusion:-

We can see that the results given by R, clearly tells us that multicollinearity has been detected in the dataset by the Chi-Square test. In the light of the given data, we do not have enough evidence to reject H_1 and hence we accept it.

Step-2: The F test

The second test is an F test done for the location of multicollinearity. Here, we have to calculate among the explanatory variables and to test for the statistical significance of these multiple correlation coefficients.

Hypothesis:

To test H_0: UR_N is not intercorrelated with (PR_N, AR_N, GR_N, DEATH_NR, DEPR_N and PGR_N) against H_1: Not H_0

That is, to test H_0: $r^2_{1.234567} = 0$ against H_1: Not H_0,

Where '**$r_{1.234567}$**' denotes the multiple correlation coefficient of UR_N based on the other 6 explanatory variables mentioned above.

Sample:

The samples, collected for the 7 explanatory variables, each are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Test Statistic:-

The F test statistic that has been used here is given by

$$\mathbf{F = (r^2_{1.234567} / 1 - r^2_{1.234567}) * ((\ n\text{-}k\ ) / (k\text{-}1))}$$

Where, 'n' is the sample size and k is the number of explanatory variables .

Here, n=27, k=7.

Distribution of the Test Statistic:-

Under H_0, the distribution of the test statistic follows $\mathbf{F_{6,20}}$, that is, a F distribution with degrees of freedom 6 and 20 respectively.

Testing Rule:-

If the p-value is found to be very small, we accept the fact that UR_N is intercorrelated with the other explanatory variables, that is, the null hypothesis is rejected and vice-versa.

NOTE: -

We have shown the testing procedure for only one explanatory variable – Transformed Unemployment Rate. Using the exact same way, we perform the testing for the other 6 explanatory variables mentioned above and found the results by performing the required calculations using R software.

Computation:-

Table-6.3: Overall Multicollinearity (MC) Diagnostics (From R software):-

Explanatory Variables	F_{obs}	Decision	Conclusion
UR_N	1.9762	Acceptance of H_0	0
PR_N	1.2451	Acceptance of H_0	0
AR_N	18.8032	Rejection of H_0	1
GR_N	16.3307	Rejection of H_0	1
DEATHR_N	1.2466	Acceptance of H_0	0

DEPR_N	**1.2364**	**Acceptance of H_0**	**0**
PGR_N	**1.1697**	**Acceptance of H_0**	**0**

Where

1 → indicates the given explanatory variable is intercorrelated with the other variables.

0 → indicates the given explanatory variable is intercorrelated with the other variables.

Decision and Conclusion:-

We can see that the results given by R, clearly tells us that the two explanatory variables are intercorrelated among them, that is, the Transformed Annual Growth Rate and the Transformed GDP Growth Rate by the F test.

Step-3: The t - test

The third test is a t – test whose aim is the detection of the explanatory variables that are causing multicollinearity. The partial correlation coefficients among the explanatory variables are calculated and their statistical significance is being tested.

Hypothesis:-

To test H_0: $r_{12.34567} = 0$ against H_1: not H_0

Where '$r_{12.34567}$' denotes the partial correlation coefficient between UR_N and PR_N eliminating the effects of the other 5 explanatory variables mentioned above.

Sample:-

The samples, collected for the 7 explanatory variables, each are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Test Statistic:-

The t test statistic that has been used here is given by

$$\mathbf{t = (r_{12.34567} * \sqrt{n - k}\ ) / \sqrt{1 - z}}$$

Where z = $r^2_{12.34567}$, 'n' is the sample size and 'k' is the number of explanatory variables.

Here, n = 27, k = 7.

Distribution of the Test Statistic:-

Under H_0, the distribution of the test statistic follows $\mathbf{t_{20}}$, that is**, a t distribution with degrees of freedom 20.**

Testing Rule:-

If the p-value is found to be very small, we accept the fact that the variables UR_N and PR_N are intercorrelated and are responsible for the problem of multicollinearity in the model, that is, the null hypothesis is rejected and vice-versa.

NOTE: - We have shown the testing procedure for only one pair of explanatory variables – Transformed Unemployment Rate and Transformed Poverty Rate. Using the exact same way, we perform the testing for the other 20 pairs taking two explanatory variables

(mentioned above) each time and found the results by performing the required calculations using R software.

Computation:-

Table-6.4: Pair-Wise t testing

Variable Pairs	**p-value**	**Decision**	**Conclusion**
(UR_N, PR_N)	**0.8764**	**Acceptance of H_0**	**0**
(UR_N, AR_N)	**0.005**	**Rejection of H_0**	**1**
(UR_N, GR_N)	**0.002**	**Rejection of H_0**	**1**
(UR_N, DEATHR_N)	**0.1562**	**Acceptance of H_0**	**0**
(UR_N, DEPR_N)	**0.3649**	**Acceptance of H_0**	**0**
(UR_N, PGR_N)	**0.3739**	**Acceptance of H_0**	**0**
(PR_N, AR_N)	**0.6018**	**Acceptance of H_0**	**0**
(PR_N, GR_N)	**0.7582**	**Acceptance of H_0**	**0**
(PR_N, DEATHR_N)	**0.3824**	**Acceptance of H_0**	**0**
(PR_N, DEPR_N)	**0.1939**	**Acceptance of H_0**	**0**
(PR_N, PGR_N)	**0.6785**	**Acceptance of H_0**	**0**
(AR_N, GR_N)	**<0.0001**	**Rejection of H_0**	**1**
(AR_N, DEATHR_N)	**0.4863**	**Acceptance of H_0**	**0**
(AR_N, DEPR_N)	**0.7573**	**Acceptance of H_0**	**0**
(AR_N, PGR_N)	**0.5251**	**Acceptance of H_0**	**0**
(GR_N, DEATHR_N)	**0.6134**	**Acceptance of H_0**	**0**
(GR_N, DEPR_N)	**0.7129**	**Acceptance of H_0**	**0**

(GR_N, PGR_N)	**0.6243**	**Acceptance of H_0**	**0**
(DEATHR_N, DEPR_N)	**0.4081**	**Acceptance of H_0**	**0**
(DEATHR_N, PGR_N)	**0.4693**	**Acceptance of H_0**	**0**
(DEPR_N,PGR_N)	**0.6034**	**Acceptance of H_0**	**0**

Where

1→ indicates the given pair of explanatory variable is responsible for multicollinearity.

0 → indicates the given pair of explanatory variable is not responsible for multicollinearity.

We can see that the results given by R, clearly shows us 3 pairs of explanatory variables are intercorrelated between them and are the root cause for it is multicollinearity, using the t test. In the light of the given data, we do not have enough evidence to reject H_1 and hence we accept it for the 3 pairs of variables. For the other 18 pairs we accept the null hypothesis, that is, these pairs are not contributing for multicollinearity in the model.

For getting more conclusive results, we need a thorough study and this study is being done with the help of partial correlation coefficients.

Table-6.5: Partial Correlation Coefficients between the explanatory variables:-

	UR_N	**PR_N**	**AR_N**	**GR_N**	**DEATHR_N**	**DEPR_N**	**PGR_N**
UR_N	**1.0000**	**-0.035**	**-0.5744**	**0.4917**	**0.3129**	**-0.2029**	**-0.1992**

PR_N	**-0.035**	**1.0000**	**0.1177**	**0.0696**	**0.1959**	**-0.2879**	**-0.0936**
AR_N	**-0.5744**	**0.1177**	**1.0000**	**0.9635**	**0.1566**	**-0.0699**	**-0.1431**
GR_N	**0.4917**	**0.0696**	**0.9635**	**1.0000**	**-0.1140**	**0.083**	**0.1105**
DEATHR_N	**0.3129**	**0.1959**	**0.1566**	**-0.1140**	**1.0000**	**0.1857**	**-0.1627**
DEPR_N	**-0.2029**	**-0.2879**	**-0.0699**	**0.083**	**0.1857**	**1.0000**	**0.1172**
PGR_N	**-0.1992**	**-0.0936**	**-0.1431**	**0.1105**	**-0.1627**	**0.1172**	**1.0000**

Where the notations have their respective meanings mentioned above.

FINDINGS:

From the t-test we have obtained 9 pairs that are causing multicollinearity. Again, we check by finding out the partial correlation coefficients and observe that out of those 3 pairs, 2 pairs **(UR_N,AR_N) ,(AR_N,GR_N) and are having high magnitudes of partial correlation coefficients underlined in Table-6.3**.

FINAL CONCLUSION FROM THE FARRAR-GLAUBER TEST:-

By critical analysis, it is found that if the explanatory variable GR_N is removed from the model, it gives a better fit than the previous case. It implies that the Transformed GDP Growth Rate is the root cause of multicollinearity.

- **FITTING OF THE REVISED MODEL:-**

Now we fit another model after removing the multicollinearity from the dataset.Hence, our new model is given by

$CR\_N_i = \beta_0 + \beta_1 * UR\_N_i + \beta_2 * PR\_N_i + \beta_3 * AR\_N_i + \beta_4 * DEATHR\_N_i + \beta_5 * DEPR\_N_i + \beta_6 * PGR\_N_i + u_i$, $i = 1(1)27$,

where β_0, β_1 , β_2 , β_3 , β_4 , $\beta_5$and β_6 are the constants involved in the regression determined by the method of least squares **based on the given data and u_i is the error term associated with the model . 'β_0' is the intercept of the regression equation. 'β_j' is called the partial correlation coefficient of CR_N on j th explanatory variable for fixed values of the remaining 6 explanatory variables, j = 1(1)6.**

We get the values of the estimates of the parameters involved in our model using the R software.

Table-6.6: Diagnostics of the revised model

Parameters	Parameter Estimates	Standard Error	T_{obs}	p-value
β_1	**-0.39222**	**0. 14141**	**-0.2774**	**0.0117**
β_j	**-0.19923**	**0.15090**	**-1.320**	**0.2017**
β_3	**-0.1223**	**0.04604**	**-2.656**	**0.0152**
β_4	**6.44802**	**4.39004**	**1.469**	**0.1574**
β_5	**-29.9772**	**12.0389**	**-2.490**	**0.0217**
β_6	**-4.18979**	**4.24998**	**-0.986**	**0.3360**

Also, the value of the coefficient of determination (r^2) comes out to be **0.4477.**

The new fitted model is:

CR_N_i =-0.00702 -0.39222 * UR_N_i - 0.19923 * PR_N_i – 0.1223 * AR_N_i + 6.44802 * $DEATHR\_N_i$ -29.9772 *$DEPR\_N_i$ - 4.18979 * PGR_N_i + $\hat{u}_i$, i = 1(1)27. (6.2)

INTERPRETATION:-

From the data calculated, it is observed out of 6 parameters, 3 parameters come out to be statistically significant. **HENCE, WE CAN INTERPRET THAT THE PROBLEM OF MULTICOLLINEARITY HAS BEEN SUCCESFULLY REMOVED.**

- **CHECKING FOR AUTOCORRELATION USING THE REVISED MODEL:-**

Since our model has been modified, now, our task is to check again, the presence of autocorrelation in the dataset. Again, using the similar testing procedure as done above, BREUSCH-GODFREY test has been performed to check by how much amount we are successful in removing the autocorrelation from the dataset using our revised model.

Sample:-

The samples, collected for the 6 explanatory variables and the response variable, each are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Testing Rule:-

If the p-value of the test is found to be very small, we accept the presence of autocorrelation in the model, that is, the null hypothesis is rejected else we accept the null hypothesis.

Computation:-

P-value of the test statistic = 0.04 (from R software).

Decision and Conclusion:-

Hence, we accept the null hypothesis. In the light of the given data, we can say that the dataset is now free from the problem of autocorrelation fully. Hence, our assumed model is now correct with respect to the autocorrelation assumption.

CHAPTER 7

❖ HETEROSCEDASTICITY IN TIME SERIES DATASET:-

Originally, heteroscedasticity means "unequal scatter". In time-series regression analysis, we usually use the term heteroscedasticity in the context of the residuals or error term. Specifically, heteroscedasticity is a systematic change in the spread of the residuals over the range of the values measured.

Above we have defined all the assumptions of Classical Linear Model, out of which one of them is the homoscedasticity or constant variance: that is, the variance of the error term remains same regardless of the changes in the values of the regressors. If heteroscedasticity is present in the dataset, then, we cannot use ordinary least square and we have to modify the model. If we still use the OLS method, the estimators will not be BLUE (Best Linear Unbiased Estimator). Standard errors are biased which leads to bias in test statistics and confidence intervals. OLS gives equal weight to all the observations but observations with larger disturbance term contains less information than observations with smaller variance term.

Hence, it is necessary to check for the presence of heteroscedasticity in the dataset.

Possible Reasons for the presence of heteroscedasticity in the dataset:-

- The error learning models: as people learn, their errors of behavior become smaller over time or the number of errors becomes more consistent.
- Companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits.
- As data collecting techniques improved, the errors are likely to decrease.
- Due to the presence of outliers in the dataset.
- The chosen regression model may be incorrect.

❖ CHECKING FOR THE PRESENCE OF HETEROSCEDASTICITY BY BREUSCH-PAGAN-GODFREY TEST:-

This is one of the most successful tests which aim at detecting the presence of heteroscedasricity. The model which is being used is given by

$CR\_N_i = \beta_0 + \beta_1 * UR\_N_i + \beta_2 * PR\_N_i + \beta_3 * AR\_N_i + \beta_4 * DEATHR\_N_i + \beta_5 * DEPR\_N_i + \beta_6 * PGR\_N_i + u_i$, $i = 1(1)27$,

where β_0, β_1 , β_2 , β_3 , β_4 , $\beta_5$and β_6 are the constants involved in the regression determined by the method of least squares **based on the given data and u_i is the error term associated with the model . 'β_0' is the intercept of the regression equation. 'β_j' is called the partial correlation coefficient of CR_N on j th explanatory variable for fixed values of the remaining 6 explanatory variables, $j = 1(1)6$.**

Assume that the error variance $\boldsymbol{\sigma^2_i}$ is described as

$\boldsymbol{\sigma^2_i = f\,(d_0 + d_1 * X_{1i} + d_2 * X_{2i} + d_3 * X_{3i} + d_4 * X_{4i} + d_5 * X_{5i} + d_6 * X_{6i}),\ i = 1(1)27,}$

That is, $\boldsymbol{\sigma^2_i}$ is some function of the nonstochastic X variables; some or all of the X_{ji} 's, j=1(1)6 can serve as the explanatory variables, where i=1(1)27. Specifically, assume that

$\boldsymbol{\sigma^2_i = d_0 + d_1 * X_{1i} + d_2 * X_{2i} + d_3 * X_{3i} + d_4 * X_{4i} + d_5 * X_{5i} + d_6 * X_{6i},\ i = 1(1)27,}$ that is, $\boldsymbol{\sigma^2_i}$ **is a linear function of the X_i 's, i=1(1)6.**

<u>Basic Test Procedures:-</u>

<u>Step-1:</u> The first step is to estimate the equation (*) by OLS and then to obtain the residuals

$\hat{u}_1, \hat{u}_2, \ldots.., \hat{u}_{27}$.

<u>Step-2:</u> The second step is to obtain $\boldsymbol{\tilde{\sigma}^2 = \Sigma_{i=1(1)27}\, \hat{u}_i^2 / 27}$.

<u>Step-3:</u> The third step is to construct variables p_i , where p_i is defined as $p_i = \boldsymbol{\hat{u}_i^2 / \tilde{\sigma}^2}$ **,i=1(1)27.**

<u>Step-4:</u> The fourth step is to regress:

$$p_i = d_0 + d_1 * X_{1i} + d_2 * X_{2i} + d_3 * X_{3i} + d_4 * X_{4i} + d_5 * X_{5i} + d_6 * X_{6i} + v_i, \qquad (7.1)$$

where v_i is the residual term of the series, i= 1(1)27.

Explained Sum of Squares (ESS) is obtained for <u>the equation (7.1)</u>.

<u>HYPOTHESIS:-</u>

To test H_0: $d_1=d_2=d_3=d_4=d_5= d_6= 0$ against H_1: Not H_0, that is, to test

H_0: σ^2_i is homoscedastic against H_1: Not H_0, i = 1(1)27.

Sample:

The samples, collected for each of the 6 explanatory variables are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Test Statistic:-

The test statistic is given by

$$\mathbf{T = ESS / 2}$$

Distribution of Test Statistic:-

Under H_0, the distribution of the test statistic is $\chi^2_{(m-1)}$ **asymptotically,** that is Chi-square distribution with degrees of freedom (m-1) where 'm' denotes the number of parameters to be estimated in the model.

Here, m=7, hence, the distribution of the test statistic is χ^2_6 **asymptotically.**

Testing Rule:-

If the p-value of the test is very small, we accept the presence of heteroscedasticity in the model, that is, the null hypothesis is rejected else we accept the null hypothesis.

Computation:-

P-value of the test statistic = 0.6901 (from R software).

Decision and Conclusion:- Since the p-value of the test is very high, in the light of the given data, we accept our null hypothesis. Hence, the dataset is free from the problem of

heteroscedasticity and our assumed model is correct with respect to the homoscedasticity assumption.

Since, all of the assumptions are being taken care of with respect to the dataset, hence, now, we can construct the model appropriate to the dataset.

The fitted model is:

CR_N_i =-0.00702 -0.39222 * UR_N_i - 0.19923 * PR_N_i – 0.1223 * AR_N_i + 6.44802 * $DEATHR\_N_i$ -29.9772 *$DEPR\_N_i$ - 4.18979 * PGR_N_i + $\hat{u}_i$, i = 1(1)27.

Observed Value = Estimated Value +Residuals

⇨ **Residuals = Observed Value - Estimated Value**

Hence, we calculate the residuals of the regression model. Our next task is to check whether the residuals follow a Normal distribution or not.

CHAPTER 8

❖ CHECKING FOR THE NORMALITY OF THE RESIDUALS:-

One of the main assumptions for fitting OLS in a linear model is that the residuals follow a Normal distribution. It is not wise to fit an OLS if this assumption is not satisfied. Since we have fitted a linear model using OLS, it is necessary for us to check whether the residuals follow a Normal distribution. If the residuals do not follow a Normal distribution, then, our fitted model is not good and we have to proceed in a different way.

❖ TESTING FOR RESIDUAL NORMALITY USING SHAPIRO-WILK TEST:-

HYPOTHESIS:-

To test H_0: $\hat{e}_i \sim N\ (0,\sigma^2)$ against H_1:Not H_0, that is, to test

H_0: Residuals are normally distributed against H_1: Not H_0, i = 1(1)27.

Sample:

The samples collected for each of the 6 explanatory variables are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Test Statistic:-

The test statistic is given by

$\mathbf{T=(\Sigma_{i=1(1)27}\ a_i * e_{(i)})^2 / \Sigma_{i=1(1)27}\ (e_i - \bar{e})^2}$, where $e_{(i)}$ pertains to the i^{th} largest value of the error terms and a_i denotes those values which are calculated using the means, variances and covariances of the e_i.

Distribution of Test Statistic:-

Under H_0, the distribution of the test statistic follows a Standard distribution.

Testing Rule:-

If the p-value of the test is very small, we accept that the errors are non-normally distributed in the model, that is, the null hypothesis is rejected **else we accept the null hypothesis.**

Computation:-

P-value of the test statistic = 0.3952 (from R software).

Decision and Conclusion:-

Since the p-value of the test is very high, we accept the null hypothesis. In the light of the given data, we can say that the residuals are distributed normally in the model. Hence, our assumed model is correct with respect to this assumption.

CHAPTER 9

GOODNESS OF FIT:-

The goodness of fit of a statistical model is often used to describe how well the statistical model fits a set of observations. The measures of goodness of fit summarize the discrepancy between the observed values and the fitted values obtained from the model that is being fitted. In regression analysis, coefficient of determination (r^2) is used to measure the goodness of fit.

In statistics, r^2 denotes the proportion of the explained variability in the model, that is, it is the proportion of the variance in the dependent variable that can be predicted from the independent variables.

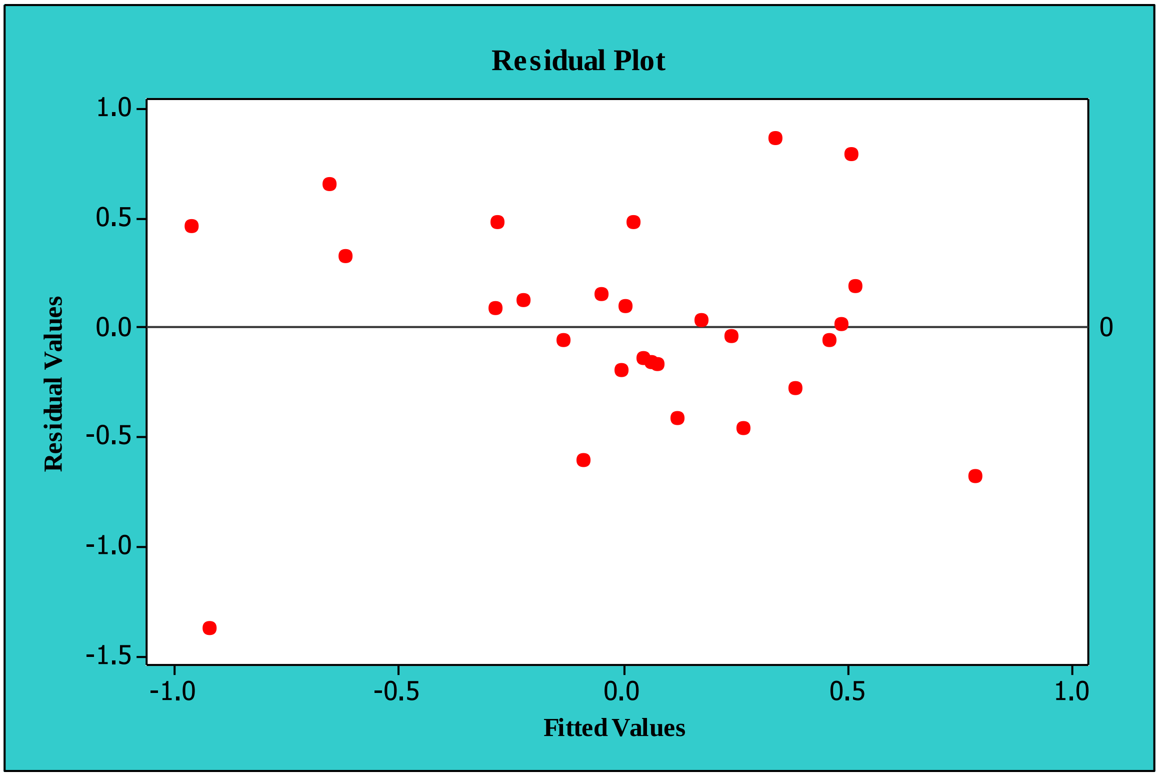
From (6.2), using R software,

We have obtained the value of $r^2 = 0.4477$.

It implies that approximately 45% of the total variance in the response is explained by the linear regression of CR_N on UR_N, PR_N, AR_N, DEATHR_N, DEPR_N and PGR_N given by (6.2).

<u>There are some basic limitations of r^2.</u> It cannot be used to determine whether the coefficient estimates and predictions are biased. For further critical analysis, we check the residual plot. A residual plot is a graph where the residuals are plotted along the y-axis and the fitted values of the response are plotted along the x-axis.

<u>Figure-9.1:</u> Residual Plot



INTERPRETATION:-

From the graph, we can observe that the residuals are more or less randomly scattered with a slight clustering around the centre. The clustering takes place may be due to the presence of some heteroscedasticity and autocorrelation which is not taken into account or explained by our model. Also, one of the explanatory variables PGR_N cannot be completely transformed into stationary series which can be one of the causes for the clustering. Overall, the residual plot is moderately good.

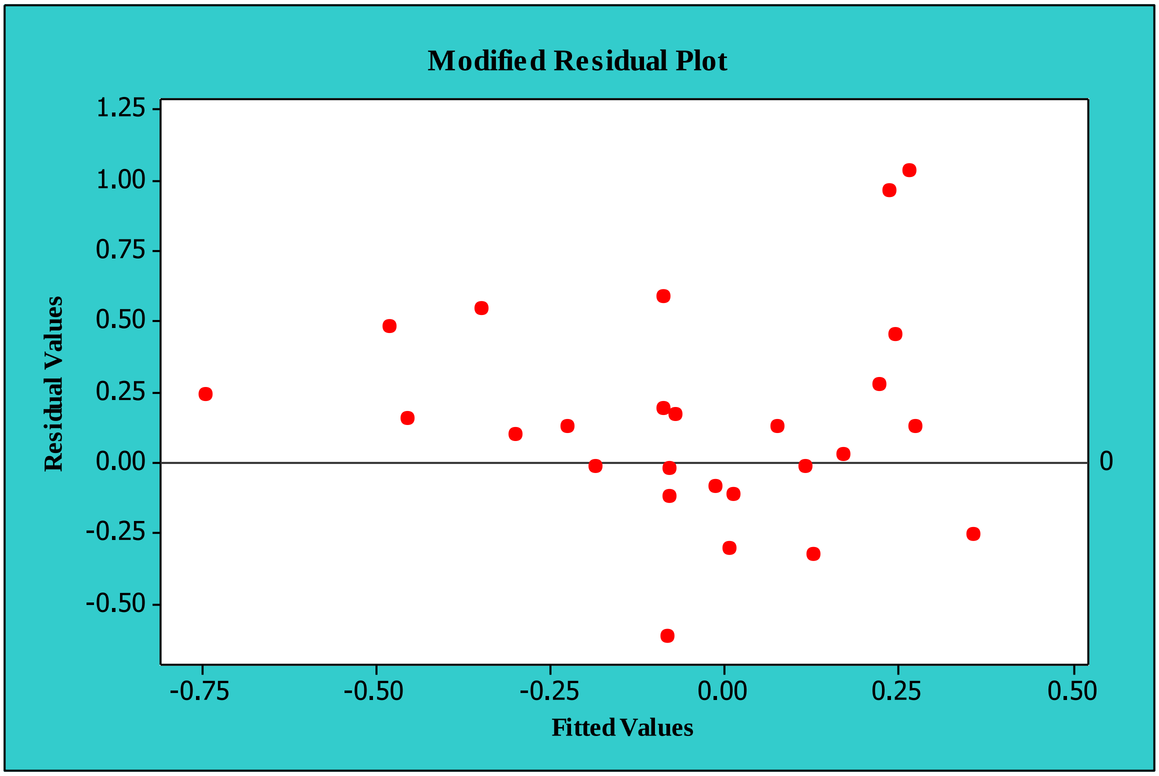
Using the standardized residual also, the point corresponding to the year 2002 (point at the bottom left corner) came out to be **-2.87552** and hence the point is suspected to be a potential outlier.

Hence, we remove the values for all the variables corresponding to the year 2002 and plot a new multiple linear regression model. The modified fittel model is:

CR_N$_i$ =-0.07108 -0.25838 * UR_N$_i$ - 0.1357 * PR_N$_i$ – 0.06959 * AR_N$_i$ + 4.62952 * DEATHR_N$_i$ – 22.8811 * DEPR_N$_i$ – 1.63167 * PGR_N$_i$ + û$_i$, i = 1(1)26.

After removing the point and again fitting the linear regression model results in a high decrease of **r^2 = 0.3362**. Hence, the point is definitely an influential point. Also, we check the residual plot for the new fitted multiple linear regression model.

Figure-9.2: Modified Residual Plot



INTERPRETATION:-

From the graph, we can clearly interpret that even after removing the potential outlier; still a slight clustering around the centre is observed. The clustering takes place may be due to the presence of some heteroscedasticity and autocorrelation which is not taken into account or explained by our model. Also, one of the explanatory variables PGR_N cannot be completely transformed into stationary series which can be one of the causes for the clustering. Overall, the residual plot is moderately good.

Hence, from the above conclusions, it is not advisable to remove the point. Hence, we keep our previous model given by equation (6.2).

CHAPTER 10

❖ HYPOTHESIS TESTS IN MULTIPLE LINEAR REGRESSION MODEL:-

If the residual terms of the multiple linear regression model satisfy the normality assumption, two types of hypothesis tests can be carried out. The tests are as follows:-

1). Test for the significance of the regression model: This test is used to check the significance of the whole regression model.

2). t test: This test is used to check the significance of the individual regression coefficients of the regression model.

Now our first task is to check for the significance of our multiple linear regression model.

❖ TESTING FOR THE SIGNIFICANCE OF THE REGRESSION MODEL:-

The test for the significance of the overall regression model is carried out using the analysis of variance technique. The test is used to check if there exists a linear statistical relationship between the response variable and at least one of the predictor variables.

HYPOTHESIS:-

To test H_0: $\beta_1= \beta_2= \beta_3= \beta_4= \beta_5= \beta_6=0$ against H_1: Not H_0, that is, it is equivalent to test,

H_0: the regression on the explanatory variables is equivalent against H_1: Not H_0.

Sample:-

The samples, collected for the 6 explanatory variables and the response variable, each are of size 27 and 27 data points correspond to 27 years from 1993-2019.

<u>Test Statistic:-</u>

The test statistic are given by

F_0 = MSR / MSE, where MSR is the regression mean square and MSE is the error mean square.

Now, MSR = SSR/k and MSE = SSE/ n – (k+1), where SSR and SSE respectively denotes the sum of squares due to regression and error sum of squares. Here, 'n' denotes the total number of observations and 'k' denotes the total number of explanatory variables.

<u>Distribution of Test Statistic:-</u>

Under H_0 , the distribution of the test statistic are

$SSR/\hat{\sigma}^2 \sim \chi^2_k$

$SSE/\hat{\sigma}^2 \sim \chi^2_{n-(k+1)}$

Hence, our required test statistic is

$F_0 = (SSR/k)/(SSE/(n-k-1)) \sim F_{k,n-(k+1)}$

Here, n = 27, k=6.

<u>Testing Rule:-</u>

We reject H_0 if the p-values obtained for at least one of the explanatory variables are very small.

<u>Computation:-</u>

The following calculations have been computed using **R software**.

Table-10.1: Analysis of Variance Diagnostics

Sources of Variation	Sum of Squares	Mean Squares	Degrees of freedom	p-value
UR_N	**0.275**	**0.275**	**1**	**0.3496**
PR_N	**0.006**	**0.006**	**1**	**0.8933**
AR_N	**1.709**	**1.709**	**1**	**0.0270**
DEATHR_N	**0.481**	**0.481**	**1**	**0.2197**
DEPR_N	**2.086**	**2.086**	**1**	**0.0158**
PGR_N	**0.291**	**0.291**	**1**	**0.3360**
Residuals	**5.998**	**0.2999**	**20**	

Decision and Conclusion:-

From the above table (10.1), we can interpret that since the p-values corresponding to the two variables AR_N and DEPR_N are very small, hence, these two are significant and hence**, we reject the null hypothesis and conclude that the regression model is significant.**

Now our second task is to check for the individual significance of the regression coefficients of the regression model using the t-test.

❖ TESTING FOR THE SIGNIFICANCE OF THE INDIVIDUAL PARAMETERS:-

The t-test is carried out to check the significance of individual regression coefficients in the multiple linear regression model. A regression model becomes more effective and useful if a significant variable is added to the model while the regression model loses its importance if an insignificant variable is added to it.

HYPOTHESIS:-

To test H_0: β_j=0 against H_1: Not H_0, j = 1(1)6.

Sample:-

The samples, collected for the 6 explanatory variables and the response variable, each are of size 27 and 27 data points correspond to 27 years from 1993-2019.

Test Statistic:-

The test statistic are given by

$t_0 = \hat{\beta}_j / SE(\hat{\beta}_j)$, where $\hat{\beta}_j$ is the estimated value of β_j and SE ($\hat{\beta}_j$) is the Standard Error of $\hat{\beta}_j$.

Distribution of Test Statistic:-

Under H_0, the distribution of the test statistic are

$t_0 = \hat{\beta}_j / SE(\hat{\beta}_j) \sim t_{n-(k+1)}$,

Where, 'n' denotes the total number of observations and 'k' denotes the total number of explanatory variables.

Here, n = 27, k=6.

Testing Rule:-

We reject H_0 if the p-value obtained is very small.

Computation:-

The following calculations have been computed using **R software**.

Table-10.2: t-test diagnostics

Parameters	**Parameter Estimates**	**Standard Error**	**T_{obs}**	**p-value**	**Decision**
β_1	**-0.39222**	**0. 14141**	**-0.2774**	**0.0117**	**1**
β_j	**-0.19923**	**0.15090**	**-1.320**	**0.2017**	**0**
β_3	**-0.1223**	**0.04604**	**-2.656**	**0.0152**	**1**
β_4	**6.44802**	**4.39004**	**1.469**	**0.1574**	**0**
β_5	**-29.9772**	**12.0389**	**-2.490**	**0.0217**	**1**
β_6	**-4.18979**	**4.24998**	**-0.986**	**0.3360**	**0**

Where

1 → indicates that the parameter is significant.

0 → indicates that the parameter is insignificant.

Conclusion:-

From the above results, we observe that 3 out of the 6 parameters, that is, (**β_1, β_3, β_5**) of the multiple linear regression model are statistically significant.

❖ REASONS FOR CHOOSING THIS MODEL:-

Although value of r^2 is moderate, the above plotted residual plot is more or less random (Perfectly random residual plot is not feasible in real time series data). Also, 3 out of the 6 parameters of the multiple linear regression model are statistically significant. Statistically significant coefficients represent the mean change in the dependent variable for a unit change in the independent variable, keeping all the other independent variables fixed. It is possible to draw valid conclusions using the above fitted model. Hence, I chose this model.

CHAPTER 11

❖ INTERPRETATION OF THE PARAMETERS:-

It is important to interpret the parameters in a model since it reflects the relative importance of the predictors in the definition of the model itself.

i). Interpretation of β_1 (the parameter associated with UR_N)

From the table-10.2, we see that the estimated value of $\boldsymbol{\beta_1}$ is **-0.39222.** If UR_N increases by 1, then, CR_N will decrease by the amount **-0.39222**, keeping all the other explanatory variables fixed. Also, the p-value came out to be **0.0117** which is very small indicating that the parameter is significant.

ii). Interpretation of β_2 (the parameter associated with PR_N)

From the table-10.2, we see that the estimated value of $\boldsymbol{\beta_2}$ is **-0.19923.** If PR_N increases by 1, then, CR_N will decrease by the amount **-0.19923**, keeping all the other explanatory variables fixed. Also, the p-value came out to be **0.2017** which is large indicating that the parameter is insignificant.

iii). Interpretation of $\boldsymbol{\beta_3}$ (the parameter associated with AR_N)

From the table-10.2, we see that the estimated value of $\boldsymbol{\beta_3}$ is **-0.1223.** If AR_N increases by 1, then, CR_N will decrease by the amount **-0.1223**, keeping all the other explanatory variables fixed. Also, the p-value came out to be **0.052** which is very small indicating that the parameter is significant.

iv). Interpretation of $\boldsymbol{\beta_4}$ (the parameter associated with DEATHR_N)

From the table-10.2, we see that the estimated value of $\boldsymbol{\beta_4}$ is **6.44802.** If DEATHR_N increases by 1, then, CR_N will increase by the amount **6.44802**, keeping all the other explanatory variables fixed. Also, the p-value came out to be **0.1574** which is large indicating that the parameter is insignificant.

v). Interpretation of $\boldsymbol{\beta_5}$ (the parameter associated with DEPR_N)

From the table-10.2, we see that the estimated value of $\boldsymbol{\beta_5}$ is **-29.9772.** If DEPR_N increases by 1, then, CR_N will decrease by the amount **-29.9772**, keeping all the other explanatory variables fixed. Also, the p-value came out to be **0.0217** which is very small indicating that the parameter is significant.

vi). Interpretation of β_6 (the parameter associated with PGR_N)

From the table-10.2, we see that the estimated value of $\boldsymbol{\beta_6}$ is **-4.18979.** If PGR_N increases by 1, then, CR_N will decrease by the amount **-4.18979**, keeping all the other explanatory variables fixed. Also, the p-value came out to be **0.3360** which is large indicating that the parameter is insignificant.

Here, we have finally developed our model and interpreting the model parameters based on the second differences of the variables. It is because if we take the original variables, non-stationarity is present and hence all the inferences will be wrong. Also, from our new fitted model based on the second differences we cannot revert back to the original variables, since it is very difficult to handle it properly. Hence, we are keeping the second differences and interpreting on the basis of it.

CHAPTER 12

❖ APPENDIX:-

```
library (quantmod)

library (tseries)

library (lmtest)

library (GGally)

library (mctest)

library (ppcor)
```

```
## Checking the Stationarity

rm (list=ls())

d= read.csv (file="C:/Users/Personal/Desktop/DISSERTATION/data.csv")

d

U= d [, 1]

P= d [, 2]

A= d [, 3]

G= d [, 4]

C= d [, 5]

D= d [, 6]

DEP=d [, 7]
```

```
PGR= d [, 8]

pp.test (U, alternative='stationary', type='Z(t_alpha)')

pp.test (P, alternative='stationary', type='Z(t_alpha)')

pp.test (A, alternative='stationary', type='Z(t_alpha)')

pp.test (G, alternative='stationary', type='Z(t_alpha)')

pp.test (C, alternative='stationary', type='Z(t_alpha)')

pp.test (D, alternative='stationary', type='Z(t_alpha)')

pp.test (DEP, alternative='stationary', type='Z(t_alpha)')

pp.test (PGR, alternative='stationary', type='Z(t_alpha)')


## Removal of Stationarity using difference method

U1=diff (U, lag=1,differences=2)

P1=diff (P, lag=1,differences=2)

G1=diff (G, lag=1,differences=2)

A1=diff (A, lag=1,differences=2)

C1=diff(C, l ag=1,differences=2)

D1=diff (D, lag=1,differences=2)
```

```
DEP1= diff (DEP, lag=1,differences=2)

PGR1=diff (PGR, lag=1,differences=2)


## Rechecking for Stationarity

pp.test (U1, alternative='stationary', type='Z(t_alpha)')

pp.test (P1, alternative='stationary', type='Z(t_alpha)')

pp.test (G1, alternative='stationary', type='Z(t_alpha)')

pp.test (A1, alternative='stationary', type='Z(t_alpha)')

pp.test (C1, alternative='stationary', type='Z(t_alpha)')

pp.test (D1, alternative='stationary', type='Z(t_alpha)')

pp.test (DEP1, alternative='stationary', type='Z(t_alpha)')

pp.test (PGR1, alternative='stationary', type='Z(t_alpha)')


## Checking for presence of Autocorrelation

data= read.csv (file="C:/Users/Personal/Desktop/IMPORTANT/DISSERTATION/
Transformed_dataset.csv")

data

UN=data [, 2]
```

```
PN=data [, 3]

AN=data [, 4]

GN=data [, 5]

CN=data [, 1]

DEATHN=data [, 6]

DEPN=data [, 7]

PGRN=data [, 8]

bgtest (CN~UN+PN+AN+GN+DEATHN+DEPN+PGRN,data=data)


## Fitting of the Model

fit= lm(CN~UN+PN+AN+GN+DEATHN+DEPN+PGRN)

summary(fit)


## Checking for the presence of Multicollinearity

ggpairs(data)

omcdiag(fit)

imcdiag(fit, method = "VIF", vif = 10)
```

```
pcor(data, method="pearson")


## Checking for the presence of Heteroscedasticity

bptest(CN~UN+PN+AN+DEATHN+DEPN+PGRN,data=data)


## Rechecking for the presence of Autocorrelation

bgtest(CN~UN+PN+AN+DEATHN+DEPN+PGRN,data=data)


## Fitting of the Revised Model

fit.n=lm(CN~UN+PN+AN+DEATHN+DEPN+PGRN)

summary(fit.n)


## Checking for the Normality of Residuals

C.f=-0.00702-0.39222*U-0.19923*P-0.12230*A+6.44802*DEATH-29.99715*DEP-4.18979*PGR

C.f

R=C-C.f

R
```

```
shapiro.test(R)



## Tests for the significance

summary(aov(CN~UN+PN+AN+DEATHN+DEPN+PGRN))
```

CHAPTER 13

❖ ACKNOWLEDGEMENT:-

Real learning comes from a practical work. I am very grateful to St. Xavier's College for providing me such an opportunity to be engaged in a project work based on my own area of interest in the field of applied statistics. I am really thankful to my supervisor for her continuous guidance, monitoring and constant support throughout the course of this project. Without her encouragement and help it would not have been possible to complete this project.

❖ BIBLIOGRAPHY:-

The book that is used in the project is:

- Basic Econometrics – Damoder N. Gujarati & Dawn C. Porter.
- Introduction to Econometrics – G.S. Maddala.

The sites that are used in the project are:

- www.macrotrends.com
- www.statista.com
- www.cdc.gov
- www.wikipedia.com
- https://statisticsbyjim.com
- https://faculty.washington.edu
- https://www.statisticshowto.com
- https://rpubs.com
- https://datascienceplus.com