STAT443 Forecasting - Assignment 4

Christopher Risi - Group B - 20239762

July 17, 2018

1 Theory

Problem 1

Consider the $MA(2)$ representation:

$$Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

Let the estimated results be as follows:

$$\hat{Y}_t = \epsilon_t + 0.5\epsilon_{t-1} - 0.2\epsilon_{t-2},$$

$$RSS = 0.375, \quad n = 150$$

Answer the following questions:

1. Show that the variance of Y_t *is approximately* 0.003225.

$$\begin{aligned}
var(Y_t) &= var(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}) \\
var(Y_t) &= var(\epsilon_t) + \theta_1^2 var(\epsilon_{t-1}) + \theta_2^2 var(\epsilon_{t-2}) \\
var(Y_t) &= \sigma^2(1 + \theta_1^2 + \theta_2^2) \\
var(Y_t) &= (0.375/150)(1 + 0.5^2 + (-0.2)^2) = 0.003225
\end{aligned} \tag{1}$$

2. Derive analytically an expression for the autocovariance function $\gamma(k)$ *pertaining to the* $MA(2)$ *representation* $Y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$.

$$\gamma(0) = \sigma^2(1 + \theta_1^2 + \theta_2^2) \tag{2}$$

$$\begin{aligned}
\gamma(1) &= E[Y_t Y_{t-1}] = E[(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2})(\epsilon_{t-1} + \theta_1 \epsilon_{t-2} + \theta_2 \epsilon_{t-3})] \\
\gamma(1) &= E[\theta_1 \epsilon_{t-1}^2 + \theta_1 \theta_2 \epsilon_{t-2}^2] \\
\gamma(1) &= \theta_1 E[\epsilon_{t-1}^2] + \theta_1 \theta_2 E[\epsilon_{t-2}^2] \\
\gamma(1) &= \sigma^2(\theta_1 + \theta_1 \theta_2)
\end{aligned} \tag{3}$$

$$\begin{aligned}
\gamma(2) &= E[Y_t Y_{t-2}] = E[(\epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2})(\epsilon_{t-2} + \theta_1 \epsilon_{t-3} + \theta_2 \epsilon_{t-4})] \\
\gamma(2) &= E[\theta_2 \epsilon_{t-2}^2] \\
\gamma(2) &= \sigma^2 \theta_2
\end{aligned} \tag{4}$$

$$\gamma(k) = 0 \quad for\ k > 2 \tag{5}$$

3. Suppose that the forecasst errors

$$\epsilon = Y_t - E_{t-1}[Y_t] = 0.04$$

and

$$\epsilon_{t-1} = Y_{t-1} - E_{t-2}[Y_{t-1}] = -0.03$$

Construct a 95% confidence interval for your forecast of Y_{t+1}.

Begin by constructing the expectation:

$$\begin{aligned} E[Y_{t+1}] &= E[\epsilon_{t+1} + 0.5\epsilon_t - 0.2\epsilon_{t-1}] \\ E[Y_{t+1}] &= E[\epsilon_{t+1}] + 0.5E[\epsilon_t] - 0.2E[\epsilon_{t-1}] \\ E[Y_{t+1}] &= 0 + 0.5(0.04) - 0.2(-0.03) = 0.026 \end{aligned} \tag{6}$$

Then calculating the variance:

$$\begin{aligned} var[Y_{t+1}] &= var[\epsilon_{t+1} + 0.5\epsilon_t - 0.2\epsilon_{t-1}] \\ var[Y_{t+1}] &= var[\epsilon_{t+1}] + 0.5^2 var[\epsilon_t] - 0.2^2 var[\epsilon_{t-1}] + zerocrosscov. \\ var[Y_{t+1}] &= \sigma^2 + 0.5^2(0) - 0.2^2(0) = (0.375/150) = 0.0025 \end{aligned} \tag{7}$$

Finally,

$$\begin{aligned} 0.95 &\approx Pr(\hat{Y}_{t+1} - 2\sqrt{var(\hat{Y}_{t+1})} < Y_{t+1} < \hat{Y}_{t+1} - 2\sqrt{var(\hat{Y}_{t+1})}) \\ 0.95 &\approx Pr(0.026 - 2\sqrt{0.0025} < Y_{t+1} < 0.026 + 2\sqrt{var(0.0025})) \\ 0.95 &\approx Pr(0.026 - 2\sqrt{0.0025} < Y_{t+1} < 0.026 + 2\sqrt{var(0.0025})) \\ 0.95 &\approx Pr(-0.074 < Y_{t+1} < 0.126) \end{aligned} \tag{8}$$

Therefore we are 95% confident that the true value of Y_{t+1} falls within the range $(-0.074, 0.126)$.

4. Prove or disprove that the $MA(2)$ *above can be approximated by an* $AR(\infty)$*. If that is the case, compute the infinite autoregressive weights* π_1, π_2, *and* π_3.

Y_t is invertible if it has an infinite autoregressive representation, $AR(\infty)$, as

$$\pi(L)Y_t = \epsilon_t$$

where $\pi(L) = 1 - \pi_1 L - \pi_2 L^2 - ...,$

where $\pi_k -> 0$ as $n -> \infty$

An $MA(2)$ is invertible only if $\Theta(r^{-1}) = 0$ produces $|r| < 1$.

$Y_t = \Theta(L)\epsilon_t, \quad \Theta(L) = 1 + 0.5L - 0.2L^2$

Define $\Theta(r^{-1}) = 1 + 0.5r^{-1} - 0.2r^{-2}$

$\Theta(r^{-1}) = 0 = r^2 + 0.5r^1 - 0.2$

Then using the quadratic formula:

$r_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-0.5 \pm \sqrt{0.5^2 + 4(0.2)}}{2}$

$r_1 = 0.262347538$ $r_2 = -0.762347538$

Since $|r_1| < 1$ and $|r_2| < 1$, then MA is invertible, therefore it has an $AR(\infty)$ representation:

$$\tilde{\pi}_k = -0.5\tilde{\pi}_{k-1} + 0.2\tilde{\pi}_{k-2} \tag{9}$$
$$\tilde{\pi}_1 = -0.5\tilde{\pi}_0 + 0.2\tilde{\pi}_{-1} = -0.5(1) + 0.2(0) = -0.5 \tag{10}$$
$$\tilde{\pi}_2 = -0.5\tilde{\pi}_1 + 0.2\tilde{\pi}_0 = -0.5(-0.5) + 0.2(1) = 0.45 \tag{11}$$
$$\tilde{\pi}_3 = -0.5\tilde{\pi}_2 + 0.2\tilde{\pi}_1 = -0.5(0.45) + 0.2(-0.5) = -0.325 \tag{12}$$

Therefore

$$-\tilde{\pi}_1 = \pi_1 = 0.5 \tag{13}$$
$$-\tilde{\pi}_2 = \pi_2 = -0.45 \tag{14}$$
$$-\tilde{\pi}_3 = \pi_3 = 0.325 \tag{15}$$

So, $Y_t = 0.5Y_{t-1} - 0.45Y_{t-2} + 0.325Y_{t-3} + ...$

5. Suppose you managed to estimate $\hat{\sigma}^2 = 0.0024$ *for the overfitted* $MA(4)$:

$$Y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \theta_3\epsilon_{t-3} + \theta_4\epsilon_{t-4}$$

.

Test $H_0 : Y_t \sim MA(2)$ *versus* $H_1 : Y_t \sim MA(4)$

The test for $H_0 : Y_t \sim MA(2)$ *versus* $H_1 : Y_t \sim MA(4)$ is given by comparing the Likelihood Ratio Test statistics to the χ^2 critical value with 2 degrees of freedom.

$$LRT_{stat} = n \cdot ln(\frac{\hat{\sigma}_R^2}{\hat{\sigma}^2}) = 150 \cdot ln(0.0025/0.0024) \approx 6.1233 \sim \chi^2_{0.95}(2)$$

The critical value for $\chi^2_{0.95}(2) \approx 5.991$ which is less than our statistic of 6.1233 therefore we reject $H_0 : \theta_3 = \theta_4 = 0$.

2 Practicum

Problem 1

Use the seasonally adjusted series you worked with on Assignment 1 to perform the following tasks: Construct a measure of the business cycle Y_t using the Trend Stationary approach and fit an $MA(q)$ process using both the $BIC(k)$ and $AIC(k)$ to estimate q. Include standard errors in brackets under your estimates of $\hat{\theta}_j$, and provide an estimate of σ. Perform a likelihood ratio test at the 5% level of

$$H_0 : Y_t \sim MA(q) \; versus \; H_1 : Y_t \sim MA(q+2)$$

where q is chosen from the $BIC(k)$. Calculate the p-value of the test statistic.

The construction of the business cycle Y_t comes from the residuals of the following model:

```
##
## Call:
## lm(formula = X_t ~ seq_along(CPE_Cons_Goods$X_t), data = CPE_Cons_Goods)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14397 -0.04027 -0.02086  0.05518  0.12311
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.207e+01  9.159e-03 1318.15   <2e-16 ***
## seq_along(CPE_Cons_Goods$X_t) 8.158e-03  8.540e-05   95.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06203 on 183 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9802
## F-statistic:  9124 on 1 and 183 DF,  p-value: < 2.2e-16
```

$$\underset{t}{\hat{X}_t} = \underset{1318.15}{\hat{\alpha}} + \underset{95.52}{\hat{\mu}}\, t = 12.07 + 0.008158$$

Next we constructed a table of $AIC(k)$ and $BIC(k)$ values for $k = 0, 1, ...19$.

q	AIC	BIC
0	-5.57	-5.57
1	-6.82	-6.81
2	-7.47	-7.44
3	-8.03	-7.98
4	-8.44	-8.37
5	-8.72	-8.63
6	-8.75	-8.64
7	-9.06	-8.94
8	-9.19	-9.05
9	-9.24	-9.09
10	-9.27	-9.10
11	-9.34	-9.15
12	-9.36	-9.15
13	-9.38	-9.15
14	-9.42	-9.18
15	-9.42	-9.16
16	-9.47	-9.19
17	-9.52	-9.22
18	-9.53	-9.22
19	-9.52	-9.19

```
## [1] "Min AIC:"
```

```
## [1] 18
```

```
## [1] "Min BIC:"
```

```
## [1] 17
```

The issue that arose however was that no matter how many values of q were added the BIC and AIC just kept chosing the next highest value of q. We were instructed by our TA to instead use a value of $q = 5$ as the best model instead. The resulting model:

```
##
## Call:
## arima(x = Yt, order = c(0, 0, 5), include.mean = FALSE)
##
## Coefficients:
##          ma1     ma2     ma3     ma4     ma5
##       1.6086  1.9782  1.9019  1.1642  0.5059
## s.e.  0.0666  0.1136  0.1234  0.1026  0.0634
##
## sigma^2 estimated as 0.0001544:  log likelihood = 546.57,  aic = -1081.13
```

$$\underset{(s.e.)}{\hat{Y}_t} = \epsilon_t + \underset{(0.0666)}{\hat{\theta}_1} \epsilon_{t-1} + \underset{(0.1136)}{\hat{\theta}_2} \epsilon_{t-2} + \underset{(0.1234)}{\hat{\theta}_3} \epsilon_{t-3} + \underset{(0.1026)}{\hat{\theta}_4} \epsilon_{t-4} + \underset{(0.0634)}{\hat{\theta}_5} \epsilon_{t-5} = \epsilon_t + 1.6086\epsilon_{t-1} + 1.9782\epsilon_{t-2} + 1.9019\epsilon_{t-3} + 1.1642\epsilon_{t-4}$$

Finally we're asked to perform a likelihood ratio test at the 5% level of our $MA(5)$ model versus an $MA(7)$:

```
## [1] "What is our test statistic?"

## [1] 67.36946

## [1] "Is our test statistic greater than our critical p-value?"

## [1] TRUE

## [1] "What is the p-value of our test statistic?"

## [1] 2.331468e-15
```

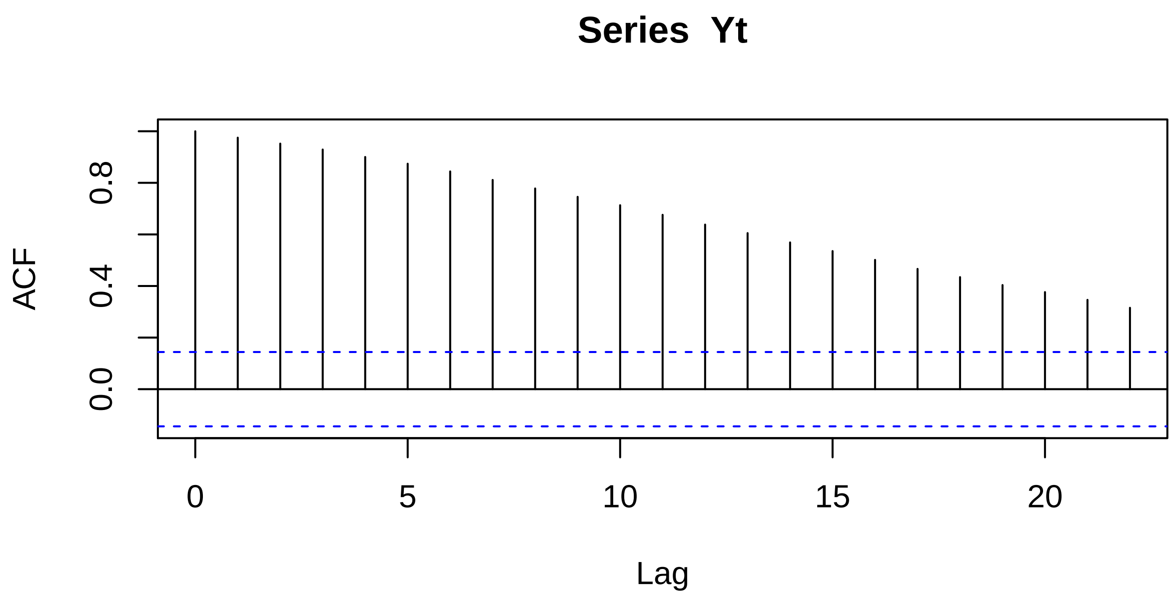
Problem 2

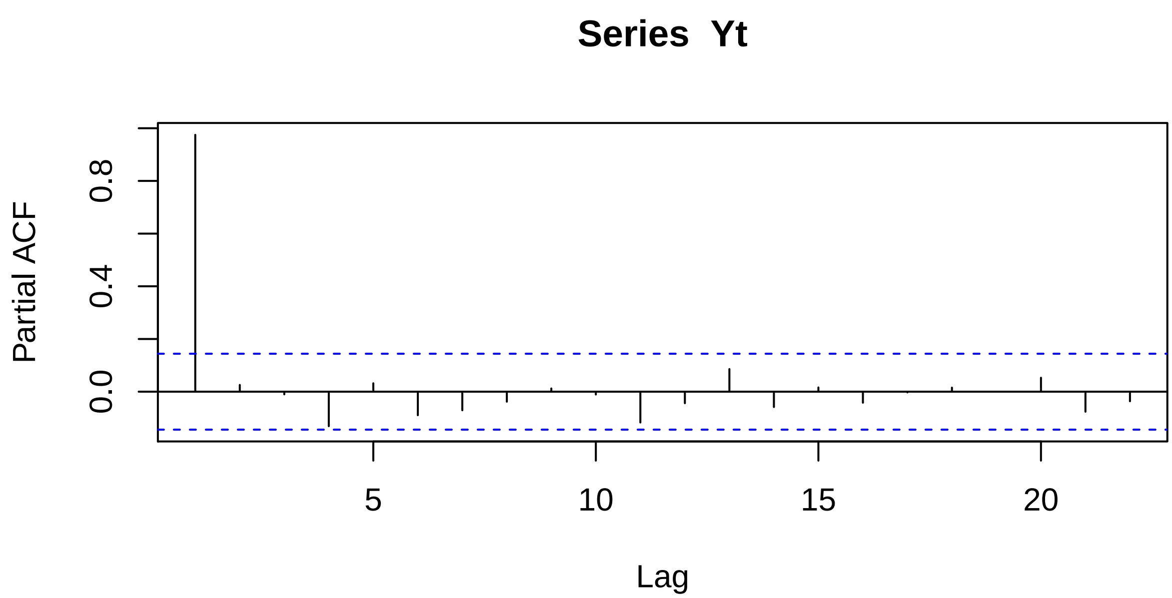
Use Box-Jenkins identification to identify and estimate an $ARMA(p, q)$ model for the Trend Stationary Y_t.

The way Box-Jenkins Identification works is by looking at the patterns in both our estimated autocorrelation function $\hat{\rho}(k)$ and our estimated partial autocorrelation function ϕ_{kk}. We can identify the most likely model by the following guide:

Model	Rho_k	Phi_kk
AR(p)	damped exponential	cut-off at k=p
MA(q)	cut-off at k=q	damped exponential
ARMA(p,q)	damped exponential	damped exponential

We can use built in R functions `acf(Yt)` and `pacf(Yt)` to visualize these two patterns.





What is immediately clear from our plots is that our Autocorrelation function is a damped exponential and our Partial Autocorrelation function has a cut-off. This indicates that we have an $AR(p)$ model, from our graph, with the dotted lines indicating our Standard Error for our Partial ACF it is clear that the cutoff should be an AR(1).

Problem 3

Using your result from (2), run the following diagnostic tests:

(a) Plot the standardized residuals $\hat{z}_t = \hat{a}_t/\hat{\sigma}$ of the fitted model and comment about whether the normal distribution appears to be appropriate.
(b) Run a formal Jarque-Bera test for normality.
(c) Run a Box-Pierce test for joint autocorrelation.
(d) Test for the presence of ARCH(6).

Fit of AR(1) Model from (2)

```
##
## Call:
## arima(x = Yt, order = c(1, 0, 0), include.mean = FALSE)
##
## Coefficients:
##          ar1
##       0.9946
## s.e.  0.0062
##
## sigma^2 estimated as 7.645e-05:  log likelihood = 612.03,  aic = -1220.06
```

$$\underset{(s.e.)}{\hat{Y}_t} = \underset{(0.0062)}{\hat{\phi}_1} Y_{t-1} = 0.9946 Y_{t-1}$$

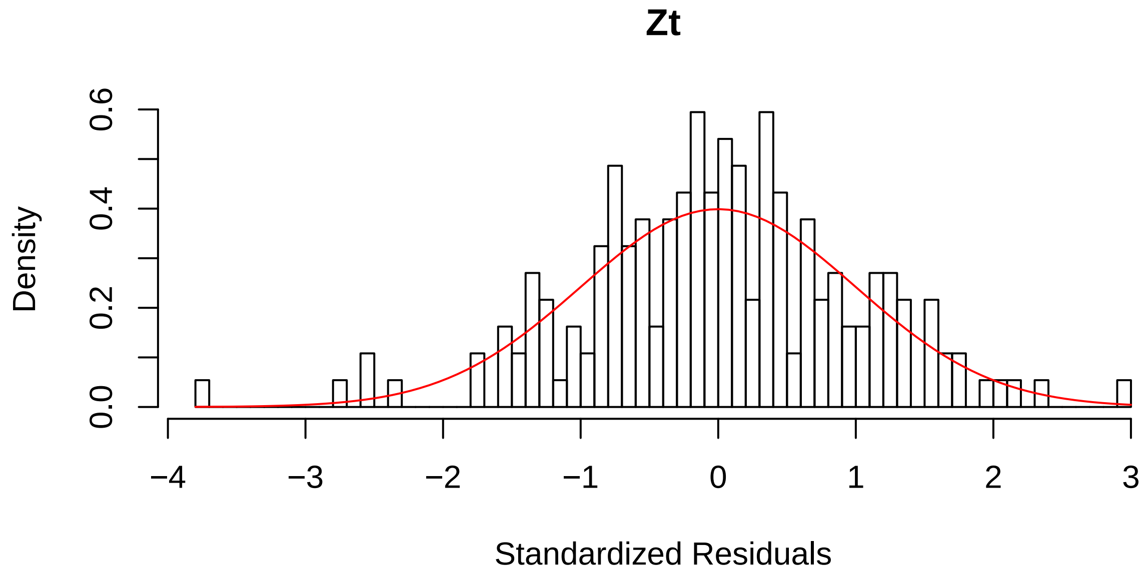
(a) Plot the standardized residuals

```
## [1] 0.0141425
```

```
## [1] 185
```

```
## [1] 185
```

```
## [1] 0.0006157278
```



From a quick visual inspection the Normal Distribution appears it might be appropriate for our fitted model. However it is a little hard to tell exactly from visual inspection, there does appear to be something off with our fit.

(b) Run a formal Jarque-Bera test for normality.

Our Jarque-Bera statistic is defined by $J_{stat} = n(\frac{\hat{k}_3^2}{6} + \frac{(\hat{k}_4-3)^2}{24}) \sim \chi^2_{0.95}(2)$

The Jarque-Bera statistic tests if our standardized residuals z_t are in fact Normally Distributed. Our null hypothesis H_0 : our residuals are normally distributed.

```
## [1] "K_3"
```

```
## [1] -0.2414272
```

```
## [1] "K_4"
```

```
## [1] 3.820804
```

```
## [1] "Jarque-Bera Statistic:"
```

```
## [1] 6.99044
```

```
## [1] "Jarque-Bera Critical Value:"
```

```
## [1] 5.991465
```

Our Jarque-Bera Statistic (6.99044) $> \chi^2_{0.95}(2)$ 5.991 so we reject our null hypothesis that our residuals are normally distributed.

(c) Run a Box-Pierce test for joint autocorrelation.

The Box-Pierce Test is to test whether or not there exists autocorrelation between ϵ_t and ϵ_{t-K} for any $k = 1, ..., M$ where $M = round_{up}(sqrt(N))$. Using the built in Box-Pierce test in R we have found a p-value of 0.000749:

```
## [1] 14
```

```
##
##  Box-Pierce test
##
## data:  residuals
## X-squared = 36.957, df = 14, p-value = 0.000749
```

Therefore we reject the null hypothesis that there does not exist correlation between some ϵ_t and ϵ_{t-K} any $k = 1, ..., 14$. Because there exists correlation between our residuals, we will move on to the ARCH test in part (d).

(d) Test for the presence of ARCH(6).

For our ARCH(6) test we are testing the null hypothesis that for our model:

$$\underset{(s.e.)}{\hat{Y}_t} = \underset{(0.0062)}{\hat{\phi}_1} Y_{t-1} = 0.9946 Y_{t-1} + a_t$$

where

$$a_t = z_t(\sigma^2 + \alpha_1 a_{t-1}^2 + ... + \alpha_6 a_{t-6}^2)^{1/2}$$

Our $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$

```
## [1] 53.54872
```

```
## [1] 12.59159
```

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

As it turns out our ARCH Test Statistical is greater than our Critical value, so we reject the null hypothesis that $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$.

Appendix Code

```r
### Data Load
library(xlsx)
library(dplyr)
CPE_Cons_Goods  <- read.xlsx("CONS_Canada.xls", sheetName = "Sheet3")
CPE_Cons_Goods <- CPE_Cons_Goods %>% mutate(X_t = log(Seasonally.Adjusted))
CPE_Cons_Goods <- CPE_Cons_Goods[complete.cases(CPE_Cons_Goods),]
model.CPE <- lm(X_t ~ seq_along(CPE_Cons_Goods$X_t) , data = CPE_Cons_Goods)
summary(model.CPE)

# Yt is the residuals of regression
Yt <- model.CPE$residuals

# q1
# function for AIC
AIC <- function(res, k, N){
  aic <-log(sum(res^2) / N)
  aic <- aic + 2 * k /N
  aic
}

BIC <- function(res, k, N){
  bic <-log(sum(res^2) / N)
  bic <- bic +  k *log(N)/N
  bic
}

# obtaining the aic and bic
aic.array <- rep(NA, 20)
bic.array <- rep(NA, 20)
N <- length(Yt)
for(ii in 0:19){
  model.arima <- arima(Yt, order = c(0, 0, ii), include.mean = FALSE)
  res.arima <- model.arima$residuals
  aic.array[ii + 1] <- AIC(res.arima, ii, N)
  bic.array[ii + 1] <- BIC(res.arima, ii, N)
}

# Puttng AIC/BIC into a table
AIC_table <- data.frame(cbind(c(0:19),aic.array,bic.array))
rownames(AIC_table) <- c()
colnames(AIC_table) <- c("q", "AIC","BIC")
round(AIC_table, digits=2)


print("Min AIC:")
which(aic.array == min(aic.array))-1
print("Min BIC:")
which(bic.array == min(bic.array))-1
```

```r
model.ma5 <- arima(Yt, order = c(0, 0, 5), include.mean = FALSE)
model.ma5


model.ma7 <- arima(Yt, order = c(0, 0, 7), include.mean = FALSE)

LR <- N * log(model.ma5$sigma2 / model.ma7$sigma2)
hyp.test <- LR > qchisq(p = 0.95, df = 2) # if true reject H_0
# or use p-value:
pvalue.LR <- 1 - pchisq(q = LR, df = 2)

print("What is our test statistic?")
LR
print("Is our test statistic greater than our critical p-value?")
hyp.test
print("What is the p-value of our test statistic?")
pvalue.LR


model <- c("AR(p)", "MA(q)", "ARMA(p,q)")
rho_k <- c("damped exponential", "cut-off at k=q", "damped exponential")
phi_kk <- c("cut-off at k=p", "damped exponential", "damped exponential")

kable( data.frame(Model = model, Rho_k = rho_k, Phi_kk = phi_kk))


# q2

# acf & pacf using R bulit-in function

acf(Yt) # plots autocorrelation function against k
pacf(Yt) # plots partial autocorrelation function against k

# can also use the method from previous assignments
# i.e acf & pacf using linear regression, not recommended.

## etc..

### We are choosing ARMA(p,q) here
# ACF damped-exponential, PACF is cut-off, so we get AR(p)



ar2 <- arima(Yt, order = c(1, 0, 0), include.mean=FALSE)
ar2
residuals <- ar2$residuals


library(ggplot2)
sigma.hat.squared <- sum(ar2$residuals^2)/length(ar2$residuals)
sigma.hat <- sqrt(sigma.hat.squared)
Zt <- residuals/sigma.hat
hist(Zt, breaks = 30, freq = F, main = "Zt", xlab = "Standardized Residuals")
```

```r
curve(dnorm(x),col = "red", add = T)


  ####### Jarque Bera Test
# can fail this test in 2 ways:
# 1. normality is rejected completely
# 2. the model is normal but didn't pass JB test because of
# the existence of significant outliers or structure breaks

#############################################################
# if an ARMA model passes all diagnostics except for JB due
# to outliers, it's still a good model
#############################################################

# find standardized residuals (same as above)
Zt <- residuals/sigma.hat
# calculate K_3, K_4,
K_3 <- (1/length(Zt))*sum(Zt^3)
K_4 <- (1/length(Zt))*sum(Zt^4)
print("K_3")
K_3
print("K_4")
K_4
# find J_Stat and J_crit.
J_Stat <- length(Zt)*(((K_3^2)/6)+((K_4-3)^2)/24)
print("Jarque-Bera Statistic:")
J_Stat
J_Crit <- qchisq(0.95,2)
print("Jarque-Bera Critical Value:")
J_Crit

#### Box-Piece test
# use built-in function:
M <- sqrt(length(residuals))
M <- 14
M

Box.test(x = residuals, type = "Box-Pierce",lag= M)
# returns p-value when run

### ARCH(6) test
## again, use standardized residuals (name them std.residuals to use this code)
std.residuals <- Zt
N <- length(std.residuals)
std.res.sq <- std.residuals^2
ARCH.model <- lm(std.res.sq[-(1:6)]~std.res.sq[-c((1:5), N)]
                     +  std.res.sq[-c(1:4,(N-1),N)] + std.res.sq[-c(1:3, (N-2):N)]
                     + std.res.sq[-c(1:2, (N-3):N)] + std.res.sq[-c(1, (N-4):N)]
                     + std.res.sq[-((N-5):N)] - 1)
R.squared <- summary(ARCH.model)$r.squared
ARCH.test.stat <- N * R.squared
ARCH.test.crit <- qchisq(0.95,6)
ARCH.Null.Hypothesis <- (ARCH.test.stat > ARCH.test.crit) # if true reject H_0
```

```
ARCH.Null.Hypothesis
```