**Stat 510 Week 3 Homework**

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

**1.** An AR(1) model is estimated to be with the usual assumptions about . There were *n* = 80 observations, , and 3. Psi-weights  to  as given by R are:

[1] −0.70000 0.49000 −0.34300 0.24010 −0.16807

A. Forecast the value of and determine a 95% prediction interval for.

x81 = 50 - 0.7 \* x80 + w81

x8081 = 50 - 0.7 \* 3 + 0 = 47.9

The standard error of the forecast error at time 81 is = sqrt(2.5 \* 1) = 1.581

The 95% prediction interval for the value at time 81 is 47.9 ± 1.581(1.96), which is 44.801 to 50.999. We are therefore 95% confident that the observation at time 81 will be between 44.801 to 50.999. If we repeated this exact process, then 95% of the computed prediction intervals would contain the true value of x at time 81.

B. Forecast the value of , the value at time 83, and determine a 95% prediction interval for .

Forecast time 82

x82 = 50 - 0.7 \* x81 + w82

x8082 = 50 - 0.7 \* 47.9 + 0 = 16.47

Forecast time 83

x83 = 50 - 0.7 \* x82 + w83

x8083 = 50 - 0.7 \* 16.47 + 0 = 38.471

Standard error is given by: = 

= sqrt(2.5 \* (1 + (-0.7)^2 + (0.49)^2)) = 2.0797

The 95% prediction interval for the value at time 83 is 38.471± 2.0797 (1.96), which is 34.395 to 42.547. We are therefore 95% confident that the observation at time 83 will be between 34.395 to 42.547. If we repeated this exact process, then 95% of the computed prediction intervals would contain the true value of

2. For this problem, use the dataset sim1week3.dat from the Week 3 folder. Download it to the directory you’ve set up for course datasets. It’s a simulated series with *n* = 240 values. Note to people who aren’t using R – the time series order in the data file is across rows.

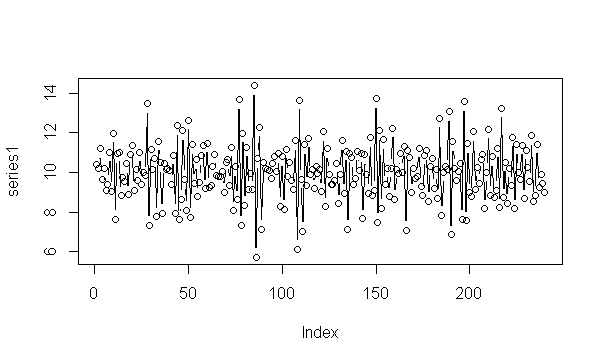
In R, change the working directory to your directory that contains the dataset. Use these commands to load the scripts (make them available) and read the data:

library(astsa)

series1 = scan ("sim1week3.dat")

These commands assume that both files are in the working directory. If not, you’ll need to specify the full path for finding the file(s) on your computer.

A. Plot the “series1” data. Comment on any important features. This command will do the trick: plot (series1, type="b")



Some features of the plot:

- There is no consistent trend and the series wanders up and down.

- There are no obvious outliers.

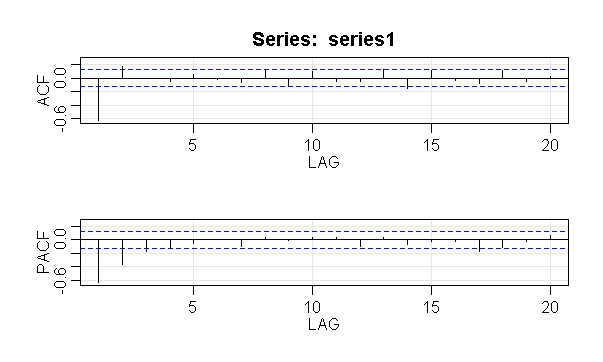
- The variance appears to be constant though it seems difficult to judge with conviction.

- Not enough information is provided about dataset to comment about seasonality

B. Examine the ACF and PACF of the data. Briefly indicate a possible model (or, models) and explain why you think this model (or models) will work.

The command acf2 (series1, 20) will give an ACF and PACF up to 20 lags.

The plot is:



One proposed model here is MA(2) model also specified as an ARIMA of order (0,0,2). This is based on the following observations:

* MA models have theoretical ACFs with non-zero values at the MA terms in the model and zero values elsewhere. We see significant values in the ACF for lags 1 and 2
* For an MA model, the theoretical PACF does not shut off, but instead tapers toward 0 in some manner. We observe this pattern for the PACF

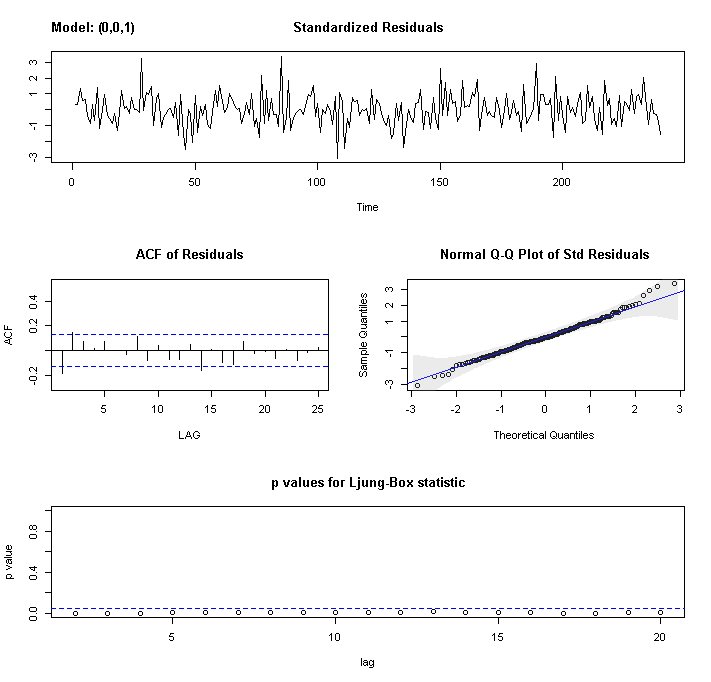
C. After the astsa library is loaded, the command

sarima (series name, order of AR, differencing order, MA order) will give the type of output described in Lesson 3.1 for this week.

For example, sarima (series1, 0, 0, 1) will fit a MA(1) to the data.

Fit an MA(1) model to the data. Using the diagnostics given in the output graph, discuss whether this seems to be a suitable model.

We get the following output



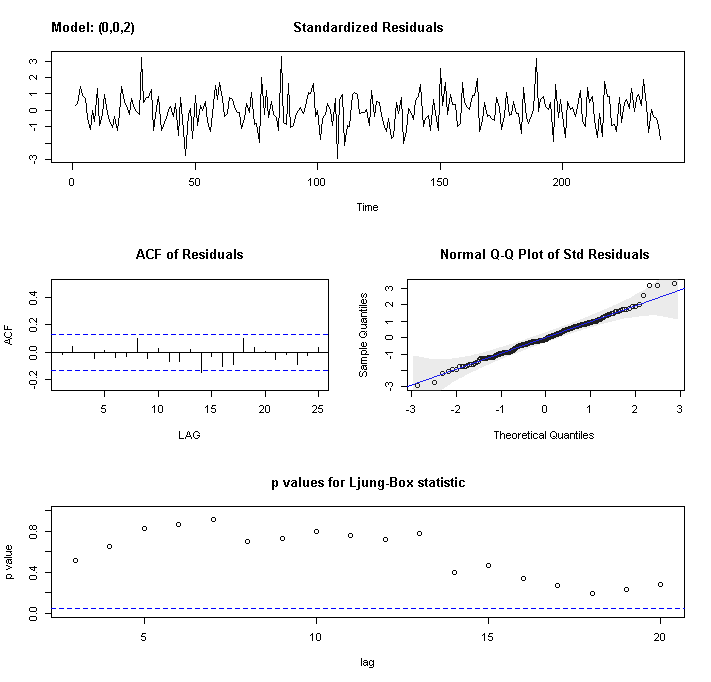
Based on this we deduce the following:

* 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 shows significant autocorrelations – not a desired result.
* The Q-Q plot is a normal probability plot – There is deviation so the assumption of normally distributed residuals is suspect.
* The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. All p-values are below the dashed blue line - That’s not a desired result.

So in summary while the model needs improvement.

D. Fit an MA(2) model to the data. Using the diagnostics given in the output graph, discuss whether this seems to be a suitable model.

The plot is:



Based on this we deduce the following:

* 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 more linear than MA(1) 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.

So in summary the model improves significantly from the MA(1).

E. Refer to the model in part D. explain whether the two MA coefficients are statistically significant or not.

We see the output:

ttable

Estimate SE t.value p.value

ma1 -0.9597 0.0612 -15.6807 0

ma2 0.2532 0.0610 4.1507 0

xmean 10.0018 0.0192 520.9480 0

MA(1) coeff (z = -0.9597/0.0612 = -15.681) is significant since |z| > 1.96

MA(2) coeff (z = 0.2532/0.0610 = 4.151) is significant since |z| > 1.96

F. Write out the estimated equation for the MA(2) model.

Use the output:

Coefficients:

ma1 ma2 xmean

-0.9597 0.2532 10.0018

s.e. 0.0612 0.0610 0.0192

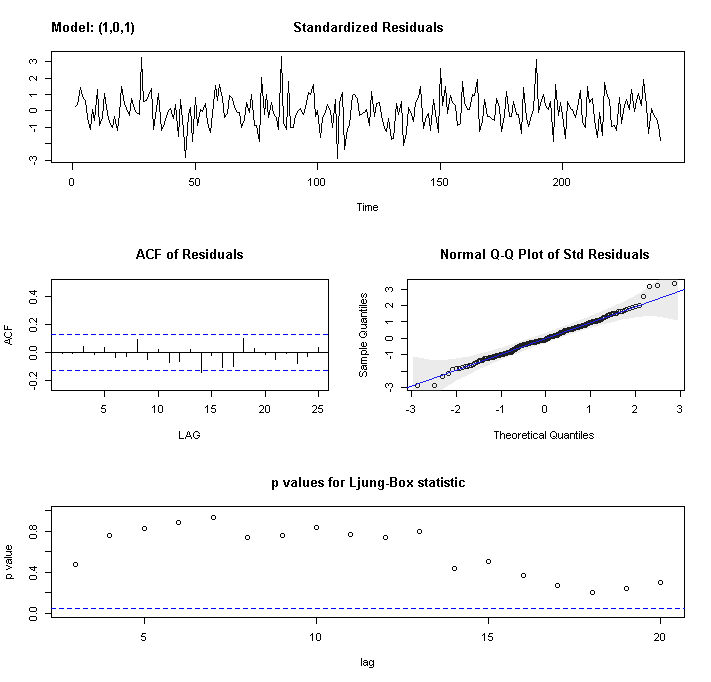
The estimated equation is:

xt = 10.0018 + wt - 0.9597 \* wt-1 + 0.2532 \* wt-2

G. Fit an ARIMA (1,0,1) to the data. This might also be called an ARMA (1,1). There’s one AR term, one MA term and no differencing.

Using the diagnostics given in the output graph, discuss whether this seems to be a suitable model.

The model gives the plot:



Based on this we deduce the following:

* 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.

H. Compare the MA(2) and ARMA(1,1) models. Does either seem better than the other? Explain.

MA(2) model gives:

Coefficients:

ma1 ma2 xmean

-0.9597 0.2532 10.0018

s.e. 0.0612 0.0610 0.0192

sigma^2 estimated as 1.014: log likelihood = -342.67, aic = 693.35

$degrees\_of\_freedom

[1] 237

$ttable

Estimate SE t.value p.value

ma1 -0.9597 0.0612 -15.6807 0

ma2 0.2532 0.0610 4.1507 0

xmean 10.0018 0.0192 520.9480 0

$AIC

[1] 1.038502

$AICc

[1] 1.047545

$BIC

[1] 0.08201004

ARIMA (1, 0, 1) gives:

Coefficients:

ar1 ma1 xmean

-0.3178 -0.6467 10.0021

s.e. 0.0763 0.0617 0.0176

sigma^2 estimated as 1.013: log likelihood = -342.55, aic = 693.1

$degrees\_of\_freedom

[1] 237

$ttable

Estimate SE t.value p.value

ar1 -0.3178 0.0763 -4.1668 0

ma1 -0.6467 0.0617 -10.4887 0

xmean 10.0021 0.0176 569.3043 0

$AIC

[1] 1.037455

$AICc

[1] 1.046497

$BIC

[1] 0.08096255

We deduce:

* The models look very similar. They have almost exactly the same values of sigma^2, AIC, AICC and BIC

I. Use the MA(2) to forecast values for the next 6 time periods past the end of the series. With astsa loaded, the command is sarima.for(series1,6, 0,0,2). The parameters being supplied, in order, are the series name, the number of times for forecasts, the AR order, the differencing and the MA order).

As an answer give, the forecasts and their standard errors (technically, the standard errors are for the forecast errors).

The output we get for prediction is:

pred

[1] 11.486156 9.548749 10.001821 10.001821 10.001821 10.001821

The output we get for se is:

se

[1] 1.006774 1.395420 1.418518 1.418518 1.418518 1.418518

J. Refer to the forecasts in part I. Explain why all forecasts after the second time after the series are the same as each other.

The forecasts after the second time after the series are the same as each other and are the same as xmean as:

Coefficients:

ma1 ma2 xmean

-0.9597 0.2532 10.0018

s.e. 0.0612 0.0610 0.0192

The wt-1 and wt-2 are 0 after the second time after the series.

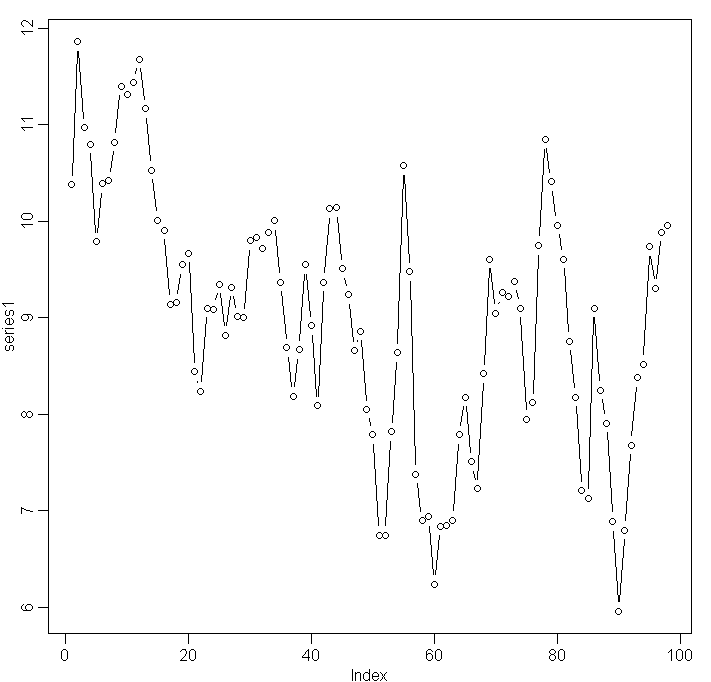
3. For this problem, use the dataset lakehuron.dat from the Week 3 folder. The series is *n* = 98 consecutive annual measures of the level of Lake Huron. The years are 1875-1972. Again, the time series order is across rows.

If you’re doing this in a new R session (compared to problem 2), start with

library(astsa) to gain access to those scripts.

A. Create a time series plot of the data. Comment on any important features.

The plot is:

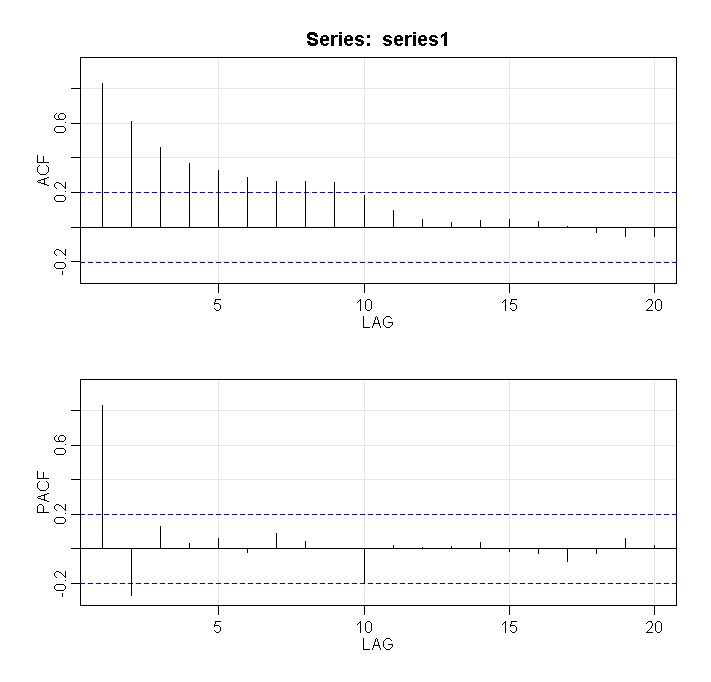


Notable features are:

* There is a downward trend, possibly a curved one.
* We have annual data and therefore deduce that there is no seasonality
* There are no obvious outliers.
* The variance across time is increasing.

B. Determine the ACF and PACF of the series. Propose a possible model (or models). Explain.

The plots are:

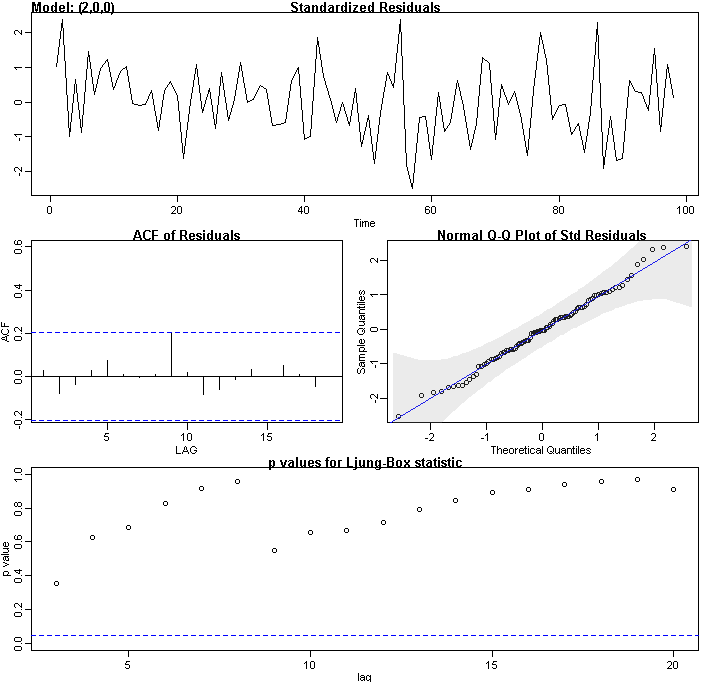


The plots indicate that we may be able to use an AR(2) model due to the following observations:

* PACF: Note that the first and second lag values are statistically significant, whereas partial autocorrelations for all other lags are not statistically significant. This suggests a possible AR(2) model for these data.
* ACF: We see that the ACF decreases to 0 as the lag h increases.

However since we also see a trend we may be able use either 1st or 2nd order difference i.e. ARIMA(2,1,0) for linear trend or ARIMA(2,2,0) for curved trend

C. Fit an AR(2) model to the series. Using the diagnostics given in the output graph, discuss whether this seems to be a suitable model.



Based on this we deduce the following:

* The time series plot of the standardized residuals mostly indicates that there’s no trend (or very little) in the residuals, no outliers, and in general, the variance is almost same across time.
* The ACF of the residuals doesn’t show significant autocorrelations except at lag 9 where it reaches the significance – a decent result.
* The Q-Q plot is a normal probability plot – The plot looks almost linear except at the ends and seemingly 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.

In summary the model seems to behave pretty well.

D. Refer to part C. Explain whether the two AR coefficients are statistically significant or not.

We see the following output:

ttable

Estimate SE t.value p.value

ar1 1.0436 0.0983 10.6185 0.0000

ar2 -0.2495 0.1008 -2.4754 0.0151

xmean 9.0473 0.3319 27.2612 0.0000

AR(1) coeff (z = 1.0436 / 0.0983= 10.62) is significant since |z| > 1.96

AR(2) coeff (z = -0.2495 / 0.1008 = -2.4752) is significant since |z| > 1.96

E. The output includes an estimate of the mean μ. For an AR(2), the intercept of the usual regression type equation is . Calculate an estimate of the intercept and then write the estimate AR(2) equation in the usual regression type fashion.

We have two methods:

Method 1:

interceptHat = 9.0473 \* (1 - 1.0436 + 0.2495) = 1.8628

xt = 1.8628 + 1.0436 xt-1 - 0.2495 \* xt-2 + wt

Method 2:

(xt−μ) = ϕ1(xt−1 − μ) + ϕ2(xt−2 − μ) + wt

xt - 9.0473 = ϕ1 \* xt-1 - ϕ1 \* 9.0473 + ϕ2 \* xt-2 – ϕ2 \* 9.0473 + wt

xt - 9.0473 = 1.0436 \* xt-1 - 1.0436 \* 9.0473 - 0.2495 \* xt-2 + 0.2495 \* 9.0473 + wt

xt = 9.0473 - 9.4418 + 2.2573 + 1.0436 xt-1 - 0.2495 \* xt-2 + wt

xt = 1.8628 + 1.0436 xt-1 - 0.2495 \* xt-2 + wt

F. The final two observations in the series are 9.89 and 9.96 (in that time order). Use these values and the equation written in part E to forecast the values for the next three times past the series. Show the calculations. (But, you can use sarima.for as explained in question 2 part I to check your answers.)

xtt+1 = 1.8628 + 1.0436 \* xt - 0.2495 \* xt-1 + wt = 1.8628 + 1.0436 \* 9.96- 0.2495 \* 9.89 = 9.7895

xtt+2 = 1.8628 + 1.0436 \* xtt+1 - 0.2495 \* xt + wt = 1.8628 + 1.0436 \* 9.7895- 0.2495 \* 9.96 = 9.5941

xtt+3 = 1.8628 + 1.0436 \* xtt+2 - 0.2495 \* xtt+1 + wt = 1.8628 + 1.0436 \* 9.5941- 0.2495 \* 9.7895 = 9.4327

The sarima output is:

pred

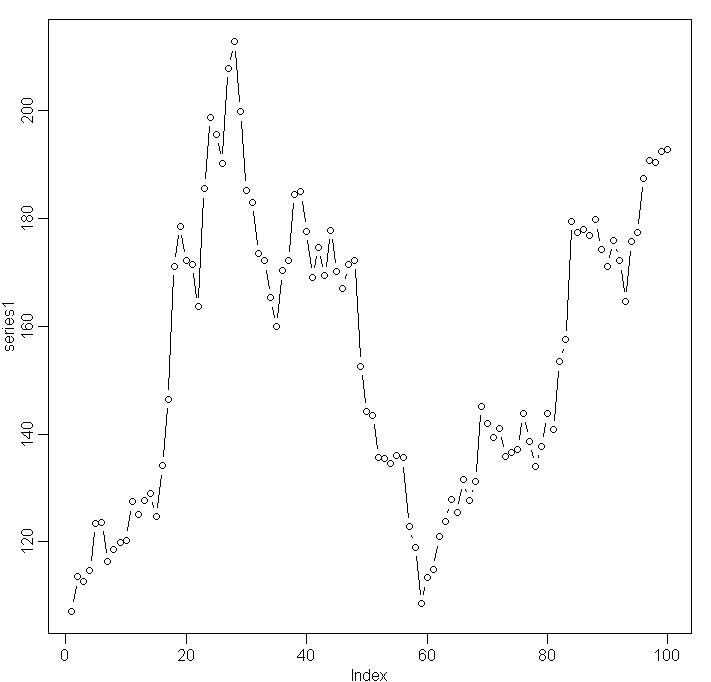
[1] 9.789559 9.594219 9.432885

*Comment: You may have thought that the time series plot showed a possible trend. This is a reasonable interpretation, though results without differencing also seem to be reasonable.*

**4**. Use the flourbuffalo.dat dataset in the Week 3 folder for this problem. The series gives index value for monthly flour prices in a Buffalo, NY commodity exchange. Again, the series time order is across rows in the dataset.

If this is a new R session, load the astsa library at the beginning of the session.

A. Plot the flour price series. Comment on any important features.

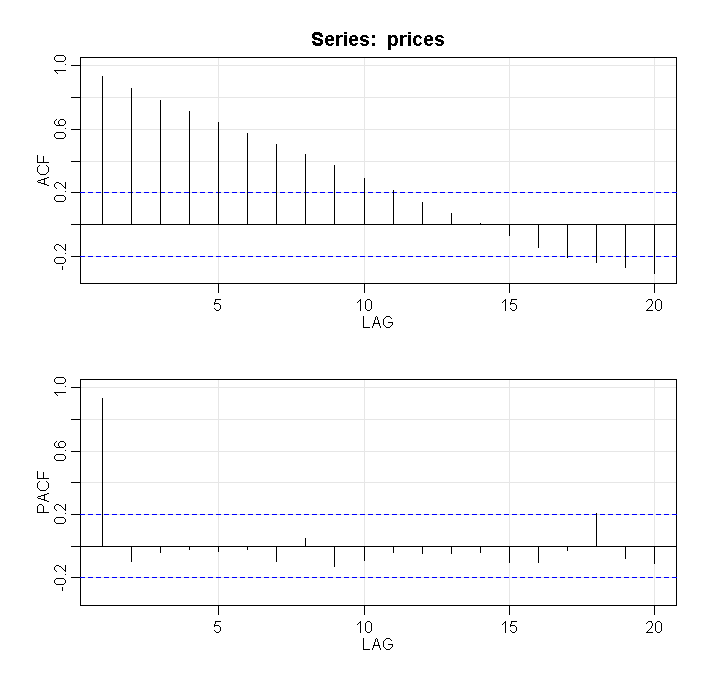


Notable features are:

* There is some trend.
* We have monthly data and there may be some seasonality
* There are no obvious outliers.
* The variance across time is not constant.

B. Determine the ACF and PACF for the series. Propose a possible model (or models). Explain.

The plot is:



The plots indicate that we may be able to use an AR(1) model due to the following observations:

* PACF: Note that only the first lag values are statistically significant, whereas partial autocorrelations for all other lags are not statistically significant. This suggests a possible AR(1) model for these data.
* ACF: We see that the ACF decreases as the lag h increases. However we see significant levels at lag 17, 18, 19 and 20. This may mean that we have to address seasonality

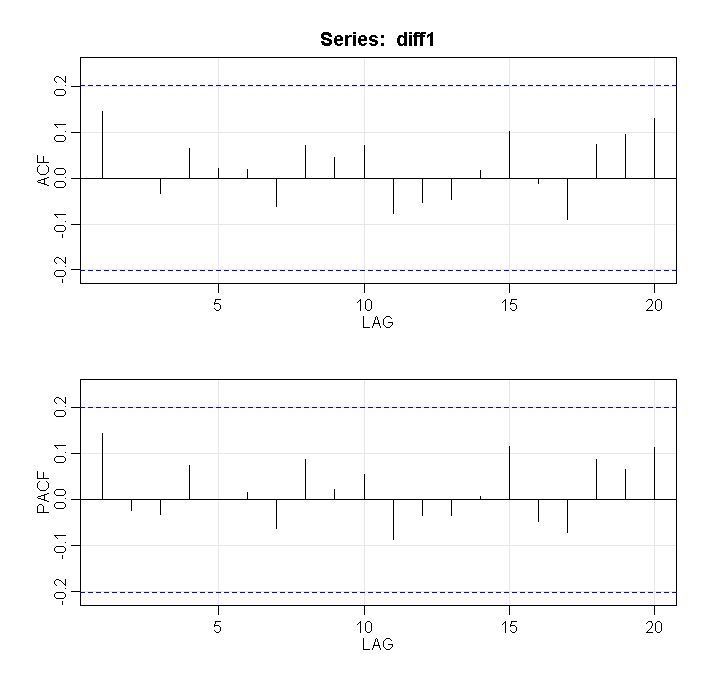
However since we also see a trend / seasonality we may be able use either 1st or 2nd order difference i.e. ARIMA(1,1,0) for linear trend or ARIMA(1,2,0) for curved trend.

C. Calculate the first difference for the series. The basic structure of the command is

New variable name = diff( series name, 1)

For example, if the series is named prices, you might use the command diff1=diff(prices,1).

Then, determine the ACF and PACF for the first differences. (For example, acf2(diff1) might work. Discuss the result.



We deduce from the above:

* Since we have taken first differences and all the autocorrelations are non-significant, so the series is called a random walk.
* A possible model for a random walk is xt = δ + xt-1 + wt.
* The data are dependent and are not identically distributed; in fact both the mean and variance are increasing through time.