W271-2 – Spring 2016 – **HW 7**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

March 30, 2016

Contents

Exercises

Question 1

1.1. **Load `hw07_series1.csv`.**

```
hw07 <- read.csv('hw07_series1.csv', header = FALSE) # CSV has no headers
names(hw07)
```

```
## [1] "V1"
```

1.2. **Describe the basic structure of the data and provide summary statistics of the series.**

```
str(hw07)
```

```
## 'data.frame':    75 obs. of  1 variable:
##  $ V1: num  10.01 10.07 10.32 9.75 10.33 ...
```

```
dim(hw07)
```

```
## [1] 75  1
```

```
# See the definition of the function in ## @knitr Libraries-Functions-Constants
desc_stat(hw07, 'Time series', 'Descriptive statistics of the time series')
```

Table 1: Descriptive statistics of the time series

	Time series
Mean	10.81
St. Dev	0.45
1st Quartile	10.48
Median	10.82
3rd Quartile	11.06
Min	9.75
Max	11.94

The data correspond to 75 observations of a single variable. No information about the time scale is given, so we'll just use an index between 1 and 75 (with `frequency = 1`). The main descriptive statistics are shonw in the Table above.

1.3. **Plot histogram and time-series plot of the series. Describe the patterns exhibited in histogram and time-series plot. For time series analysis, is it sufficient to use only histogram to describe a series?**

```
hw07.ts <- hw07[, 1]
```

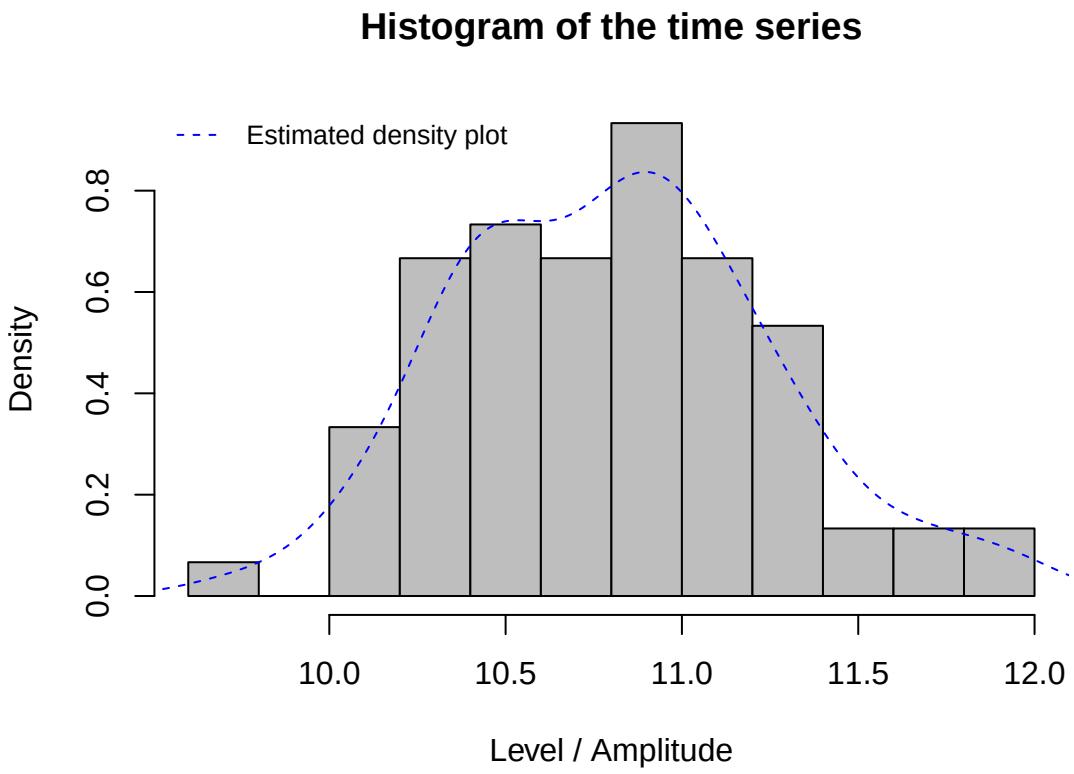

Figure 1: Histogram of the data

Let's get some descriptive statistics about the distribution of the data, that complement the histogram above:

Table 2: Descriptive statistics about normality of the time series

	Time series
skewness	0.276
skew.2SE	0.498
kurtosis	-0.153
kurt.2SE	-0.140
normtest.W	0.987
normtest.p	0.627

The series has an *excess kurtosis* (kurtosis minus 3, the value for a normal distribution) which is negative, which indicates a platykurtic distribution (thinner tails than a normal one). But that excess kurtosis is close to zero, and not statistically significant (the parameter `kurt.2SE` would have to be greater than 1; see the documentation of `stat.desc` for more information). Likewise, the *skewness* is positive, which indicates

a right-skewed distribution (with a heavy right tail), but again is not far from zero and not statistically significantly different than it. A Shapiro-Wilk test is also non-significant ($p = 0.627$), so we cannot reject the hypothesis that the distribution of the series is normal (though its size—75—is a bit reduced).

For time series analysis (unlike cross-sectional data analysis), the histogram alone is not enough to describe a series because it tells us nothing about the dynamics of it; e.g., this one in particular lets us know that there was one time period where the value of the time series was below 10, but not when that happened (in the 4th time period, as shown in the next Figure).

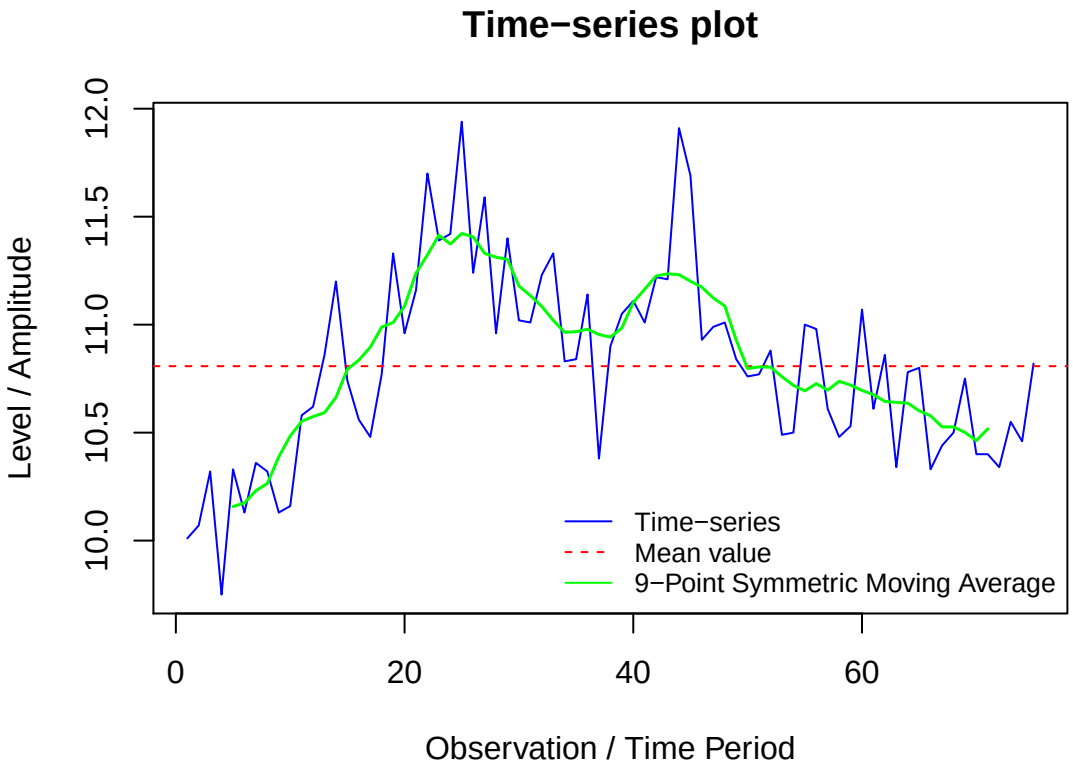

Figure 2: Time-series plot of the data

The time-series plot has been plotted together with the mean value and a 9-point symmetric moving average, that exhibits an increasing trend during the first 30 time periods or so, then a small decline for about 10 time periods, another increasing trend for a few time periods, and finally a decreasing trend during the last 30 time periods or so (this one not so linear; the decrease is steeper at the beginning, mainly due to the effect of a pike in observations 44 and 45); the smoother yields no values for the first and last 4 observations, but it seems the trend becomes positive again at the end of the observation period. Overall, we can say the time series is persistent.

1.4. **Plot the ACF and PACF of the series. Describe the patterns exhibited in the ACF and PACF.**

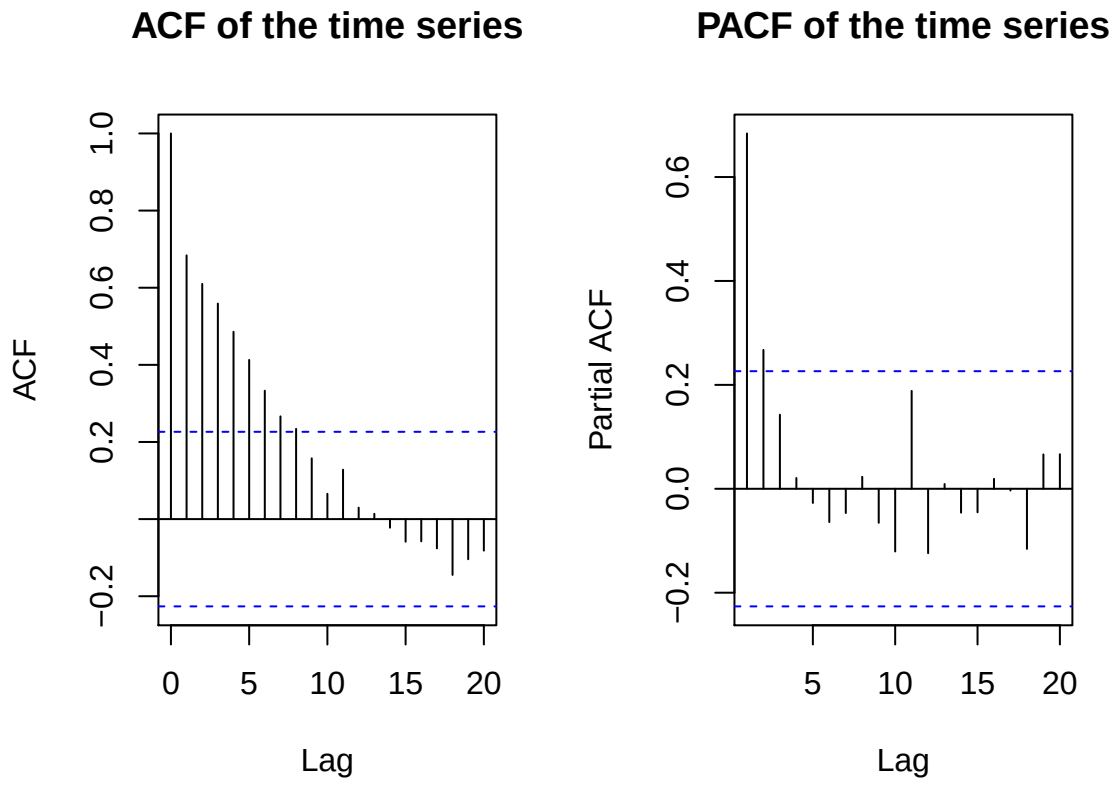

Figure 3: Autocorrelation and partial autocorrelation graphs

The correlogram has a wave-like shape that resembles that of a shrinking cosine function (typical of an AR(2) process). It decreases relatively slowly (the first autocorrelation not statistically significantly different from zero corresponds to the 9th lag). As for the PACF, it drops off relatively abruptly at lag 2 (another indication of an AR(2) process).

1.5. **Estimate the series using the `ar()` function.**

```
(hw07.arfit <- ar(hw07.ts, method = "mle"))
```

```
##
## Call:
## ar(x = hw07.ts, method = "mle")
##
## Coefficients:
##      1       2
## 0.4959  0.3042
##
## Order selected 2  sigma^2 estimated as  0.0917
```

1.6. **Report the estimated AR parameters, the order of the model, and standard errors.**

Order of the model:

```
hw07.arfit$order # order of the AR model with lowest AIC
```

```
## [1] 2
```

Estimated AR parameters:

```
hw07.arfit$ar # parameter estimates
```

```
## [1] 0.4959087 0.3041799
```

Other parameters of the estimated AR model:

```
hw07.arfit$aic # AICs of the fit models (differences vs. best model)
```

```
##           0           1           2           3           4           5
## 52.66714788  5.02320403  0.00000000  0.06211148  1.88648551  3.87456769
##           6           7           8           9          10          11
##  5.46602327  7.12088537  9.38090030 10.65835341 11.77778745 11.28242189
##          12
## 11.50016867
```

```
hw07.arfit$x.mean; mean(hw07.ts) # mean of the fit model and the data
```

```
## [1] 10.75694
```

```
## [1] 10.80773
```

```
hw07.arfit$var.pred # prediction variance
```

```
## [1] 0.09169526
```

```
hw07.arfit$asy.var.coef # asymptotic Covariance matrix
```

```
##              [,1]         [,2]
## [1,]  0.011625209 -0.007950168
## [2,] -0.007950168  0.011625209
```

As shown in the output of the code above, models of other orders (especially the one of order 3) have similar AIC values. The last parameter shown above is the asymptotic covariance matrix of the coefficient estimates, so the square root of the elements in the diagonal of that matrix are the standard errors. Together with them, we can also estimate the confidence interval of the coefficient estimates:

```
Parameters <- cbind(hw07.arfit$ar,
                    sqrt(diag(hw07.arfit$asy.var.coef)),
                    matrix(sapply(c(-2,2), function(i)
                      hw07.arfit$ar + i * sqrt(diag(hw07.arfit$asy.var.coef))),
                      ncol = 2))
```

Table 3: Coefficients, SEs, and 95% CIs of the estimated AR(2) model

	Coefficient	SE	95% CI lower	95% CI upper
lag 1	0.4959087	0.1078203	0.2802682	0.7115492
lag 2	0.3041799	0.1078203	0.0885393	0.5198204

Both coefficients are significant (the CI does not include zero in both cases).

Question 2

2.1. **Simulate a time series of lenght 100 for the following model. Name the series x.**

$$\mathbf{x_t} = \frac{\mathbf{5}}{\mathbf{6}}\mathbf{x_{t-1}} - \frac{\mathbf{1}}{\mathbf{6}}\mathbf{x_{t-2}} + \omega_\mathbf{t}$$

This is an AR(2) model with coefficients $5/6 = 0.833$ and $-1/6 = -0.167$ respectively.

```
set.seed(12345) # Fix a seed to get same results every time
x <- arima.sim(model = list(ar = c(5/6, -1/6), ma = 0), n = 100)
```

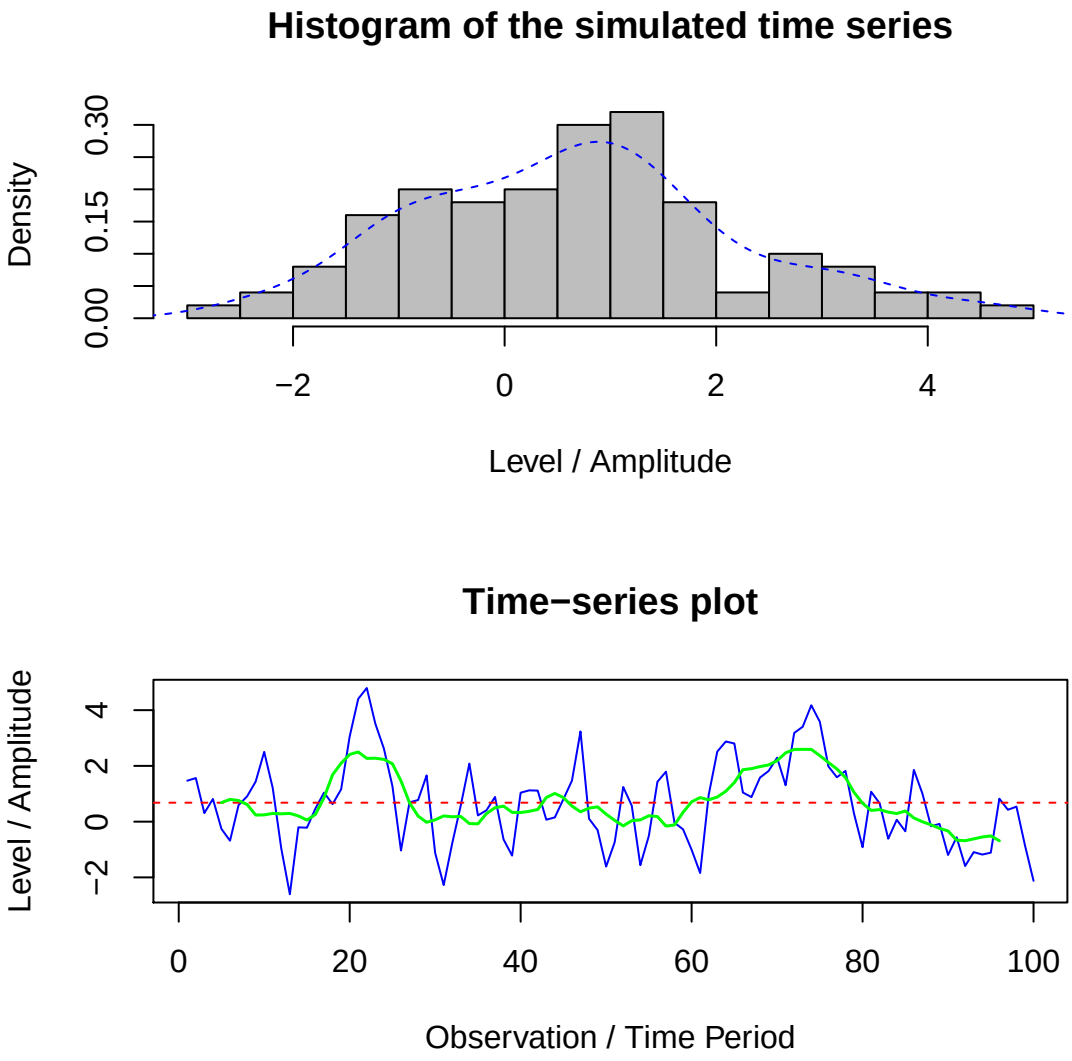

Figure 4: Histogram and time-series plot of the simulated time series

(The Figure above have the same legends than Figures 1 and 2, Question 1.)

2.2. **Plot the correlogram and partial correlogram for the simulated series. Comments on the plots.**

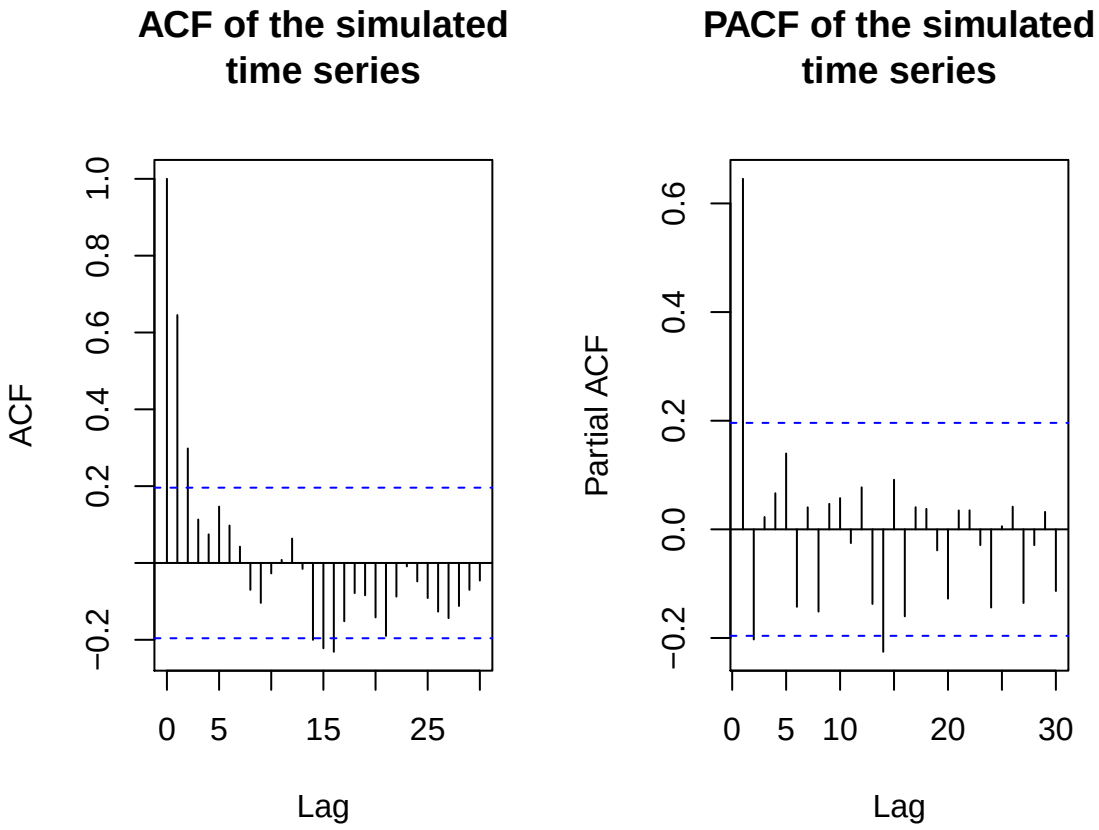

Figure 5: Autocorrelation and partial autocorrelation graphs

As in Question 1, the correlogram has a wave-like shape (that slightly resembles that of a shrinking cosine function). It decreases relatively quickly (though this changes in other simulations—if other seed is used).

The PACF drops off relatively abruptly at lag 1 or 2, depending on the simulation. For this particular simulation the partial autocorrelation for lag 2 is on the verge of being significantly different from zero; in other simulations it is not significantly different from zero. This is partly due to the limited size of the simulated series: if $\rho_k = 0$, its 95% confidence interval is $-\frac{1}{n} \pm \frac{2}{\sqrt{n}}$; for $n = 100$, that is $[-0.210, 0.190]$... which includes the "true" model coefficient for lag 2 ($-1/6 = -0.167$).

2.3. **Estimate an AR model for this simulated series. Report the estimated AR parameters, standard errors, and the order of the AR model.**

An AR model of order 2 is estimated, with coefficients quite close to the "true" ones (0.833 and -0.167). The result depends on the series that we simulated (i.e., on the seed that we used to generate the simulated series): for many other seeds an AR(1) model is estimated.

```
x.arfit <- ar(x, method = "mle")
# Estimated AR parameters
x.arfit$ar
```

```
## [1]  0.8037362 -0.2148526
```

```
# Standard errors
sqrt(diag(x.arfit$asy.var.coef))
```

```
## [1] 0.09631491 0.09631491
```

```
# Order of the AR model (with lowest AIC)
x.arfit$order
```

```
## [1] 2
```

2.4. **Construct a 95% confidence intervals for the parameter estimates of the estimated model. Do the "true" model parameters fall within the confidence intervals? Explain the 95% confidence intervals in this context.**

In the answer to the previous question we already estimated the SE of the parameter estimates.

```
Parameters <- cbind(x.arfit$ar, sqrt(diag(x.arfit$asy.var.coef)),
                    matrix(sapply(c(-2,2), function(i)
                      x.arfit$ar + i * sqrt(diag(x.arfit$asy.var.coef))),
                      ncol = 2))
```

Table 4: Coefficients, SEs, and 95% CIs of the estimated AR(2) model

	Coefficient	SE	95% CI lower	95% CI upper
lag 1	0.8037362	0.0963149	0.6111064	0.9963660
lag 2	-0.2148526	0.0963149	-0.4074824	-0.0222228

Based on Table 4 above (see the last two columns), the "true" model parameters (0.8333 and -0.1667) fall within the confidence intervals.

As we explained in **2.3**, what happens in many of the simulations (when using a different seed) is that the best fitted model is an AR(1) so there is no estimate for the coefficient for lag 2). In any case, if we repeat the simulation over and over, the "true" model parameters will fall within the CIs about 95% of the time (at

least the coefficient of lag 1; and also the coefficient of lag 2 if we "force" our fitted model to be AR(2)[1]).

2.5. **Is the estimated model stationary or non-stationary?**

For an AR model (x_t depends only on white noise and previous values of it) to be stationary, *all* the roots of its *characteristic equation* $\Phi(B) = 0$ must exceed unity in absolute value (i.e., they must be outside the unit circle). In this case, $\Phi(B) = 1 - 0.8037B - -0.2149B^2$ (rather than $\Phi(B) = 1 - 0.8333B + 0.1667B^2$, because we are asked about the estimated model, not the "true" one), so let's check the roots of the characteristic equation:

```
# True model
pol_true <- c(1, -5/6, 1/6); (roots <- polyroot(pol_true))
```

```
## [1] 2+0i 3+0i
```

```
Mod(roots); all(Mod(roots) > 1)
```

```
## [1] 2 3
```

```
## [1] TRUE
```

```
# Fitted model
(pol_fitted <- c(1, -x.arfit$ar)); (roots <- polyroot(pol_fitted))
```

```
## [1]  1.0000000 -0.8037362  0.2148526
```

```
## [1] 1.870436+1.075092i 1.870436-1.075092i
```

```
Mod(roots); all(Mod(roots) > 1)
```

```
## [1] 2.157395 2.157395
```

```
## [1] TRUE
```

All the roots are outside the unit circle, so the estimated model (as well as the "true" model) is stationary.

> Actually, this check is not even necessary: had the estimated model not been stationary, the `ar()` function would have returned an error.

---

[1] That would have to be done using the `arima()` function rather than `ar()`: `arima(x, order = c(2,0,0))`.

2.6. **Plot the correlogram of the residuals of the estimated model. Comment on the plot.**

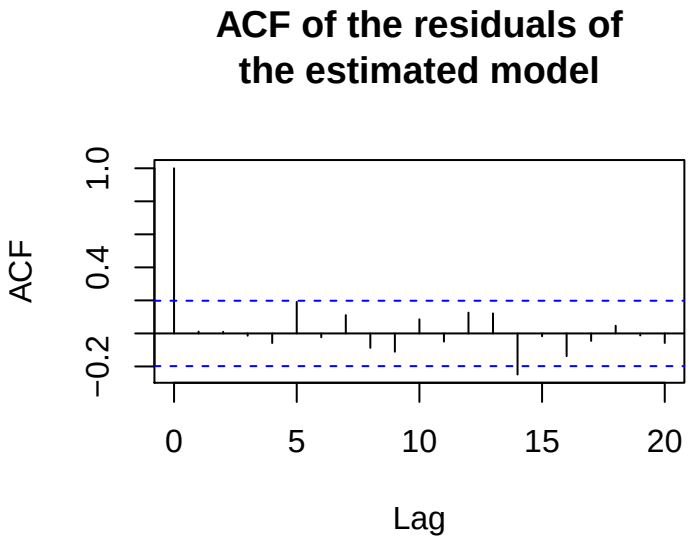

Figure 6: Autocorrelation graph of the residuals of the estimated model

None of the autocorrelations (with the exception of $\rho_0 = 1$, of course) is statistically significantly different from zero, which indicates that the residuals could be white noise (at least we have no evidence against that hypothesis). A Q-Q plot indicates that the distributions of the residuals is close to normal, which strengthens the idea that the residuals might be (gaussian) white noise, and indicates that the estimated model (an AR(2) stochastic process) is—as expected, because we know how the data were generated—a good fit for the series.

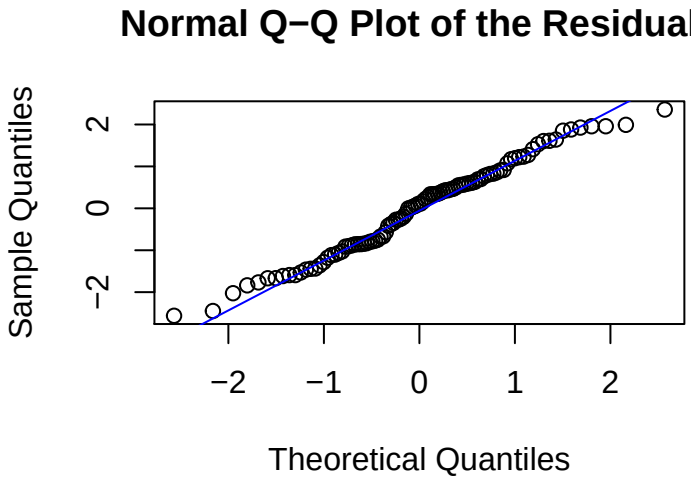

Figure 7: Normal Q-Q plot of the residuals of the estimated model