RISHIRAJ SUTAR 221392

MRINMOY SAHA 221352

BIPLAB NASKAR 221299

MANJEET CHAUDHARY 221346

PRAVIN RAMAN BHARTI 221375

**UNDER THE SUPERVISION**

**OF**

**DR. AMIT MITRA**

Time series analysis of electric production ip index

**1: INTRODUCTION:**

The analysis of time series data has become an increasingly valuable tool in understanding and predicting trends in various domains, ranging from economics to finance, climate science to healthcare.

Electric production is a critical component of the industrial sector, serving as a foundational indicator of economic activity and energy demand. The Industrial Production Index, a key economic indicator published by the Federal Reserve, measures the overall performance of the U.S. industrial sector, including manufacturing, mining, and electric production. Within this broader index, the Electric Production IP Index specifically tracks the output and performance of the electric utilities sector.

In this project, we intend to:

* Impute the missing values in the data using interpolation method.
* Detect the presence of Trend and Seasonality in the data.
* Remove the systematic components in order to obtain the residuals.
* Check the Stationarity of the residuals using Dicky-Fuller test.
* Fit an ARMA model to the data.
* Forecast the future outcomes using the model.

# **2: DEALING WITH MISSING VALUES:**

There are a total of 8 missing values in the data and we need to estimate these values using a suitable technique.

In Time Series Analysis, we cannot impute the missing values with global mean or median because the time series data might have some seasonality or trend and therefore, these methods can cause biasness to the data. Rather we would be using the interpolate function for imputations. Linear Interpolation simply means to estimate a missing value by connecting dots in a straight line in increasing order. In short, It estimates the unknown value in the same increasing order from previous values. The default method used by Interpolation is Linear. So, while applying it, we need not specify it.

**3: CHECKING THE RANDOMNESS OF THE DATA:**

First, we need to check for the randomness of our data. For this purpose, we use ‘Turning Point Test’.

H0: Purely Random Series.

against

H1: Non-Random Series.

We say Yi is a turning point if Yi > Yi-1 and Yi > Yi+1 or Yi<Yi-1 and Yi < Yi+1.

For our problem the turning point is

T=

Here the test statistic is given by:

Under Ho:

Z = ~ N(0,1)

where, E(T) = and V(T)

Test criterion:

We reject H0 at level of significance α if |Zobs|>Zα/2

Here we get |Z|= 13.412 > 1.96 (Z0.025) and hence, our null hypothesis is rejected at 5% level of significance.

Thus, our data is not random and there is presence of deterministic components in our model.

# **4: VERIFYING THE PRESENCE OF TREND AND SEASONALITY:**

## 4.1: Verifying presence of Trend in the model:

Here we wish to verify whether there is presence of trend in our model or not.

For this purpose, we will conduct the **Relative Ordering** test which involves

H0: There is no trend in the model.

Against

H1: There is trend in the model.

R: Number of discordant pairs.

E(R) =

R>E(R): indicates falling trend.

R<E(R): indicates rising trend.

Kendell’s T, the rank correlation coefficient is given as:

T =

Under H0:

E(T)= 0,

Var(T)=

Here our test statistic is:

Z = ~ N(0, 1)

Test criterion:

Here H0 is rejected if |Zobs|>Zα/2 at level of significance α.

Value of R= 11890, E(R) = 34503 which indicates rising trend in our model.

|Zobs|= 9.922 > 1.96(Z0.025), and hence our null hypothesis is rejected.

We conclude that trend is present in our model.

## 4.2: Verifying the presence of Seasonality:

**Friedman's non-parametric test** is utilized to examine the presence of seasonality in de-trended data.

: No seasonality

against

: Seasonality is present

For a specific month, the occurrence of higher ranks (ranking for a given particular month and year) indicates seasonal peaks. In the absence of any seasonality in the monthly data, the ranks assigned to the monthly values should exhibit a random distribution. To establish the test statistic, the calculation of monthly totals () is a requirement.

Here, with "r" representing the month and "c" denoting the year,

we assume a chi-square distribution

X ∼ χ2(r−1) under the null hypothesis.

Test criterion:

Here H0 is rejected if Xobs > χ2(r−1).(α)

Consequently, we reject the null hypothesis, leading to the conclusion that seasonality is indeed present in the data.

In Our case r = 12

c = 30

Observed value of chi-square = 244 > 19.675 i.e., χ2(r−1); α

Thus, we conclude that seasonality is present in the model.

## **5: THE PERIOD OF SEASONALITY:**

In order to eliminate the seasonality from the data, we first need to determine the period of seasonality.

Let, us consider the model:

*Yi = mt + St + et*

*where* mt : Trend Component

St : Seasonal Component

Suppose ‘d’ be the period of seasonality in the data, then St-d = St = St-d and

▽dYt = mt – mt-d + et – et-d

Therefore, the new series ▽dYt is free from the Seasonality component.

Now, the task is to test the presence on seasonality in ▽dYt using Friedman’s Non-Parametric test for seasonality.

In the testing procedure, if we consider d = 12, the value of the Friedman’s Non-Parametric test statistic is 13.26 which is less than the critical point. Thus, we accept H0 and can conclude that the time series ▽12Yt is free from the seasonality component.

Now, the next task is to estimate and eliminate the deterministic components, namely, Trend and Seasonality from the data.

**6: ESTIMATION AND ELIMINATION OF TREND AND SEASONALITY FROM THE TIME SERIES DATA:**

We have considered the Additive Model: *Yi = mt + St + et*

were*,* mt: Trend Component

St: Seasonal Component

Assumptions:

E(et) = 0

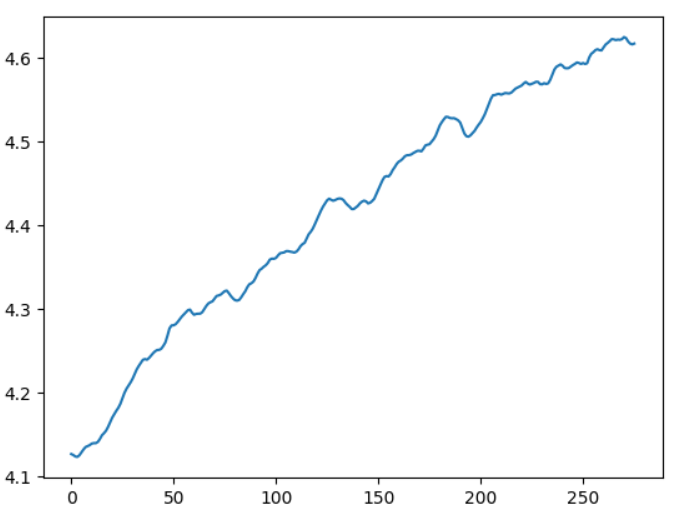
Var(et) = σ2 (finite)

Our main objective here is to estimate the trend and the seasonal components and eventually eliminate them from the underlying data in order obtain the irregular component. Since, the Trend is not constant for a particular year, we would use the Rapid Trend Method.

**STEP 1:**

Since, we have found out that the period of seasonality (d) is 12, we would obtain a rough estimate of the trend using the MA filter; filter window such that the seasonal component is eliminated and the noise is dampened. Therefore, the corresponding estimate of trend is given by:

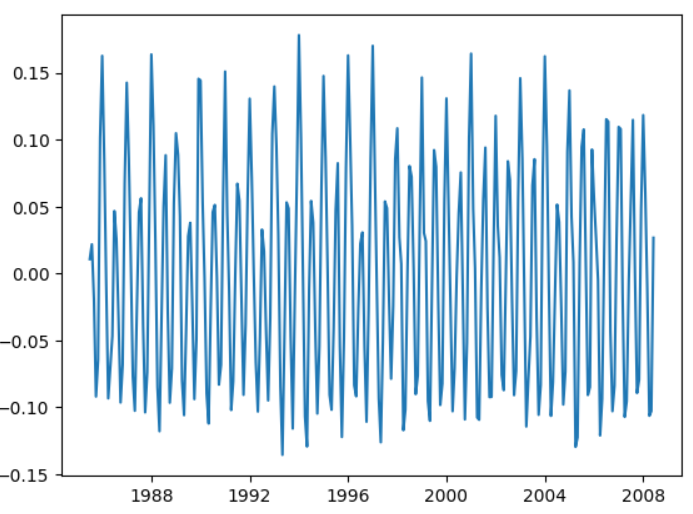
= + ( )



Trend line

**STEP 2:**

Subtract the trend estimates from the actual data to obtain the detrended data



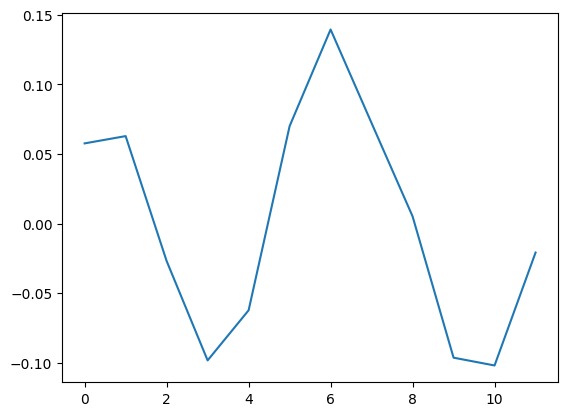
Detrended Data

**STEP 3:**

Combine the averages (say wk) of the deviations { } over J = 24 years

The estimates of the seasonal factors are given as:

; k = { 1, 2 , …. , d}

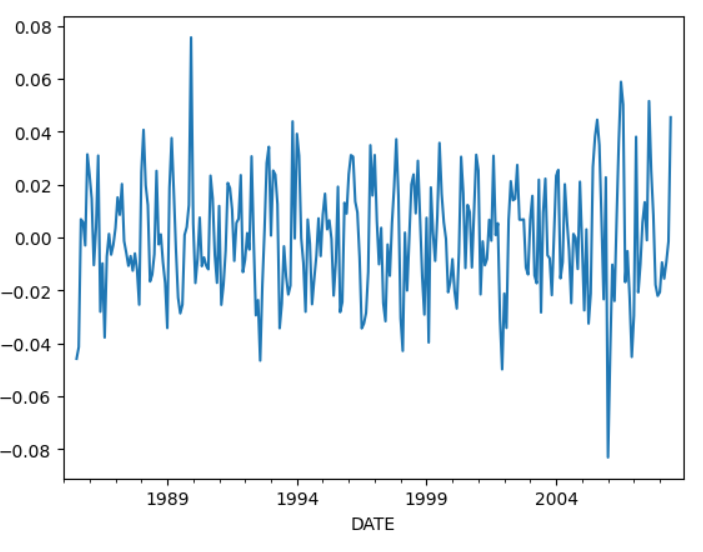


seasonal indices

**STEP 4:**

De-seasonalise the data and repeat the procedures until the trend as well as the seasonal component have been completely eliminated from the data.

So, according to the test procedure we first de-trend and de-seasonalise our data and then recheck if the new time series data contains trend of seasonality. After performing the testing, we find that the deterministic components are completely eliminated and we are left with only the random error component.



**7: ANALYSIS OF THE IRREGULAR COMPONENTS**

# 7.1: Test for stationarity: augmented dickey-fuller test.

We need to ensure that the irregular component obtained previously is stationary or not then we can proceed towards fitting of our model. For this purpose, we need to perform the Augmented Dicky Fuller test (ADF).

An Augmented Dickey–Fuller test (ADF) is actually an augmented version of the Dickey–Fuller test used for large and more complicated version of time series models. It actually involves test for a unit root in the sample of time series. The ADF statistic is generally a negative number in the test. With the increase in negativity the rejection of hypothesis becomes fairly strong that is there is a unit root at some confidence level.

Here the unit root test is performed under the null hypothesis(H0) ϒ= 0 against the alternative hypothesis(H1) ϒ< 0.

So, the test statistic is:

DFτ =

## **Computation:**

Dickey-Fuller Test Statistic: -7.57

Lag order: 15

p-value: 2.72 x e-11

Since, p-value < 0.05, we reject Ho, thereby concluding that the data obtained after removal of trend and seasonality, is stationary.

Now to further use the data for forecasting we have to, at first we need to model the data properly. We have to check in which of the category does the data belong.

## 7.2: Test for White Noise:

We would apply Ljung-Box test for detecting the WHITE NOISE of irregular components.

Here, H0: data is independently distributed

Against

H1: serial correlation is present in the data

Under the Null hypothesis, the test statistic is given as:

where n is the sample size and is the sample auto correlation at lag k and h is the number of lags used in the test statistic.

Test criterion:

Reject H0 if, = 18.31

Qobs = 84.094

p-value = 3.519e-14

Since, p-value < 0.05, we reject Ho at 5% level of significance, thereby concluding that the data is not a white noise process.

## **8: MODEL BUILDING:**

8.1: MA MODEL:

An MA (Moving Average) process is a statistical model used in time series analysis to describe the behaviour of a time series data set. The MA process is characterized by its mathematical structure, which involves the weighted average of past white noise (random) values to generate the observed data points in a time series.

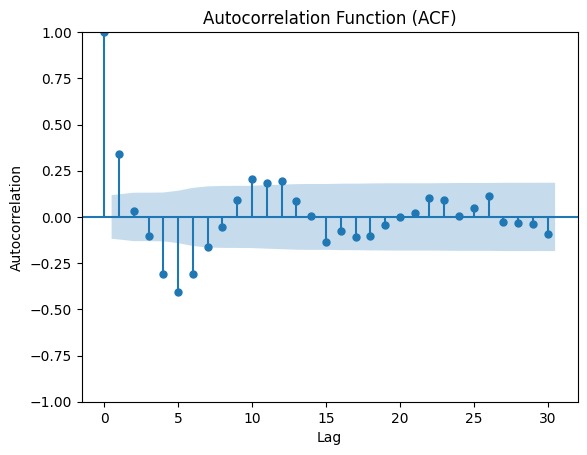
**The general form of a stationary MA(q) model:**

where µ is the mean of the series, the θ1, θ2,..., θq are the parameters of the model and ɛ1 , ɛ2 , … ɛt-q are the white noise.

are white noise error terms.

**Determining the order of the MA Model:**

To determine the order of the MA model, we would plot the Autocorrelation Function (ACF) of the irregular component.



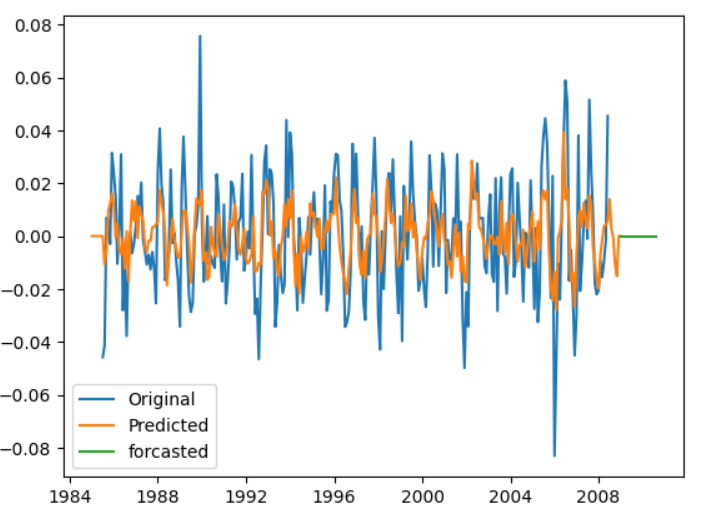
From the plot, we observe that there are spikes at lag 1 to lag 5 and the spikes are non-significant from thereafter. So, we would fit a MA model of order 2 to order 5 and would select the model for which, the AKAIKE INFORMATION CRITERION (AIC) is minimum.

The mathematical form of AIC is given as: AIC =

Where, K = (q + 1) is the number of model parameters and Ln(L) is the log-likelihood of the model.

|  |  |
| --- | --- |
| MODEL | AIC |
| MA (1) | -1354 |
| MA (2) | -1353 |
| MA (3) | -1351 |
| MA (4) | -1393 |
| MA (5) | -1409 |
| MA (6) | -1394 |

From the above table, we can observe that the AIC is minimum for MA(5) model.



We can observe that the model has approximated the irregular component quite well.

8.2: AR MODEL:

An AR (Auto-Regressive) model is a type of statistical model used in time series analysis. It is designed to describe and forecast a time series by relating the current value of the series to its past values. The primary idea behind an AR model is that the current value of a time series is a linear combination of its past values and a white noise error term. The term "autoregressive" indicates that the process regresses on its own past values.

**The general form of a stationary AR(p) model:**

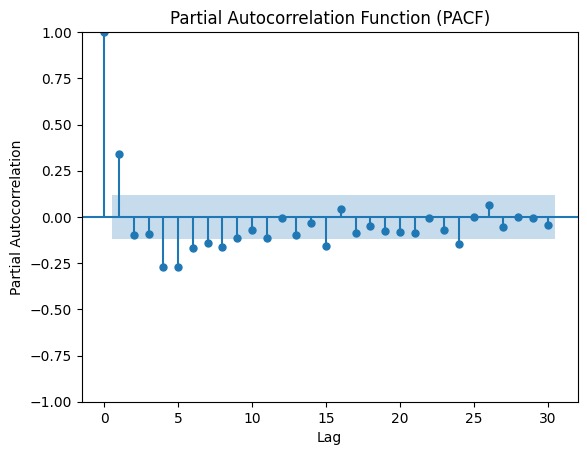
Where, φ₁, φ₂, ..., φp are the autoregressive coefficients that represent the weights applied to the past values of the time series (lags)

Xt-1 , Xt-2 , ..., Xt-p are the lagged values of the time series at time points preceding "t" by 1, 2, ..., p time periods.

t is a white noise error term at time "t."

**Determining the order of the AR Model:**

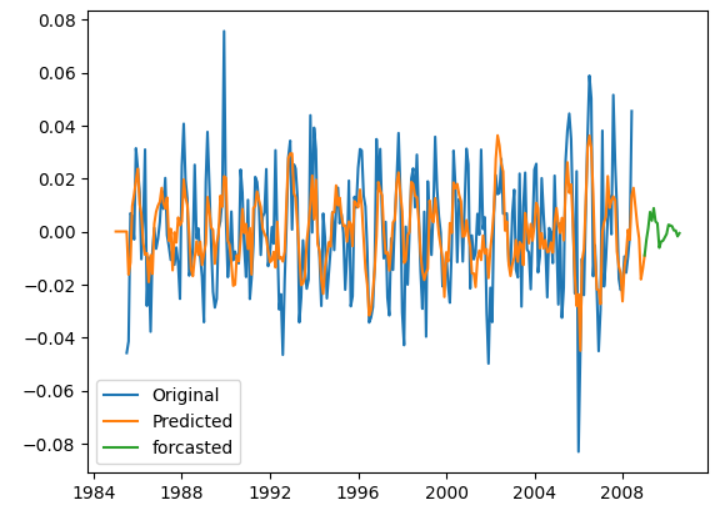
If we observe the ACF plot, we would see that the spikes dampens after the lag 5 and decreases gradually as the lag increases. This is a good indication that the data follows a AR process. However, to determine the order of the process, we would look into the PACF plot.



From the plot, we observe that there are spikes at lag 2 to lag 5 and the spikes are non-significant from thereafter. So, we would fit a AR model of order 2 to order 5 and would select the model for which, the AKAIKE INFORMATION CRITERION (AIC) is minimum.

|  |  |
| --- | --- |
| MODEL | AIC |
| AR (1) | -1148 |
| AR (2) | -1203 |
| AR (3) | -1209 |
| AR (4) | -1215 |
| AR (5) | -1214 |
| AR (6) | -1212 |

From the above table, we can observe that the AIC is minimum for AR(4) model. Therefore, we would fit a AR(4) model to our data.



8.3: ARMA MODEL

An ARMA (Auto-Regressive Moving Average) model is a time series forecasting method used in statistics and econometrics to analyse and predict data points that are collected sequentially over time. It combines two fundamental components: autoregressive (AR) and moving average (MA) processes.

**The general form of a stationary ARMA(p , q) model:**

Where, φ₁, φ₂, ..., φp , θ1, θ1,..., θq are the parameters of the model

ɛ1 , ɛ2 , … ɛt-q are the white noise.

Xt-1 , Xt-2 , ..., Xt-p are the lagged values of the time series at time points preceding "t" by 1, 2, ..., p time periods

By looking at the ACF and PACF, we observe that the spikes are beyond the acceptance band up to a certain lag. After that, the spikes dampen gradually as the lag increases. This is an indication that the optimum model may be ARMA.

**Determining the order of the ARMA Model:**

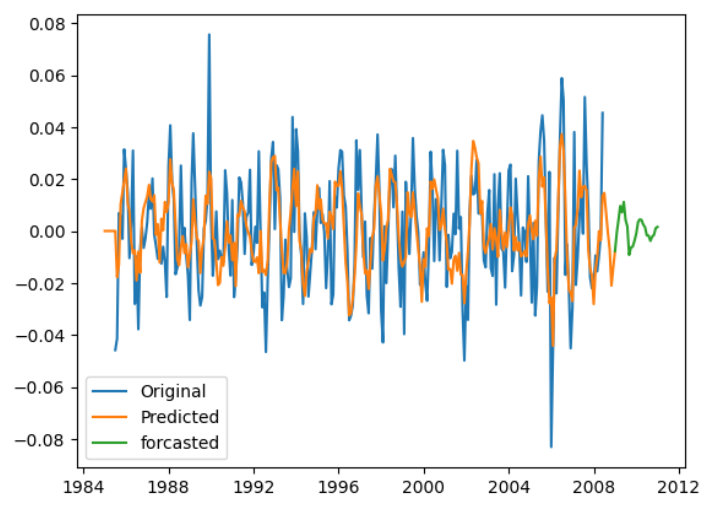
Here, for the different values of p and q, we would compute the AIC and consider that model for which the corresponding AIC is minimum.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Q  P | 3 | 4 | 5 | 6 |
| 3 | -1410 | -1418 | -1410 | -1412 |
| 4 | -1412 | -1420 | -1423 | -1405 |
| 5 | -1414 | -1421 | -1422 | -1409 |
| 6 | -1409 | -1415 | -1421 | -1410 |

From the above table, we observe that ARMA(4 , 5) model has the minimum AIC = -1423

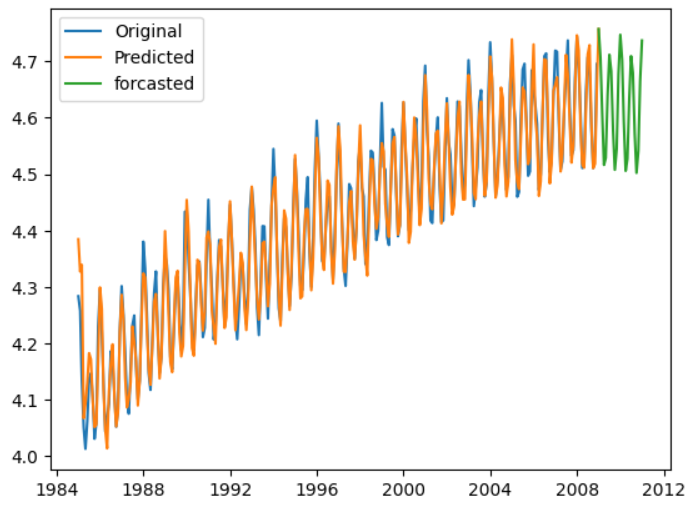
Hence, our estimated value of p = 4

and estimated value of q = 5



8.4: FINAL MODEL:

Since, ARMA(4 , 5) has the lowest AIC value, we would consider it as the final model and forecasting would be done based on this model.



**9: FORECASTING**

One of the primary objectives of building a model for a time series is to be able to forecast the values for that series at future time points. We also want to assess the precision of the forecast. Therefore, using this model, let us forecast the next 10 month’s Electric Production IP Index.

|  |  |  |
| --- | --- | --- |
|  | ACTUAL | PREDICTED |
| 1 | 116.83 | 116.36 |
| 2 | 104.42 | 110.57 |
| 3 | 98.56 | 100.67 |
| 4 | 88.19 | 91.45 |
| 5 | 87.53 | 92.58 |
| 6 | 97.23 | 103.48 |
| 7 | 103.91 | 111.19 |
| 8 | 105.74 | 108.04 |
| 9 | 94.88 | 97.19 |
| 10 | 89.29 | 90.66 |

**10: GOODNESS OF FIT TEST:**

Now, we shall perform the goodness of fit test, in order to access how well the forecasted values fits a set of actual observations. [Pearson's chi-square test](https://en.wikipedia.org/wiki/Pearson%27s_chi-square_test) uses a measure of goodness of fit which is the sum of differences between observed and [expected outcome](https://en.wikipedia.org/wiki/Expected_value) frequencies (that is, counts of observations), each squared and divided by the expectation.

Here, H0: the fit is good

against

H1: not a good fit

The test statistic is given as:

(Under Ho)

Reject null hypothesis if Χobs2 > χn-1;α at α level of significance

Decision:

Since, χobs2 = 1.758 < χn-1;α = 16.918., we accept Ho at 5% level of significance, thereby concluding that the fit is good.

**10: CONCLUSION:**

After considering different standard tests and procedures we eliminate the trend and seasonality from the data and we turn our focus on model fitting we consider several models for this purpose using AIC value.

Both ACF and PACF plots show significant spikes at the first few lags and then drops off which suggests an ARMA model. Also we confirm the same using AIC value.

Later we find that the predicted values using this model is close to actual values. The goodness of fit test suggest the same argument. So in all respect we see that the ARMA model is the best fit for our data.

�2=∑�=1�(��−��)��2