Forecasting the Real Gross Domestic Product

CSC425 Time Series Analysis

Prof. Raffaella Settimi

Naga Venkateshwarlu Yadav Dokku

Weiwei Yao

Table of Contents

Abstract

The national income and product accounts (NIPAs) produced by the Bureau of Economic Analysis. The NIPAs feature several widely followed measures of aggregate U.S. economic activity, including gross domestic product (GDP), gross domestic income (GDI), personal income, and personal saving among others. GDP covers the goods and services produced by labor and property located in the United States and is thus consistent with key economic indicators of employment, productivity, and industry output. The change also facilitated comparisons of economic activity in the United States with that in other countries. Gross domestic product (GDP), the featured measure of U.S. output, is the market value of the goods and services produced by labor and property located in the United States. Because the labor and property are located in the United States, the suppliers— that is, the workers and, for property, the owners—may be either U.S. residents or residents of the rest of the world. Our goal in this project is to observe and predict the real gross domestic product is the inflation adjusted value of the goods and services produced by labor and property located in the United States.

Goal: To identify the best fit model that represents real gross domestic product.

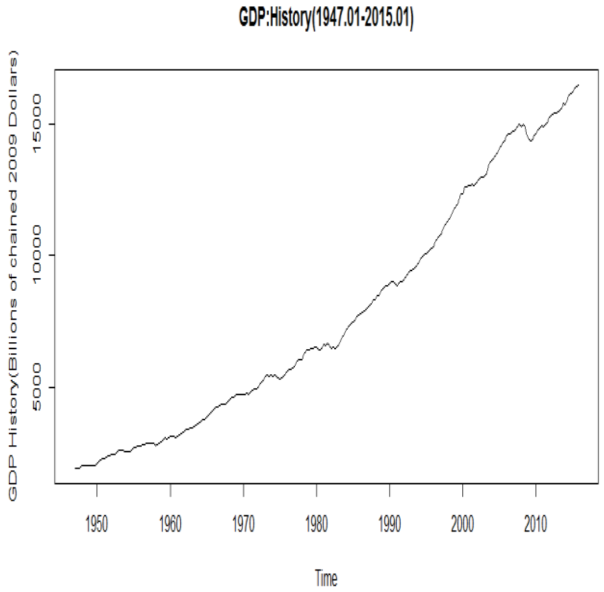
Exploratory Analysis

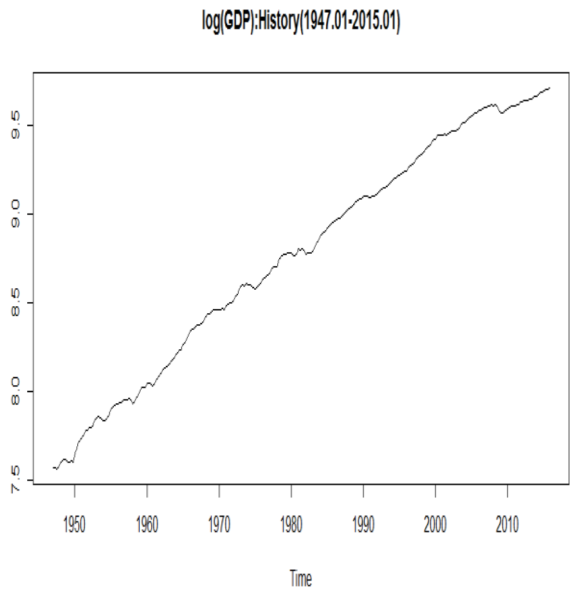
Check Data Distribution

We calculated the basic statistics of our data, which showed that the Skewness is 4.51 and the Kurtosis is -1.15. It means our data is not coming from a normal distribution but skewed and with heavy tails, which could also be concluded from the distribution plots above. So we decided to apply a log transformation here.

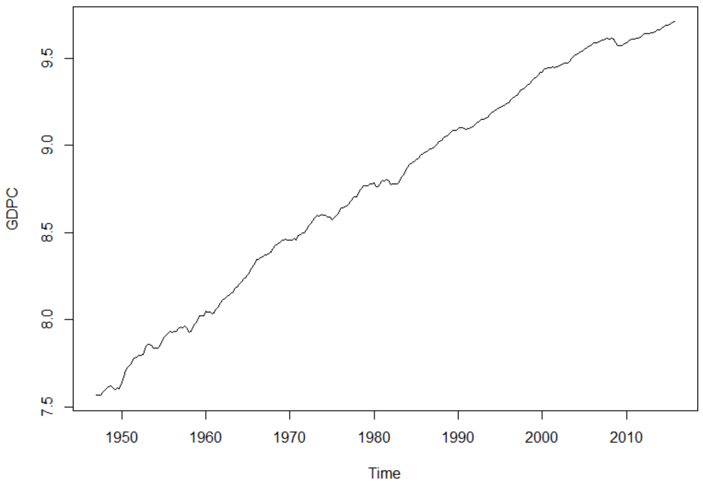
Apply Log Transformation

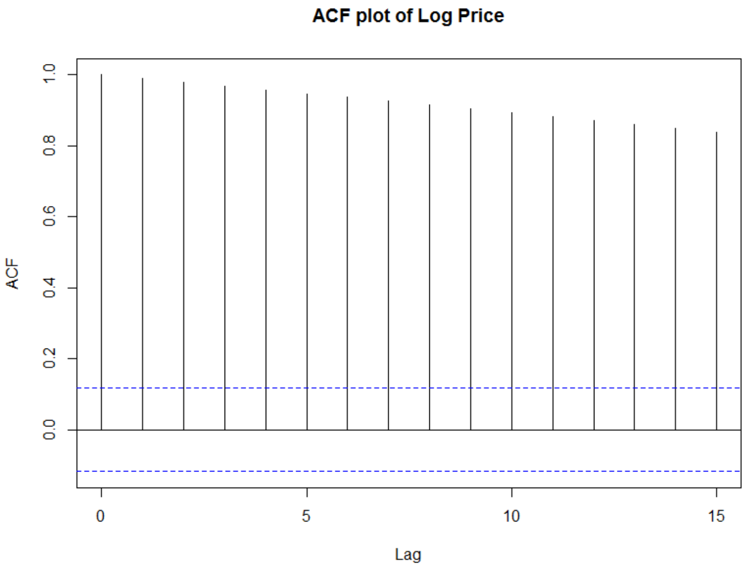
After using log, the Skewness is reduced to -0.21 and the Kurtosis is -1.19, which is better than the original data. The histogram below shows that the distribution of the log data is close to normal.





Check Time Series





The time series plot shows a continuous upward trend for GDPC. The means increased over time. Besides, ACF plot decays very slowly and the first lag in PACF plot is very close to 0, all of which suggests that this time series is a non-stationary process. First difference is needed here.

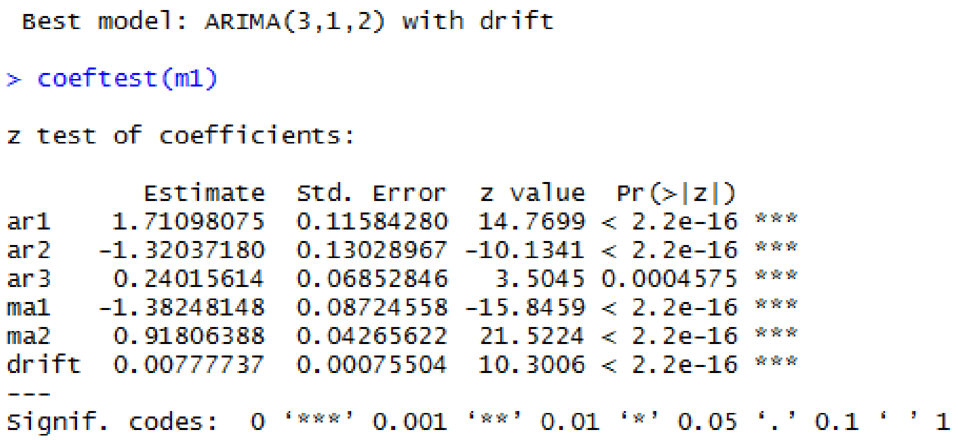
Model Selection

In model selection, the goal is to select the one, among a set of candidate models that represents the closest approximation to the underlying process in some defined sense. For this we have different statistical criterions for the selection of best models. In report we go to consider Akaike information criterion (AIC) Bayesian information criterion (BIC). The smaller the value of the AIC and BIC criterion, the better is the estimation.

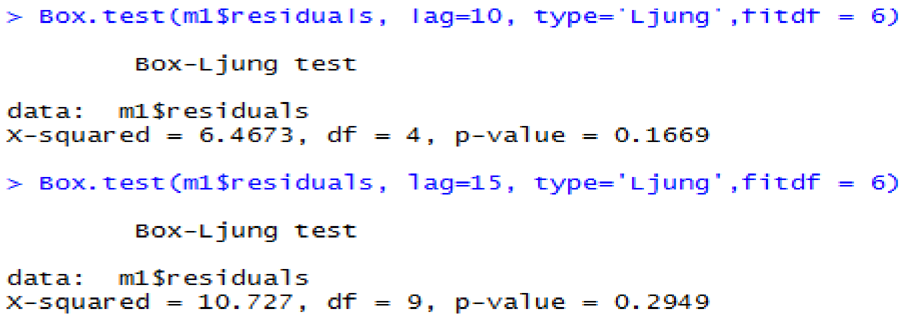
Also for the detection of any auto correlation existing in between the residuals of the data we have variety of diagnostic tests which are used in time series and econometrics such as Ljung-Box Q Pierce test

Approach I: ARIMA(3,1,2) with drift

This model is the best model suggested by auto.arima ('aic').



```
 Best model: ARIMA(3,1,2) with drift

> coeftest(m1)

z test of coefficients:

         Estimate  Std. Error  z value  Pr(>|z|)    
ar1    1.71098075  0.11584280  14.7699 < 2.2e-16 ***
ar2   -1.32037180  0.13028967 -10.1341 < 2.2e-16 ***
ar3    0.24015614  0.06852846   3.5045 0.0004575 ***
ma1   -1.38248148  0.08724558 -15.8459 < 2.2e-16 ***
ma2    0.91806388  0.04265622  21.5224 < 2.2e-16 ***
drift  0.00777737  0.00075504  10.3006 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

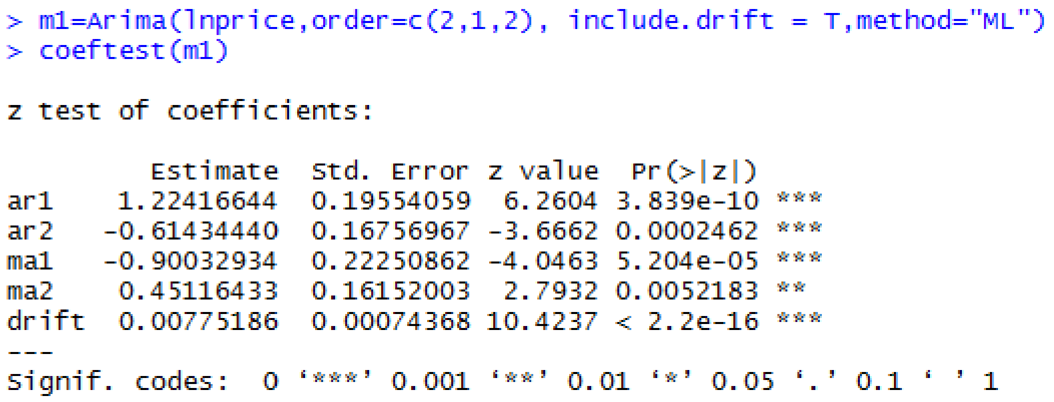


```
> Box.test(m1$residuals, lag=10, type='Ljung',fitdf = 6)

        Box-Ljung test

data:  m1$residuals
X-squared = 6.4673, df = 4, p-value = 0.1669

> Box.test(m1$residuals, lag=15, type='Ljung',fitdf = 6)

        Box-Ljung test

data:  m1$residuals
X-squared = 10.727, df = 9, p-value = 0.2949
```

In sum, this model is an adequate fit for this data. It not only has all significant variables, but also has independent residuals, which means this model captures the dynamic processes in the data.

Approach II: ARIMA(2,1,2) with drift

Other than the models suggested by auto.arima. We also tried a lot of different models and compare how well they fit in the data. The ARIMA(2,1,2) with drift model stands out to be another adequate model here.



```
> m1=Arima(lnprice,order=c(2,1,2), include.drift = T,method="ML")
> coeftest(m1)

z test of coefficients:

        Estimate   Std. Error z value  Pr(>|z|)    
ar1    1.22416644  0.19554059  6.2604 3.839e-10 ***
ar2   -0.61434440  0.16756967 -3.6662 0.0002462 ***
ma1   -0.90032934  0.22250862 -4.0463 5.204e-05 ***
ma2    0.45116433  0.16152003  2.7932 0.0052183 ** 
drift  0.00775186  0.00074368 10.4237 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

However, the residuals are not independent in this model as lag 30 also exceeds the blue dot line in residual plot

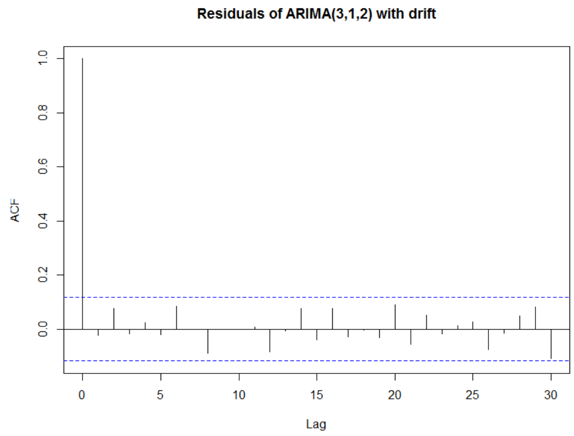
Final Model: ARIMA(3,1,2) with drift

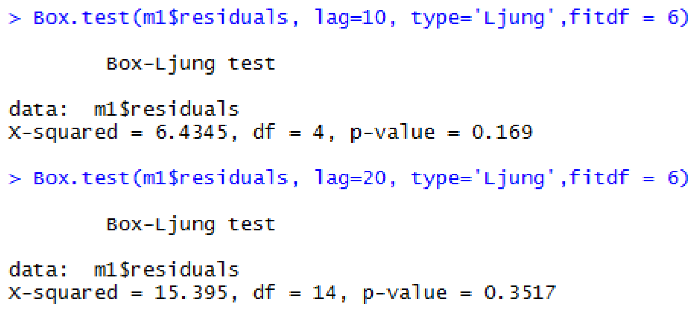
After comparing all the models we have, we decided to use ARIMA(3,1,2) with drift as our final model. This parameters are all significant at 5% level and the residuals are independent and white noises.

Model Expression:

$(1-1.71B+1.32B^2-0.24B^3)(1-B)Log(X_t) = 0.007 + (1+1.38B-0.91B^2)*a_t$

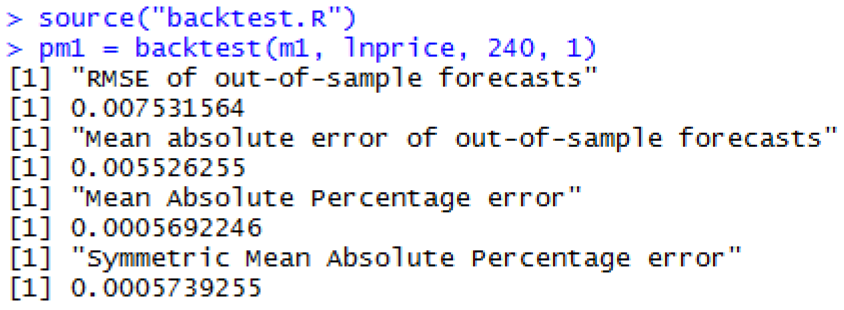
Model Diagnostics


Residuals of ARIMA(3,1,2) with drift



```
> Box.test(m1$residuals, lag=10, type='Ljung',fitdf = 6)

        Box-Ljung test

data:  m1$residuals
X-squared = 6.4345, df = 4, p-value = 0.169

> Box.test(m1$residuals, lag=20, type='Ljung',fitdf = 6)

        Box-Ljung test

data:  m1$residuals
X-squared = 15.395, df = 14, p-value = 0.3517
```

As we can see from the plots above, the residuals of our final model are independent and white noises.



```
> source("backtest.R")
> pm1 = backtest(m1, lnprice, 240, 1)
[1] "RMSE of out-of-sample forecasts"
[1] 0.007531564
[1] "Mean absolute error of out-of-sample forecasts"
[1] 0.005526255
[1] "Mean Absolute Percentage error"
[1] 0.0005692246
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.0005739255
```

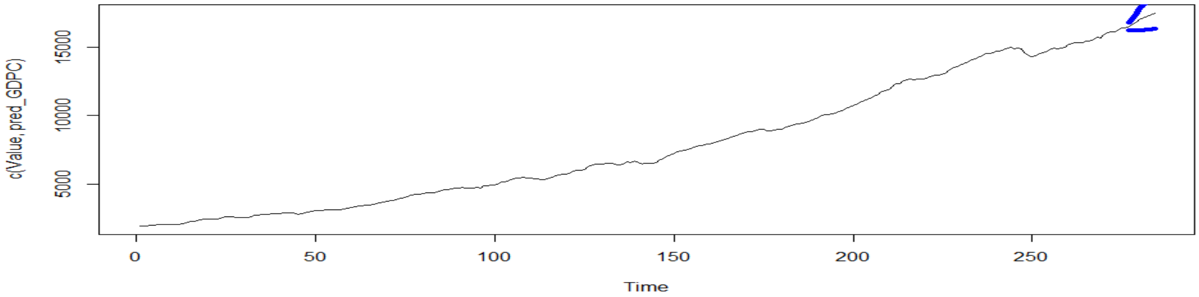
Besides, the backtesting statistics are very good. This model is the best model for our data.

Model Forecasts

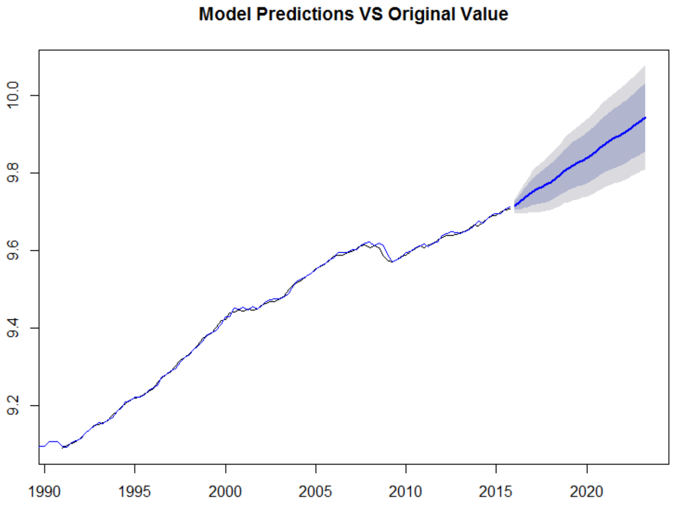
Using the exponential form, the real GDP was forecast 20 periods ahead. The equation generated by the regression analysis is as follows:

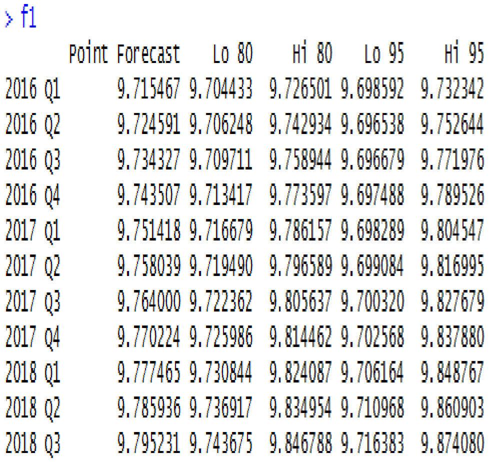
$$(1-1.71B+1.32B^2-0.24B^3)(1-B)Log(X_t) = 0.007 + (1+1.38B-0.91B^2)*a_t$$

The graphical representation of this equation:



The model ARIMA(3,1,2) with drift is used to forecast the real GDP for the period of Feb 2015 to Feb 2020.





```
> f1
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2016 Q1       9.715467 9.704433 9.726501 9.698592 9.732342
2016 Q2       9.724591 9.706248 9.742934 9.696538 9.752644
2016 Q3       9.734327 9.709711 9.758944 9.696679 9.771976
2016 Q4       9.743507 9.713417 9.773597 9.697488 9.789526
2017 Q1       9.751418 9.716679 9.786157 9.698289 9.804547
2017 Q2       9.758039 9.719490 9.796589 9.699084 9.816995
2017 Q3       9.764000 9.722362 9.805637 9.700320 9.827679
2017 Q4       9.770224 9.725986 9.814462 9.702568 9.837880
2018 Q1       9.777465 9.730844 9.824087 9.706164 9.848767
2018 Q2       9.785936 9.736917 9.834954 9.710968 9.860903
2018 Q3       9.795231 9.743675 9.846788 9.716383 9.874080
```

Conclusion

The GDP is among the most studied economic statistics, particularly since the beginning of the 20th century. Drawing on some of the experience gained through those studies, in this project we have attempted to identify some of the trends in current research, highlight the implications of changes in the real GDP, On the basis of the overall findings of this project, it can be concluded that in case of real gross domestic product ARIMA (3, 2, 1) with drift model add value significantly to the forecasting values and provide a 20 quarter look-ahead forecast of the GDP. According to forecast values that the real Gross Domestic Product will experience continuous upward growth over next 20 quarter as real GDP series appears to be an exponential. Further investigation and refining of the model should lend more credibility to that forecast as well as shrink the range of possible outcomes, i.e. the range covered by the 95% confidence interval.

Appendix (Code)

```
library(zoo)
library(tseries)
library(fBasics)
library(lmtest)
library(forecast)

#read table and create time series
D=read.table('GDPC1.csv',header=T,sep=',')
head(D)
value=D[,2]
price = ts(value, start=c(1947,1), freq=4)
lnprice=log(price)
plot(lnprice,type="l", xlab="Time", ylab="GDPC")
basicStats(lnprice)
acf(coredata(lnprice), plot=T, lag=15)
pacf(coredata(lnprice), plot=T, lag=15)

#check Distribution
hist(lnprice, xlab="Log GDPC", prob=TRUE, main="Histogram")
xfit<-seq(min(lnprice),max(lnprice),length=40)
yfit<-dnorm(xfit,mean=mean(lnprice),sd=sd(lnprice))
lines(xfit, yfit, col="blue", lwd=2)
qqnorm(lnprice)
qqline(lnprice,col='blue',lwd=2)

#model selection
```

```
m1=auto.arima(lnprice,ic="bic",max.P = 8,max.Q = 8,trace = T)

coeftest(m1)

acf(coredata(m1$residuals),lag=30, main='Residuals of ARIMA(1,1,0) with drift')

Box.test(m1$residuals, lag=7, type='Ljung',fitdf = 2)


m1=auto.arima(lnprice,ic="aic",max.P = 8,max.Q = 8,trace = T)

coeftest(m1)

acf(coredata(m1$residuals),lag=30,main='Residuals of ARIMA(3,1,2) with drift')

Box.test(m1$residuals, lag=10, type='Ljung',fitdf = 6)

Box.test(m1$residuals, lag=15, type='Ljung',fitdf = 6)


#final model

m1=Arima(lnprice,order=c(2,1,2), include.drift = T,method="ML")

coeftest(m1)

acf(coredata(m1$residuals),lag=30, main='Residuals of ARIMA(2,1,2) with drift')

Box.test(m1$residuals, lag=10, type='Ljung',fitdf = 5)

Box.test(m1$residuals, lag=20, type='Ljung',fitdf = 5)

polyroot(c(1,-m1$coef[1:2]))


#backtesting

source("backtest.R")

pm1 = backtest(m1, lnprice, 240, 1)




#model forecast

pred=forecast.Arima(m1)

plot(forecast.Arima(m1, h=10), include =100, main='Model Forecasts')


# plot predicted values and 95% prediction bounds in blue
```

```
pred_GDPC=exp(pred$mean)

plot.ts(c(price,pred_GDPC))

lines(c(rep(NA, length(price)), exp(pred$lower[,2])), col='blue', lwd=5)

lines(c(rep(NA, length(price)), exp(pred$upper[,2])),col='blue',lwd=5)


f1=forecast.Arima(m1,h=30)

plot(f1, include=100,main='Model Predictions VS Original Value')

lines(ts(c(f1$fitted, f1$mean), frequency=4,start=c(1947,1)), col="blue")
```