# Homework 5
# STAT 5511
# Charles R. Doss
# Solution

The usual formatting rules:

- Your homework (HW) should be formatted to be easily readable by the grader.
- You may use knitr or Sweave in general to produce the code portions of the HW. However, the output from knitr/Sweave that you include should *be only what is necessary to answer the question*, rather than just any automatic output that R produces. (You may thus need to avoid using default R functions if they output too much unnecessary material, and/or should make use of `invisible()` or `capture.output()`.)
  - For example: for output from regression, the main things we would want to see are the estimates for each coefficient (with appropriate labels of course) together with the computed OLS/linear regression standard errors and p-values. If other output is not needed to answer the question, it should be suppressed!
- Code snippets that directly answer the questions can be included in your main homework document; ideally these should be preceded by comments or text at least explaining what question they are answering. Extra code can be placed in an appendix.
- All plots produced in R should have appropriate labels on the axes as well as titles. Any plot should have explanation of what is being plotted given clearly in the accompanying text.
- Plots and figures should be *appropriately sized*, meaning they should not be too large, so that the page length is not too long. (The arguments fig.height and fig.width to knitr chunks can achieve this.)
- **Directions for "by-hand" problems:** In general, credit is given for (correct) shown work, not for final answers; so show **all** work for each problem and explain your answer fully.

```
options(warn=-1)
```

1. Shumway and Stoffer (4th ed.), question 3.11,

   **Solution:**

   **3.11**

   (a) By invertibility, we can write $W_{n+1} = \sum_{j=0}^{\infty} \pi_j X_{n+1-j}$ with $\pi_j = (-\theta)^j$, so that $X_{n+1} = -\sum_{j=1}^{\infty}(-\theta)^j X_{n+1-j} + W_{n+1}$. Thus, taking conditional expectations, $\tilde{X}_{n+1} = -\sum_{j=1}^{\infty}(-\theta)^j X_{n+1-j}$. Then we have that the MSE is

   $$E(X_{n+1} - \tilde{X}_{n+1})^2 = E(-\sum_{j=1}^{\infty}(-\theta)^j X_{n+1-j} + W_{n+1} + \sum_{j=1}^{\infty}(-\theta)^j X_{n+1-j})^2 = EW_{n+1}^2 = \sigma^2.$$

   (b) Truncating yields

   $$\tilde{X}_{n+1}^n := -\sum_{j=1}^{n}(-\theta)^j X_{n+1-j}. \tag{1}$$

   Thus the MSE is

   $$E\left[\sum_{j=n+1}^{\infty} -(-\theta)^j X_{n+1-j} + W_{n+1}\right]^2 = E\left[-(-\theta)^{n+1}\sum_{j=n+1}^{\infty}(-\theta)^{j-(n+1)} X_{n+1-j} + W_{n+1}\right]$$
   $$= E\left[-(-\theta)^{n+1}W_0 + W_{n+1}\right]^2 = \sigma^2(1+\theta^{2(n+1)}).$$

   Since $|\theta| < 1$, when $\theta^{2n}$ decreases exponentially in $n$ so for moderate or large values of $n$ it is nearly 0.

For the next two questions, download the file `HW5dat.rsav` from Canvas and `load(``HW5dat.rsav'')`. It has two time series objects, `dat1` and `dat2`, which you will analyze in the next two questions.

The analysis for each dataset should begin on a new page and should have as label the name of the dataset (dat1 or dat2). Your job is to fit the best ARIMA$(p, d, q)$ model to each dataset that you can. Your output should be in the following format. (Points will be deducted if it is not.)

- The analysis for each question/dataset should begin on a new page and should have as label the name of the dataset.
- On the first page you should state on which page each question begins. [Note: One simple way to automate this in LaTeXis to use `\section{}` to start each dataset and then include a `\tableofcontents`. You could also use `\label{}` and `\pagref{}` commands.]
- On the first page of output for each problem, you should first have a summary (labeled "Summary") that provides the model chosen, parameter estimates, standard errors, and p-values in that model. Specify explicitly if you exclude a constant term. For example, "I chose an ARIMA$(1, 2, 3)$ model including a constant/intercept term. The parameter estimates were ...". If you believe the data cannot distinguish between two (or more) models you should describe both (all) of them in this manner here.
- After the summary, should be an explanation (labeled "Explanation"). Provide a clear explanation of why you selected the model you selected. Refer to the output of your analysis, which will be below. The model selection and diagnostic techniques we have discussed in class can be discussed here. You do not need to (and should not) provide an exhaustive list of all possible models, but should rather provide explanation for which models were reasonable contenders (and why), and which model (or models) were the best out of those contenders (and why).
- After the explanation is the "Output" you refer to in your explanation. (The output may be plots or output from various commands.) All of it should be clearly formatted, and labeled or described. You do not need to provide exhaustive output from every command you have run, but you should include enough to justify all the arguments you make in your summary.

Finally, *please refer to the original/raw (untransformed) time series as $X_t$ in your descriptions and as xx in your code.*

2. Analyze `dat1`.

**Solution:**

Please note: For the "Output" section of the answers below, we have provided more output than most students will include, for pedagogical reasons. Student "output" will generally be somewhat shorter.

**Solution:**

**Summary:** I choose the model ARIMA(1,0,2) excluding a constant, $\phi(B)X_t = \theta(B)W_t$, where $\phi(B) = 1 - \phi_1 B$ and $\theta(B) = 1 + \theta_1 B + \theta_2 B^2$. See the table below for the parameter estimates, standard errors, and p-values.

Table 1: Parameter estimates, standard errors, and p-values for dat1

| Coefficient | Estimate | Standard Error | P-value |
|---|---|---|---|
| $\phi_1$ | 0.5867 | 0.0471 | 0.0000 |
| $\theta_1$ | 0.9070 | 0.0546 | 0.0000 |
| $\theta_2$ | 0.2542 | 0.0549 | 0.0000 |

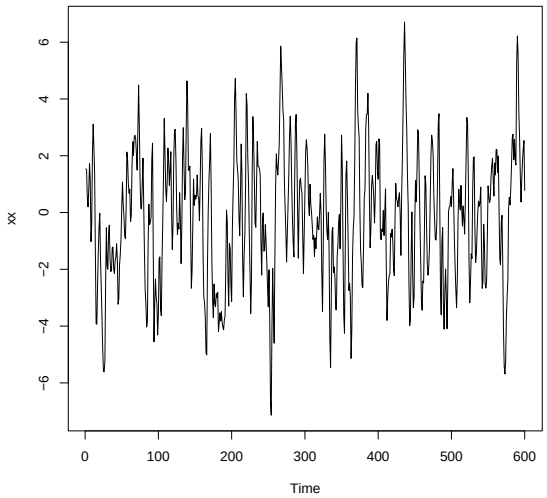
Note: ARIMA(3,0,1) excluding the constant and ARIMA(4,0,0) excluding the constant are also acceptable.

**Explanation:** We first check the data plot, the sample ACF plot and the sample PACF plot, from which we conclude that the data are stationary, getting $d = 0$.

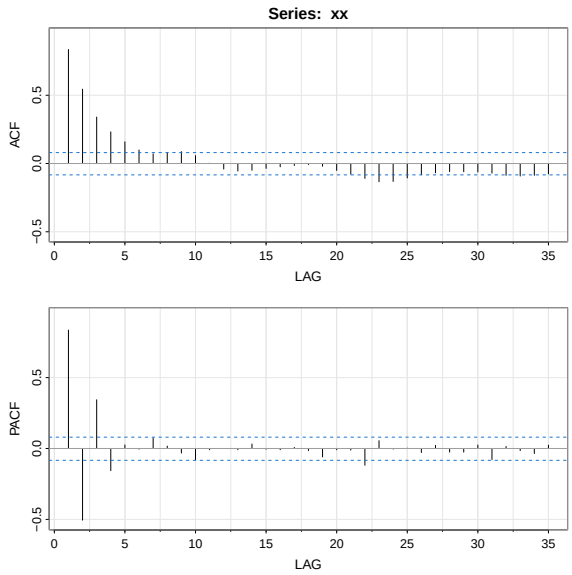
The sample ACFs for lags 1, 2, 3, 4 and 5 are all relatively large and we try guessing $q = 0, 1, 2, 3, 4, 5$. The sample PACFs for lags 1, 2, 3 and 4 are all relatively large and we try guessing $p = 0, 1, 2, 3, 4$. By comparing the AICc and BIC values, we get ARIMA(1,0,2) excluding the constant. The diagnostic plots all support this model.

**Output:**

```
load("HW5dat.rsav")
library(astsa)
xx=dat1
nn=length(xx)
plot.ts(xx)
```




```
invisible(acf2(xx))
```




From the sample ACF plot, we see that most points lie within the 95% confidence interval. We assume stationarity for the data and take $d = 0$. The sample PACF has large spikes at lags 1, 2, 3 and 4, so we try $p = 0, 1, 2, 3, 4$. The sample ACF has large spikes at lags 1, 2, 3, 4 and 5, so we try $q = 0, 1, 2, 3, 4, 5$.

```
maxAR<-4+1
maxMA<-5+1
fits_all<-vector("list",length=maxAR)
AICmin<-BICmin<-Inf
for (ii in 1:maxAR){
  fits_all[[ii]]<-vector("list",length=maxMA)
  for (jj in 1:maxMA){
    fits_all[[ii]][[jj]]<-sarima(xx,p=ii-1,d=0,q=jj-1,no.constant=FALSE, details=FALSE)
    if(fits_all[[ii]][[jj]]$AICc<AICmin) AICmin<-fits_all[[ii]][[jj]]$AICc
    if(fits_all[[ii]][[jj]]$BIC<BICmin) BICmin<-fits_all[[ii]][[jj]]$BIC
  }
}
#look at AICs and BICs
print("check AICc's")

## [1] "check AICc's"

for (ii in 1:maxAR){
  for(jj in 1:maxMA){
    print(paste0(ii-1,"",jj-1))
    print(((fits_all[[ii]][[jj]])$AICc - AICmin) * nn)
  }
}

## [1] "00"
## [1] 978.0999
## [1] "01"
## [1] 360.678
## [1] "02"
```

```
## [1] 96.53504
## [1] "03"
## [1] 20.50228
## [1] "04"
## [1] 11.37281
## [1] "05"
## [1] 8.219487
## [1] "10"
## [1] 261.6708
## [1] "11"
## [1] 16.50688
## [1] "12"
## [1] 0.02648861
## [1] "13"
## [1] 2.060205
## [1] "14"
## [1] 1.610372
## [1] "15"
## [1] 3.478157
## [1] "20"
## [1] 86.5334
## [1] "21"
## [1] 3.78256
## [1] "22"
## [1] 2.060242
## [1] "23"
## [1] 3.376939
## [1] "24"
## [1] 3.529663
## [1] "25"
## [1] 5.49845
## [1] "30"
## [1] 13.01408
## [1] "31"
## [1] 0
## [1] "32"
## [1] 1.917694
## [1] "33"
## [1] 2.569418
## [1] "34"
## [1] 2.384855
## [1] "35"
## [1] 0.489483
## [1] "40"
## [1] 0.1155028
## [1] "41"
## [1] 1.651422
## [1] "42"
## [1] 3.62537
## [1] "43"
## [1] 1.971962
```

```
## [1] "44"
## [1] 3.795836
## [1] "45"
## [1] 1.858773

print("check BICc's")

## [1] "check BICc's"

for (ii in 1:maxAR){
  for(jj in 1:maxMA){
    print(paste0(ii-1,"",jj-1))
    print(((fits_all[[ii]][[jj]])$BIC - BICmin) * nn)
  }
}

## [1] "00"
## [1] 964.9432
## [1] "01"
## [1] 351.9048
## [1] "02"
## [1] 92.13858
## [1] "03"
## [1] 20.47579
## [1] "04"
## [1] 15.70947
## [1] "05"
## [1] 16.91243
## [1] "10"
## [1] 252.8976
## [1] "11"
## [1] 12.11042
## [1] "12"
## [1] 0
## [1] "13"
## [1] 6.396863
## [1] "14"
## [1] 10.30332
## [1] "15"
## [1] 16.52049
## [1] "20"
## [1] 82.13694
## [1] "21"
## [1] 3.756071
## [1] "22"
## [1] 6.3969
## [1] "23"
## [1] 12.06988
## [1] "24"
## [1] 16.572
## [1] "25"
```

```
## [1] 22.88325
## [1] "30"
## [1] 12.98759
## [1] "31"
## [1] 4.336658
## [1] "32"
## [1] 10.61064
## [1] "33"
## [1] 15.61176
## [1] "34"
## [1] 19.76966
## [1] "35"
## [1] 22.20978
## [1] "40"
## [1] 4.452161
## [1] "41"
## [1] 10.34437
## [1] "42"
## [1] 16.66771
## [1] "43"
## [1] 19.35676
## [1] "44"
## [1] 25.51614
## [1] "45"
## [1] 27.90758
```

AICc hits (1,2), (1,4), (3,1), (3,2), (3,5), (4,0), (4,1) and (4,5).

BIC hits (1,2).

I select (1,2) based upon ICs and now check if we should include the constant.

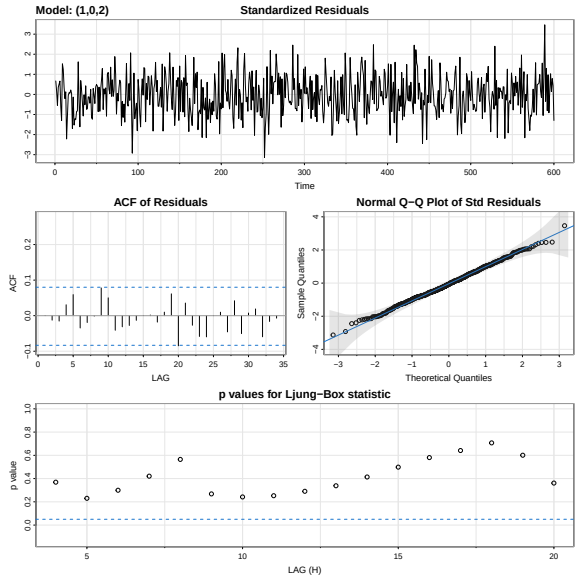
```
#including a constant
sarima(xx, 1,0,2,no.constant=F,details=F)
```

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     ma1     ma2    xmean
##       0.5841  0.9084  0.2555  -0.1779
## s.e.  0.0473  0.0546  0.0549   0.2142
##
## sigma^2 estimated as 1.024:  log likelihood = -859.56,  aic = 1729.12
##
## $degrees_of_freedom
## [1] 596
##
```

```
## $ttable
##       Estimate     SE t.value p.value
## ar1     0.5841 0.0473 12.3602  0.0000
## ma1     0.9084 0.0546 16.6477  0.0000
## ma2     0.2555 0.0549  4.6568  0.0000
## xmean  -0.1779 0.2142 -0.8304  0.4066
##
## $AIC
## [1] 2.881869
##
## $AICc
## [1] 2.881981
##
## $BIC
## [1] 2.91851

#excluding a constant"
sarima(xx, 1,0,2,no.constant=T,details=F)

## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     ma1     ma2
##       0.5867  0.9070  0.2542
## s.e.  0.0471  0.0546  0.0549
##
## sigma^2 estimated as 1.025:  log likelihood = -859.9,  aic = 1727.81
##
## $degrees_of_freedom
## [1] 597
##
## $ttable
##     Estimate     SE t.value p.value
## ar1   0.5867 0.0471 12.4579       0
## ma1   0.9070 0.0546 16.6209       0
## ma2   0.2542 0.0549  4.6310       0
##
## $AIC
## [1] 2.87968
##
## $AICc
## [1] 2.879747
##
## $BIC
## [1] 2.908993
```

We exclude the constant and take a look at the residual diagnostics of this model. (Note: It would be best to examine diagnostic plots for several of the IC contenders but we do not print the results here.)

```
m=capture.output(sarima(xx, 1,0,2,no.constant=T))
```




The residuals' ACF plot and the plot of p-values for Ljung-Box statistics show there is no significant correlation between residuals. The QQ-plot shows the assumption of normality of the residuals is reasonable. All the plots support our model.

The final model we select is ARMA(1,2) excluding the constant.

3. Analyze `dat2`.

**Solution:**

Please note: For the "Output" section of the answers below, we have provided more output than most students will/should include, for pedagogical reasons. Student "output" will generally be somewhat shorter and more focused/succinct.

**Solution:**

**Summary**: I choose ARIMA(3,1,0) excluding a constant. The model can be written as $\phi(B)(1-B)X_t = W_t$, where $\phi(B) = 1-\phi_1 B-\phi_2 B^2-\phi_3 B^3$. See the tables below for the parameter estimates, standard errors, and p-values.

Table 2: ARIMA(3,1,0) parameter estimates, standard errors, and p-values for dat2

| Coefficient | Estimate | Standard Error | P-value |
|---|---|---|---|
| $\phi_1$ | 0.4979 | 0.0560 | 0 |
| $\phi_2$ | 0.2463 | 0.0614 | 1e-04 |
| $\phi_3$ | -0.2304 | 0.0561 | 1e-04 |

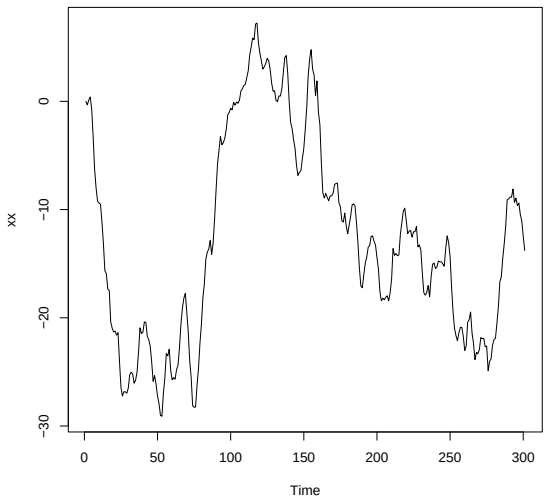
Note: ARIMA(0,1,2) excluding the constant is also acceptable.

**Explanation:** We first check the data plot, the sample ACF plot and the sample PACF plot, from which we conclude that the data are not stationary. Then we try a first difference. The sample ACF plot for the differenced data has large spikes at lags 1 and 2, and the sample PACF plot for the differenced data has large spikes at lags 1, 2 and 3. We now try fitting the ARIMA$(p, 1, q)$ model for which we try $p = 1, 2, 3$ and $q = 1, 2$. By looking at the p-value of the coefficients, we continue trying smaller values of $p$ and $q$ until all coefficients are significant. I choose ARIMA(3,1,0) excluding the constant based upon the IC values. The residual diagnostic plots support this model.

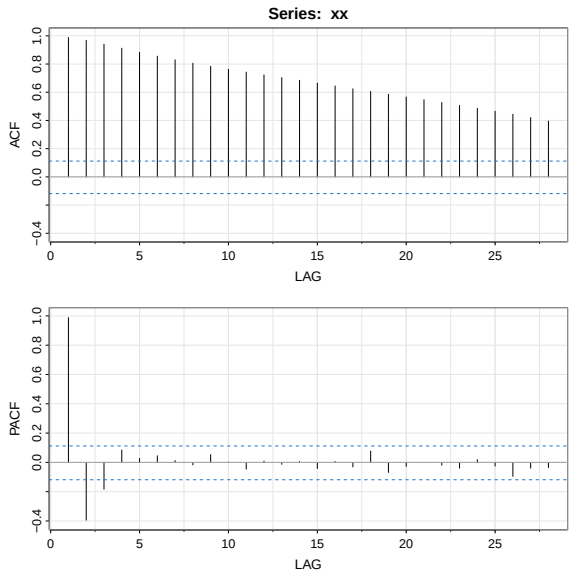
Note: The two models with all coefficients being significant are ARIMA(3,1,0) excluding the constant and ARIMA(0,1,2) excluding the constant. The model ARIMA(3,1,0) has a smaller AIC while ARIMA(0,1,2) has a smaller BIC. Residual diagnostic plots for both models are acceptable.

**Output:**

```
xx=dat2
nn=length(xx)
plot.ts(xx)
```

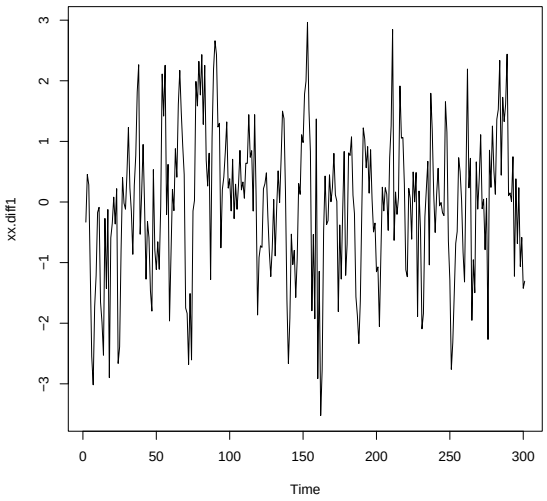



```
invisible(acf2(xx))
```

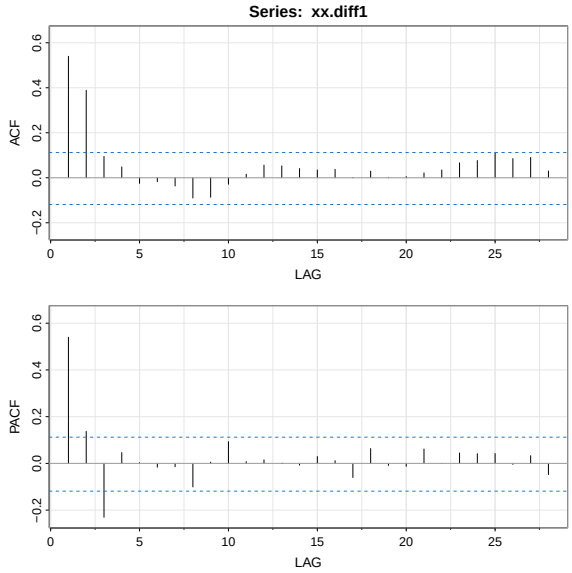



From the ACF plot we see that the data are not stationary. We try a first difference.

```
xx.diff1=diff(xx)
plot.ts(xx.diff1)
```




```
invisible(acf2(xx.diff1))
```




Now most of the points in the sample ACF plot are within the 95% confidence interval and we can conclude the differenced data are stationary. The sample ACF plot for the differenced data has large spikes at lags 1 and 2, and the sample PACF plot for the differenced data has large spikes at lags 1, 2 and 3. We try fitting an ARIMA$(p, 1, q)$ model where $p = 0, 1, 2, 3$ and $q = 0, 1, 2$ for the original data.

```
maxAR<-3+1
maxMA<-2+1
fits_all<-vector("list",length=maxAR)
AICmin<-BICmin<-Inf
for (ii in 1:maxAR){
  fits_all[[ii]]<-vector("list",length=maxMA)
  for (jj in 1:maxMA){
    fits_all[[ii]][[jj]]<-sarima(xx.diff1,p=ii-1,d=0,q=jj-1,no.constant=FALSE, details=FALS
    if(fits_all[[ii]][[jj]]$AICc<AICmin) AICmin<-fits_all[[ii]][[jj]]$AICc
    if(fits_all[[ii]][[jj]]$BIC<BICmin) BICmin<-fits_all[[ii]][[jj]]$BIC
  }
}
#look at AICs and BICs
print("check AICc's")

## [1] "check AICc's"

for (ii in 1:maxAR){
  for(jj in 1:maxMA){
    print(paste0(ii-1,"",jj-1))
    print(((fits_all[[ii]][[jj]])$AICc - AICmin) * nn)
  }
}

## [1] "00"
## [1] 120.042
## [1] "01"
## [1] 58.71611
```

```
## [1] "02"
## [1] 3.14246
## [1] "10"
## [1] 18.26533
## [1] "11"
## [1] 17.15524
## [1] "12"
## [1] 2.091955
## [1] "20"
## [1] 14.46103
## [1] "21"
## [1] 5.972569
## [1] "22"
## [1] 1.908571
## [1] "30"
## [1] 0
## [1] "31"
## [1] 1.431746
## [1] "32"
## [1] 3.44066
```

```
print("check BICc's")
```

```
## [1] "check BICc's"
```

```
for (ii in 1:maxAR){
  for(jj in 1:maxMA){
    print(paste0(ii-1,"",jj-1))
    print(((fits_all[[ii]][[jj]])$BIC - BICmin) * nn)
  }
}
```

```
## [1] "00"
## [1] 109.5352
## [1] "01"
## [1] 51.89833
## [1] "02"
## [1] 0
## [1] "10"
## [1] 11.44756
## [1] "11"
## [1] 14.01278
## [1] "12"
## [1] 2.61093
## [1] "20"
## [1] 11.31857
## [1] "21"
## [1] 6.491544
## [1] "22"
## [1] 6.074957
## [1] "30"
```

```
## [1] 0.5189747
## [1] "31"
## [1] 5.598132
## [1] "32"
## [1] 11.24029
```

AICc hits: (2,2), (3,0) and (3,1).

BIC hits: (0,2) and (3,0).

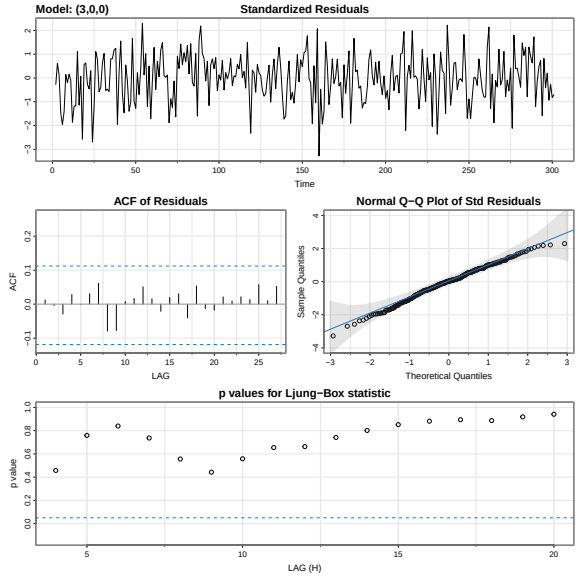
I select (3,0) based upon ICs and now check if we should include the intercept.

```
#including a constant
sarima(xx.diff1, 3,0,0,no.constant=F,details=F)

## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     ar2      ar3    xmean
##       0.4975  0.2460  -0.2307  -0.0511
## s.e.  0.0560  0.0614   0.0561   0.1171
##
## sigma^2 estimated as 0.9797:  log likelihood = -422.87,  aic = 855.75
##
## $degrees_of_freedom
## [1] 296
##
## $ttable
##       Estimate     SE t.value p.value
## ar1     0.4975 0.0560  8.8830  0.0000
## ar2     0.2460 0.0614  4.0096  0.0001
## ar3    -0.2307 0.0561 -4.1163  0.0000
## xmean  -0.0511 0.1171 -0.4364  0.6629
##
## $AIC
## [1] 2.85249
##
## $AICc
## [1] 2.852942
##
## $BIC
## [1] 2.91422

#excluding a constant
sarima(xx.diff1, 3,0,0,no.constant=T,details=F)

## $fit
```

```
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = xmean, include.mean = FALSE, transform.pars = trans, fixed = fixed,
##     optim.control = list(trace = trc, REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     ar2      ar3
##       0.4979  0.2463  -0.2304
## s.e.  0.0560  0.0614   0.0561
##
## sigma^2 estimated as 0.9803:  log likelihood = -422.97,  aic = 853.94
##
## $degrees_of_freedom
## [1] 297
##
## $ttable
##     Estimate     SE t.value p.value
## ar1   0.4979 0.0560  8.8893   0e+00
## ar2   0.2463 0.0614  4.0120   1e-04
## ar3  -0.2304 0.0561 -4.1095   1e-04
##
## $AIC
## [1] 2.846458
##
## $AICc
## [1] 2.846728
##
## $BIC
## [1] 2.895841
```

We exclude the constant and take a look at the residual diagnostics of this model. (Note: It would be best to examine diagnostic plots for several of the IC contenders but we do not print the results here.)

```
m=capture.output(sarima(xx.diff1,3,0,0,no.constant=T))
```




The residuals' ACF plot and the plot of p-values for Ljung-Box statistics show there is no significant correlation between residuals. The QQ-plot shows the assumption of normality of the residuals is reasonable. All the plots support our model.