南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

时间序列分析小组报告

题　　目：基于乘法季节模型的纪念品月销量的时间序列分析报告

姓　　名：邓春丽 11711432
刘润祺 11711331
陆昕子 11711234

组　　别：第十九组

2019 年 12 月 27 日

The Time Series Analysis of souvenir sales based on Multiplicative Seasonal ARIMA Model

Deng Chunli 11711432 Liu Runqi 11711331 Lu Xinzi 11711234
Southern University of Science and Technology

December 27, 2019

Abstract

In this report, we use the R statistical software to carry out some analyses that are common in analyzing time series data. With the data of monthly sales for a souvenir shop at a beach resort town in Queensland, Australia, from January 1987 to December 1993, we initially give an overall description and analysis of the data selected, as well as a preliminary judgment of this time series. In Section 2, we start the work of model specification, and with reliable verification and specific pattern recognition, we specify the time series as ARIMA(0,1,1)$\times(0,1,1)_{12}$ model. The following thing we do is to estimate the parameters in the model as effectively as possible in Section 3. Also, a very important thing is model diagnostics, we mainly do related tests about residuals in this part. Lastly, forecasting, which is vitally important, is to use the model we fit to predict future sales. We also see how well our model fits the real data and find a certain degree of error.

Keywords: Time Series Analysis, Model Fitting, Multiplicative Seasonal ARIMA(SARIMA), Model Specification, Parameter Estimation, Model Diagnostics, Forecasting.

1 Data

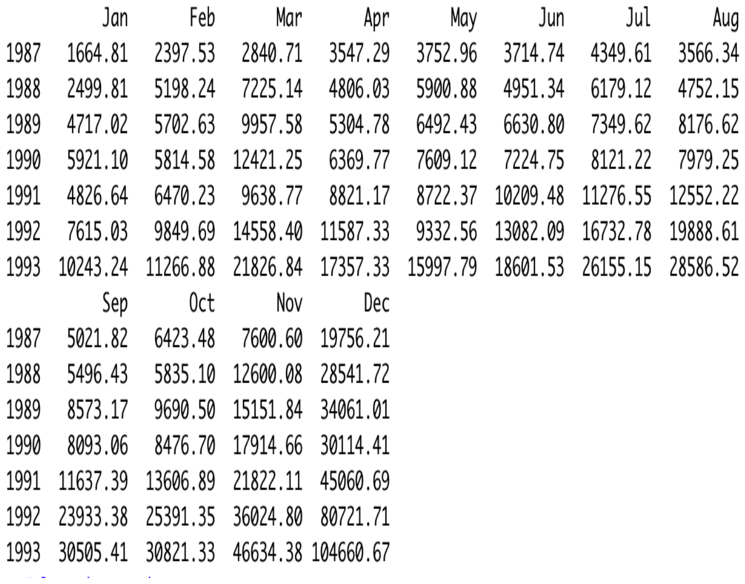
1.1 Background

With the rapid upgrade of the tourism market, the development of domestic and foreign tourism has reached a new level. When traveling to a place, we always want to bring back some local souvenirs as a record of beautiful memories. After all, souvenirs play a role as the carrier of the unique culture of the tourist attractions, and as reported, there is a big market for tourist souvenirs. Facing the new situation of competition and industry's change and challenge, especially in tourism developed countries and regions, tourist souvenir market shares account for 40% to 60% in the whole industry chain, including accommodation, travel, shopping, and entertainment, there is no doubt that its consumption potential is very large. Therefore, this paper selects the monthly sales data of a souvenir shop located in a seaside resort in Queensland from January 1987 to December 1993 and carries out a series of analyses on the time series data based on our learned knowledge and available tools, to build a reasonable and effective model to fit the data and predict the future trend.

1.2 Data Source

Before formally determining the data, we spent a lot of time searching reliable data on different websites, and we preprocessed some data to find the most suitable data. Lastly, we visited the Time Series Data Library, which is made available by Rob Hyndman (http://robjhyndman.com/TSDL/), and found the data in Wheelwright and Hyndman, 1998 (http://robjhyndman.com/tsdldata/data/fancy.dat). The data is accessible online by opening the link.

1.3 Data Cutout



```
           Jan       Feb       Mar       Apr       May       Jun       Jul       Aug
1987   1664.81   2397.53   2840.71   3547.29   3752.96   3714.74   4349.61   3566.34
1988   2499.81   5198.24   7225.14   4806.03   5900.88   4951.34   6179.12   4752.15
1989   4717.02   5702.63   9957.58   5304.78   6492.43   6630.80   7349.62   8176.62
1990   5921.10   5814.58  12421.25   6369.77   7609.12   7224.75   8121.22   7979.25
1991   4826.64   6470.23   9638.77   8821.17   8722.37  10209.48  11276.55  12552.22
1992   7615.03   9849.69  14558.40  11587.33   9332.56  13082.09  16732.78  19888.61
1993  10243.24  11266.88  21826.84  17357.33  15997.79  18601.53  26155.15  28586.52
           Sep       Oct       Nov       Dec
1987   5021.82   6423.48   7600.60  19756.21
1988   5496.43   5835.10  12600.08  28541.72
1989   8573.17   9690.50  15151.84  34061.01
1990   8093.06   8476.70  17914.66  30114.41
1991  11637.39  13606.89  21822.11  45060.69
1992  23933.38  25391.35  36024.80  80721.71
1993  30505.41  30821.33  46634.38 104660.67
```

Figure 1.1 Data Observation

http://robjhyndman.com/tsdldata/data/fancy.dat

1.4 Time Series Diagram

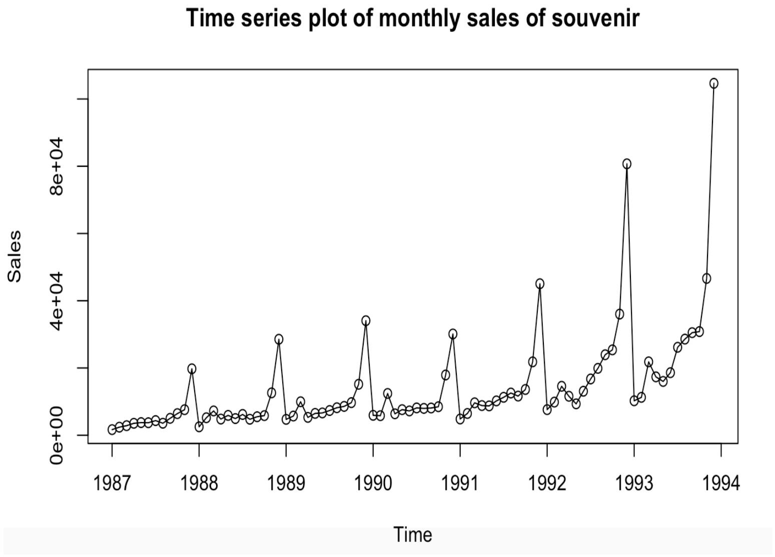


Figure 1.2 Time Series Diagram

Figure 1.2 above shows the monthly sales data for the souvenir store from January 1987 to December 1993. It is found that the monthly distribution trend of souvenir sales in the souvenir store is nearly the same every year, and the annual sales volume has been increasing since 1987, although it fluctuated between 1990 and 1991, and increased significantly in the last two years.

Considering only the upward trend in the figure, we can see that the time series is not stationary, from which we will initially establish a non-stationary model.

In the plot, the upward trend of sales volume is not stable in the past eight years from 1987 to 1994, in this case, it appears that an additive model is not appropriate for describing this time series, since the size of the seasonal fluctuations and random fluctuations seem to increase with the level of the time series. Thus, we may need to transform the time series in order to get a transformed time series that can be described using an additive model.

We take the logarithm of the original series and get the following diagram.

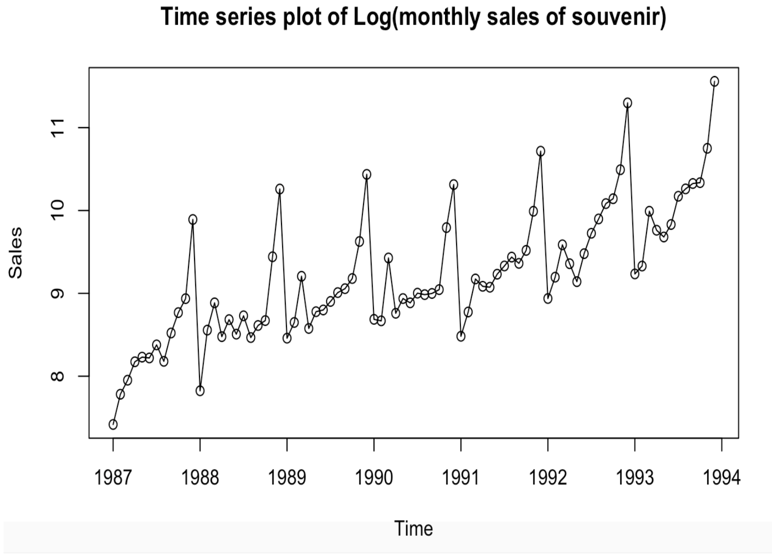


Figure 1.3 Time Series Diagram of Log-data

Here we can see that the time series diagram obtained after taking the logarithm has a more stable upward trend, since the size of the seasonal fluctuations and random fluctuations in the log-transformed time series seem to be roughly constant over time, and do not depend on the level of the time series.

Thus, we transform the sample data by logs to model this series.(We call it as logged data in the next content.). And the log-transformed time series can probably be described using an additive model.

2 Model Specification

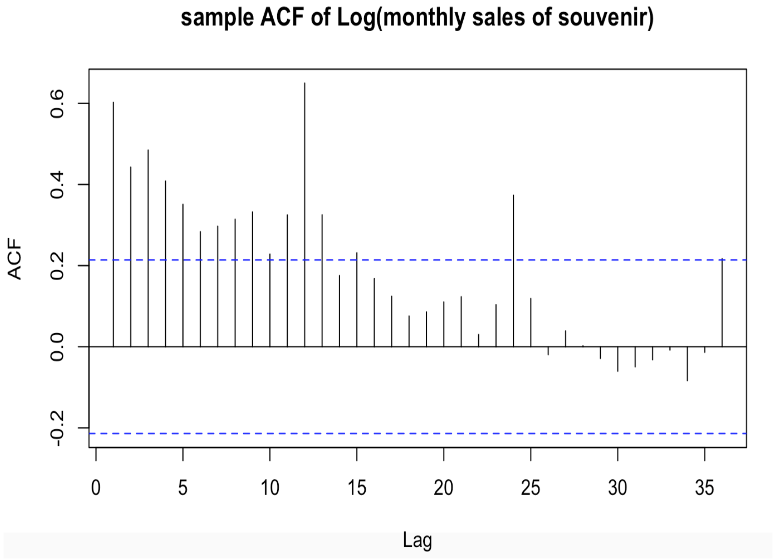


Figure 2.1 Sample acf of log-data

The above figure (figure 2.1) shows the sample autocorrelation of the time series. As shown in the figure, the seasonal autocorrelation is very significant. So the current model is not ideal. Moreover, the higher lines in the figure also present the seasonality, so we need to include other correlations.

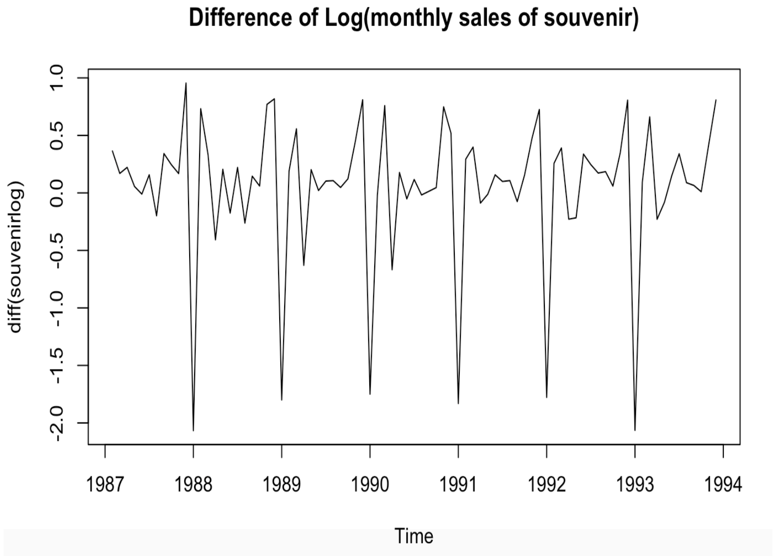


Figure 2.2 Diagram of logged souvenir sales after first difference.

Initially, we have the first difference of this time series, and the figure above (Figure 2.2) shows the time series diagram of souvenir sales after the first difference. As you can see, the general upward trend has now disappeared but the strong seasonality is still present. Perhaps a simpler model can be established by using the time series obtained by the seasonal difference method.

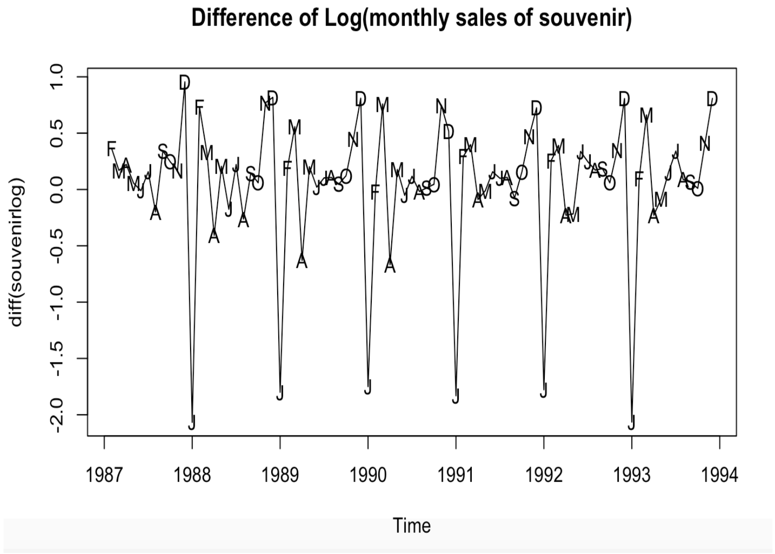


Figure 2.3 Time Series Diagram of Difference Log-data with lables

Therefore, we add the label of months to the original first-order difference graph, so that we could clearly see that the lowest point of each year in the graph was in January and the highest point was in December.

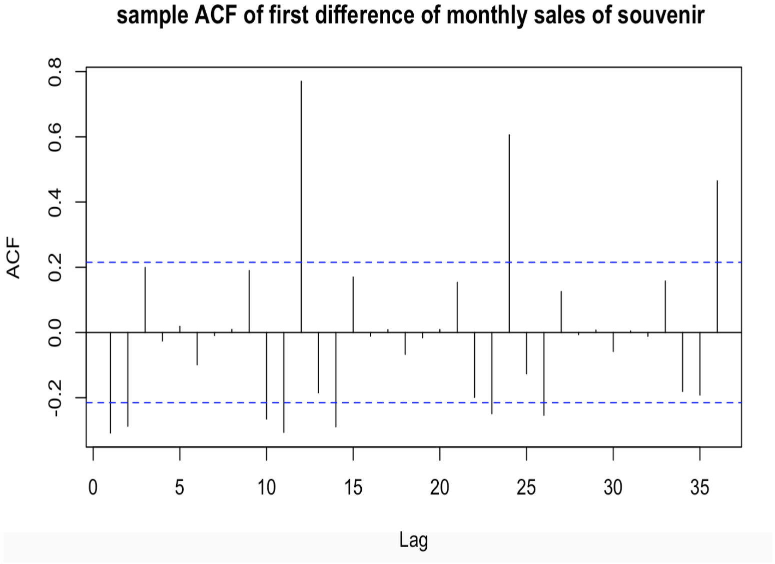


Figure 2.4 ACF Diagram of Difference Log-data

After that, we obtain the ACF diagram of the sample after the first difference, (Figure 2.4)which shows obvious and strong autocorrelation at the lags of 12, 24 and 36 orders, so we considered the seasonal model.

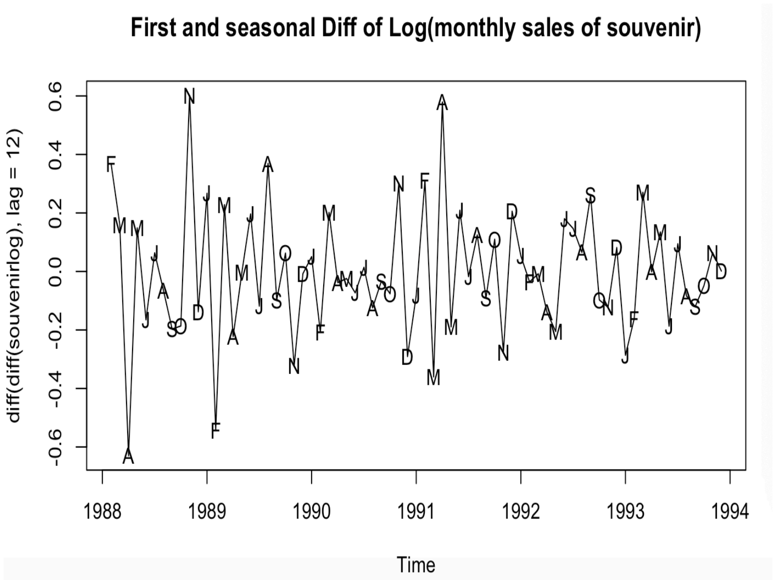
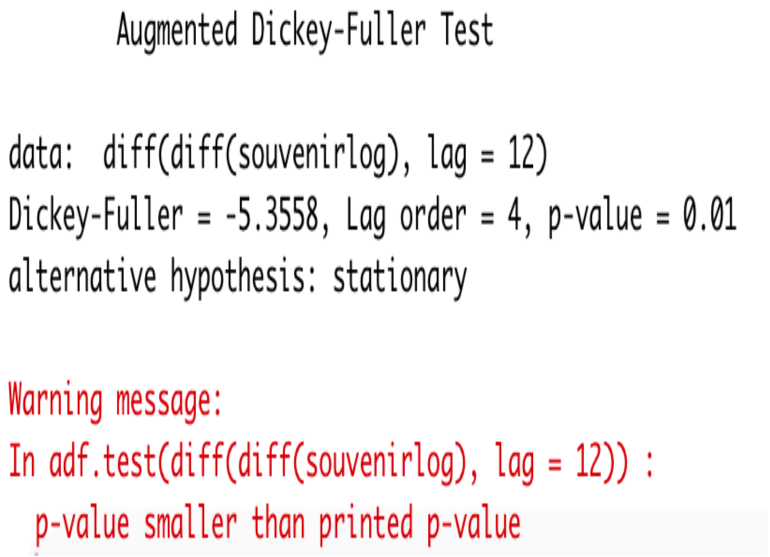


Figure 2.5 Diagram of the first-order and seasonal difference Log-data

As the above figure shows, after taking the first-order and seasonal difference with lag 12, it appears that most, if not all, of the seasonality is gone now. After removing the seasonality, we find that the series seems to be stationary. To verify the assumption, we turn to the unit root test (Figure 2.6). The P-value smaller than 0.01, we can reject the null hypothesis that there is a unit root in a non-stationary time series, so the series is stationary after the first and seasonal difference of logged sales series.



```
        Augmented Dickey-Fuller Test

data:  diff(diff(souvenirlog), lag = 12)
Dickey-Fuller = -5.3558, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(diff(diff(souvenirlog), lag = 12)) :
  p-value smaller than printed p-value
```

Figure 2.6 Augmented Dickey-Fuller Test

3 Model Fitting

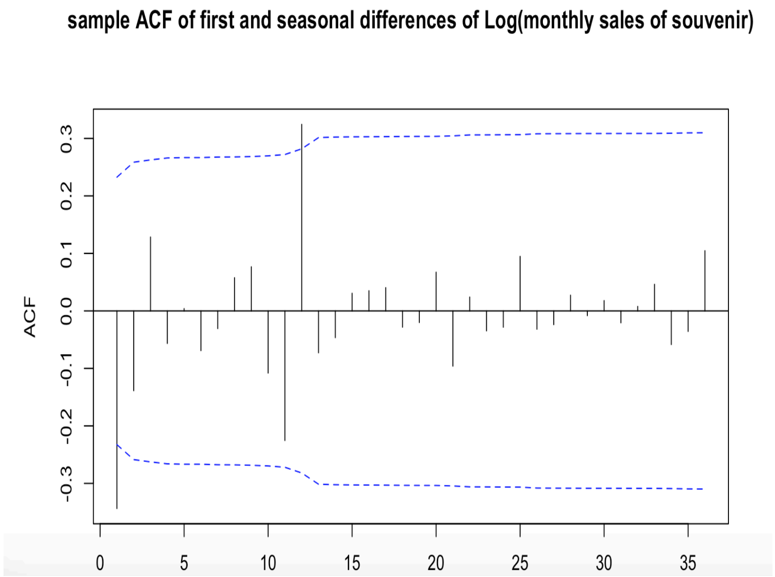
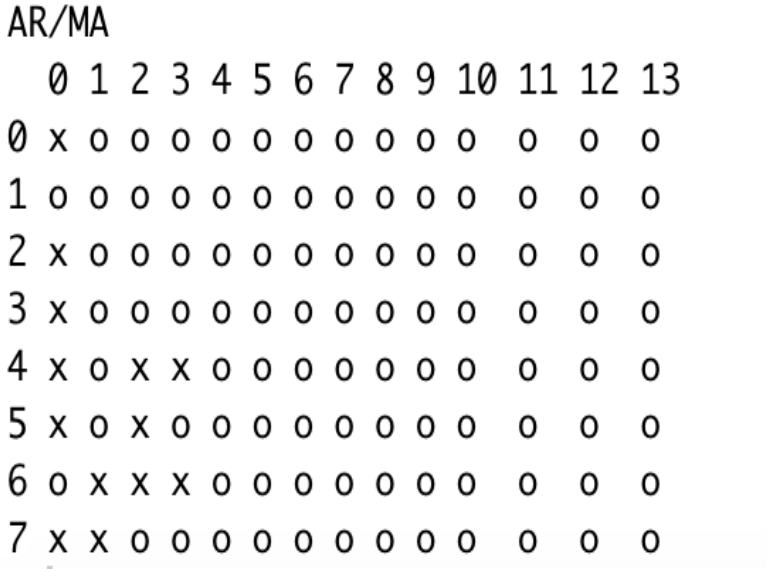


Figure 3.1 Augmented Dickey-Fuller Test

It is conformed by the above figure that the time series has little autocorrelation except at lag 1 and lag 12. Since we consider the model

ARIMA(p,d,q)×$(P,D,Q)_{12}$, the spike at lag 1 and lag 12 suggests q=1 and Q=1.



```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  o  o  o
1 o o o o o o o o o o o  o  o  o
2 x o o o o o o o o o o  o  o  o
3 x o o o o o o o o o o  o  o  o
4 x o x x o o o o o o o  o  o  o
5 x o x o o o o o o o o  o  o  o
6 o x x x o o o o o o o  o  o  o
7 x x o o o o o o o o o  o  o  o
```

Figure 3.2 EACF screenshot

According to the EACF, we consider ARMA(0,1) or ARMA(1,1) for the time series after first and seasonal difference of logged sales series.

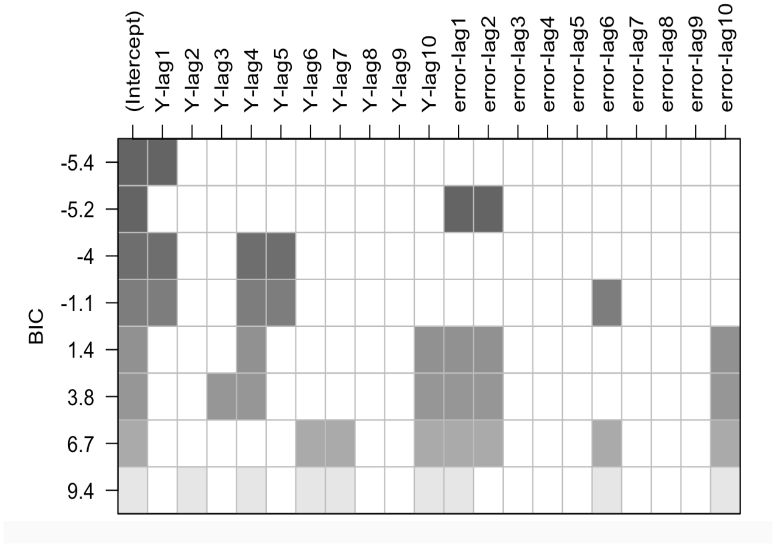
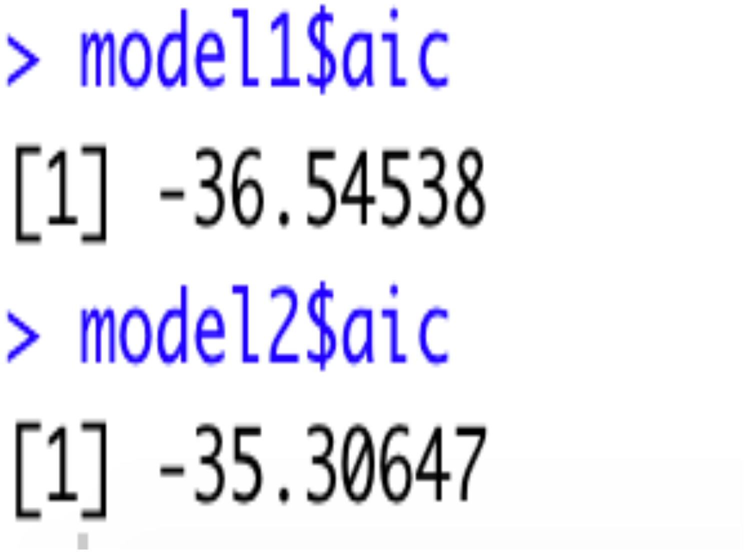


Figure 3.3 BIC chart

We want to minimize the BIC and the deeper color, the greater the probability we have. According to the BIC figure, ARMA(1,0) for the time series after the first and seasonal difference of logged sales series is suggested. However, based on the previous analysis, ARMA(0,1) is preferred. Comparing these two BIC values, which are -5.4 and -5.2, they are quite similar. So, the ARMA(0,1) model is also possible. Here, since q=1, we do not consider the ARMA(1,0) model.

Based on the first-order and seasonal difference, as a result, for the logged time series, multiplicative season model ARIMA(0,1,1)$\times(0,1,1)_{12}$ and ARIMA(1,1,1)$\times(0,1,1)_{12}$ are likely to be the model we want.



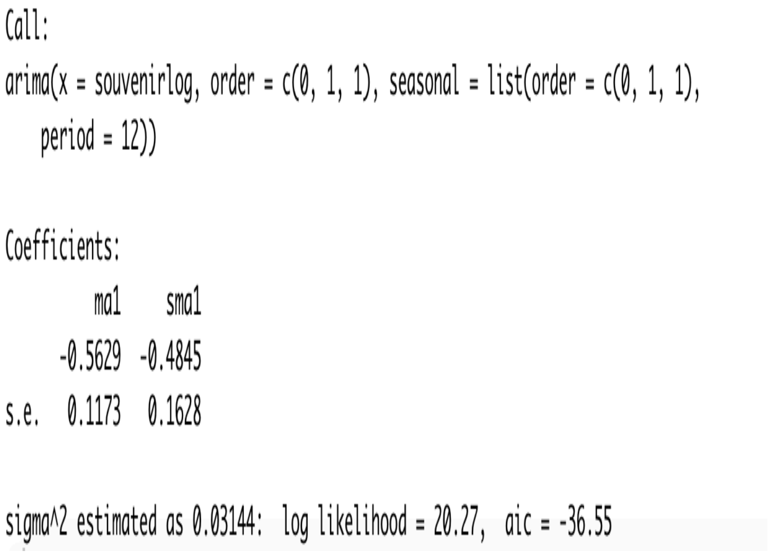
```
> model1$aic
[1] -36.54538
> model2$aic
[1] -35.30647
```

Figure 3.4 AIC comparison

Then, we calculate the Akaikes's Information Criterion (AIC). The AIC for ARIMA(0,1,1)$\times(0,1,1)_{12}$ is smaller(which means better), so we consider specifying the multiplicative, seasonal ARIMA(0,1,1)$\times(0,1,1)_{12}$.

Finally, we get ARIMA(0,1,1)$\times(0,1,1)_{12}$ for log(souvenir).

After establishing a tentative seasonal model for a specific time series, the next thing to do is to estimate the parameters in the model as effectively as possible.



```
Call:
arima(x = souvenirlog, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1),
    period = 12))

Coefficients:
          ma1     sma1
      -0.5629  -0.4845
s.e.   0.1173   0.1628

sigma^2 estimated as 0.03144:  log likelihood = 20.27,  aic = -36.55
```

Figure 3.5 Parameter Estimation

From the above results, we find all the coefficients in model ARIMA(0,1,1)×$(0,1,1)_{12}$ are significant.

4 Model Diagnostics

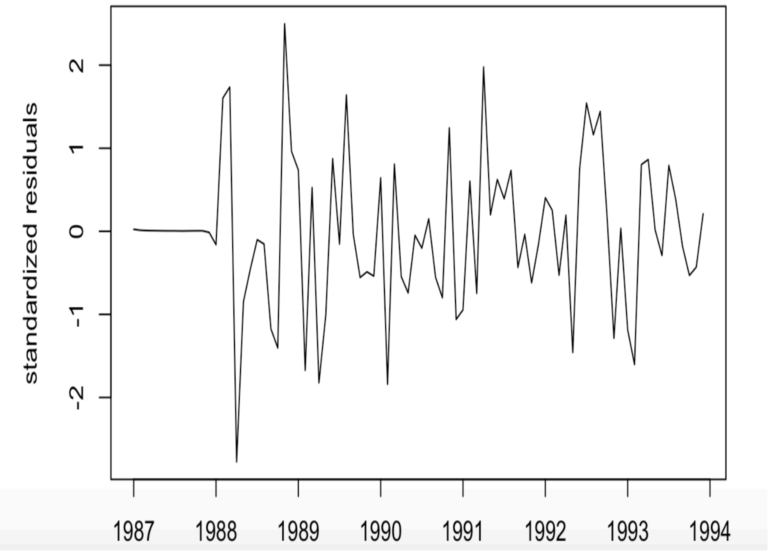


Figure 4.1 Plot of residuals

We can see from Figure 4.1, other than some strange behavior in the middle of the series, this plot does not suggest any major irregularities with the model.

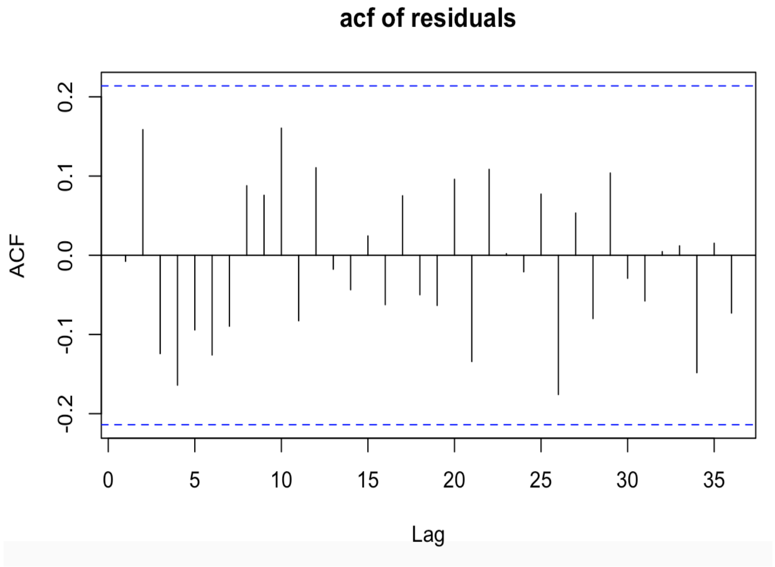


Figure 4.2 ACF Plot of residuals

Without marginal significance at any lag, the model seems to have captured the essence of the dependence in the series.

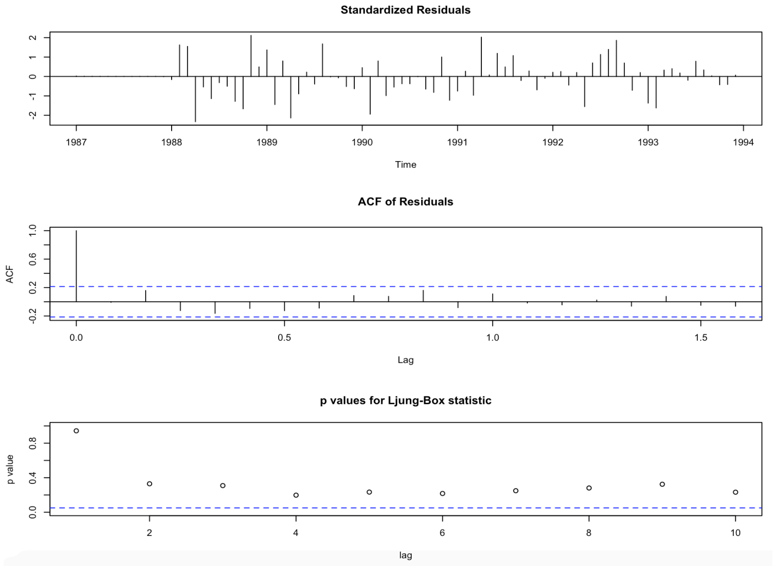
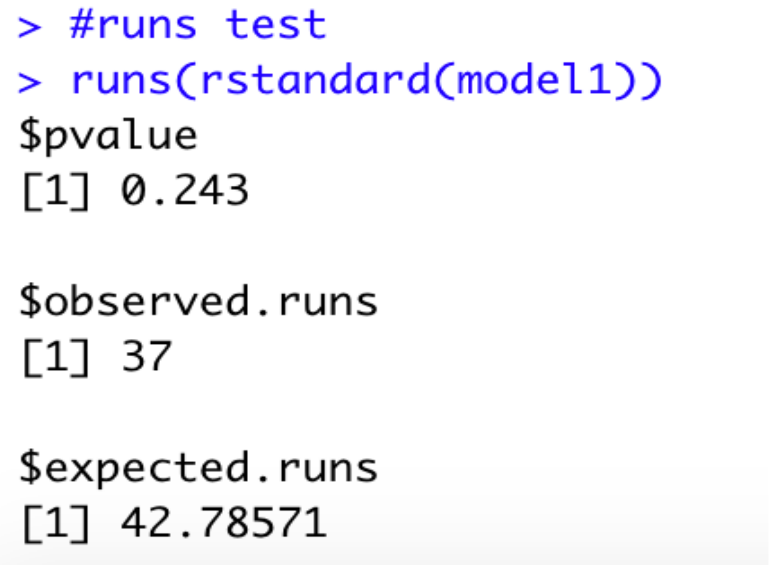


Figure 4.3 Ljung-Box Test

Ljung-Box Tests: to test whether the error terms are correlated.

H_0: the error terms are uncorrelated.

According to the good results above, especially all the points of p-value are above the 0.05-line, we can say that the model has captured the dependence in the time series.



```
> #runs test
> runs(rstandard(model1))
$pvalue
[1] 0.243

$observed.runs
[1] 37

$expected.runs
[1] 42.78571
```

Figure 4.4 Run Test

The Run test is a method to check whether the occurrence of data in a sequence is independent of the order. Here, we want to check the independence of standardized residuals. We get the p-value greater than 0.05, thus we don't reject H_0 that the standardized residuals series are independent white noise.

To check the normality of the error terms, we plot the histogram of the residuals and its QQ plot, as well as doing the Shapiro-Wilk normality test.

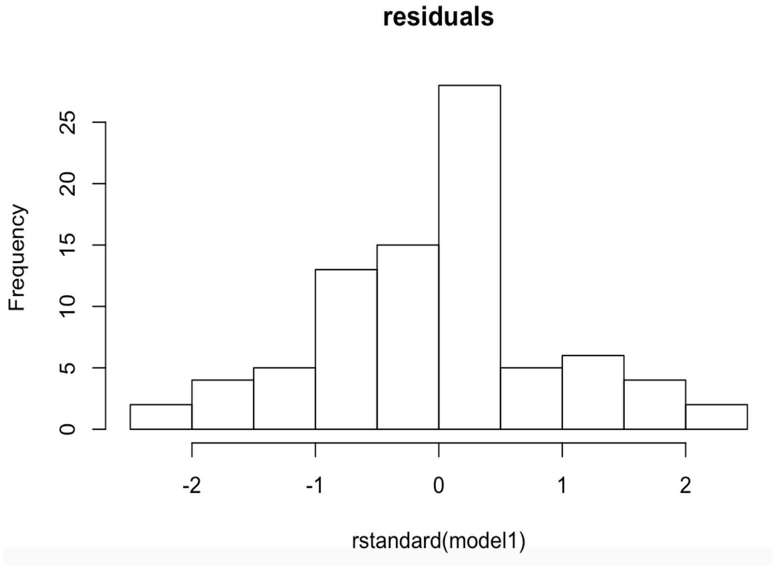


Figure 4.5 The histogram of the residuals

From the histogram above, which is somewhat “bell-shaped”, we can get some possibility of ”normality”.

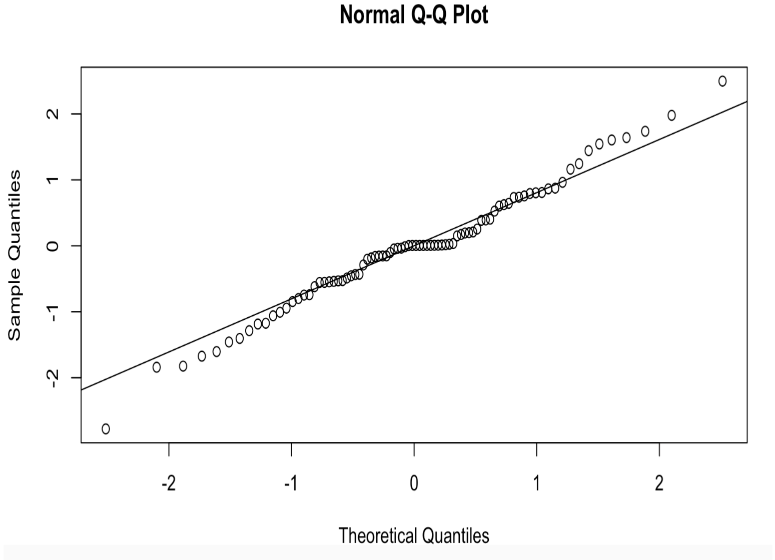
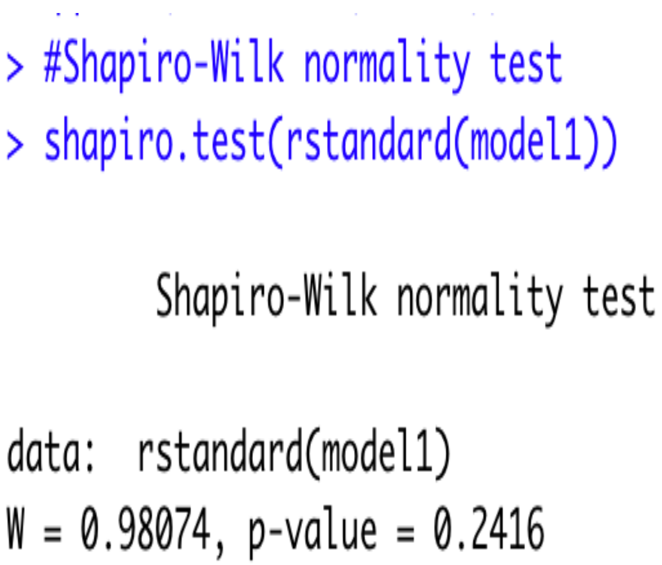


Figure 4.6 Q-Q plot

Normality is certainly not rejected, since most of the points are still on the line or in the region where the line is.



```
> #Shapiro-Wilk normality test
> shapiro.test(rstandard(model1))

        Shapiro-Wilk normality test

data:  rstandard(model1)
W = 0.98074, p-value = 0.2416
```

Figure 4.7 Shapiro-Wilk normality test

Shapiro-Wilk normality test, also known as the W test, mainly tests whether the research object (here the error terms) conforms to the normal distribution.

The calculated results showed that the W statistic (0.98074) is close to 1 and the p-value is significantly greater than 0.05, so we could not reject its normal distribution.

5 Forecasting with R

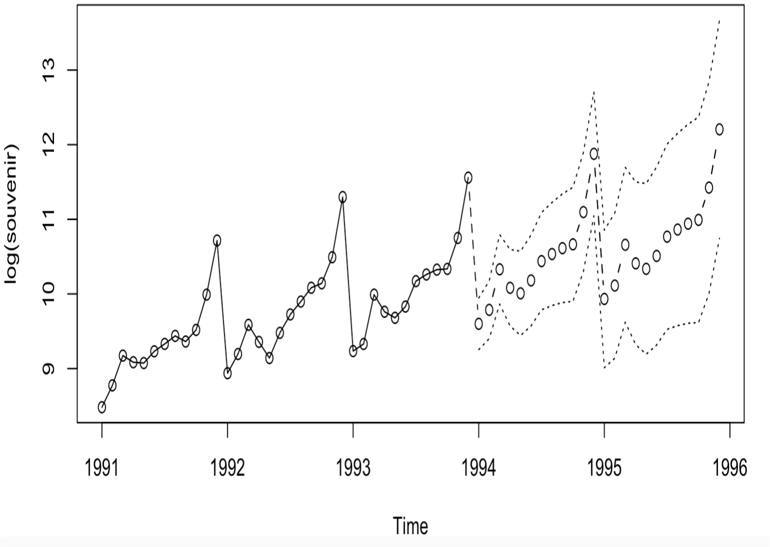


Figure 5.1 Forecasting for a lead time of two years

We do forecasting with prediction limits, the above figure 5.1 presents the forecasts and 95% forecast limits for a lead time of two years for the ARIMA(0,1,1)$\times(0,1,1)_{12}$ model that we fit for the souvenir sales data.

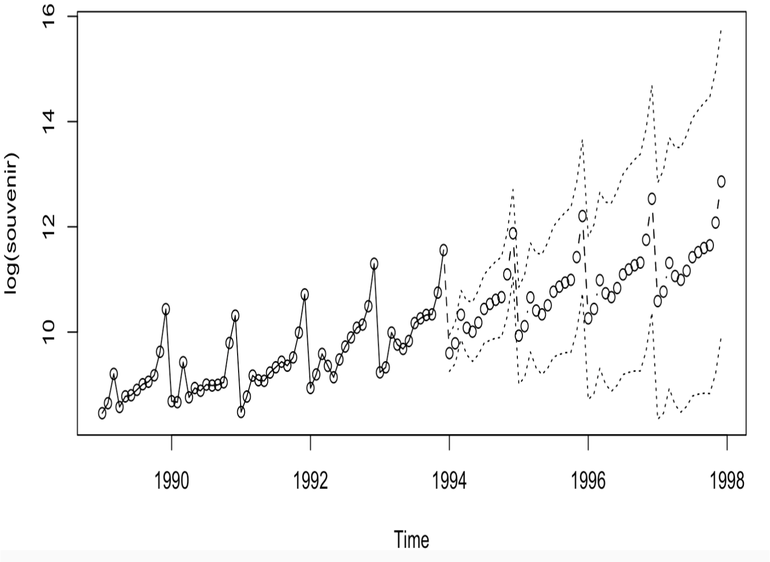


Figure 5.2 Forecasting for four years

The plot above displays the last two years of observed data and forecasts out four years.
At this lead time, it is easy to see that the forecast limits are getting wider, as there is more uncertainty in the forecasts.

6 Conclusion Remakes

We use helpful methods for specifying models and for efficiently estimating the parameters in the model we get. Also, by model diagnostics, we test the goodness of fit of a model. 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 times. Of equal importance is the assessment of the precision of those forecasts. Therefore, we consider the calculation of forecasts and their properties, which help us get a deeper understanding of the model fitting and diagnostics, at the same time, we realize their practical application in real life.

7 Rcode

```
souvenir <- scan("http://robjhyndman.com/tsdldata/data/fancy.dat")

###Data

library(TSA)
library(tseries)
souvenirts <- ts(souvenir, frequency=12, start=c(1987,1))
souvenirts

#plot time series
plot(souvenirts,type='o',ylab="Sales",
      main="Time series plot of monthly sales of souvenir")
souvenirlog=log(souvenirts) #take log
plot(souvenirlog,type='o',ylab="Sales",
      main="Time series plot of Log(monthly sales of souvenir)")
```

```
#plot acf
acf(as.vector(souvenirlog),lag.max = 36,
    main=”sample␣ACF␣of␣Log(monthly␣sales␣of␣souvenir)”)
###


### Model specification

#plot the time series after first difference
plot(diff(souvenirlog),
     main=”Difference␣of␣Log(monthly␣sales␣of␣souvenir)”)
points(y=diff(souvenirlog),x=time(diff(souvenirlog)),
       pch=as.vector(season(diff(souvenirlog))))

#plot acf after first difference
acf(as.vector(diff(souvenirlog)),lag.max = 36,
  main=”sample␣ACF␣of␣first␣difference␣of␣monthly␣sales␣of␣souvenir”)

#seasonal diff

#plot the time series after first and seasonal differences
plot(diff(diff(souvenirlog),lag=12),type=’l’,
  main=”First␣and␣seasonal␣Diff␣of␣Log(monthly␣sales␣of␣souvenir)”)
points(diff(diff(souvenirlog),lag=12),
       x=time(diff(diff(souvenirts),lag=12)),
       pch=as.vector(season(diff(diff(souvenirts),lag=12))))
#ADF unit root test
adf.test(diff(diff(souvenirlog),lag=12))

#plot acf after first and seasonal differences
acf(as.vector(diff(diff(souvenirlog),lag=12)),
    lag.max = 36,ci.type=”ma”,
    main=”sample␣ACF␣of␣first␣and␣seasonal␣differences␣of
␣␣␣␣Log(monthly␣sales␣of␣souvenir”)

#consider specifying the multiplicative seasonal ARIMA model
#for log(souvenir)
###
```

```
eacf(as.vector(diff(diff(souvenirlog),lag=12)))
plot(armasubsets(y=diff(diff(souvenirlog),lag=12),
                 nar=10,nma=10,ar.method="ols"))
###Model fitting

model1=arima(souvenirlog,order=c(1,1,0),
             seasonal = list(order=c(0,1,1),period=12))
model2=arima(souvenirlog,order=c(1,1,1),
             seasonal = list(order=c(0,1,1),period=12))
model1$aic
model2$aic

#Therefore, model1 is better, consider seasonal
#ARIMA(0,1,1)*(0,1,1) for log(souvenir)

###


###Diagnostic checking

model1
#plot of residuals
plot(rstandard(model1),ylab="standardized␣residuals")
#acf plot of residuals
acf(as.vector(rstandard(model1)),lag.max = 36,main="acf␣of␣residuals")
#Ljung-Box test
tsdiag(model1)

##Check the normality of the residuals
#plot the histagram
hist(rstandard(model1),main="residuals")
#qqplot
qqnorm(rstandard(model1))
qqline(rstandard(model1))
#Shapiro-Wilk normality test
shapiro.test(rstandard(model1))
#Ljung-Box test
```

```
Box.test(rstandard(model1), type='Ljung-Box')
#runs test
runs(rstandard(model1))
#Residuals have normality
###

###Forecasting

#Forecast next 2 years
plot(model1,n1=c(1991,1),n.ahead = 24,ylab="log(souvenir)")
#Forecast next 4 years
plot(model1,n1=c(1989,1),n.ahead = 48,ylab="log(souvenir)")
```