**Stat 510 Week 4 Homework**

**Submit answers in a Word document or a pdf to the designated ANGEL Drop Box in the Drop Boxes folder.**

**1**. Problem 3.36 of the book (p. 170) asks students to sketch the ACF of a seasonal ARIMA (0,1)×(1,0)12 model with the non-seasonal MA = .5 and the seasonal AR . (They left out the differencing location, which means there is no differencing.)

Here’s how you can do it in R:

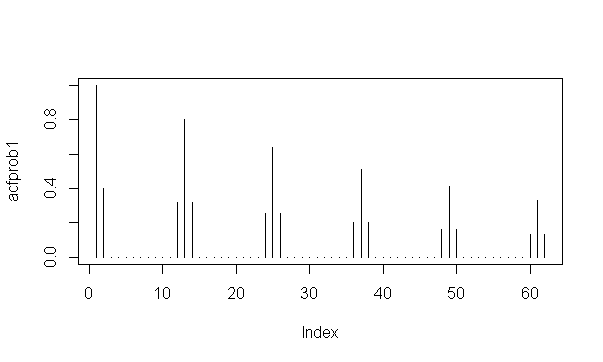
acfprob1=ARMAacf(ar = c(0,0,0,0,0,0,0,0,0,0,0,.8), ma=c(.5),lag.max=61)

plot(acfprob1, type="h")

The first command asks for 61 lags. The ar list gives coefficients that are 0 up to the 12th lag where the coefficient = .8. The ma list gives the information that the first lag coefficient = .5. In the plot command, the type = “h” causes projections from the value to the axis as we usually do in an ACF.

a. Use R to sketch the ACF in this situation. You may either type the plot command in R rather than copy and paste or you must paste as plain text. R doesn’t like Words’ quote marks. Give the ACF graph as the answer to this part.

The plot is as follows:

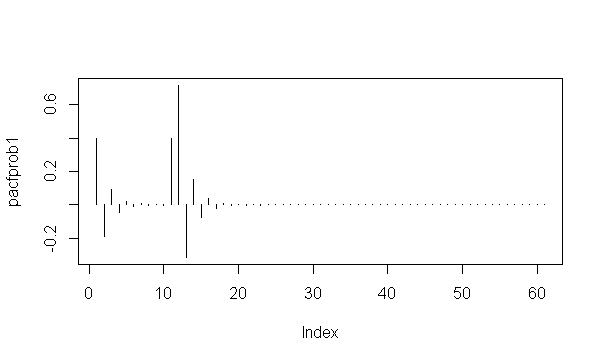


b. To get the PACF of this model, modify the commands in the following way:

pacfprob1=ARMAacf(ar = c(0,0,0,0,0,0,0,0,0,0,0,.8), ma=c(.5),lag.max=61, pacf=TRUE)

plot(pacfprob1, type="h")

Use these commands to get the PACF of the model. Give the PACF as the answer to this problem.



c. Write a brief description of the ACF and PACF patterns for this model.

We have a model represented as ARIMA(p, d, q) × (P, D, Q)S

ARIMA(0, 0, 1) × (1, 0, 0)12 (non-seasonal MA(1) and seasonal AR(1) model

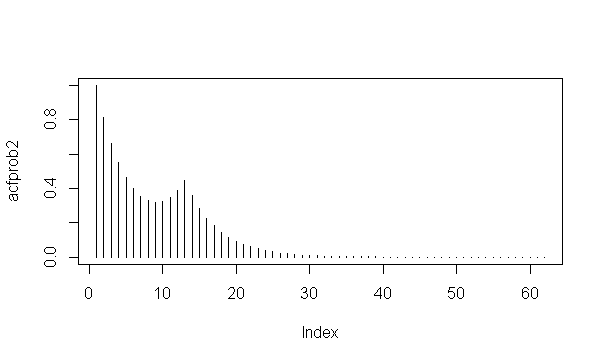
The non-seasonal components are: MA(1): θ(B) = 1 + θ1B

The seasonal components are: Seasonal AR(1): Φ(B12) = 1 - Φ1B12

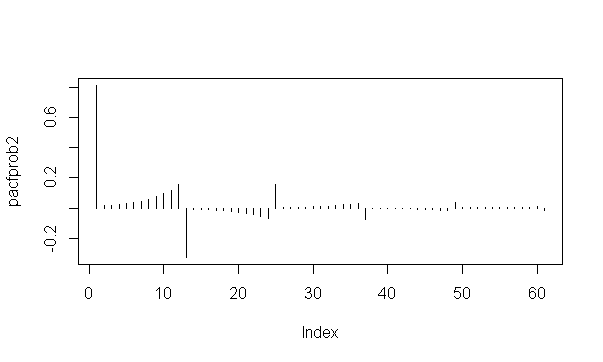
The plots have:

* Spikes in the ACF (at low lags) followed by a cut-off indicate non-seasonal MA terms. This is reinforced by tailing off of the PACF for low lags
* Spikes in the PACF (at seasonal lags of 11, 12 and 13) indicated possible seasonal AR terms.

d. Now consider the ARIMA (1,0)×(0,1)12 model with non-seasonal AR and seasonal MA =.5. Graph the ACF of this model. (Reverse the seasonality options in the acfprob1 = command.) Give the graph as the answer to this part.



e. Refer to part d. Graph the PACF for the model. Give the graph as the answer to this part.



f. Write a brief description of the ACF and PACF patterns in parts d and e.

We have a model represented as ARIMA(p, d, q) × (P, D, Q)S

ARIMA(1, 0, 0) × (0, 0, 1)12 (seasonal MA(1) and non-seasonal AR(1) model

The seasonal components are: MA(1): θ(B12) = 1 + θ1B12

The non-seasonal components are: AR(1): Φ(B) = 1 - Φ1B

The plots have:

* The PACF shows a clear spike at lag 1 followed by insignificant values about lag 11. This is accompanied by a tapering pattern in the early lags of the ACF. A non-seasonal AR(1) is indicated.
* We look at what’s going on around lags 12, 24, and so on. In the ACF, there’s a rise of spikes around lag 12 and then continued tapering off. The PACF tapers in multiples of S; that is the PACF has significant lags at 12, 24, 36 and so on. A seasonal MA(1) component is indicated.

**2.** This is a more structured version of problem 3.39 in the book (p. 170) which asks students to fit a model of their choice to the log transformed Johnson and Johnson earnings data series. In Homework 1, we saw that this series of 84 quarterly measures of the earnings of this company had both trend and seasonality and that a log transform made the trend linear and the variance stable.

The data are in the dataset jj.dat. (Might already be on your computer from Homework 1, If not, it’s in the Datasets folder.)

Load the astsa library, read the dataset, and create the log transform with something like:

library(astsa)

earnings=scan("jj.dat")

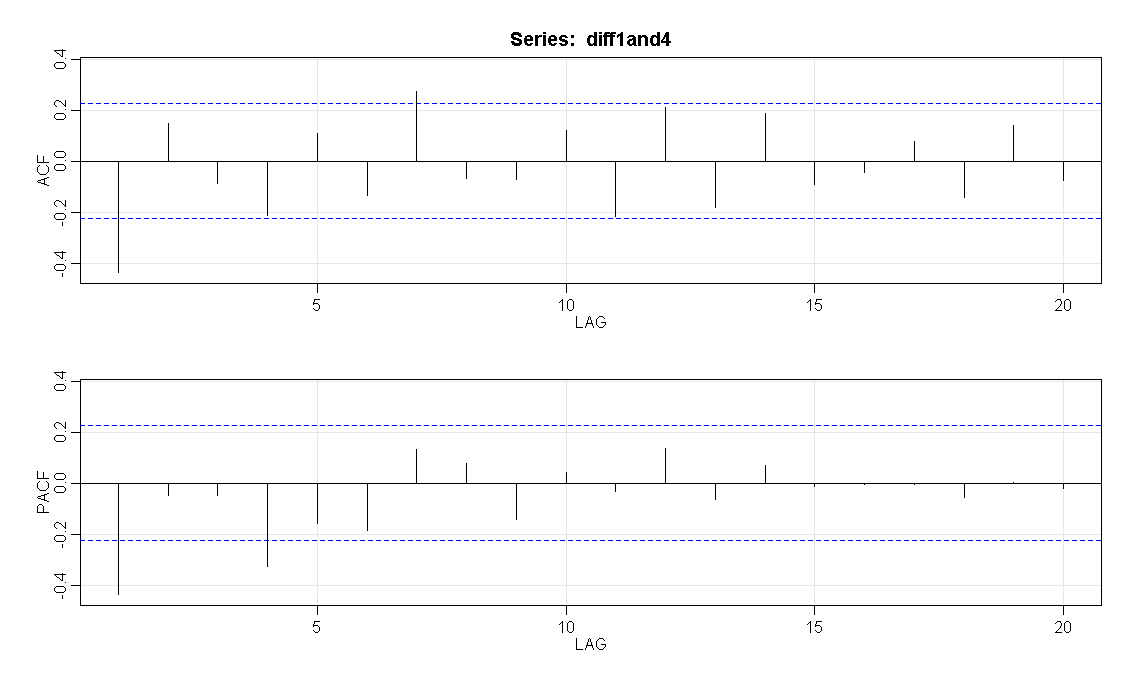
earnings = log(earnings)

Then, because we have linear trend and seasonality (quarterly effects), apply both first and fourth differences. Something like:

diff1 = diff(earnings,1)

diff1and4=diff(diff1,4)

a. Use acf2 to determine the ACF and PACF of the **differenced** data. Be sure that you supply the differenced series (diff1and4) as a parameter of the acf2 command. Give the ACF and PACF of the differenced data.



b. Consider the ACF and PACF values for the first two or three lags. Explain why they suggest either an MA(1) or AR(1) as a possible non-seasonal component.

In the first 2/3 lags, we observe:

* The PACF has a spike at (1) and then cuts off almost entirely till lag 4. The ACF is not very clear and can be seen as tailing off and cutting off.
  + Relying on the PACF though it points to an AR(1) model
  + The ACF is not giving a clear signal. In itself with only a significant value at lag 1 it points to an MA(1) model.

c. Examine the ACF and PACF at lags 4, 8 and 12. What possibilities are suggested for a seasonal component of the model?

In the seasonal lags at 4, 8 and 12, we observe

* The PACF has a spike at lag4 indicating a seasonal AR(1) model. However there is no clear cut off
* The ACF also gives mixed signal. There is almost a significant value at lag 4, 7, 11 and 12 but there is no clear tapering or cutting off. The graph can be interepreted as either MA(2) or AR(1) seasonal model

d. The ACF and PACF don’t actually give a clear pointer to the best model in this case, but here are some possibilities (Note that this list is not all-inclusive):

ARIMA (1,1,0)×(0,1,1)4

ARIMA (0,1,1)×(0,1,1)4

ARIMA (0,1,1)×(1,1,0)4

ARIMA (1,1,0)×(1,1,0)4

Assuming that you’ve loaded the astsa library, these commands will estimate the four models above.

sarima ( name of log series, 1,1,0,0,1,1,4)

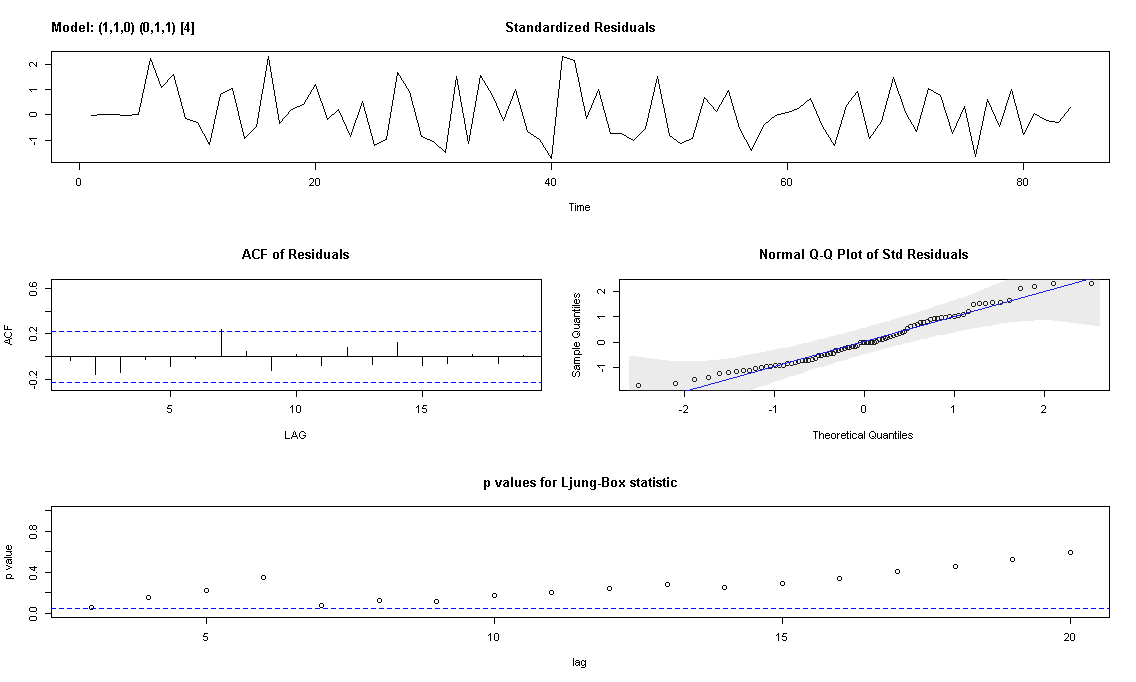
sarima ( name of log series, 0,1,1,0,1,1,4)

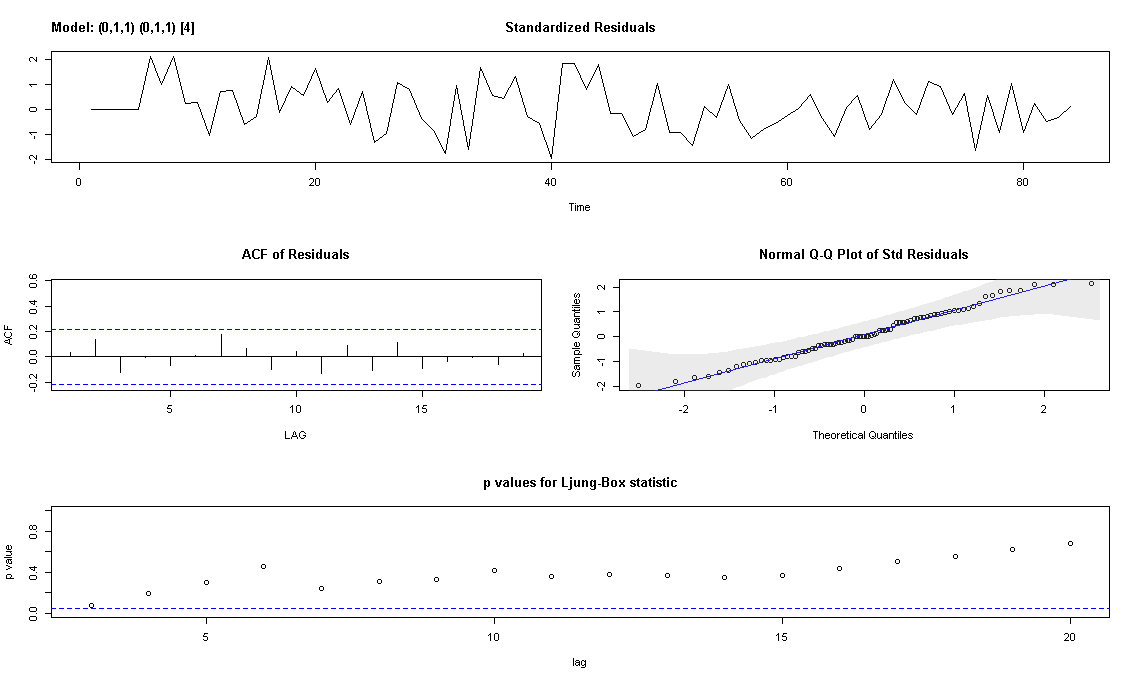
sarima ( name of log series, 0,1,1,1,1,0,4)

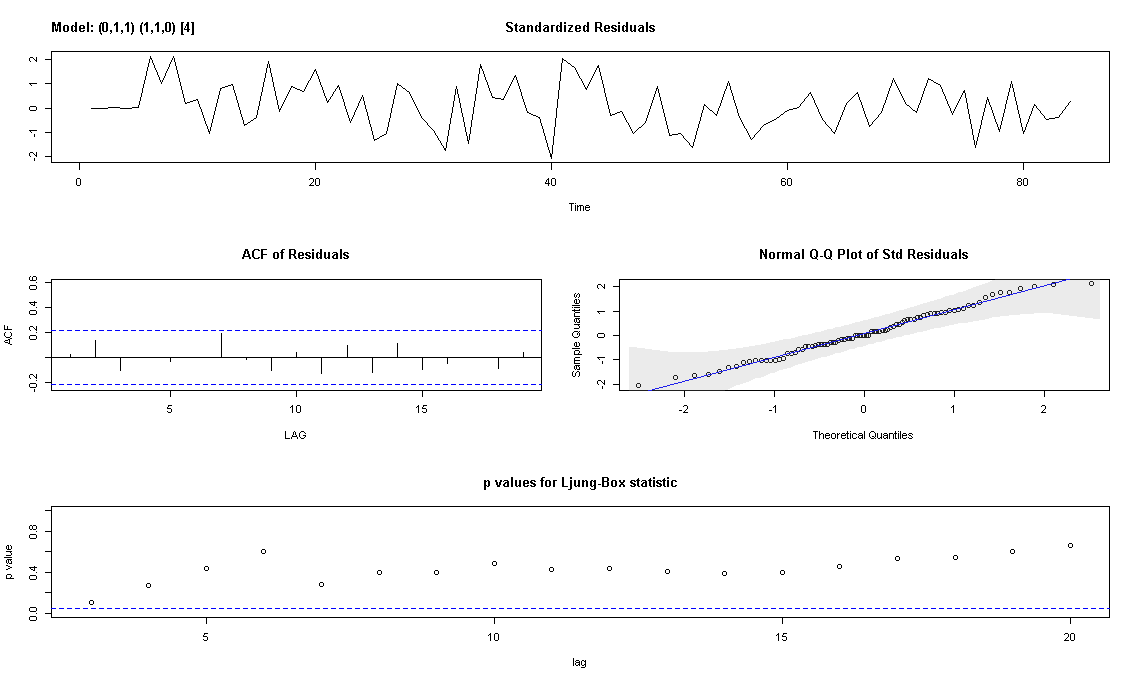
sarima ( name of log series, 1,1,0,1,1,0,4)

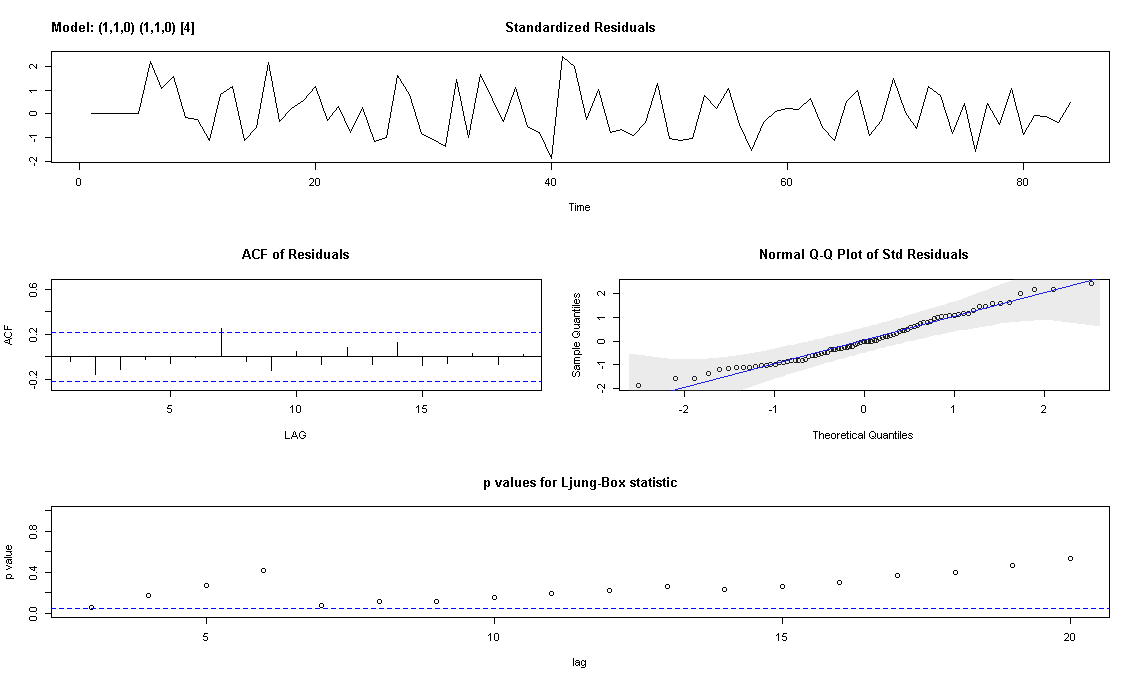
Estimate the four models above. For each give the variance estimate and a brief description of the model diagnostics. (You don’t have to give all the graphs.)

|  |  |
| --- | --- |
| **Model** | **Notes** |
| Non Seasonal AR with diff  Seasonal MA with diff  ARIMA (1,1,0)×(0,1,1)4 | Variance: 0.008472  $AIC [1] -3.723399  $AICc [1] -3.696018  $BIC [1] -4.665522  Diagnostics:   * The ACF at lag 7 is significant – not desirable result * Normality assumption is suspect * Ljung statistic is not desirable for lags 1 and 7 |
| Non Seasonal MA with diff  Seasonal MA with diff  ARIMA (0,1,1)×(0,1,1)4 | Variance: 0.007931  $AIC [1] -3.789403  $AICc [1] -3.762022  $BIC [1] -4.731526  Diagnostics:   * No significant ACF – desirable result * Normality assumption has improved from model 1 * Ljung statistic is not desirable for lags 1 * Overall very similar to model 1 |
| Non Seasonal MA with diff  Seasonal AR with diff  ARIMA (0,1,1)×(1,1,0)4 | Variance: 0.007913  $AIC [1] -3.791572  $AICc [1] -3.764191  $BIC [1] -4.733696  Diagnostics:   * No significant ACF – desirable result * Normality assumption has improved from model 1 * Ljung statistic is desirable for all lags |
| Non Seasonal AR with diff  Seasonal AR with diff  ARIMA (1,1,0)×(1,1,0)4 | Variance: 0.008468  $AIC [1] -3.723852  $AICc [1] -3.696471  $BIC [1] -4.665975  Diagnostics:   * Significant ACF at lag 7 – not desirable result * Normality assumption is suspect * Ljung statistic is not desirable for lags 1 and 7 |









e. Which of the four models in part d do you think is the best model? Explain.

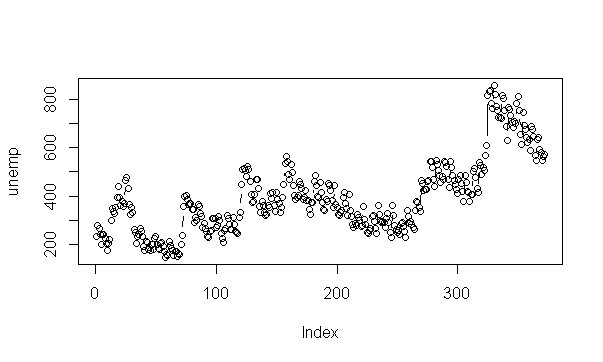
The models have very similar outputs but I think we should choose model 3 - Non Seasonal MA with diff and Seasonal AR with diff - ARIMA (0,1,1)×(1,1,0)4 because of the following:

* Excluded model 1 and 4 based on the diagnostics for the residuals. In both of these we see some significant ACF at lag 7. We also find that for both the Ljung statistic is not desirable for lags 1 and 7
* To choose between model 2 and model 3, I chose the one with lower values of AIC, AICc and BIC

**3.** This is a more structured version of problem 3.37 on page 170 of the book. The problem asks students to determine a model for a monthly series of number of unemployed persons in the United States. There are n = 372 data values, the monthly numbers for 1948-1978.

The data are in the file unemp.dat in the Datasets folder.

There is an upward trend because these are numbers of people rather than rates, so the number increases as the population size increases.



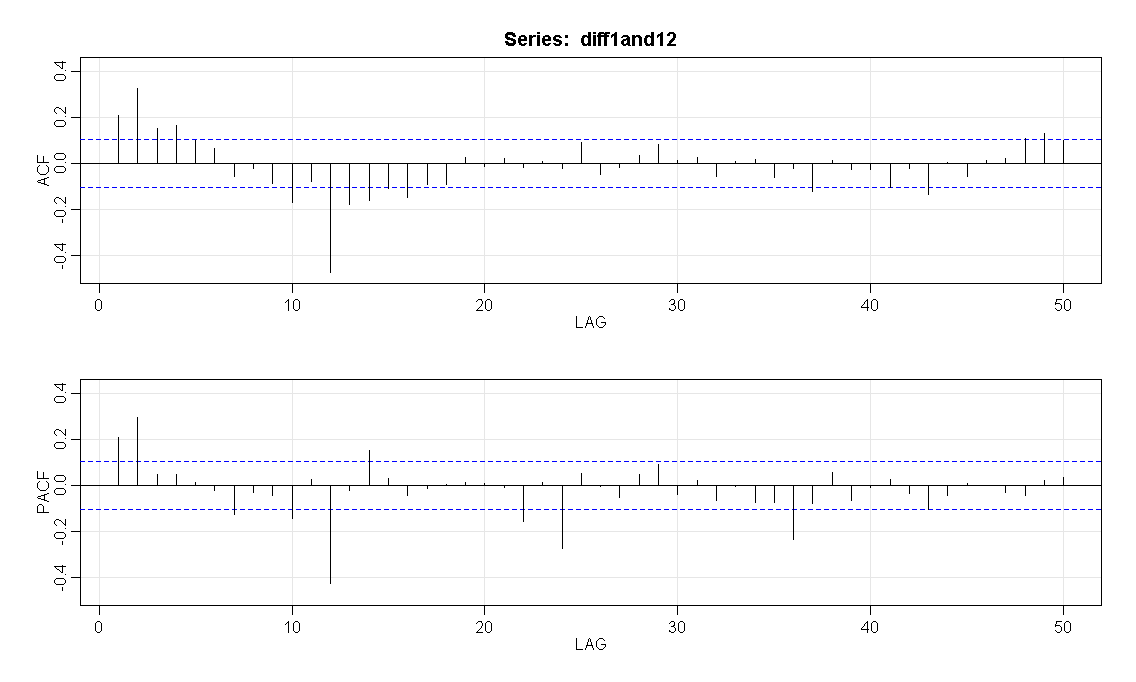
There also is seasonality. Unemployment numbers are affected by such things as seasonal agricultural production, holiday seasons, and so on.

a. Calculate a variable that is the result of applying a 1st difference and then a 12th difference. Something like the following should do the trick:

diff1=diff(x,1)

diff1and12=diff(diff1,12)

Then determine the ACF and PACF of the differenced data series (diff1and12). Be sure to use the differenced series, not the original data for this step.



(i) Discuss why the ACF and PACF suggest that an AR(2) may be good for the non-seasonal component. (Look at what’s happening for the first few lags in the ACF and PACF.)

In the first few lags we see the following aspects that suggest an AR(2) model:

* The PACF is significant for the first 2 lags and then is cut off. This is in alignment with the expected behavior from an AR(2) series
* The ACF indicates both a trailing off and a sinusoidal behavior converging to 0, which are also in alignment with the expected behavior from an AR(2) series

(ii) Discuss why the ACF and PACF suggest that a seasonal MA(1) may be good for the seasonal component. Look at what’s happening around lags 12, 24, and so on.

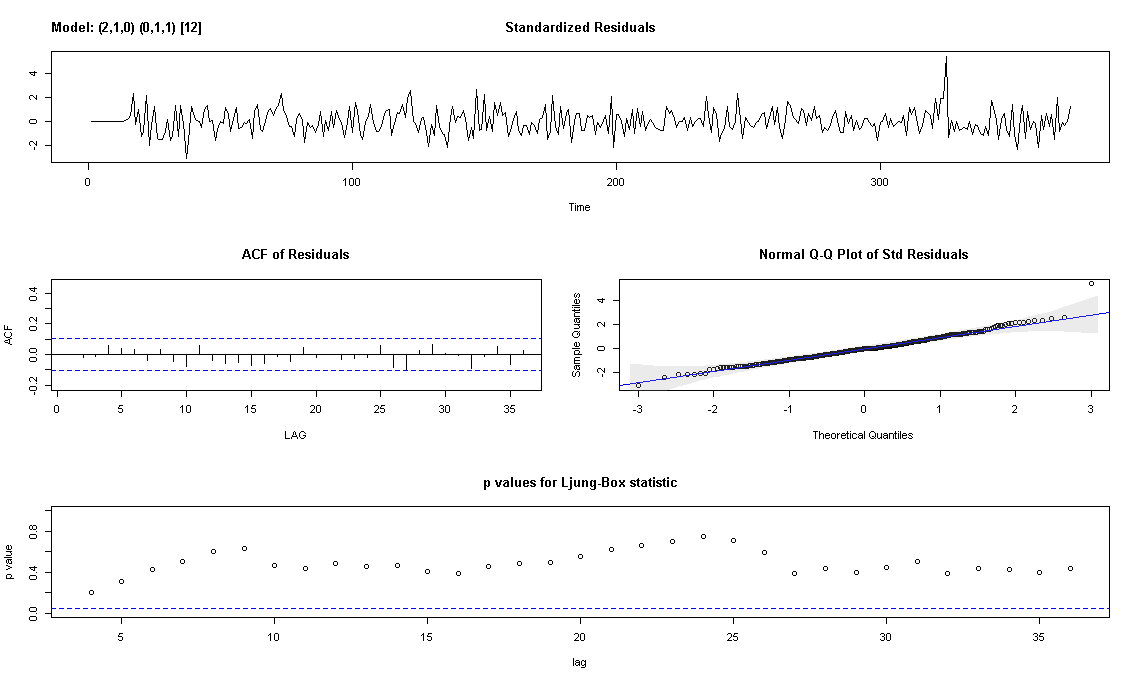
In the lags at 12, 24 we see the following aspects that suggest an MA(1) model:

* Spike in the ACF at first seasonal lag (lag 12) indicates a seasonal MA(1) term
* We also see that the PACF tails off at the seasonal lags of 12, 24, 36, 48 etc

Both of the above point towards a seasonal MA(1) model

b. Fit an ARIMA (2, 1, 0,)× (0,1,1)12 model to the original series. In the sarima command, specify the original series, not the differenced series because the differencing will be specified in the command itself. Discuss whether the diagnostics for the model suggest that the model is suitable. (Look at ACF of residuals, Ljung-Box-Pierce tests, and so on.)

We get the following:



The diagnostics have the following features:

* The time series plot of the standardized residuals mostly indicates that there’s no trend in the residuals, no outliers, and in general, no changing variance across time.
* The ACF of the residuals doesn’t show significant autocorrelations – a good result.
* The Q-Q plot is a normal probability plot – The plot looks almost linear and the assumption of normally distributed residuals holds.
* The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. All p-values are above the dashed blue line - That’s a good result.

c. Use the model fit in part b to forecast the unemployment values for the next 12 months following the end of the series. One possible command structure is

sarima.for (name of data series, 12, 2,1,0,0,1,1,12)

Give the 12 forecasts as the answer to this problem.

The forecasts are:

$pred

Time Series:

Start = 373

End = 384

Frequency = 1

**[1] 676.4664 685.1172 653.2388 585.6939 553.8813 664.4072 647.0657 611.0828 594.6414**

**[10] 569.3997 587.5801 581.1833**

$se

Time Series:

Start = 373

End = 384

Frequency = 1

[1] 21.20465 32.07710 43.70167 53.66329 62.85364 71.12881 78.73590 85.75096

[9] 92.28663 98.41329 104.19488 109.67935

