Notes

This is a .ipynb file created in Google Colab .However, all of the code for this assignment is in R language.

```
[98] install.packages('TSA')
     install.packages('stats')
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

Warning message:
"package 'stats' is a base package, and should not be updated"
```

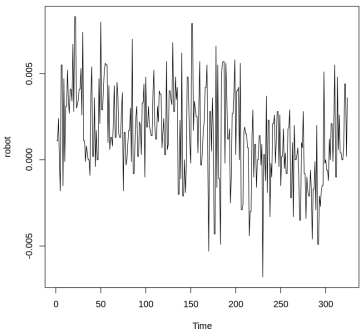
```
[ ] library('TSA')
    library('stats')
```

```
Attaching package: 'TSA'


The following objects are masked from 'package:stats':

    acf, arima


The following object is masked from 'package:utils':

    tar
```

```
[95] require(graphics)
```

2. (15%) Recall the dataset "robot" firstly introduced in TSA HW06.

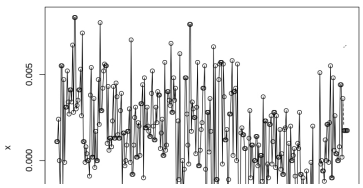
```
[59] rawRobotData <- read.csv(file = 'TSA HW06.robot.csv')
     robotData = ts(rawRobotData)
     plot(robotData)
```

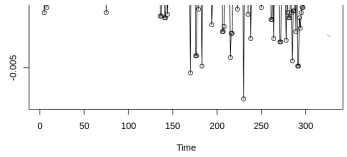


(a) Use IMA(1, 1) to forecast five values ahead and calculate the 95% confidence intervals.

```
[72] model <- arima(robotData, order = c(0,1,1))
     model
     # plot the prediction
     plot(model, n.ahead=5)$pred
     # plot the upper prediction limit
     plot(model, n.ahead=5)$upi
     # plot the lower prediction limit
     plot(model, n.ahead=5)$lpi
```

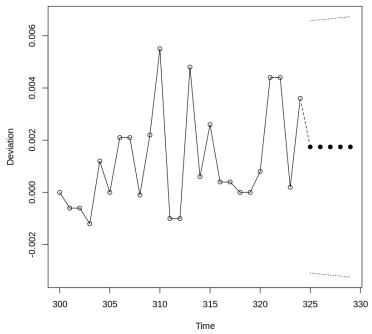
```
Call:
arima(x = robotData, order = c(0, 1, 1))

Coefficients:
          ma1
      -0.8713
s.e.   0.0389

sigma^2 estimated as 6.069e-06:  log likelihood = 1480.95,  aic = -2959.9
A Time Series:
0.00174267189481582 · 0.00174267189481582 · 0.00174267189481582 · 0.00174267189481582 · 0.00174267189481582
A Time Series:
0.00657134425328359 · 0.00661118310258771 · 0.00665069858738487 · 0.00668989845620757 · 0.00672879015299119
A Time Series:
-0.00308600046365195 · -0.00312583931295607 · -0.00316535479775323 · -0.00320455466657593 · -0.00324344636335955
```





(b) Display the actual values, the five forecasts and the 95% confidence intervals of the five forecasts, all in one graph. What do you observe?

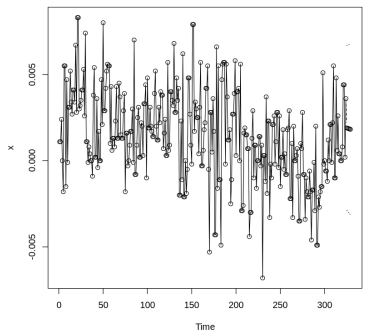
```
[61] plot(model,n1=300,n.ahead=5,ylab='Deviation',pch=19)
     abline(h=coef(model)[names(coef(model))=='intercept'])
```



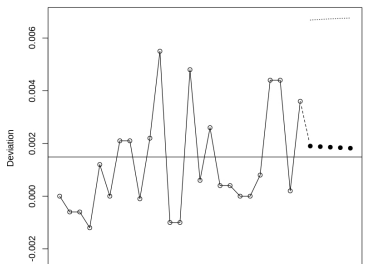
Use ARMA(1, 1) to forecast five values ahead and calculate the 95% confidence intervals. Compare the results with those in (a), what do you observe?

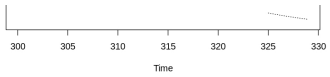
```
[62] model <- arima(robotData, order = c(1,0,1))
     model
     # plot the prediction
     plot(model, n.ahead=5)$pred
     # plot the upper prediction limit
     plot(model, n.ahead=5)$upi
     # plot the lower prediction limit
     plot(model, n.ahead=5)$lpi
```

```
Call:
arima(x = robotData, order = c(1, 0, 1))

Coefficients:
         ar1      ma1  intercept
      0.9472  -0.8062     0.0015
s.e.  0.0309   0.0609     0.0005

sigma^2 estimated as 5.948e-06:  log likelihood = 1489.3,  aic = -2972.61
A Time Series:
0.00190134843058665 · 0.00187944389457476 · 0.00185869511989287 · 0.00183904112445712 · 0.00182042414381558
A Time Series:
0.00668147288597162 · 0.00670688136785437 · 0.00672819325712715 · 0.00674597162041198 · 0.00676069967622052
A Time Series:
-0.00287877602479832 · -0.00294799357870484 · -0.0030108030173414 · -0.00306788937149774 · -0.00311985138858936
```




```
[63] plot(model,n1=300,n.ahead=5,ylab='Deviation',pch=19)
     abline(h=coef(model)[names(coef(model))=='intercept'])
```





Compare the results with those in (a), what do you observe?

The prediction, upper limit prediction and lower prediction limit of the ARMA(1,1) model is only marginally higher than those of IMA(1,1) model across the board.

Both the IMA(1,1) and ARMMA(1,1) yield nearly identical forecast.

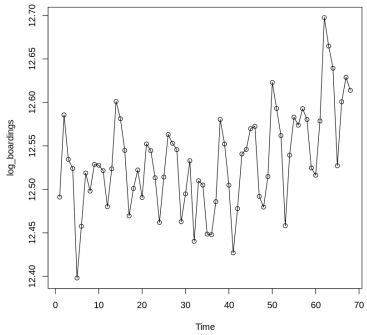
Discussions for question 2

- In the IMA(1,1) model, it can be observed that the forecast vales remain constant. This is because the model only has the MA components. MA components cannot forecast using future noise values. Thus, the forecast is consant.
- However, in the ARMA(1,1) model, the addition of the AR part allows the future values to be takekn into consideration. As a result, a small downward trend in the forecast can be observed.

3. (15%) The dataset “boardings” contains the monthly number of passengers who boarded light rail trains and buses in Denver, Colorado, from August 2000 to March 2006.

(a) Plot the time series and tell your observation if there exists seasonality and if the series is stationary.

```
[64] rawBoradingsData <- read.csv(file = 'TSA HW08.boardings.csv')
     boardingsData = ts(rawBoradingsData)
     plot(boardingsData)
     points(boardingsData)
```



Comments

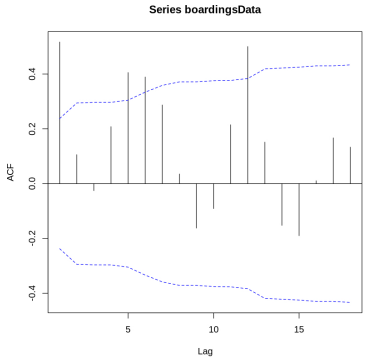
Based on the graph, there is a **clear seasonality** dislpayed. The data shows the log_boardings value that **peaks every 12 data points and reaches a relative minimum 3 data points following the peak**.

Considering that this is a boardings data, it can be deduced that each time point represents a month, and the peak occurs during the holiday period.

In the big picture, the plot shows a gradually upward trend as time progresses, suggesting that the data **might be nonstationary.**

(b) Plot the sample ACF and see what are the significant lags?

```
[65] acf(boardingsData, ci.type='ma')
```



Comments

The ACF plot shows that there is significant correlation at lags 1, 5, 6, and 12.

(c) Fit the data with ARMA(0,3)×(1,0)12, evaluate if the estimated coefficients{$\theta 1, \theta 2, \theta 3, \phi 12$} are significant. Hint: you need to check the associated standard errors “s.e.” to the estimated coefficients to know if the coefficients are significant, via hypothesis testing.

```
[66] model = arima(boardingsData, order = c(0,0,3), seasonal= list(order = c(1,0,0), period = 12));
     model
```

```
Call:
arima(x = boardingsData, order = c(0, 0, 3), seasonal = list(order = c(1, 0,
    0), period = 12))

Coefficients:
         ma1     ma2     ma3    sar1  intercept
      0.7290  0.6116  0.2950  0.8776    12.5455
s.e.  0.1186  0.1172  0.1118  0.0507     0.0354

sigma^2 estimated as 0.0006542:  log likelihood = 143.54,  aic = -277.09
```

Comments

Approximately 95% of the observations should fall within plus/minus 2*standard error of the regression from the regression line. This condition is met by all the coefficients.

Thus, all coefficients $\theta 1, \theta 2, \theta 3, \phi 12$ are significant.

Discussions for question 3

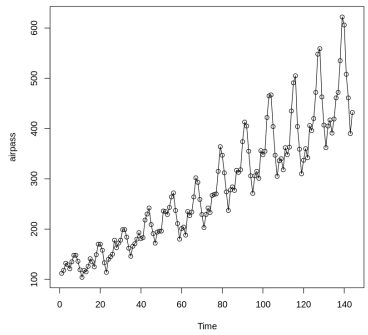
- Plotting the ACF, as always, help identify the significant lags that should be paid attention too when formulating the seasonal ARIMA model.
- For the chosen 95% confidence, a = 0.05, if p-value is less than 0.05, then the null hypothesis that the data are normally distributed is rejected. If the p-value is greater than 0.05, then the null hypothesis is not rejected. Using this fact, the significance of each parameter can also be used to detmine the significance of each coefficients.

4. (30%) The monthly airline passengers, first investigated by Box and Jenkins in 1976, is considered as the classic time series dataset (see “TSA HW08.airpass.csv”).

(a) Plot the time series in its original scale and the log-transformed scale. Do you think making the log-transformation is appropriate?

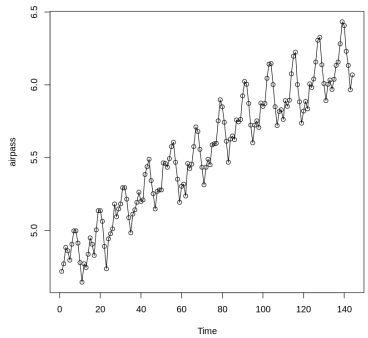
Original Scale

```
[67] rawAirpassData <- read.csv(file = 'TSA HW08.airpass.csv')
     airpassData = ts(rawAirpassData)
     plot(airpassData)
     points(airpassData)
```



Logarithmic-transformed Scale

```
[68] plot(log(airpassData))
     points(log(airpassData))
```



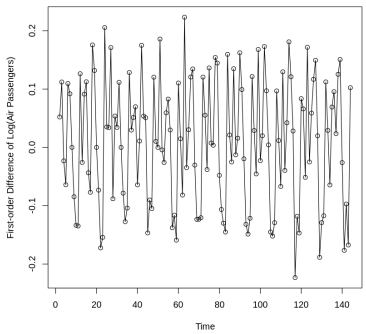
Comments

The log-transformed model displays seasonal behaviors and upward trends just like the original one. However, it displays the upward trend at a uch more linear or consistent manner.

(b) Make the first-order difference over the “log-transformed” data. What do you observe?

```
[69] plot(
       diff(log(airpassData)),
```

```
      type = 'o',
      ylab='First-order Difference of Log(Air Passengers)'
    )
```

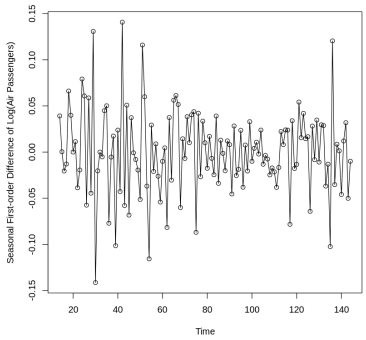


Comments

From the plot, it can be observed that the seasonality is lost as the peaks and troughs become inconsistent. For instance, the number of periods taken to reach the next peak is no longer constant.

(c) Make a seasonal difference of the resulted series in (b), what do you observe?

```
[74] plot(
       diff(diff(log(airpassData)), lag = 12),
       type = 'o',
       ylab='Seasonal First-order Difference of Log(Air Passengers)'
     )
```

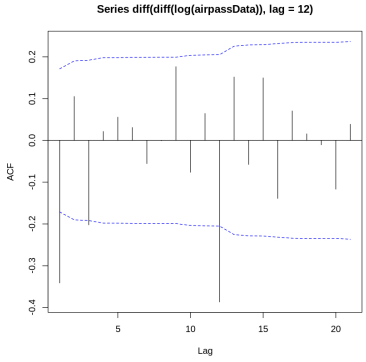


Comments

The seasonal difference of the log-transformed data makes the lack of seasonality even clearer. The data points with high difference appears random in magnitude and the time in which it occurs.

(d) Plot the sample ACF of the resulted series in (c), explain what you see.

```
[75] acf(diff(diff(log(airpassData)), lag = 12), ci.type='ma')
```



Comments

From the plot, it can be inferred that at lag 1 and 12, there exist significance correlation.

It can also be seen that at lag = 3, the ACF is greater than the threshold, but only marginally.

Thus, only lag = 1 and 12 are worth investigating.

(e) Fit an ARIMA(0,1,1) × (0,1,1)12 model to the log-transformed series. Diagnose the residuals of this model,including the sample ACF and the normality test.

Fitting an ARIMA(0,1,1) × (0,1,1)12 model

```
[102] model = arima(log(airpassData), order = c(0,1,1), seasonal= list(order = c(0,1,1), period = 12));
      model
```

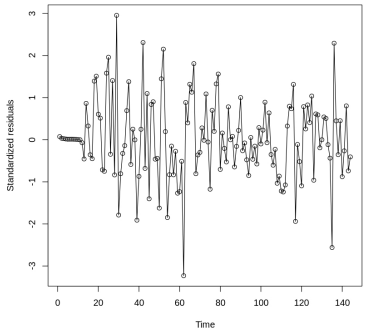
```
Call:
arima(x = log(airpassData), order = c(0, 1, 1), seasonal = list(order = c(0,
    1, 1), period = 12))

Coefficients:
          ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348:  log likelihood = 244.7,  aic = -485.4
```

Comments

Judging from the standard error of ma1 and sam1 coefficients, both the seasonal and nonseasonal parameters are significant.

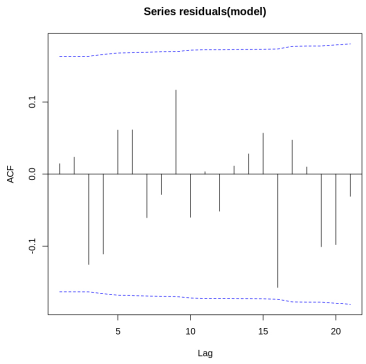
Diagnosis of the residuals

```
[121] plot(rstandard(model),ylab='Standardized residuals',type='o')
```



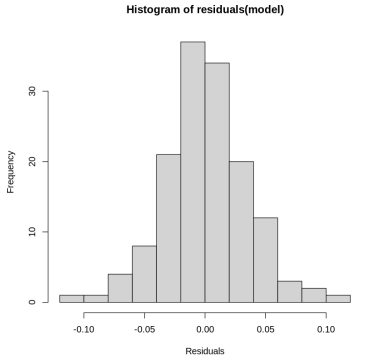
ACF of the log-transformed data

```
[124] acf(residuals(model), ci.type='ma')
```



Normality Test

```
[125] hist(residuals(model),xlab = 'Residuals')
```



Shapiro Test

```
[107] shapiro.test(residuals(model))
```

```
        Shapiro-Wilk normality test
```

```
data:  residuals(model)
W = 0.98637, p-value = 0.1674
```

Comments

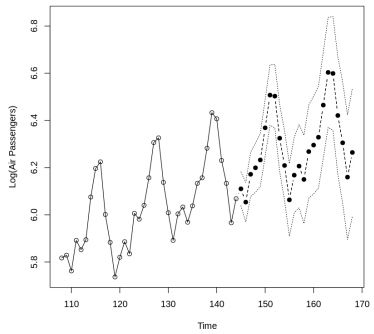
The standard residuals plot, and the ACF of residuals plot show that there are no apparent concerns with the model. There are no outliers observed and the correlation between the residuals are not found significant in both the joint and individual cases.

In terms of the normality, the histogram of the residuals show a normal distribution shaped result. This can be confirmed with the Shapiro test.

For the chosen 95% confidence, a = 0.05, if p-value is less than 0.05, then the null hypothesis that the data are normally distributed is rejected. If the p-value is greater than 0.05, then the null hypothesis is not rejected. In this case, p value is higher than 0.05, meaning that the Shapiro Walk test suggests that the residuals are normally distributed.

(f) Make forecasts for “two” years based on the model in (e). The confidence intervals shall be included.

```
[112] plot(model,n1=108,n.ahead=24,pch=19,ylab='Log(Air Passengers)')
```



Discussions for question 4

- In question 4, it is shown that the logarithmic transformation can assist in the formulation of seasonal models by making the trend more clearly defined.
- By doing the first-order difference equation, it can be more clearly seen whether the seasonality of data actually exists compared to just looking at the shape of the data's plot directly.
- Plotting the ACF, as always, help identify the significant lags that should be paid attention too when formulating the seasonal ARIMA model.
- Analyzing the resituals is a good method of identifying whether the seasonal ARIMA model is effective in modelling the data. By identifying that the residuals themselves has no significance and are normally distributed, we can prove the validity of the SARIMA model.