# MSBA 6430: Practice Exam

*Danny Moncada (monca016)*

*April 16, 2020*

```
suppressWarnings(suppressPackageStartupMessages({
library(TSA)
library(ggplot2)
library(dplyr)
library(forecast)
library(tseries) # ADF Test for Stationarity
library(igraph)
}))
```

**Question 1**: Suppose a time series {Yt} t>= 1 follows

$$Y_t = e_t - 0.5e_{t-12}$$

a) What is the name and order of the model for $Y_t$ t>= 1 (e.g. AR(1), MA(3), etc.)?

** The name and order of the model is an MA(12) or SMA(1)[12].

b) Find the mean function E($Y_t$) and the variance Var($Y_t$).

Xuan Bi:

** E(Yt) = 0 - 0.5 x 0 = 0
** Var(Yt) = 1 + (-0.5)^2 * 1 = 1.25

Me:
** The mean function E($Y_t$) is 0.
** The variance function Var($Y_t$) is $1 + \theta_1^2 = 1.25$.

c) Find lag-1 auto-covariance Y1 = Cov($Y_t$, $Y_{t-1}$), and lag-12 auto-covariance Y12 = Cov($Y_t$, $Y_{t-12}$).

Xuan Bi:

** Cov (Yt, Yt-1) = C (et - 0.5et-12, et-1-0.5et-13) = 0
** Cov(et - 0.5et-12, et-12 - 0.5et-24)

** The auto-covariance Y1 Cov($Y_t$, $Y_{t-1}$) = -$\theta_1$ = 0
** The auto-covariance Y12 Cov($Y_t$, $Y_{t_12}$) = -0.5

d) Is {$Y_t$} t>= 1 stationary? Why?

Xuan Bi:

Yes. None of E, Var, Cov, are functions of t.

** Yes, {$Y_t$} is stationary because:
*** the mean function E($Y_t$) is constant over time and is zero.
*** the auto-covariance function Cov($Y_t$, $Y_{t-k}$) = $\gamma_k$ which is also constant over time.

**Question 2**: Suppose we observe the following R output.

```
# Call:
# arima(x = Y, order = c(2, 0, 1), method = "ML")

# Coefficients:
#           ar1       ar2      ma1  intercept
#       0.5259    0.4364   -0.8407     0.0500
# s.e.  0.0332    0.02394   0.0245     0.1269
#
# sigma^2 estimated as 0.9566:  log likelihood = -1397.23, aic = 2802.46
```

a) What is the model? (Hint: recall that R assumes positive sign in front of both AR and MA coefficients. Please round up to 2 digits.)

Xuan Bi: (don't forget et!!!!!)

** (Yt - 0.05) = 0.53(Yt-1 - 0.05) + 0.44(Yt-2 - 0.5) + et - 0.84et-1

** The model is an ARIMA(2,0,1) with this format:

$$(Y_t - 0.05) = 0.53 \cdot (Y_{t-1} - 0.05) + 0.44 \cdot (Y_{t-2} - 0.05) + e_t - 0.84 \cdot e_{t-1}$$

b) Supposed we observed $Y_1$, $Y_2$,..., $Y_{100}$. What are the forecasting values $\hat{Y}_{101}$ and $\hat{Y}_{102}$? (The final result shouldn't contain any $e_t$ or $\hat{e_t}$))

Xuan Bi:

** Yˆ101 - 0.05 = 0.53(Y100 - 0.05) + 0.44(Y99 - 0.05) + eˆ101 - 0.84e$^{100}$ ** Y101 - 0.05 = 0.53(Y100 - 0.05) + 0.44(Y99 - 0.05) + 0 - Y100 - Yˆ100 (borrow result from R)

** Yˆ102 - 0.05 = 0.53(Yˆ101 - 0.05) + 0.44(Y100 - 0.05) + 0 - 0

** $\hat{Y}_{101}$ - 0.05 = 0.53·($Y_{100}$ - 0.05) + 0.44·($Y_{99}$ - 0.05) - 0.84·$e_{t-1}$

** $\hat{Y}_{102}$ = 0.05 + 0.53·($\hat{Y_{101}}$ - 0.05) + 0.44·($Y_{100}$ - 0.05) - 0.84·$e_{t-1}$

** Based on R output, $\hat{Y}_{101}$ = -1.105 and $\hat{Y}_{102}$ = -1.069

c) For the same data, suppose your colleague proposes another model below. What is her model? Is it stationary? Compared with our model above, which one is better and why?

```
# Call:
# arima(x = Y, order = c(0, 1, 1), method = "ML")

# Coefficients:
#           ma1
#       -0.9470
# s.e.   0.0109
#
# sigma^2 estimated as 1.162:  log likelihood = -1493.85, aic = 2989.7
```

Xuan Bi:

** Not stationary, because of differencing.
** Our model is better, smaller AiC. If their model was better, it would be higher.

** Our colleague's model is an ARIMA(0,1,1) model with this format:

$$Y_t = Y_{t-1} + et - 0.95 \cdot e_{t-1}$$

** When comparing the AiC from our first model (2802.46) and the AiC from our colleague's model (2989.7), we see that our AiC is smaller.

**Question 3**: Suppose you collect a set of data {Yt} t>= 1 following

$$Y_t = \emptyset Y_{t-1} + e_t$$

but we are not sure if $\emptyset = 1$ or $\emptyset < 1$.

a) What's the model name if $\emptyset = 1$?

Xuan Bi:
** a) Random walk.

** b) Not stationary.

** c) ARIMA(0, 1, 0)

** d) Fit a trend which does not exist. Predicted values keep increasing with t, while random walk has mean 0 and can decrease anytime. Regression diagnosticstocheck residual correlation; QQ plot; and ADF test.

** The model name when phi is equal to 1 is a random walk.

b) Suppose we conduct the ADF test with R, and get the following result. What conclusion shall we make?

```
#      Augmented Dickey-Fuller Test

# data: y
# Dickey-Fuller = -2.4591, Lag order = 21, p-value = 0.3839
# alternative hypothesis: stationary
```

** The p-value is not low enough to reject the null hypothesis, so we cannot conclude that the process is stationary. We would have to conclude that this process is a random walk.

c) Following question (b), the model is an ARIMA($p$, $d$, $q$). Then what are the values of $p$, $d$, $q$?

- The value of $p$ is 0, the value of $d$ is 1, and the value of $q$ is 0.

d) Based on the conclusion in (b), what's the consequence of fitting this model with a linear regression? How do we check if such regression is spurious? (Hint: think about linear model assumptions and regression diagnostics)

** We can simulate an arbitrary sample path from a pure random walk, and then run a regresson against the process that formed that model. If we see that the random walk we just simulated seems to "explain" the process, then we know that this is spurious regression. We can also perform a residual analysis using a residual plot and QQ plot. If the residuals look like white noise, then we can conclude the process has a linear trend; if we clearly see dependency among consecutive residuals, then it is not white noise and the process is a random walk.

**Question 4**: Suppose we collect a set of quarterly sales data {Yt}t>=1, and 4 dummy variables corresponding to 4 quarters. And we fit with linear regression as below.

```
# Call:
# lm(formula = y ~ 0 + ., data = Quarterly_Sales)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -4.0211 -0.9670 -0.0028 0.9986 4.5368
#
# Coefficients:
#     Estimate  Std. Error t value Pr(>|t|)
# Spring  1.3545   0.1370  9.890   < 2e-16 ***
# Summer  0.3247   0.1370  2.371    0.0182 *
# Fall   -1.4127   0.1370  -10.315 < 2e-16 ***
# Winter -0.7207   0.1370  -5.262 2.34e-07 ***
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 1.37 on 396 degrees of freedom
# Multiple R-squared:  0.3749,   Adjusted R-squared:  0.3686
# F-statistic: 59.38 on 4 and 396 DF, p-value: 2.2e-16
```

Xuan Bi:

a) Yt = 1.35(t==Spring) + 0.32(t==Summer) - 1.41(t==Fall) - 0.72(t==Winter) + et (this is not deterministic model, )
** They are all significant.

b) MA(1)

c) et + 0.77et-1

Yt = 1.35(t==Spring) + 0.32(t==Summer) - 1.41(t==Fall) - 0.72(t==Winter) + et + 0.77et-1

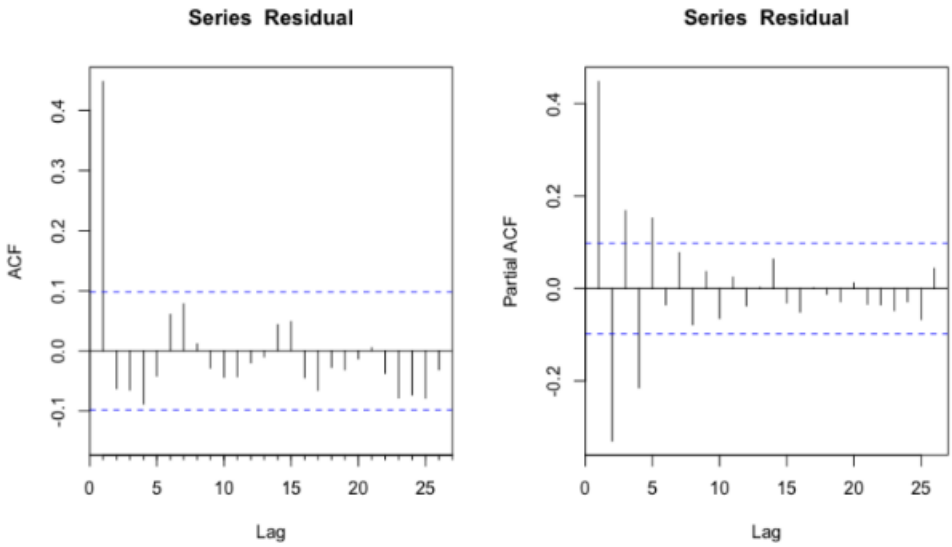
d) E(Yspring) = 1.35 + E(et) + 0.77*E(et-1) (we cancel out the et and et-1 to 0s) Final answer: 1.35

e) Please write down the model formula for $Y_t$. Is each quarter's effect significant?

** The model formula is (1 - $B^4$)(1 - $B$)$Y_t$ = $c$ + (1 - B)(1 - $B^4$)$e_t$

** A coefficient is significant if its magnitude is at least twice as large as its standard error; in this instance, all of the coefficients are significant.

b) Suppose we conduct ACF and PACF on the model residuals, and observe the following. What's the suggested model for the residuals (e.g., AR(1))?




** The suggested model for the residuals is MA(1).

c) After fitting the residuals, we find that the model coefficient is 0.77 in R. What's the overall model for the time series now? Suppose we are in Winter, then what's the expected value (mean) of sales forecast for the upcoming Spring?

** The overall model for the time series $Y_t =$ .

** The expected value (mean) of sales forecast for the upcoming Spring is _____.

**Question 5**: Suppose you observe the following adjacency matrix

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    0    2    0    0    0    0
## [2,]    0    0    2    0    0    0
## [3,]    0    0    0    2    0    0
## [4,]    2    0    0    0    1    1
## [5,]    0    0    0    1    0    1
## [6,]    0    0    0    1    1    0
```

Xuan Bi:

a) Yes, it is weighted, and directed.

b) Same as the R output. I can draw this out by hand.

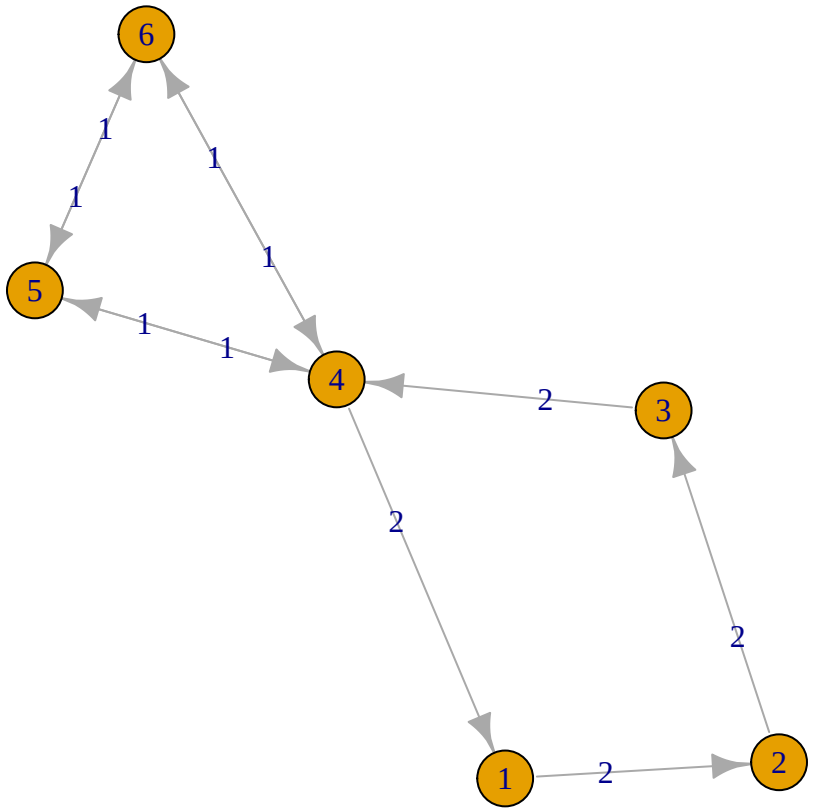
c) If there is an arrow on both sides, then we have a connection in a undirected graph.

d) Is this network weighted? directed?

** Yes, it is weighted because the values are 2 and not 1; it is a directed graph.

b) Please draw the network

```
g <- graph_from_adjacency_matrix(A, weighted = TRUE)
par(mar = c(0, 0, 0, 0)); plot(g, edge.label = E(g)$weight)
```




c) If we convert the network to an undirected one, what's the size of the maximum clique?

** The size of the maximum clique is 3.
** We look at bottom of the graph, with 4, 5, 6... if we convert it to an undirected graph, then these three vertices are connected together and thus the maximum clique is 3.
** Cliques are subsets of vertices, all adjacent to each other, which is also called subgraphs.

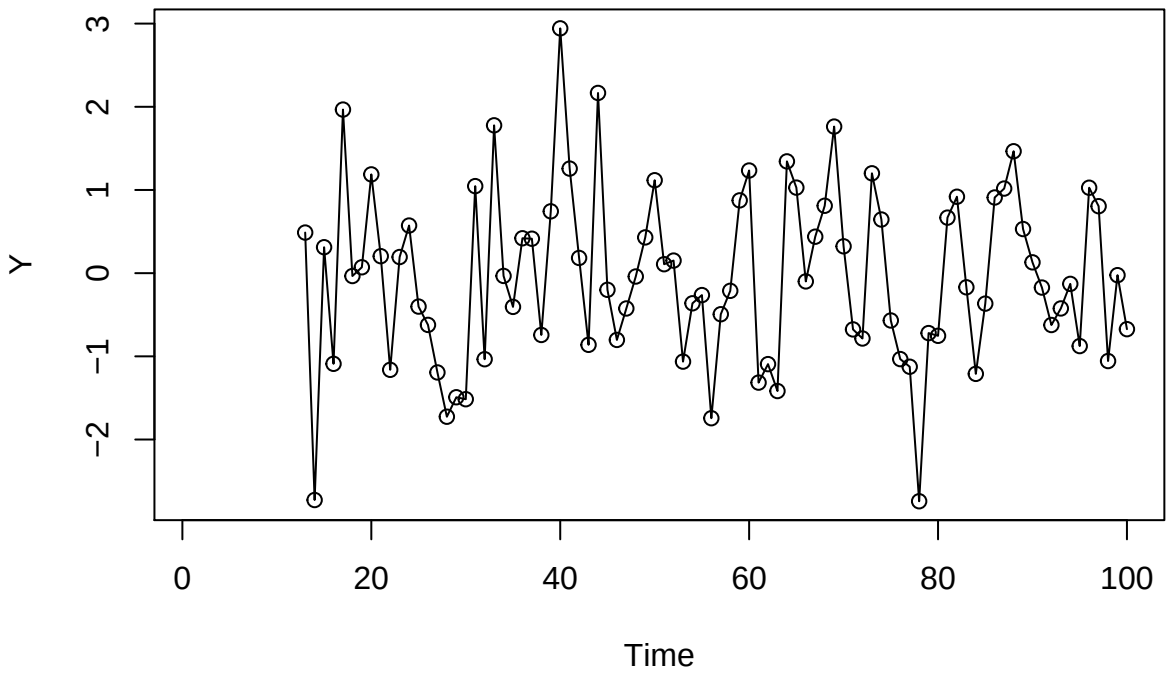
d) Suppose node 2's sender-specific latent factor is p2 = (0, 0.4, 0.6) and node 1's receiver-specific latent factor is q1 = (0.7, 0.1, 0). How likely will node 2 send a signal invitation to node 1?

** Likelihood equation = (0 x 0.7) + (0.4 x 0.1) + (0.6 x 0)
*** 0.04 is your probability.

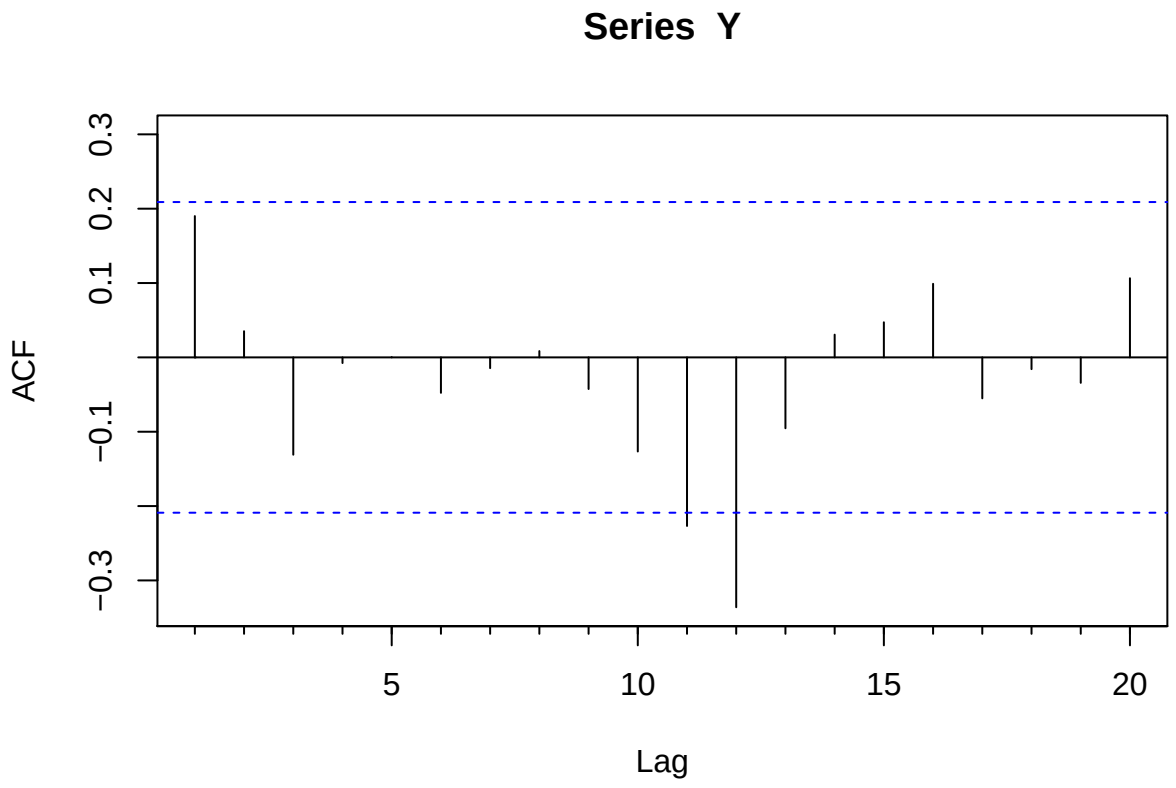
THAT'S ALL FOLKS!

```
##############################################################################################
##############################################################################################
# Code for Q1:

# a)
set.seed(666)
e = rnorm(100)
Y=ts(e-0.5*zlag(e, 12))
plot(Y, type = 'o')
```




```
Acf(Y)
```




```
# b)
# Use R to check the mean
mean(Y[13:100]) # -0.04
```

```
## [1] -0.04347753
```

```
var(Y[13:100]) # 1.09 - close!
```

```
## [1] 1.094065
```

```
# c)
## Come back to this one during exam review
cov(Y[13], Y[14])
```

```
## [1] NA
```

```
Y[13]
```

```
## [1] 0.4879981
```

```
Y[14]
```

```
## [1] -2.727333
```

```r
cov(Y[13], Y[25])
```

```
## [1] NA
```

```r
# d)
adf.test(Y[13:100])
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  Y[13:100]
## Dickey-Fuller = -3.8836, Lag order = 4, p-value = 0.01866
## alternative hypothesis: stationary
```

```r
# Yes

# Code for Q2:

# b)
set.seed(666);
y2 = arima.sim(model=
                 list(order=c(2, 0, 1), ar=c(0.5259, 0.4364), ma=-0.8407, sd=1),
                  n=10000)
auto.arima(y2)
```

```
## Series: y2
## ARIMA(2,0,1) with non-zero mean
##
## Coefficients:
##          ar1     ar2      ma1    mean
##       0.5243  0.4387  -0.8490  0.0672
## s.e.  0.0106  0.0093   0.0079  0.0406
##
## sigma^2 estimated as 0.997:  log likelihood=-14172.9
## AIC=28355.8   AICc=28355.8   BIC=28391.85
```

```r
y2[99]
```

```
## [1] 0.07392018
```

```r
y2[100]
```

```
## [1] -0.09339175
```

```r
# Mathematically calc Y101
0.05 + (0.53*-0.09339175) - 0.5 + (0.44*0.07392018 - 0.05) - 0.84
```

```
## [1] -1.356973
```

```r
# Use R to assist here.
y2_pred <- forecast(y2[0:100], h=2)
y2_pred
```

```
##     Point Forecast     Lo 80     Hi 80     Lo 95    Hi 95
## 101      -1.105703 -2.483957 0.2725511 -3.213561 1.002155
## 102      -1.069320 -2.447671 0.3090299 -3.177325 1.038684
```

```r
yhat_101 <- -1.105703
yhat_102 <- -1.069320

# Code for Q4:

library(TSA)
data(tempdub)
month. <- season(tempdub)
seasonal_lm <- lm(tempdub ~ month.-1)

auto.arima(tempdub)
```

```
## Series: tempdub
## ARIMA(0,0,0)(2,1,0)[12]
##
## Coefficients:
##          sar1     sar2
##       -0.5403  -0.3078
## s.e.   0.0906   0.0937
##
## sigma^2 estimated as 17.25:  log likelihood=-376.58
## AIC=759.17   AICc=759.35   BIC=767.81
```

```r
# Code for Q5:
# c)
# Calculates the size of the largest clique(s).
clique_num(g)
```

```
## Warning in clique_num(g): At cliques.c:1087 :directionality of edges is
## ignored for directed graphs
```

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

```r
# d)
likelihood = 0 * 0.7 + 0.4 * 0.1 + 0.6 * 0
likelihood
```

```
## [1] 0.04
```