Predicting Housing Index in United States Using Time Series Models

Abstract :

This project aims to conduct a time series analysis of the United States Housing Price Index using data that was originally obtained from the Federal Housing Finance Agency (FHFA). The method of maximum likelihood is used to estimate the parameters and to forecast the housing price index for this future. This dataset has data from January 1991 to April 2013. Through the analysis, it is revealed that the ARI(1,1) model explains the trend of Housing Price Index(HPI) more adequately when compared to other simple time series models. Through this model, it is evident how slowly and constantly the HPI increases with time under assumption of the economic conditions, housing inventory, landscape availability and government policies remain constant in the future. This model can be useful for accessing house value by the owners to increase and decrease the values of the house, analyzing real estate investments, policy making and urban planning to monitor house market stability, making informed buying and selling decisions, and for making mortgage and lending decisions.

Introduction :

Data :

The present study used the HPI data released by the Federal Housing Finance Agency (FHFA). The HPI data is collected monthly and segregated as monthly, quarterly and yearly bases. Though this data is originally obtained from FHFA, the original data contains a lot of information regarding various factors such as geographical location, HPI_flavour, HPI_type which can give a way additional information to the analysis. But only a condensed form of this data is taken from the data repositories of University of Texas Tech, which gives a more pre-processed version of the original data. This data contains information about only HP indexes in a time frame of January 1991 to April 2013 i.e one observation from each month from 1991 to 2013. For this analysis only the first 80 samples of the dataset are used i.e data from January 1991 to July 1997. In the end, using the model, the next ten observations are predicted. The following section includes 1) Model Selection, 2) Model Fitting and Diagnostics, 3)Forecasting and 4) Discussions. All statistical analysis are conducted using R

Model Selection:

First when we get the data set, it is necessary to plot the scatter plot of the data to check if there is any change in the variance in the data. The main goal of plotting this graph is to ensure our model has constant variance. If the data does not have constant variance, then we need to perform box cox transformation on the data for the purpose of having constant variance. When plotting the Housing Price Index(HPI) graph, the variance in each interval of time is constant as there are no big oscillations and data points hanging in together in the graph. Hence we do not have to perform box-cox transformations. The scatterplot of Housing Price Index is shown in Figure 1

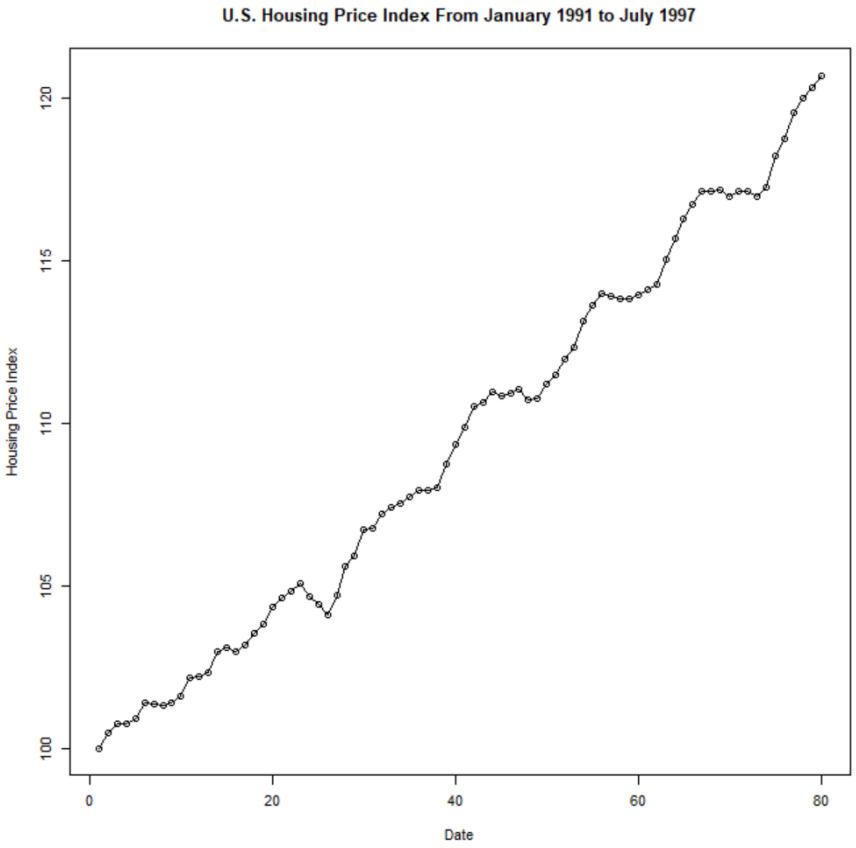


Figure 1 : Scatterplot of Housing Price Index

From the graph, it is depicted that there is an increasing trend as the time changes, which says the process is non-stationary. To confirm this let us check the ACF graph. The ACF graph of the data shows that the autocorrelations of the data are slowly decreasing to zero, which is a sign of non-stationarity. To build a time series model we need to have a stationary process. Hence we do differencing on our data to convert it into a stationary process.

From figure 2 it is evident that the process does not have a trend, which makes it stationary.

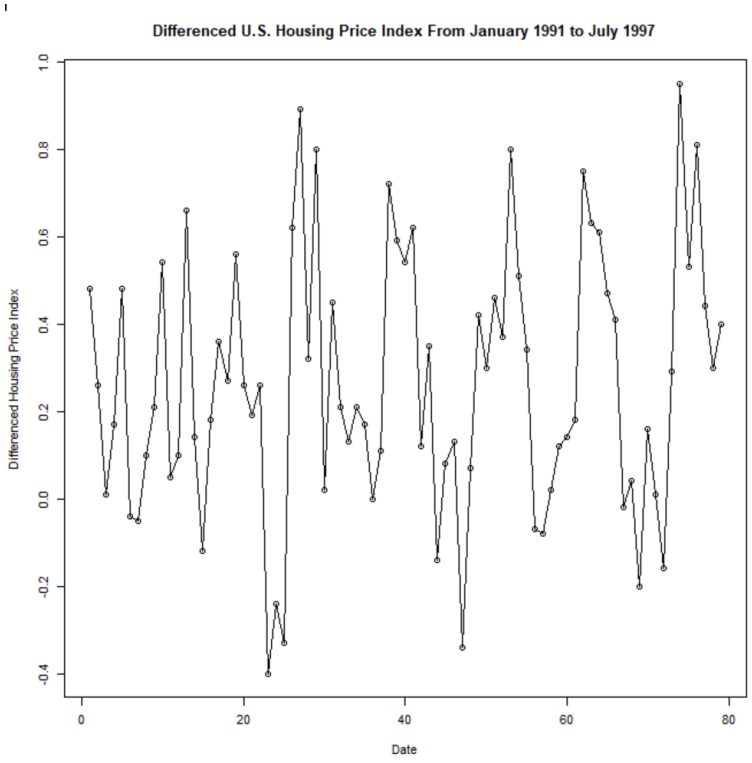


Figure 2: Differenced Scatterplot of Housing Price Index

To confirm stationary of our process, let us perform Augmented Dickey-Fuller (ADF) test, which tells us if the process is stationary or not

Null Hypothesis H_0 : The time series has unit roots(non-stationary)

Alternative Hypothesis H_a : The time series does not have unit roots(stationary)

From the ADF test we see that the p_value is 0.01, which is less than the significance level 0.05

Hence we reject the null hypothesis and state that the process is stationary.

Let us also examine the Autocorrelation Function(ACF) and Partial Autocorrelation(PACF) function graphs of the data.

In the figure we see that the autocorrelation functions are random and are slightly having a sinusoidal trend as the time changes, which indicates it can be an Auto Regressive(AR) model. The AR model's ACF will be a sinusoidal pattern as the time changes. In the PACF figure, we see that it is slightly similar to the ACF graph, but there is a sudden cut off in autocorrelation after lag 1 which again suggests that the process can be an AR model. These two graphs confirm the data is stationary. As our model is stationary after the first difference our d = 1. d is a parameter that gives information about the number of differences the data needs in order to be stationary. Now we need to check which model can explain our data. In order to know this we need to find p and q values. p value indicates the order of AR model and q values indicates the order of MA model. From our PACF graph in Figure, it is clear that p can be 1 as after lag 1 the autocorrelation cuts off to zero. The ACF graph does not show any characteristics of the MA model for this data.

The autocorrelations of the data are as follows:

```
Autocorrelations of series 'diff(data)', by lag

     1      2      3      4      5      6      7      8      9     10
 0.378  0.148 -0.048 -0.296 -0.328 -0.461 -0.322 -0.083  0.090  0.164
```

The partial autocorrelation of the data are as follows:

```
Partial autocorrelations of series 'diff(data)', by lag

     1      2      3      4      5      6      7      8      9     10
 0.378  0.006 -0.124 -0.283 -0.147 -0.333 -0.134 -0.012  0.009 -0.118
```

From these values we can say that there is a strong positive correlation for both ACF and PACF at lag 1 as the values at lag 1 is greater than $\pm\frac{2}{\sqrt{80}}(\pm 0.22)$. To confirm the orders of our model i.e to check which order of a MA or AR model our data belongs to, Extended Autocorrelation Function(EACF) graphs must be examined.

From the EACF graph, we can see there is a possibility of a lot of models. ARI(1,1) and IMA(1,1) can be our candidate models which we can study. There are other models we can study from this EACF graph, but first we can fit the lower parameter models to make models less parsimonious. We can also see the Bayesian information criterion(BIC) graph to see what subsets of models can explain the data. The models with low BIC values are better. From the BIC graph, we get the same information as in the EACF graph. It shows darker shades in the cell of diff.lag1 and error.lag1 which indicates ARI(1,1) model and IMA(1,1) model.

Model Fitting and Diagnostics :

Now that we have two candidate models to build our data on, we can fit these models

ARI(1,1):

The models which become AR(1) after their first differencing are known as ARI(1,1) models. The general form of ARI(1,1) model in backshift notation is given as $(1 - \phi B)(1 - B)Y_t = e_t$ or $\nabla Y_t = \nabla Y_{t-1} + e_t$ where e_t is white noise and ∇Y_t is difference of values at Y_t and Y_{t-1}.

As this is an ARI(1,1) model, we can use three methods to fit the model. The methods used in fitting the model are:

i. Method of Moments

ii. Conditional Least Squares

iii. Maximum Likelihood Estimate.

But in this analysis, we use Maximum Likelihood Estimation to calculate the parameter estimate as it is widely used over most of the time series models and also the method of moments is

inefficient in calculating the estimates for MA models. Henc to maintain uniformity in calculating estimates of all models we use Maximum Likelihood. The results of parameter estimate using Maximum Likelihood is mentioned below:

The parameter estimate of ARI(1,1) model is denoted by $\widehat{\phi}$ and is equal to 0.6486 and the standard error of the estimate is 0.0858. We now check the significance of these estimates.To know the significance of the estimates we need to calculate the confidence interval of the estimates at 0.05 significance level.

95% of confidence interval = $\widehat{\phi} \pm z_{\alpha/2}$*SE

$= 0.6486 \pm 1.96*0.0858$

$= 0.6486 \pm 0.1681$

$= 0.6486 - 0.1681, 0.6486 + 0.1681$

$= [0.4805015, 0.816753]$

The confidence intervals do not contain zero, which means that the parameter estimates are significantly different from zero. Hence we can consider this model for our further analysis. Our ARI(1,1) model is

$$Y_t - Y_{t-1} = 0.6486(Y_{t-1} - Y_{t-2}) + e_t$$

$$\nabla Y_t = 0.6486 \nabla Y_{t-1} + e_t$$

Next we need to check if the model is following independence, normality and is actually adequate or not. Hence we need to perform a diagnostics test on the residuals of the model.

Independence Test:

To perform independence test we use runs test

Null Hypothesis H_0 : All the error terms are independent

Alternate Hypothesis H_a : All the error terms are not independent

p-value = 0.151

Significance level $\alpha = 0.05 < 0.151$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

$\therefore$ We can say error terms are independent of each other. ARI(1,1) follows independence assumption

Normality Test:

To perform normality test we use shapiro.tests

Null Hypothesis H_0 : All the error terms are normally distributed

Alternate Hypothesis H_a : All the error terms are not normally distributed

Here the W value is 0.9351 and the corresponding p-value = 0.4065

Significance level $\alpha = 0.05 < 0.6788$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

$\therefore$ We can say error terms are normally distributed.ARI(1,1) follows normality assumption

When seeing the graphical representation of the residuals using histogram and qq-plot, it is evident that the histogram is normally distributed and in the qq-plot most of the points are on the line qq-line. These two graphs provide strong evidence of normality along with the shapiro test.

Model Adequacy Test:

To test the model adequacy, we use Ljung-Box test

Null Hypothesis H_0 : The ARI(1,1) model is appropriate

Alternative Hypothesis H_a : The ARI(1,1) model is not appropriate.

The p-value is 0.1247, which is greater than the critical value 0.05.

Hence we fail to reject the null hypothesis

$\therefore$ The ARI(1,1) model is appropriate

The graphical representation of the Ljung-box test shown in Figure shows that the residuals of the model are randomly distributed. From the first graph all the standardized residuals are within [-3,3] range and are also random. The sample ACF of residuals plot shows that residuals are approximately uncorrelated, behaving like a white noise . In the Ljung-box statistics most of the residuals are above the margin of error, suggesting there are no outliers.

Our next candidate model is IMA(1,1) model.

IMA(1,1):

To estimate the parameters of the IMA(1,1) model we use the maximum likelihood estimate as it is an effective way of calculating the estimates of MA models. The general form of IMA(1,1) model in backshift notation is given as $(1 - B)Y_t = (1 - \theta B)e_t$ or $\nabla Y_t = e_t - \theta e_{t-1}$

The parameter estimate of IMA(1,1) model is denoted by $\hat{\theta}$ and is equal to 0.5070 and the standard error of the estimate is 0.0879. We now check the significance of these estimates.To know the significance of the estimates we need to calculate the confidence interval of the estimates at 0.05 significance level.

95% of confidence interval = $\hat{\theta} \pm z_{\alpha/2}$*SE

$= 0.5070 \pm 1.96*0.0879$

$= 0.5070 \pm 0.1722$

$= 0.5070 - 0.1722, 0.5070 + 0.1722$

$= [0.3346534, 0.6793362]$

The confidence intervals do not contain zero, which means that the parameter estimates are significantly different from zero. Hence we can perform a diagnostics test on this. Our IMA(1,1) model can be written as

$$\nabla Y_t = e_t - 0.5070 e_{t-1}$$

Independence Test:

Null Hypothesis H_0 : All the error terms are independent

Alternate Hypothesis H_a : All the error terms are not independent

p-value = 0.267

Significance level $\alpha = 0.05 < 0.267$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

$\therefore$ We can say error terms are independent of each other. IMA(1,1) follows independence assumption

Normality Test:

Null Hypothesis H_0 : All the error terms are normally distributed

Alternate Hypothesis H_a : All the error terms are not normally distributed

Here the W value is 0.9351 and the corresponding p-value = 0.3178

Significance level $\alpha = 0.05 < 0.3178$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

$\therefore$ We can say error terms are normally distributed.IMA(1,1) follows normality assumption

The histogram of this model is uniformly distributed and in the qq-plot most of the plots lie on the qq-line which again suggests a good evidence for normality.

Model Adequacy Test:

Null Hypothesis H_0 : The IMA(1,1) model is appropriate

Alternative Hypothesis H_a : The IMA(1,1) model is not appropriate.

The p-value is 0.01132, which is less than the critical value 0.05.

Hence we reject the null hypothesis

$\therefore$ The IMA(1,1) model is not appropriate

From the model diagnostics of both the models we conclude that we take ARI(1,1) for our further analysis as it passes all the diagnostic tests and its parameter estimates are significantly different from zero. The BIC and the EACF graphs also suggest that the ARI(1,1) is a better model. The BIC and AIC values of ARI(1,1) are low. The IMA(1,1) even though it passed normality and independence, it does not pass the Ljung-box test which states that IMA(1,1) is not appropriate for our data. The summary of the Model selection is given below:

	ARI(1,1)	**IMA(1,1)**
Parameter Estimates	0.6486	0.5070
Standard Error	0.0858	0.0879
Parameter Significance	Significant	Significant
Independence	p_value : 0.151	p_value : 0.267
Normality	p_value : 0.6788	p_value : 0.3178
Ljung-Box	Fail to reject	Reject
AIC	37.89	53.24

Model Overfitting:

Now that we have taken ARI(1,1) from our model diagnostics, we now try to see if we fit the model with a slightly more number of parameters, to check if our model explains the data any differently. To overfit any ARIMA(p,d,q) process, we chose either ARIMA(p+1,d,q) or ARIMA(p,d,q+1). Hence we take different models that are one order bigger than our original ARI(1,1). So our overfit models are ARIMA(1,1,1) and ARI(2,1).

Overfitting Model ARIMA(1,1,1):

Generally ARIMA(1,1,1) model is expressed as $(1 - \phi B)(1 - B)Y_t = (1 - \theta B)e_t$

The parameter estimates of ARIMA(1,1,1) are given below:

	$\widehat{\phi}$	$\widehat{\theta}$
Parameter Estimate	0.7844	-0.2377
Standard Error	0.1019	0.1566
Confidence Interval	[0.5846475 ,0.9841789]	[-0.5446648, 0.0692134]

Two condition to choose overfitted model over the original model are:

i. When additional parameters are added to the model,there should be a significant change in the standard error and the estimate of the already existing parameters in the model.

ii. The parameter estimates of the new model should be significantly different from zero.

From the table we can see that there is a drastic change in $\widehat{\phi}$ and SE of $\widehat{\phi}$ but we see that the confidence interval of $\widehat{\theta}$ is not significantly different from zero. Hence we do not go further with this model.

Overfitting Model 2 ARI(2,1) :

The parameter estimates of ARI(2,1) are given below: The ARI(2,1) model is expressed as

$(1 - \phi_1 B - \phi_2 B^2)(1 - B)Y_t = e_t$

	$\widehat{\phi_1}$	$\widehat{\phi_2}$
Parameter Estimates	0.5426	0.1643
Standard Error	0.1108	0.1112
Confidence Interval	[0.32546945, 0.7597369]	[-0.05371892, 0.3822518]

Again in the table, that the confidence interval of $\widehat{\phi_2}$ is not significantly different from zero. Hence we do not go further with this model.

After evaluating various models and considering the potential for overfitting, we have determined that the most appropriate and parsimonious model for the given data is the first-order autoregressive integrated model ARI(1,1).

Forecasting:

From the entire dataset we only used 80 samples for model building. For forecasting we are going to predict the next 10 values i.e housing price index from 1997-09-01 to 1998-06-01.

Using ARI(1,1) model our prediction of the testing data with their predicted interval is given below:

Date	Observed	Predicted	Lower PI	Upper PI	MMSE
1997-09-01	120.47	120.959	120.366	121.553	0.303
1997-10-01	120.88	121.128	119.984	122.271	0.584
1997-11-01	121.04	121.237	119.559	122.915	0.856
1997-12-01	121.1	121.308	119.129	123.486	1.111
1998-01-01	121.22	121.354	118.712	123.995	1.348
1998-02-01	122.13	121.383	118.314	124.452	1.566
1998-03-01	123	121.403	117.939	124.867	1.767
1998-04-01	123.89	121.415	117.585	125.246	1.954
1998-05-01	124.81	121.423	117.251	125.596	2.129
1998-06-01	126	121.429	116.936	125.921	2.292

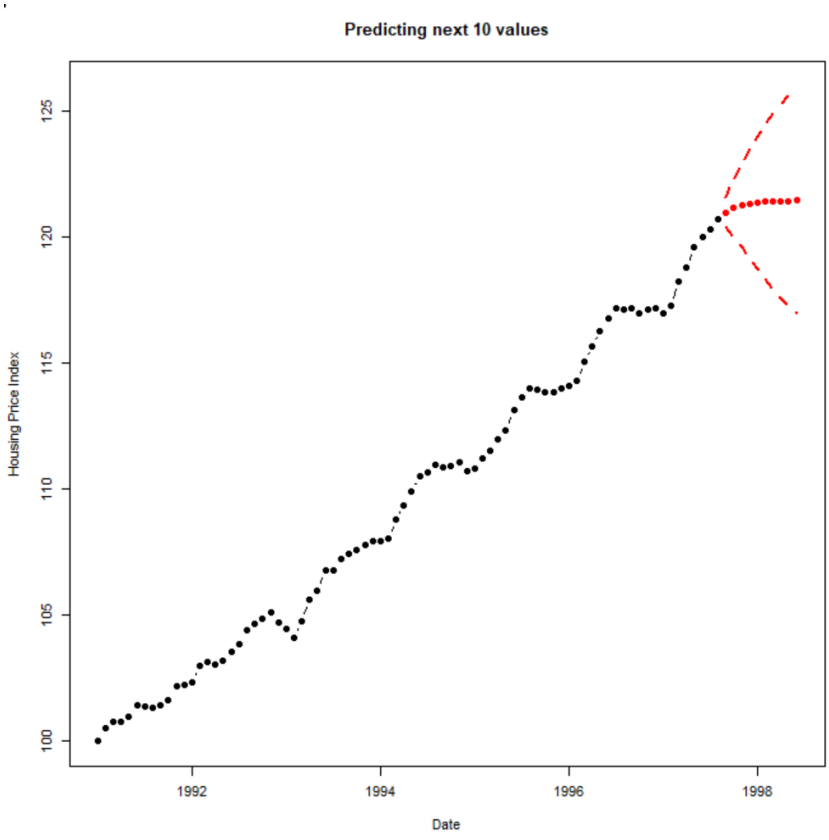
The results of the predicted values are very close to the original values suggesting that the model is doing a good job in predicting the housing index. After lag 1 we see most of the values are the

same which is actually explaining how ARI(1,1) model prediction works. In the ARI(p,d) model all the predictions will converge to the mean after the p lags. Hence we see the most of the values after lag 1 are nearly 121. Hence our final model for the data Housing Index price prediction is

$$Y_t - Y_{t-1} = 0.6486(Y_{t-1} - Y_{t-1}) + e_t$$

$$\nabla Y_t = 0.6486 \nabla Y_t + e_t$$

The graph of predicting the future values is given in Figure:



Figure

According to the predicted values, the housing index price in the United States in 1997-09-01 is 120.47 which is very near to the original values. Even though the model is doing a good job predicting these values, it is to be noted that the model is trained on only 80 samples of the data, so as the time increases the model just generalizes the predicted value. Hence it is necessary to train the model with a huge amount of data. This data might also change depending on various factors, so there may be cases where the predicted values are very different from the original data. Hence to build a more robust model, we need to consider different factors like location of the place, increase in prices of different commodities and many other factors that directly or indirectly affect the housing price index.

Conclusion :

In this analysis, we applied time series forecasting techniques to predict the Housing Price Index (HPI) for the United States. By examining the historical HPI data and considering various time series models, the ARI(1,1) model, a first-order autoregressive integrated model, is the most suitable for capturing the underlying dynamics of the housing market.

Using the ARI(1,1) model, predictions for the next 10 months are generated, starting from September 1997 and extending until September 1999. The forecasted values provide an indication of the expected trajectory of the Housing Price Index during this period. The predictions are accompanied by lower and upper prediction intervals, which offer a range of plausible values for the HPI based on the model's uncertainty.

The forecasted results suggest a continuation of the overall upward trend in the Housing Price Index, indicating a positive outlook for the U.S. housing market during the specified time frame. However, it is important to interpret these predictions cautiously and consider the limitations of

the analysis which includes limited data, geography of the place and other factors that might influence the housing price index.

References

Case, K. E., & *Shiller*, R. J. (2003). Is there a bubble in the housing market? *Brookings Papers on Economic Activity*, *2003* (2), 299-362.

Green, R., & *Hendershott*, P. H. (1996). Age, housing demand, and real house prices. *Regional Science and Urban Economics*, *26* (5), 465-480.

Rapach, D. E., & *Strauss*, J. K. (2009). Differences in housing price forecastability across US states. *International Journal of Forecasting*, *25* (2), 351-372.

Glaeser, E. L., *Gyourko*, J., & *Saks*, R. E. (2005). Why have housing prices gone up? *American Economic Review*, *95* (2), 329-333.

Cryer, J. D., & *Chan*, K. S. (2008). *Time series analysis: with applications in R*. Springer.

Appendix

Figure 1 : Housing Price Index Original Data

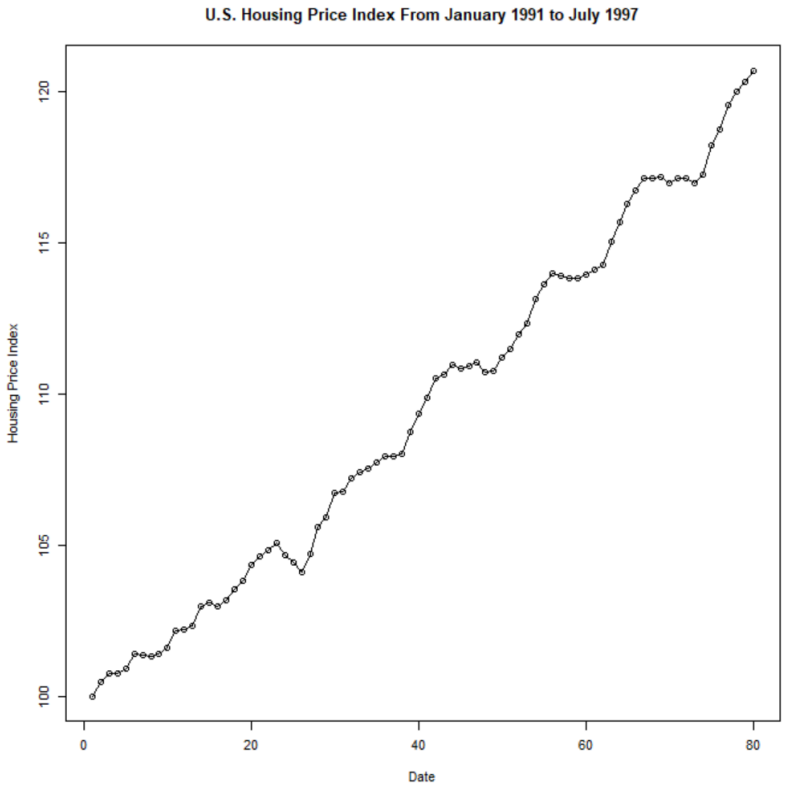


Figure 2 : Sample ACF of Original Data

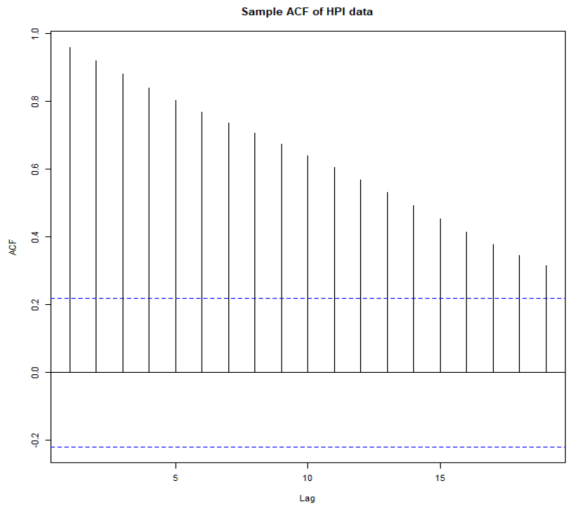


Figure 3 : Sample PACF of Original Data

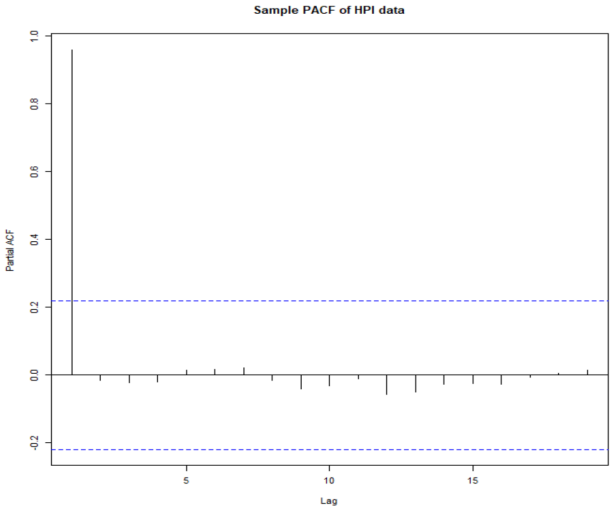


Figure 3 : Differenced HPI Data

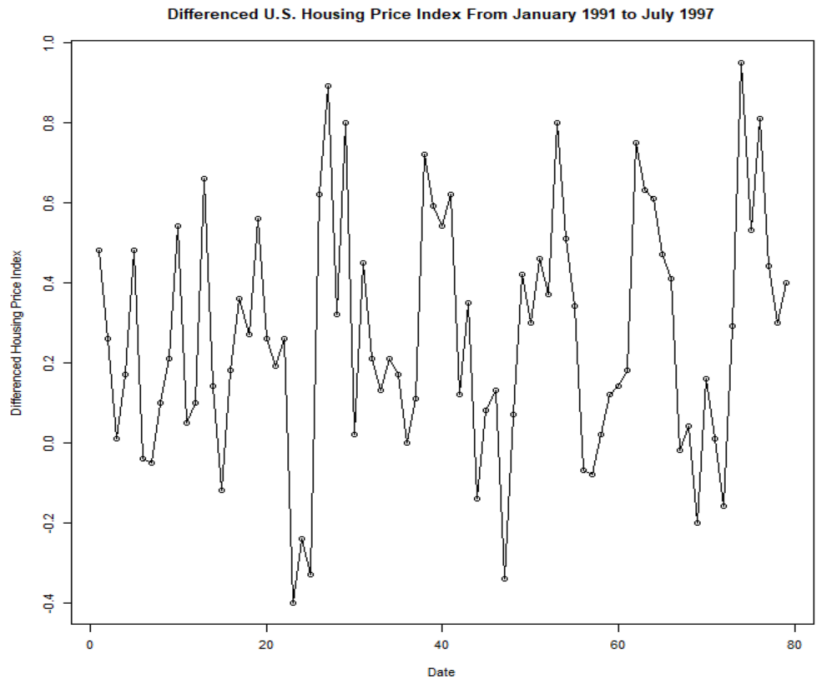


Figure 4 : Sample ACF of Differenced Data

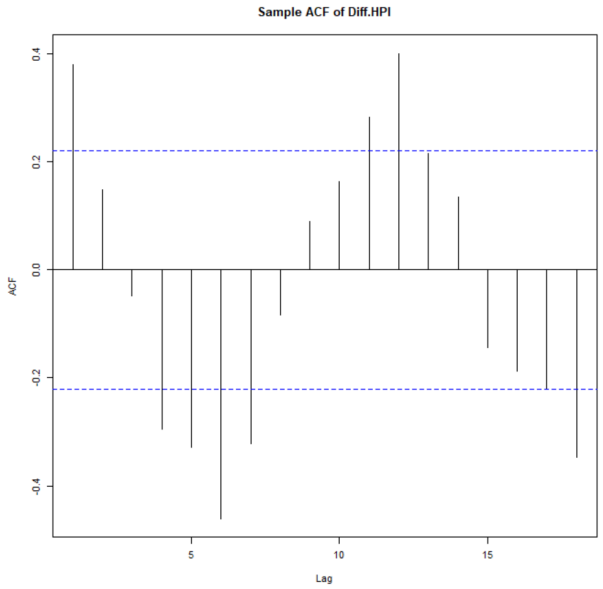


Figure 5 : Sample PACF of Differenced Data

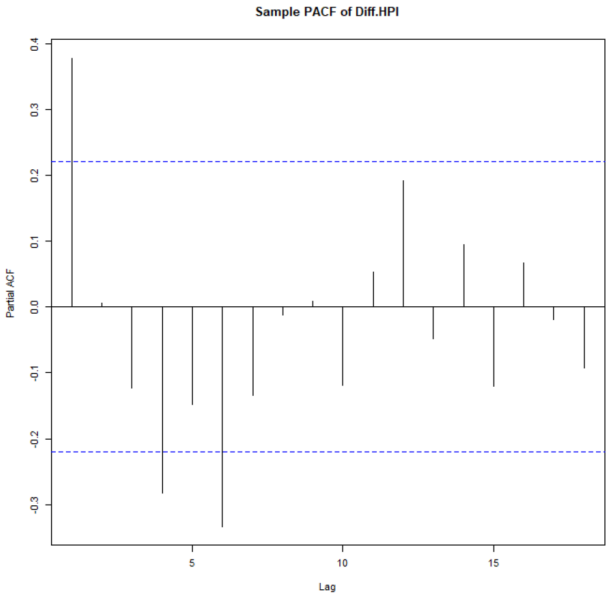


Figure 6 : BIC

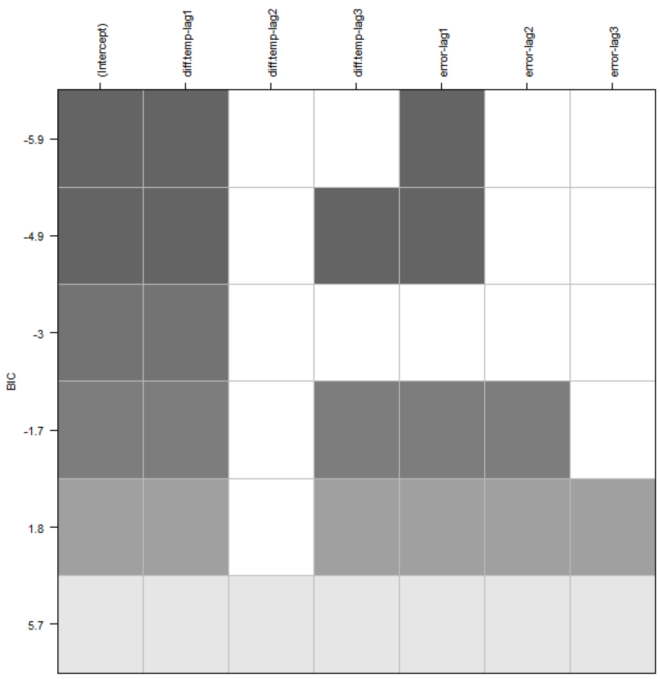


Figure 7 : Normality Assumption test with Residuals - ARI(1,1) Model

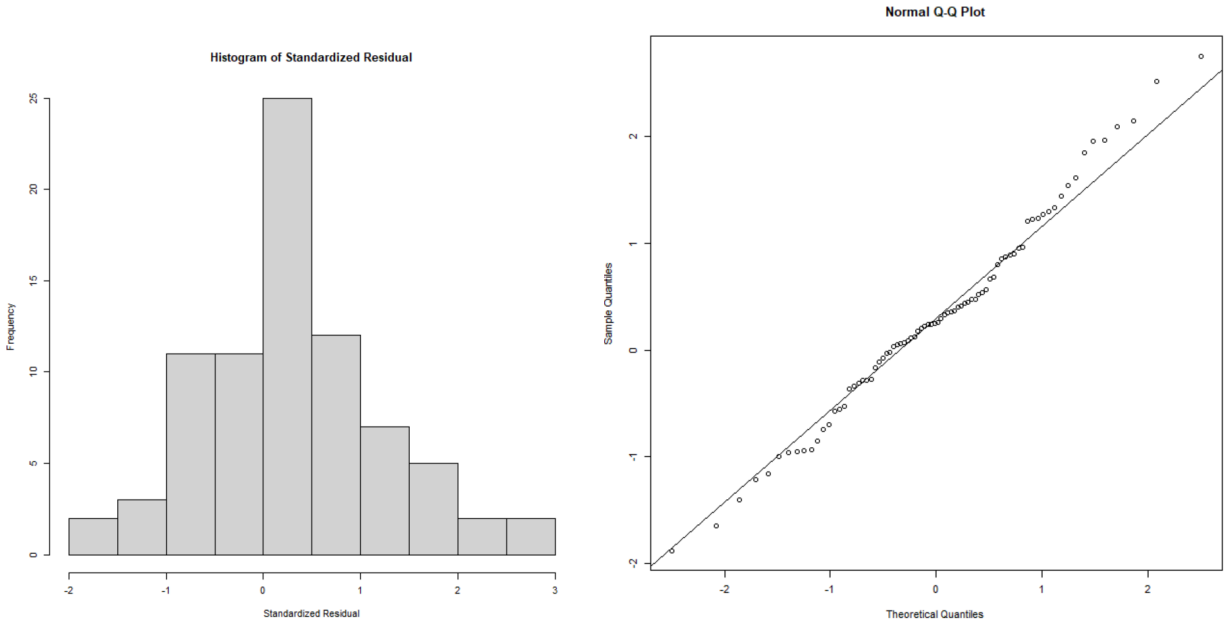


Figure 8 : Plot for testing Independence Assumption with Residual - ARI(1,1) Model

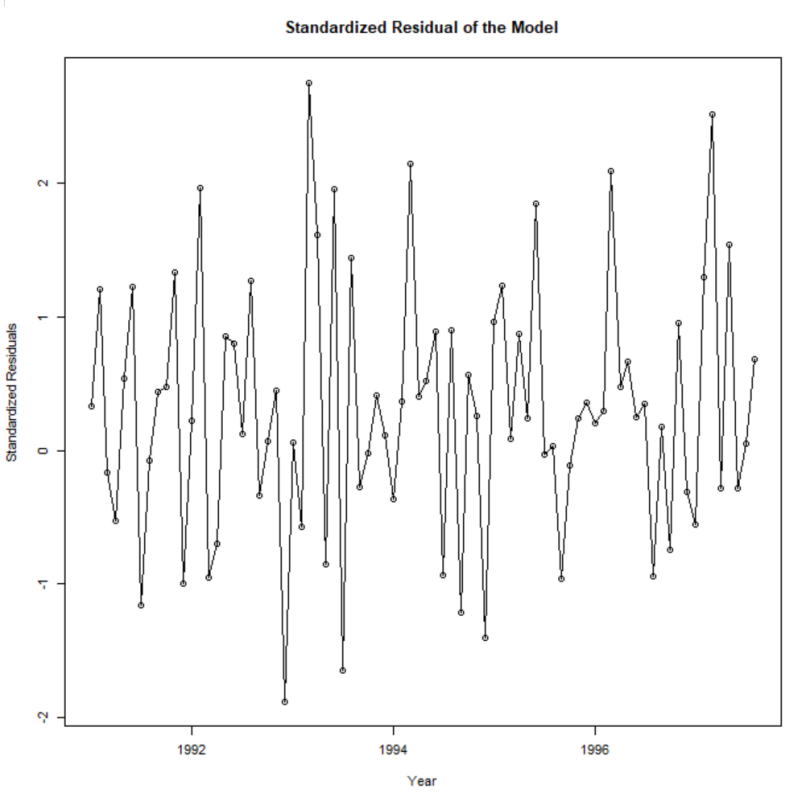


Figure 9 : Ljung-Box Test - ARI(1,1) Model:

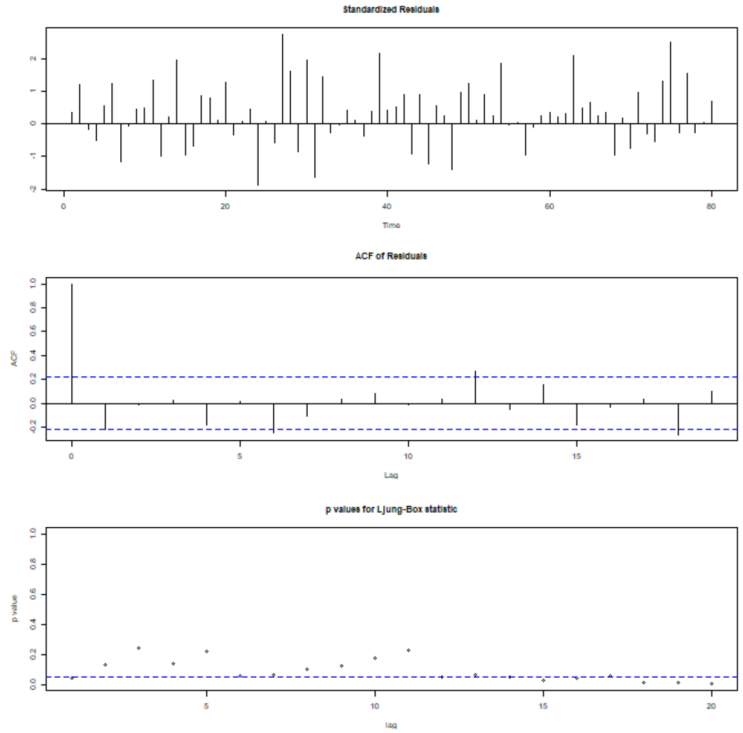


Figure 10 : Normality test plot : IMA(1,1)

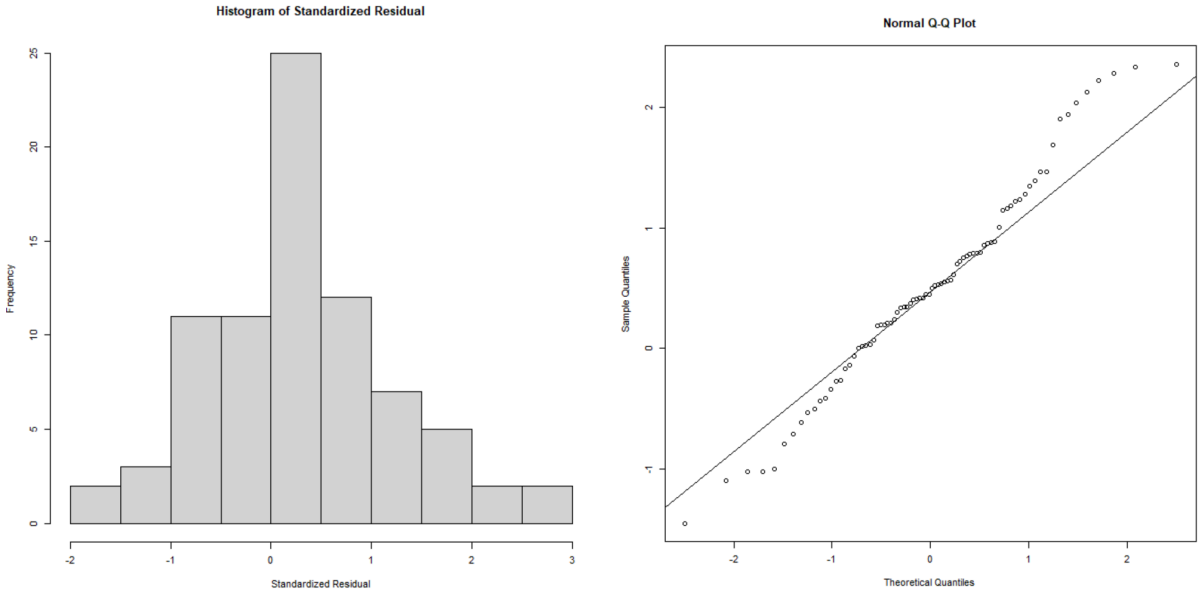


Figure 11: Independence test plot : IMA(1,1)

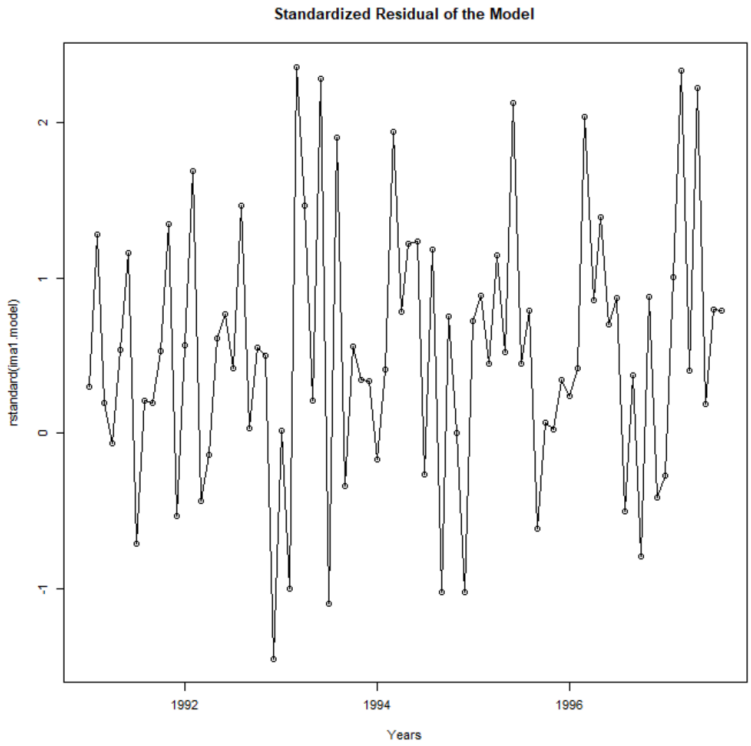


Figure 12 : Ljung-Box Test : IMA(1,1)

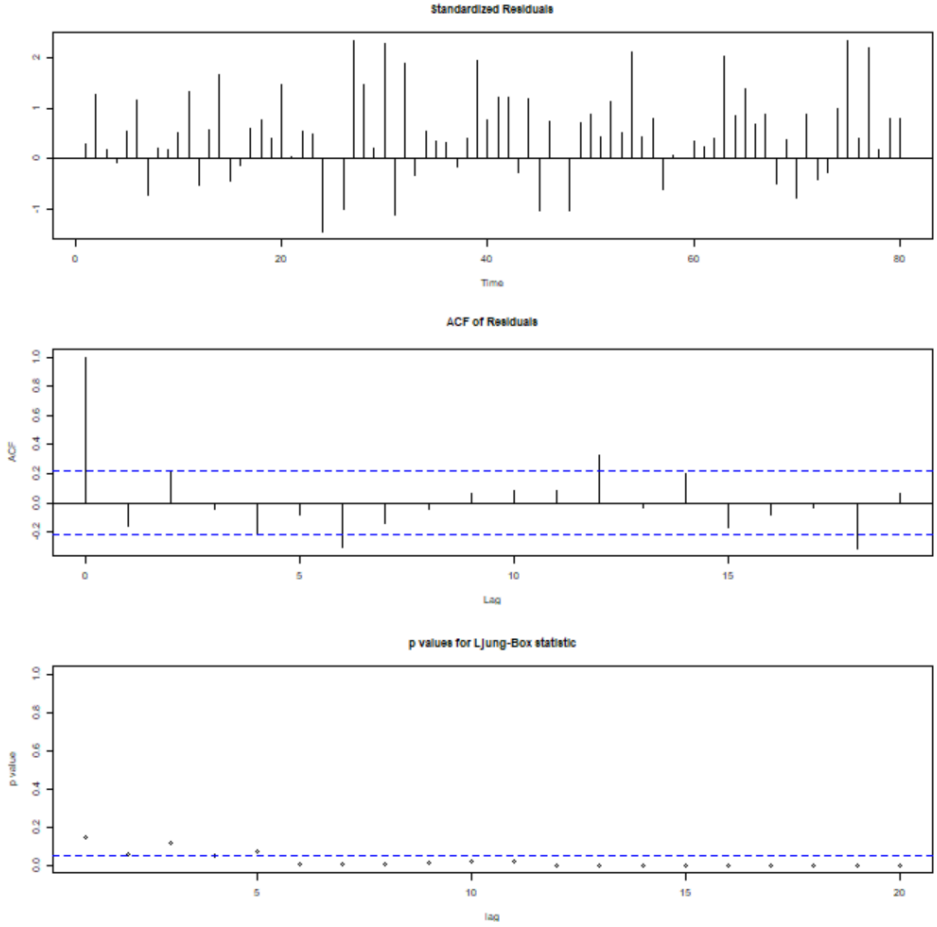
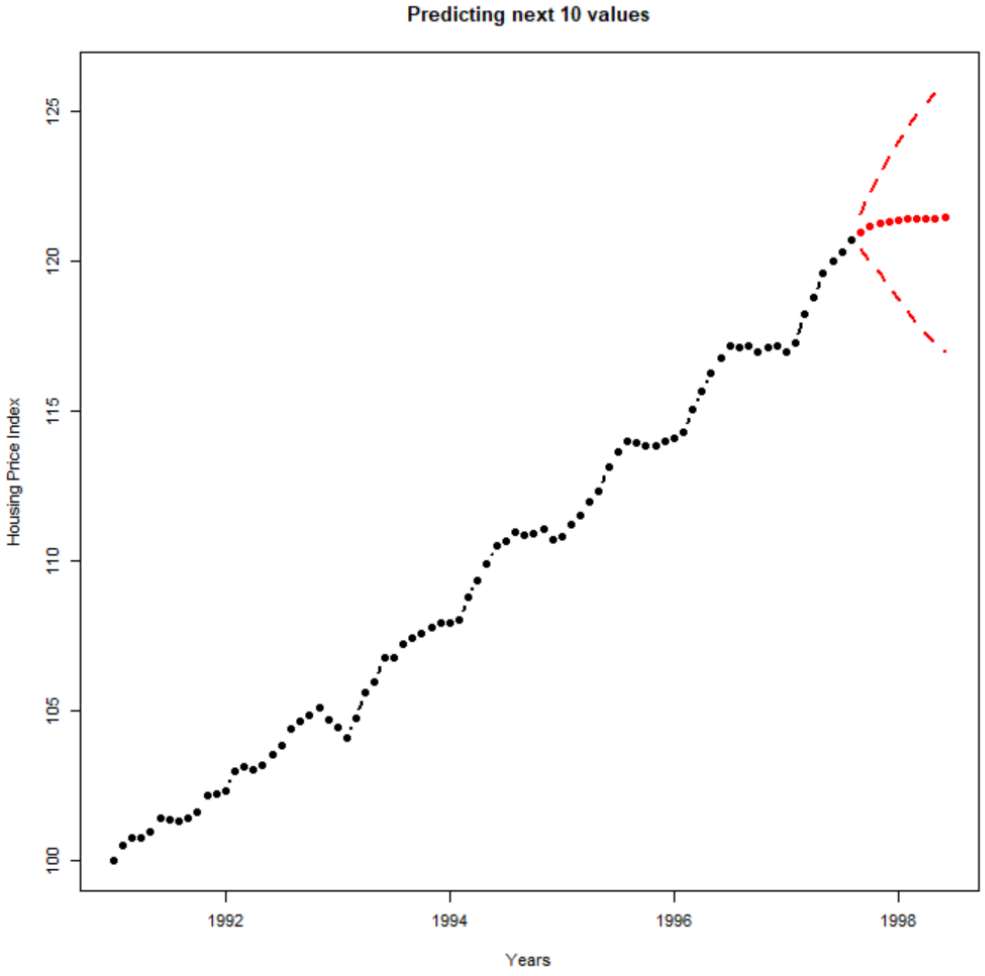


Figure 13 : Prediction next 10 Housing Price Index from 1997-09-01 to 1998-06-01



Output 1: Augmented Duckey-Fuller Unit Root test with the original HPI data

```
> adf.test(data)

        Augmented Dickey-Fuller Test

data:  data
Dickey-Fuller = -1.7979, Lag order = 5, p-value = 0.6607
alternative hypothesis: stationary
```

Null Hypothesis H_0 : The time series has unit roots(non-stationary)

Alternative Hypothesis H_a : The time series does not have unit roots(stationary)

From the ADF test we see that the p_value is 0.6607, which is greater than the significance level 0.05

Hence we fail to reject the null hypothesis and state that the process is not stationary.

Output 2: Augmented Duckey-Fuller test Unit Root test with the differenced HPI data

```
> adf.test(diff(data))

        Augmented Dickey-Fuller Test

data:  diff(data)

Dickey-Fuller = -5.4629, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Null Hypothesis H_0 : The time series has unit roots(non-stationary)

Alternative Hypothesis H_a : The time series does not have unit roots(stationary)

From the ADF test we see that the p_value is 0.01, which is less than the significance level 0.05

Hence we reject the null hypothesis and state that the process is stationary.

Output 3: Sample EACF:

```
> eacf(diff(data))
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o x x x x o o o x  x  o  o
1 o o o o o o o o o o o  x  o  o
2 o o o o o o o o o o o  o  o  x
3 x o o o o o o o o o o  o  o  o
4 x x o o o o o o o o o  o  o  o
5 x x o x x o o o o o o  o  o  o
6 x o o o x o o o o o o  o  o  o
7 x o o x o o o o o o o  o  o  o
```

Output 4: ARI(1,1) Model Fitting

```
> ar1.model<-arima(data,order=c(1,1,0),method='ML')

> ar1.model#-1412

Call:

arima(x = data, order = c(1, 1, 0), method = "ML")

Coefficients:

         ar1

      0.6486

s.e.  0.0858
sigma^2 estimated as 0.09158:  log likelihood = -17.94,  aic = 37.89
```

- **Significance of Estimates:**

```
> confint(ar1.model)#Significant
            2.5 %    97.5 %
ar1 0.4805015 0.816753
```

As this interval does not include zero, the AR parameter estimate is significantly different from 0

Output 5 : Normality test for ARI(1,1)

```
> shapiro.test(rstandard(ar1.model))#Pass

      Shapiro-Wilk normality test

data:  rstandard(ar1.model)
W = 0.98821, p-value = 0.6788
```

To perform normality test we use shapiro.tests

Null Hypothesis H_0 : All the error terms are normally distributed

Alternate Hypothesis H_a : All the error terms are not normally distributed

Here the W value is 0.9351 and the corresponding p-value = 0.4065

Significance level $\alpha = 0.05 < 0.6788$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

$\therefore$ We can say error terms are normally distributed.ARI(1,1) follows normality assumption

Output 6 : Independence test for ARI(1,1)

```
> runs(rstandard(ar1.model))#Pass
$pvalue
[1] 0.151

$observed.runs
[1] 43

$expected.runs
[1] 36.775

$n1
[1] 27

$n2
[1] 53

$k
```

```
[1] 0
```

Null Hypothesis H_0 : All the error terms are independent

Alternate Hypothesis H_a : All the error terms are not independent

p-value = 0.151

Significance level $\alpha = 0.05 < 0.151$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

∴ We can say error terms are independent of each other. ARI(1,1) follows independence assumption

Output 7 : Ljung-Box Test for ARI(1,1)

```
> Box.test(residuals(ar1.model),lag=10,type="Ljung-Box",fitdf = 1)#Pass

        Box-Ljung test

data:  residuals(ar1.model)
X-squared = 13.935, df = 9, p-value = 0.1247
```

Null Hypothesis H_0 : The ARI(1,1) model is appropriate

Alternative Hypothesis H_a : The ARI(1,1) model is not appropriate.

The p-value is 0.1247, which is greater than the critical value 0.05.

Hence we fail to reject the null hypothesis

∴ The ARI(1,1) model is appropriate

Output 8: IMA(1,1) Model Fitting

```
> ima1.model#-1399

Call:

arima(x = data, order = c(0, 1, 1), method = "ML")

Coefficients:

          ma1

      0.5070

s.e.  0.0879
```

```
sigma^2 estimated as 0.1116:  log likelihood = -25.62,  aic = 53.24
```

➢ **Significance level of IMA(1,1)**

```
> confint(ima1.model)#Significant
```

```
          2.5 %    97.5 %
ma1 0.3346534 0.6793362
```

As this interval does not include zero, the MA parameter estimate is significantly different from 0

Output 9 : Normality Test of IMA(1,1) Model:

```
> shapiro.test(rstandard(ima1.model))#Pass
```

```
        Shapiro-Wilk normality test

data:  rstandard(ima1.model)
W = 0.98191, p-value = 0.3178
```

Null Hypothesis H_0 : All the error terms are normally distributed

Alternate Hypothesis H_a : All the error terms are not normally distributed

Here the W value is 0.9819 and the corresponding p-value = 0.3178

Significance level $\alpha = 0.05 < 0.3178$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

$\therefore$ We can say error terms are normally distributed.IMA(1,1) follows normality assumption

Output 10 : Independence Test for IMA(1,1) Model:

```
> runs(rstandard(ima1.model))#Pass
$pvalue
[1] 0.267

$observed.runs
[1] 33

$expected.runs
[1] 28.9

$n1
[1] 18
```

```
$n2
[1] 62

$k
[1] 0
```

Null Hypothesis H_0 : All the error terms are independent

Alternate Hypothesis H_a : All the error terms are not independent

p-value = 0.267

Significance level $\alpha = 0.05 < 0.267$

The significance level is less than the p-value.

Hence we fail to reject null hypothesis

∴ We can say error terms are independent of each other. IMA(1,1) follows independence assumption

Output 11 : Ljung-Box Test for IMA(1,1) Model:

```
> Box.test(residuals(ima1.model),lag=10,type="Ljung-Box",fitdf = 1)#Fail

      Box-Ljung test

data:  residuals(ima1.model)
X-squared = 21.316, df = 9, p-value = 0.01132
```

Null Hypothesis H_0 : The IMA(1,1) model is appropriate

Alternative Hypothesis H_a : The IMA(1,1) model is not appropriate.

The p-value is 0.011, which is less than the critical value 0.05.

Hence we reject the null hypothesis

∴ The IMA(1,1) model is not appropriate

Output 13 : Forecasting with ARI(1,1) Model:

```
> hi.ar1.pred <- predict(ar1.model, n.ahead = 10)
> round(hi.ar1.pred$pred, 3)
Time Series:
Start = 81
End = 90
Frequency = 1
 [1] 120.959 121.128 121.237 121.308 121.354 121.383 121.403 121.415
 [9] 121.423 121.429
> round(hi.ar1.pred$se, 3)
Time Series:
Start = 81
End = 90
Frequency = 1
 [1] 0.303 0.584 0.856 1.111 1.348 1.566 1.767 1.954 2.129 2.292
> lower.pi <- hi.ar1.pred$pred - qnorm(0.975, 0, 1) * hi.ar1.pred$se
> upper.pi <- hi.ar1.pred$pred + qnorm(0.975, 0, 1) * hi.ar1.pred$se
> # Generate monthly dates for the prediction period
> start_date <- as.Date("1997-09-01")
> end_date <- as.Date("1998-06-01")
> monthly_dates <- seq(start_date, end_date, by = "month")
> # Create a data frame with monthly predictions and intervals
> predictions_df <- data.frame(Date = monthly_dates,
+                              Prediction = round(hi.ar1.pred$pred, 3),
+                              Lower_PI = round(lower.pi, 3),
+                              Upper_PI = round(upper.pi, 3))
> # Display the data frame
> print(predictions_df)
         Date Prediction Lower_PI Upper_PI
1  1997-09-01    120.959  120.366  121.553
2  1997-10-01    121.128  119.984  122.271
3  1997-11-01    121.237  119.559  122.915
4  1997-12-01    121.308  119.129  123.486
5  1998-01-01    121.354  118.712  123.995
6  1998-02-01    121.383  118.314  124.452
7  1998-03-01    121.403  117.939  124.867
8  1998-04-01    121.415  117.585  125.246
9  1998-05-01    121.423  117.251  125.596
10 1998-06-01    121.429  116.936  125.921
```

Output 13 : Forecasting with IMA(1,1) Model:

```
> hi.ar1.pred <- predict(ima1.model, n.ahead = 10)
> round(hi.ar1.pred$pred, 3)
Time Series:
Start = 81
End = 90
Frequency = 1
 [1] 120.834 120.834 120.834 120.834 120.834 120.834 120.834 120.834
 [9] 120.834 120.834
> round(hi.ar1.pred$se, 3)
Time Series:
Start = 81
End = 90
Frequency = 1
 [1] 0.334 0.604 0.786 0.934 1.061 1.174 1.277 1.373 1.462 1.547
> lower.pi <- hi.ar1.pred$pred - qnorm(0.975, 0, 1) * hi.ar1.pred$se
> upper.pi <- hi.ar1.pred$pred + qnorm(0.975, 0, 1) * hi.ar1.pred$se
> # Generate monthly dates for the prediction period
> start_date <- as.Date("1997-09-01")
> end_date <- as.Date("1998-06-01")
> monthly_dates <- seq(start_date, end_date, by = "month")
> # Create a data frame with monthly predictions and intervals
> predictions_df <- data.frame(Date = monthly_dates,
+                              Prediction = round(hi.ar1.pred$pred, 3),
+                              Lower_PI = round(lower.pi, 3),
+                              Upper_PI = round(upper.pi, 3))
> # Display the data frame
> print(predictions_df)
         Date Prediction Lower_PI Upper_PI
1  1997-09-01    120.834  120.179  121.489
2  1997-10-01    120.834  119.650  122.018
3  1997-11-01    120.834  119.293  122.375
4  1997-12-01    120.834  119.004  122.664
5  1998-01-01    120.834  118.755  122.913
6  1998-02-01    120.834  118.533  123.135
7  1998-03-01    120.834  118.330  123.338
8  1998-04-01    120.834  118.143  123.525
9  1998-05-01    120.834  117.968  123.700
10 1998-06-01    120.834  117.803  123.865
```