Lab 2

Glenn (Ted) Dunmire

November 13, 2015

Question 1: Concepts

1.1 (1 points) Define the term stochastic process.

A stochastic process is a statistical phenomenon that evolves over time according to probabilistic laws– a stochastic process generates a single time series in the same way a random variable generates a single number.

1.2 (2 points) Define the term time series. What is the difference between a stochastic process and a time series?

A time series is a set of observations generated sequentially in time. A time series arises as the result of a stochastic process and is a realization of the stochastic process with specific values over a time period.

1.3 (3 points) In your own words, discuss the mean function, variance function, and measure of dependency structure in the context of time series (week 8 - 10) and compare them with those we studied in classical linear model (week 1 - 7).

Prevously, in the classic linear model, there was no concept of time when we discussed the mean. With a time series, we add the complexity of time. It no longer makes sense to just report a mean if the mean is changing over time. Now, the mean function is $\mu_x(t) = E[x_t] = \int_{-\infty}^{+\infty} x_t f_t(x_t) dx_t$. The expectation is over the set of all time series that could be generated by the stochastic process. If a process is stationary in the mean, that means the mean is constant and we can drop the time parameter and write the mean as μ, similar to what we studied in a classic linear model.

The variance function is $\sigma_x^2(t) = E(x_t - \mu_x(t))^2 = \int_{-\infty}^{+\infty} (x_t - \mu_x(t))^2 f_t(x_t) dx_t$. Variance also may depend on time, but it's impossible to estimate a separate variance because we only have a single time series. For this reason, it's popular to assume that the series is stationary in variance meaning that the variance function is constant across time.

In classical models, we measure linear dependency using correlation or covariance. In a time series, we are interested in dependency between different random variables in the same series. In order to do this, we use autocovariance and autocorrelation. Typically, we think of autocovariance and autocorrelation being a function of the lag in a time series.

1.4 (2 points) Define strict and weak statonarity

Strict stationarity is when the joint distributions for all time periods are the same. Formally it is if

$$F(x_t 1, ..., x_t n)$$

and

$$F(x_t 1 + m, ..., x_t n + m)$$

are equal for all (t1, . . ., tn) and m.

Weak stationarity is when a time series is stationary in mean and variance and if its autocovariance depends on the lag (k) and can be written as

$$\gamma^k$$

1.5 (3 points) Give an example of a time series in real life. Describe the series. Evaluate (not empirical work is needed) whether or not the series can be modeled using the class of autoregressive models?

An example of a time series in real life would be monthly bookings at Walt Disney World Resort from 1980-2015. This time series would fluctuate seasonally and would likely have an upward trend, especially given more rooms are added over time. This time series is not likely to be stationary in the mean, and thus wouldn't be good for an autoregressive model without transformations. To use autoregressive techniques on a series like this, the series would first need to be made stationary by removing trends and seasonality.

1.6 (4 points) In your own words, define and describe partial autocorrelation function (PACF). Why is it not enough just to autocorrelation functon (ACF) to capture the dependency of a series?

The partial autocorrelation function partials out variance explained due to shorter lags. For example, a partial autocorrelation function for the 3rd lag is controlling for the 2nd lag and 1st lag. You must look at the partial autocorrelation function to capture the unique dependency of the series attributable to only the lag of interest and not the lags before it.

Question 2

Determine if each of the following models is stationary

2.1 $z_t = 0.95z_{t-1} + \omega_t$

Reorganize in terms of backward operator

$\omega_t = 1 - 0.95B$

set ω_t equal to 0

$0 = 1 - 0.95B$

Solve for B

$B = 1/0.95 = 1.05$

As the root is greater than 1, the AR model is stationary

Alternatively you can use the polyroot function to determine the roots:

```
polyroot(c(1,-.95))
```

```
## [1] 1.052632+0i
```

This gives the same result as above.

2.2 $z_t = 0.8z_{t-1} + 0.3z_{t-2} + \omega_t$

Reorganize in terms of backward operator

$\omega_t = 1 - 0.8B - 0.3B^2$

Use polyroot

```
polyroot(c(1,-0.8,-0.3))
```

```
## [1]  0.9274433+0i -3.5941100-0i
```

As one root is not greater than 1 in absolute value, the AR model is not stationary

2.3 $z_t = -0.5z_{t-1} + 0.5z_{t-2} + \omega_t$

Reorganize in terms of backward operator

$\omega_t = 1 + 0.5B - 0.5B^2$

Use polyroot

```
polyroot(c(1,0.5,-0.5))
```

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

As one root is not greater than 1 in absolute value, the AR model is not stationary

2.4 $z_t = z_{t-1} + 0.4z_{t-2} + \omega_t$

Reorganize in terms of backward operator

$\omega_t = 1 - B - 0.4B^2$

Use polyroot

```
polyroot(c(1,-1,-0.4))
```

```
## [1]  0.7655644+0i -3.2655644-0i
```

As one root is not greater than 1 in absolute value, the AR model is not stationary

2.5 $z_t = -0.5z_{t-1} - 0.25z_{t-2} + \omega_t$

Reorganize in terms of backward operator

$\omega_t = 1 + 0.5B + 0.25B^2$

Use polyroot

```
polyroot(c(1.0,0.5,0.25))
```

```
## [1] -1+1.732051i -1-1.732051i
```

As both roots are greater than 1 in absolute value, the AR model is stationary

Question 3

3.1 Load the series series1.csv

```
#Read Series Data

series <- read.table("series.csv", quote="\"", comment.char="")
#series <- read.table("//vivica/Documents/MIDS/W271/w271_F15_week10_lab2/series.csv", quote="\"", comme

#Reference libraries
library(ggplot2)
library(gridExtra)
library(grid)
library(car)
```

3.2 Describe the basic structure of the data and provide summary statistics of the series

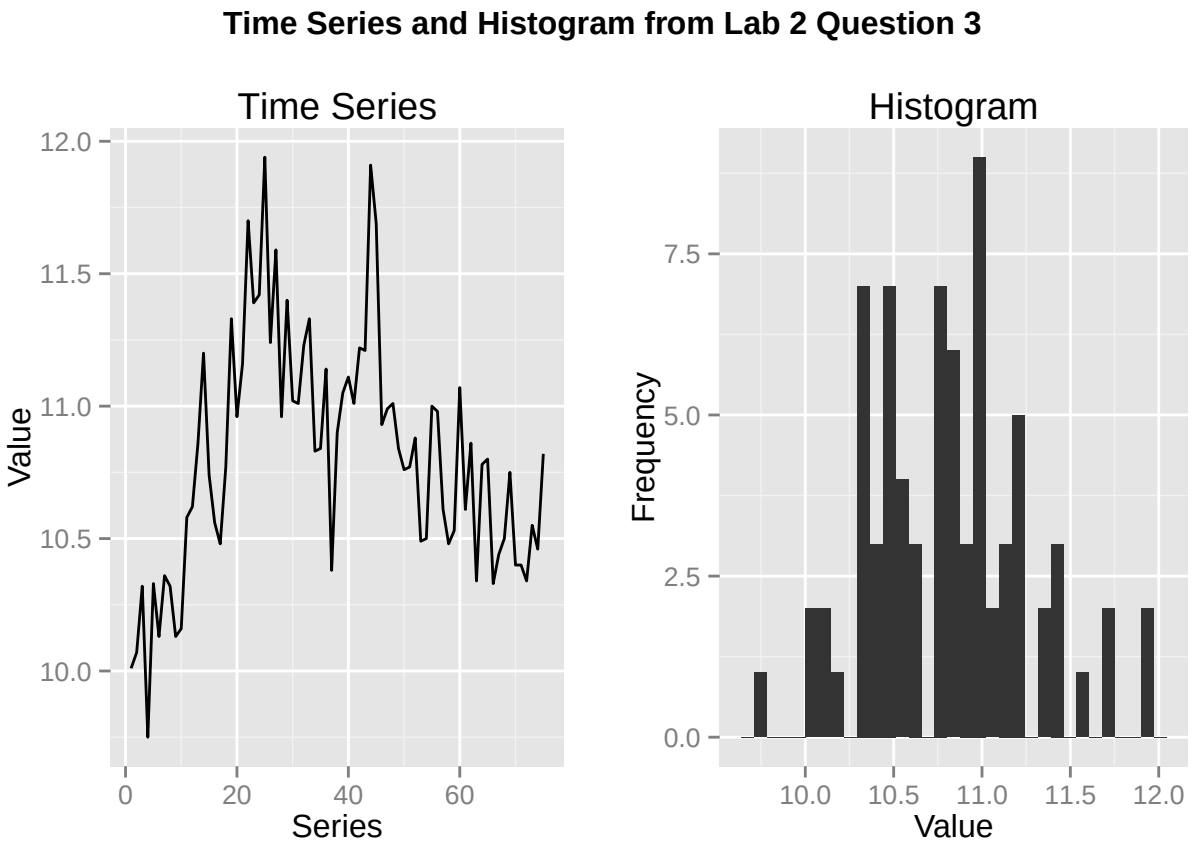
```
summ = summary(series)
sd = sd(series$V1)
head(series)
```

```
##      V1
## 1 10.01
## 2 10.07
## 3 10.32
## 4  9.75
## 5 10.33
## 6 10.13
```

There are 75 observations in the series, of 1 variable. The mean of the series is Mean :10.81 while the standard deviation is 0.4474781. The minimum and maximum of the series are Min. : 9.75 and Max. :11.94 , respectively. The 25th, 50th and 75th quantiles are 1st Qu.:10.48 , Median :10.82 and 3rd Qu.:11.06 , respectively. We have displayed the first 5 observations using the 'head' command to show in general what the series looks like.

3.3 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?

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



The time series shows an increasing trend from 0 through 25 and after that there is a decreasing trend until the end of the dataset. However, around time period 39 there is a large drop and then a significant spike

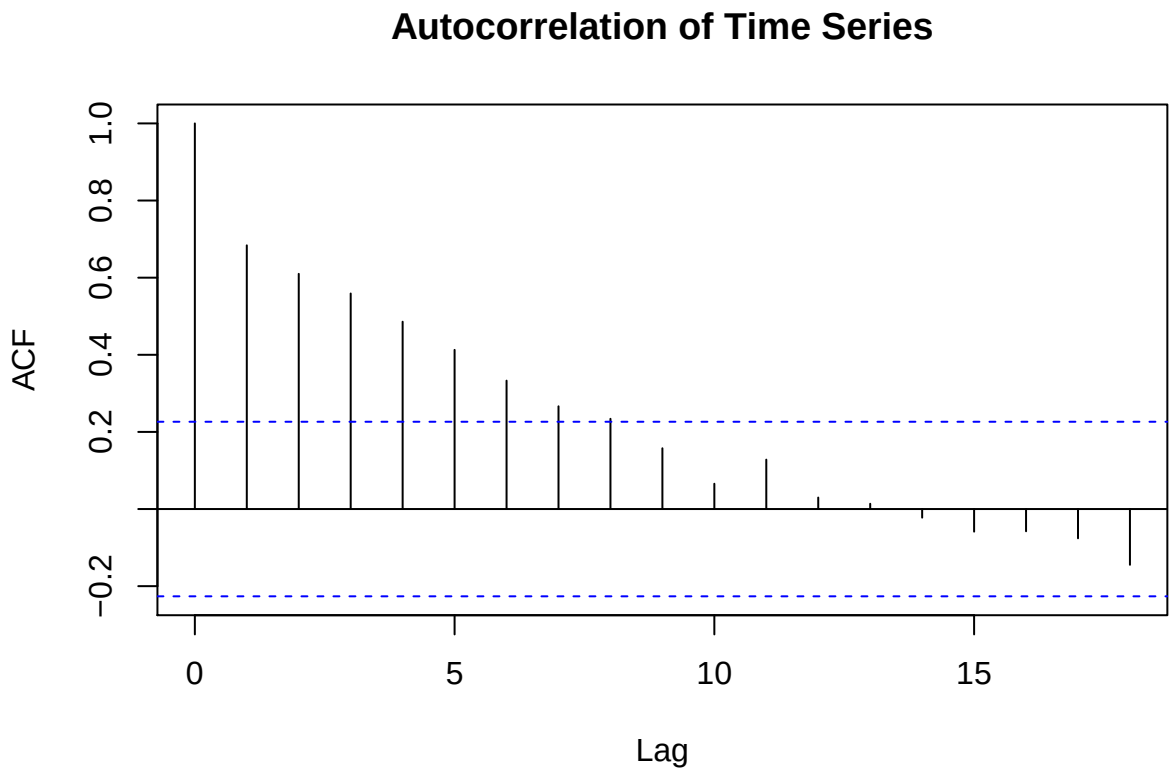
persisting through time period 45 (approximately). Then the series continues to trend downward. There is seasonality throughout the entire dataset, exhibited by frequent peaks and troughs.

The distribution is clearly not normal and appears to be platykurtic, as there are many frequent points around the mean.

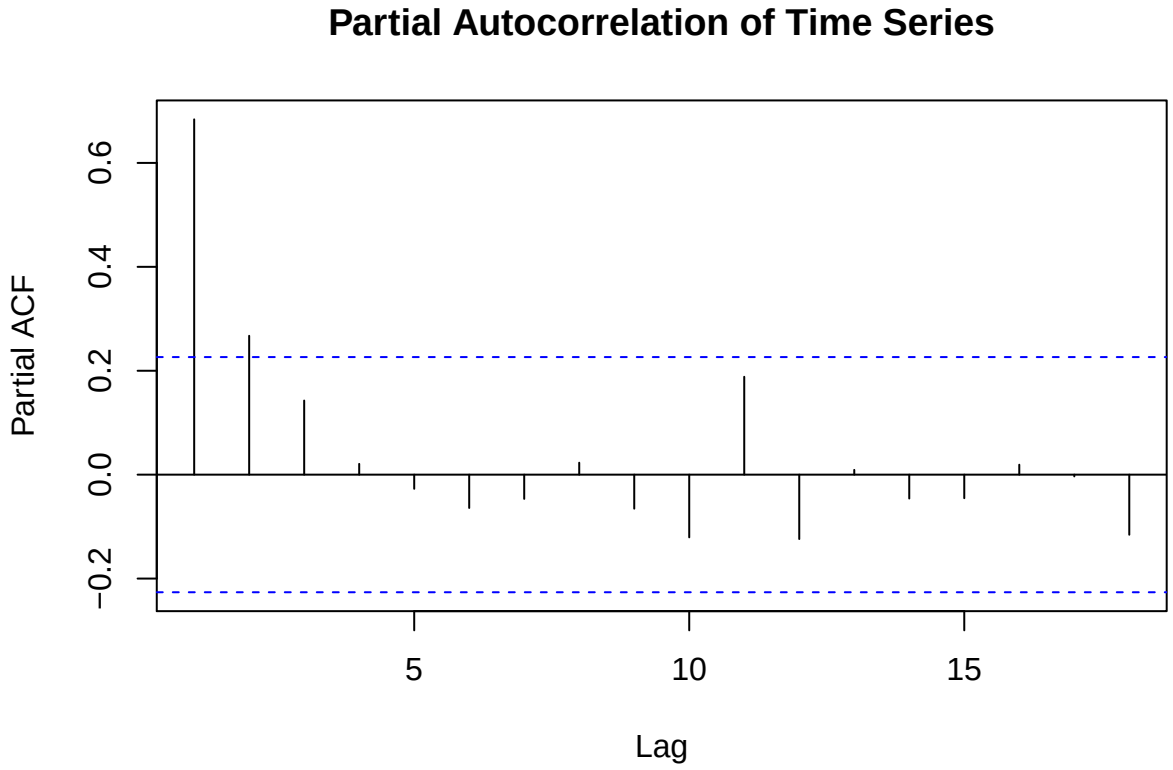
In general with time series it is not sufficient to examine only the histogram. The histogram will tell us the number of times a particular value occurs in the dataset. However, quite often with time series we are interested in the trend or changes in the observations through time. A histogram does not give us any insight into how the values will change over time, for example, in detecting seasonality. A histogram is thus certainly valuable, but not sufficient.

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

```
#Plot the ACF of the time series
acf(series, main="")
title("Autocorrelation of Time Series")
```




```
#Plot the PACF of the time series
pacf(series, main="")
title("Partial Autocorrelation of Time Series")
```



The autocorrelation of the time series indicates that the series is highly dependent on its previous lag at a statistically significant level through the 8th lag. The autocorrelation function does not fluctuate around zero, which probably indicates that the parameters of the series are positive. Based on how long it takes for the function to decay, this suggests there is a moderate level of persistence in the time series.

The partial autocorrelation of the time series indicates that the series is only significantly dependent on previous values through the second lag. This means that after controlling for the shorter lags, only the first two lags remain statistically significant.

3.5 Esimate the series using the maximum likelihood method option of ar() functions

```
arfit1 <- ar(df$V1, method = "mle")
```

```
## Warning in arima0(x, order = c(i, 0L, 0L), include.mean = demean): possible
## convergence problem: optim gave code = 1
```

```
arfit1
```

```
##
## Call:
## ar(x = df$V1, method = "mle")
##
## Coefficients:
##      1       2
## 0.4959  0.3042
##
## Order selected 2  sigma^2 estimated as  0.0917
```

3.6 Report the estimated AR parameters, the order of the model and standard errors

The order of the model is 2 while the estimated parameters are 0.4959087, 0.3041799. The standard errors are 0.1078203, 0.1078203

3.7 Estimate the series using the Ordinary Least Square option of ar() functions

```
arfit2 <- ar(df$V1, method = "ols")
arfit2
```

```
##
## Call:
## ar(x = df$V1, method = "ols")
##
## Coefficients:
##      1       2       3       4
## 0.4039  0.1788  0.0946  0.0901
##
## Intercept: 0.02404 (0.03443)
##
## Order selected 4  sigma^2 estimated as  0.08373
```

3.8 Report the estimated AR parameters, the order of the model and standard errors

The order of the model is 4 while the estimated parameters are 0.4039403, 0.178792, 0.0946442, 0.0900817. The standard errors are 0.1154417, 0.1234757, 0.1230623, 0.1134199

3.9 Estimate the series using the Yule-Walker Equations option of ar() functions

```
arfit3 <- ar(df$V1, method = "yw")
arfit3
```

```
##
## Call:
## ar(x = df$V1, method = "yw")
##
## Coefficients:
##      1       2
## 0.5011  0.2673
##
## Order selected 2  sigma^2 estimated as  0.1017
```

3.10 Report the estimated AR parameters, the order of the model and standard errors

The order of the model is 2 while the estimated parameters are 0.5011075, 0.2672509. The standard errors are 0.1135645, 0.1135645

3.11 Are these estimates the same? If so, derive the forumla to justify your answer. If not, please explain. How does the function ar() choose the best AR model

```
#Confidence Interval first parameter
arfit1$ar[1] + c(-2, 2) * sqrt(arfit1$asy.var.coef)[1,1]
## [1] 0.2802682 0.7115492
arfit2$ar[1] + c(-2, 2) * arfit2$asy.se.coef$ar[1]
## [1] 0.1730569 0.6348237
arfit3$ar[1] + c(-2, 2) * sqrt(arfit3$asy.var.coef)[1,1]
## [1] 0.2739784 0.7282365
```

```
#Confidence interval second parameter
arfit1$ar[2] + c(-2,2) * sqrt(arfit1$asy.var.coef)[2,2]
## [1] 0.08853934 0.51982039
arfit2$ar[2] + c(-2, 2) * arfit2$asy.se.coef$ar[2]
## [1] -0.06815945  0.42574336
arfit3$ar[2] + c(-2, 2) * sqrt(arfit3$asy.var.coef)[2,2]
## [1] 0.04012179 0.49437991

#Confidence interval third parameter
0
## [1] 0
arfit2$ar[3] + c(-2, 2) * arfit2$asy.se.coef$ar[3]
## [1] -0.1514804  0.3407687
0
## [1] 0

#Confidence interval Fourth parameter
0
## [1] 0
arfit2$ar[4] + c(-2, 2) * arfit2$asy.se.coef$ar[4]
## [1] -0.1367582  0.3169216
0
## [1] 0
```

The order of these models are not necessarily the same. The MLE and Yule-Walker methods give a 2 order model, while the OLS method gives a 4 order model. That being said, the standard errors of the estimates themselves tend to be large, making the confidence interval fall in the range of +/- about 0.20, making it a range of about 0.40. The confidence intervals are displayed above, showing a lot of overlap between estimates. However, this does not indicate the estimates are the same. In fact, these three methods of fitting the model use different ways to calculate the estimates, the Yule-Walker equations are different than the OLS method. Therefore we suggest the model estimates are probably not the same.

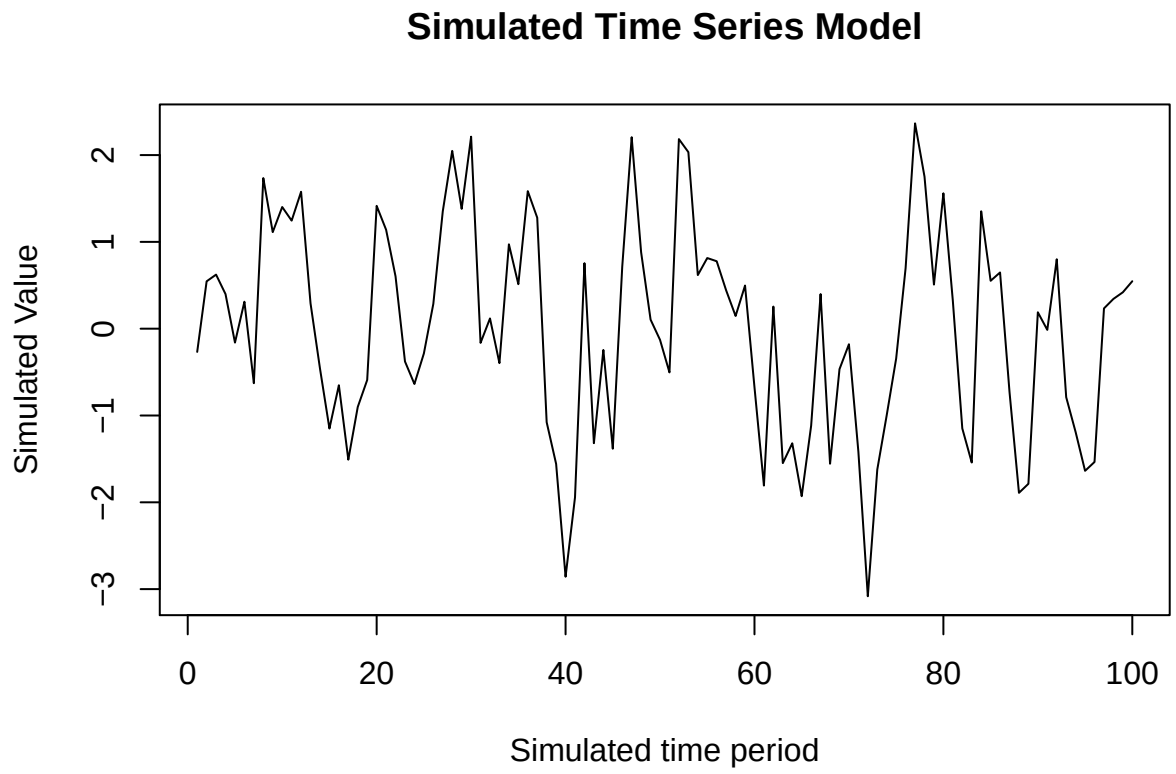
In general the ar() function chooses the model with the lowest AIC. For example, using our arfit1 model we can show the AIC for the various possible models is 52.6671479, 5.023204, 0, 0.0621115, 1.8864855, 3.8745677, 5.4660233, 7.1208854, 9.3809003, 10.6583534, 11.7777874, 11.2824219, 11.5001687. The 2nd order model is shown as 0 because this hd the lowest AIC of all the models. The other AIC's are shown as the difference between the lowest AIC and the AIC for that order model. That is, the 3rd order model shows the difference between the 3rd order model AIC - 2nd order model AIC.

Question 4

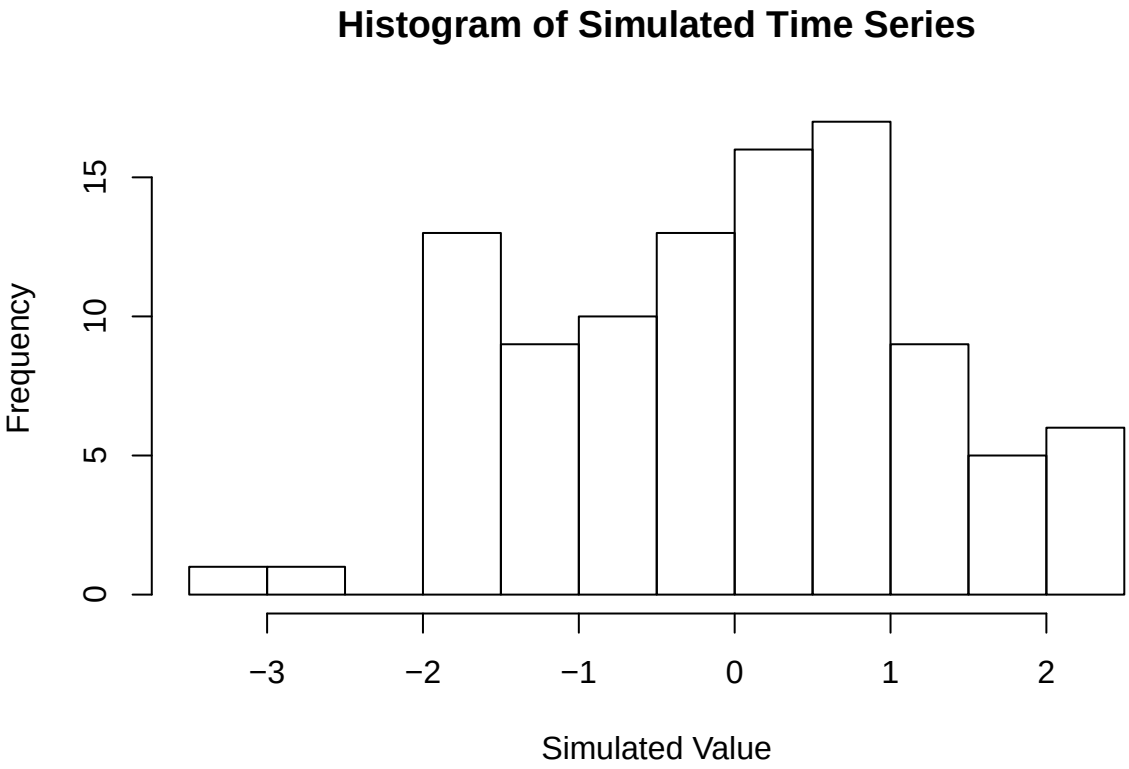
4.1 Simulate a time series of length 100 for the following model. Name the series x.

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

```
set.seed(100)
x <- arima.sim(n = 100, list(ar = c(5/6, -1/6), ma=0))
plot.ts(x, xlab = "Simulated time period", ylab = "Simulated Value", main = "Simulated Time Series Model
```

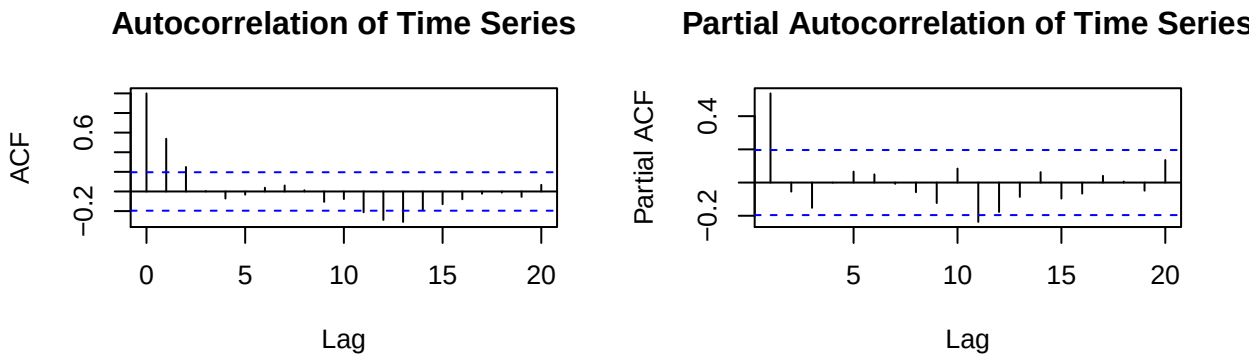



```r
hist(x, xlab = "Simulated Value", main = "Histogram of Simulated Time Series")
```



4.2 Plot the correlogram and partial correlogram for the simluation series. Comment on the plots

```r
par(mfrow = c(2,2))
#Autocorrelation Function
acf(x, main="")
title("Autocorrelation of Time Series")

#Partial Autocorrelation Function
pacf(x, main="")
title("Partial Autocorrelation of Time Series")
```



The autocorrelation of the time series indicates that the series is highly dependent on its previous lag at a statistically significant level through the 2nd lag. The switching signs in the autocorrelation suggests a negative parameter (which we know is correct from the model given).

The partial autocorrelation of the time series indicates that the series is only significantly dependent on previous values through the first lag. This means that after controlling for the shorter lags, only the first lag remains statistically significant. This would suggest an AR(1) model would be appropriate to estimate this series.

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

```r
arfit <- ar(x, method = "mle")
```

The order of the model is 1 while the estimated parameter is 0.5332786. The standard error is 0.0843038

4.4 Construct a 95% confident intervals for the parameter estimates of the estimated model. Do the 'true' mode parameters fall within the confidence intervals? Explain the 95% confidence intervals in this context

```r
CI <- arfit$ar + c(-2, 2) * sqrt(arfit$asy.var.coef)
```

This confidence interval for the estimated parameter is 0.364671, 0.7018862. To recall, the true parameters of this model are 0.8333333 and -0.1666667. Neither of these parameters are contained in the confidence interval obtained for the simulated AR(1) model. In the classic linear model, we check if the confidence contains 0 because if the confidence interval does contain 0 we know that the test statistic will not be statisically significant (assuming a 0.05 level). In this case, we are interested in determining if our model's estimated parameter contains the true parameter, to see how well our model estimates the true series. However, in this case our model does not contain the true parameters so we would conclude (at the 0.05 level) our models estimate is statisitcally significantly different from the true parameter.

4.5 Is the estimated model stationary or not station?

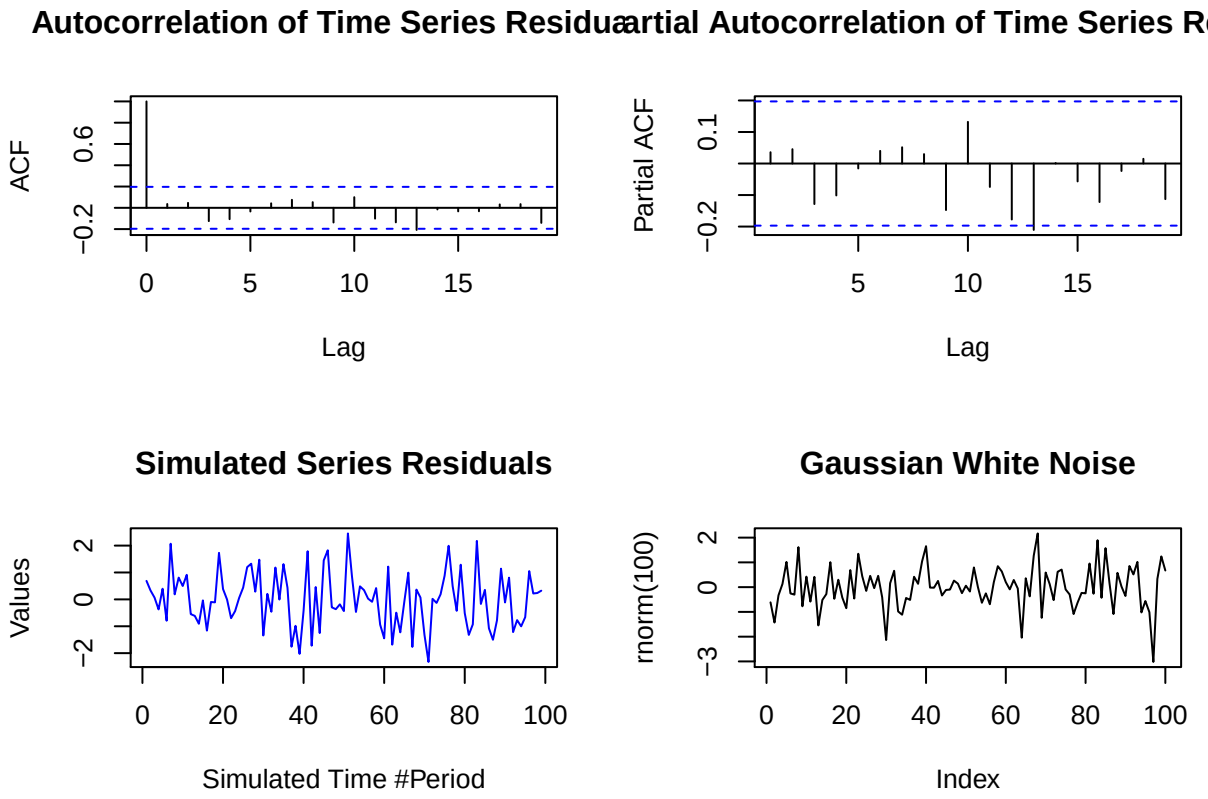
```
polyroot(c(1, -arfit$ar))
```

```
## [1] 1.875192+0i
```

The root is greater than 1 in absolute value, therefore this model is stationary.

4.6 Plot the correlogram of the residuals of the estimated model. Comment on the plot

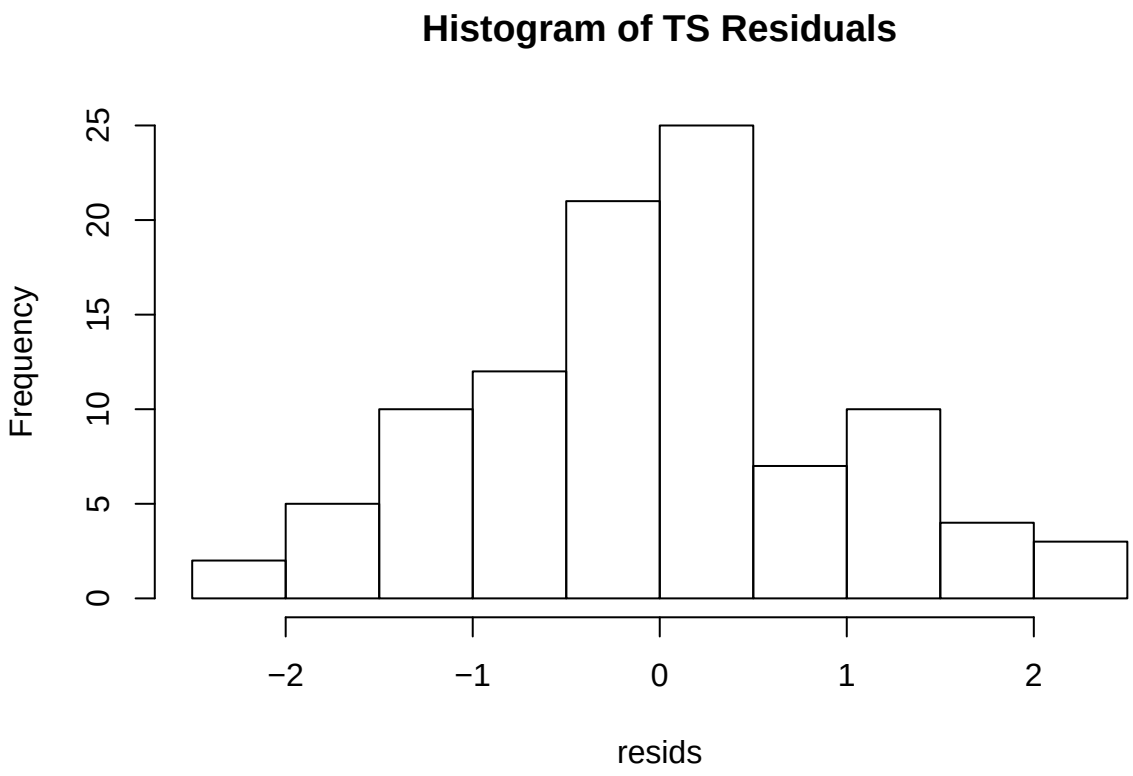
```
resids <- arfit$resid[2:100] #subsetting to remove NA
par(mfrow = c(2,2))
#Autocorrelation Function of Residuals
acf(resids, main="Autocorrelation of Time Series Residuals")

#Partial Autocorrelation Function of Residuals
pacf(resids, main = "Partial Autocorrelation of Time Series Residuals")

#Time Series of Residuals
plot.ts(resids, main="Simulated Series Residuals",col="blue", ylab="Values", xlab="Simulated Time #Peri

#White noise for comparison
plot(rnorm(100), type="l", main="Gaussian White Noise")
```



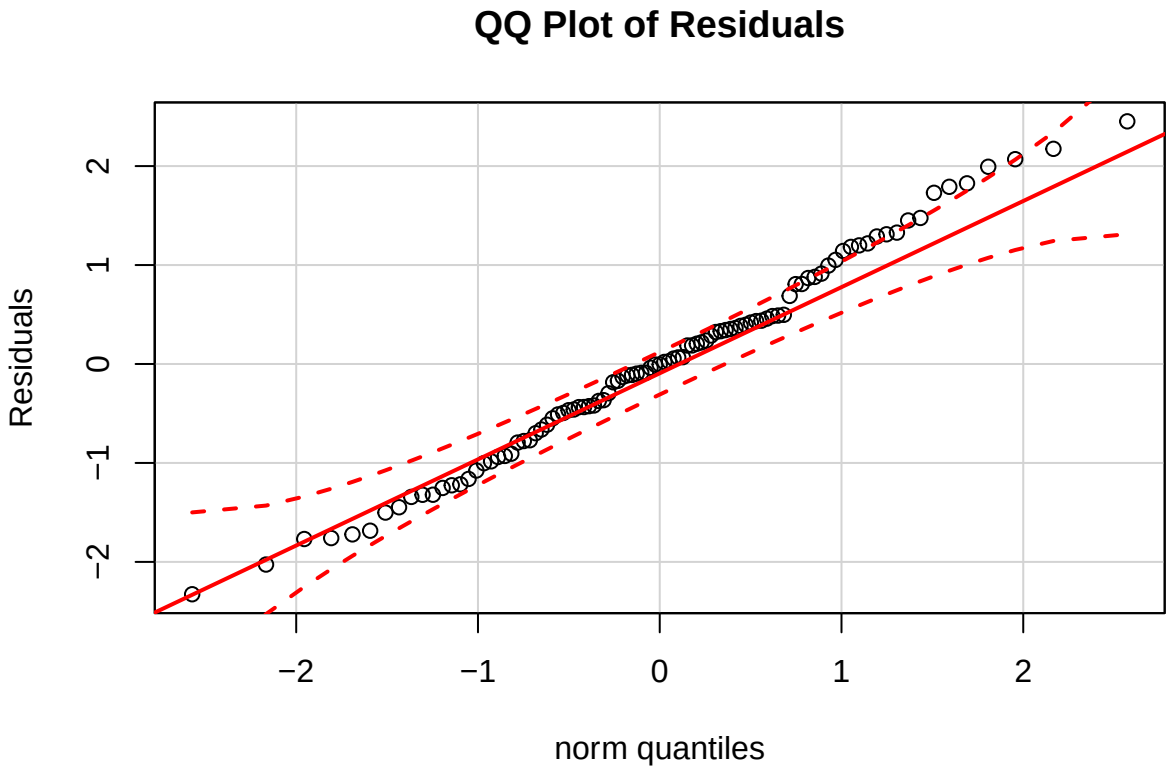
The autocorrelation function does not really show any autocorrelation.

At a glance, it's a little hard to tell the white noise and the residuals apart. The simulated residuals display more flucuations than the white noise, but there is not an obvious trend. Based on that, and the lack of autocorrelation from above, it is probably reasonable to say the evidence supports the hypothesis that the residuals resemble white noise.

```r
#Histogram of Residuals
hist(resids, main = "Histogram of TS Residuals")
```




```r
#QQ Plot of residuals
qqPlot(resids, main="QQ Plot of Residuals", ylab="Residuals")
```



The histogram does not look particularly normally distributed. However, the QQ plot is mostly normal, following the line with some exceptions. The QQ plot shows a bit of a trend in the residuals, as values fluctuate below and above the y = x line especially in the positive direction. This is probably enough evidence to be cautious about concluding the residuals are normally distributed.