--- Timothy 11/02

library(tseries)

library(forecast)

# international-airline-passengers.csv contains airline passenger totals indexed by month

data\_dir <- '/Users/timmy/Documents/school/year-4/fall/pstat175\_survival\_analysis/final\_project/data'

data\_path <- file.path(data\_dir, 'international-airline-passengers.csv')

airline <- read.csv(data\_path)

colnames(airline) <- c('month', 'passengers')

# splitting into train (first 90%) and test (last (10%)) set

train\_indices <- c(1:132) # leaving out last year

train\_airline <- airline[train\_indices,]

test\_airline <- airline[-train\_indices,]

# plotting time series

data <- train\_airline$passengers

plot(ts(data))

# plot shows seasonality, an upward trend, and increasing variance over time

par(mfrow=c(2,1))

# apply lag-12 difference on log-transformed data

diff\_data <- diff(log(data), lag = 12)

plot(ts(diff\_data))

abline(lm(diff\_data ~ as.numeric(1:length(diff\_data))))

# apparent downward trend

var(diff\_data) # 0.004036354

# lag-1 difference on previously differenced data

diff2\_data <- diff(diff\_data, lag = 1)

plot(ts(diff2\_data))

abline(lm(diff2\_data ~ as.numeric(1:length(diff2\_data))))

# horizontal line -> no trend

var(diff2\_data) # 0.002028451

# ACF and PACF

par(mfrow=c(2,1))

acf(diff2\_data)

pacf(diff2\_data)

--- Sarah 11/19

\_\_\_ changes/additions as of 11/25

# look at ACF and PACF for seasonal and non-seasonal preliminary model building

par(mfrow = c(1,2))

acf(diff2\_data, lag.max = 11)

pacf(diff2\_data, lag.max = 11)

acf(diff2\_data, lag.max = 48)

pacf(diff2\_data, lag.max = 48)

# model building

# we know it will be SARIMA, we also know d = D = 1 and s = 12

# seasonal part: (1,1,0)

# non-seasonal part: lets look at MA(1) and AR(1)

# use AIC first

AICc <- numeric()

for(p in 0:3) {

for (q in 0:3) {

AICc <- c(AICc, sarima(diff2\_data, p, 1, q, 1,1,0, 12, details = FALSE)$AICc) } }

AICc <- matrix(AICc, nrow = 4, byrow = TRUE)

rownames(AICc) <- c("p=0", "p=1", "p=2", "p=3")

colnames(AICc) <- c("q=0", "q=1", "q=2", "q=3")

AICc

AICc <- data.frame(AICc)

aicc <- setNames(AICc, c("q=0", "q=1", "q=2", "q=3"))

# lowest model -> p = q =3: dismiss by parsimony

# next 3: p=q=1, p=3 q=1, p=0 q=2

# use BIC

BIC <- numeric()

for(p in 0:1) {

for (q in 0:1) {

BIC <- c(BIC, sarima(diff2\_data, p, 1, q, 1, 1, 0, 12, details = FALSE)$BIC) } }

BIC <- matrix(BIC, nrow = 2, byrow = TRUE)

rownames(BIC) <- c("p=0", "p=1")

colnames(BIC) <- c("q=0", "q=1")

BIC

BIC <- data.frame(BIC)

bic <- setNames(BIC, c("q=0", "q=1"))

# from BIC <- p = q = 1 is smallest, p = 0, q = 1 next smallest

# both criteria gave us the same models so we will test SARIMA (1,1,1,1,1,0) and SARIMA (0,,1,1,1,1,0)

# fit and estimation based on MLE method

# model 1: SARIMA (1,1,1,1,1,0)

sarima1 <- arima(diff2\_data, order = c(1,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

Sarima1

plot.roots(NULL, polyroot(c(1, -0.3499)), main = "roots for ar")

plot.roots(NULL, polyroot(c(1, -0.9999)), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6788)), main = "roots for sar")

# test diagnostics, looks good

resid1 <- residuals(sarima1)

hist(resid1)

qqnorm(resid1)

qqline(resid1)

shap1 <- matrix(c(shapiro.test(resid1)$statistic, shapiro.test(resid1)$p.value))

shap1

acf(resid1)

pacf(resid1)

# model 2:SARIMA (0,1,1,1,1,0) # GOOODDD MODEL!!!!

sarima2 <- arima(diff2\_data, order = c(0,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

sarima2

plot.roots(NULL, polyroot(c(1, -0.1900, -0.7802, -0.4018)), main = "roots for ar")

plot.roots(NULL, polyroot(c(1, -1.1043, 1.1043, -0.9999)), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -1.0000)), main = "roots for sma")

# test diagnostics for model2

resid2 <- residuals(sarima2)

hist(resid2)

qqnorm(resid2)

qqline(resid2)

shap2 <- matrix(c(shapiro.test(resid2)$statistic, shapiro.test(resid2)$p.value))

shap2

# 1st model better according to box test

# 1st model lower aic

# 2nd model better qq plot and histogram plot

# 2nd model less parameters

# BUT 2nd model has a unit root

Box.test(resid1, type = "Box-Pierce")

Box.test(resid1, type = "Box-Pierce")

Box.test(resid1, type = "Ljung-Box")

shapiro.test(resid1)

# all tests pass because p > 0.05

# lets go with 1st model

#model 3: sarima(3,1,1,1,1,0), gives same forecast as 1

sarima3 <- arima(diff2\_data, order = c(3,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

resid3 <- residuals(sarima3)

hist(resid3)

qqnorm(resid3)

qqline(resid3)

shap3 <- matrix(c(shapiro.test(resid3)$statistic, shapiro.test(resid3)$p.value))

shap3

acf(resid3)

pacf(resid3)

Box.test(resid3, type = "Ljung-Box")

Sarima3

#model 3 is the best, no unit roots!

# forecasting using model 3

# sarima.for(data, 12, 3,1,1,1,1,0,12)

trans <-log(data)

# based on final model with transformation

fit = arima(trans, order = c(3,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

pred1 <-predict(fit, n.ahead = 12)

u.pred1 <- pred1$pred + 2\*pred1$se

u.pred1

l.pred1 <- pred1$pred - 2\*pred1$se

ts.plot(trans, xlim = c(1, length(trans) + 12))

lines(u.pred1, col = "blue", lty = "dashed")

lines(l.pred1, col = "blue", lty = "dashed")

points((length(trans)+1):(length(trans)+12), pred1$pred, col = "red")

# based on original

pred2 <-exp(pred1$pred)

u.pred2 <- exp(u.pred1)

l.pred2 <- exp(l.pred1)

data3 <-ts(data)

ts.plot(data3, xlim = c(1, length(data3)+12))

lines(u.pred2, col = "blue", lty = "dashed")

lines(l.pred2, col = "blue", lty = "dashed")

points((length(trans)+1):(length(trans)+12), pred2, col = "red")

# zoom in on original data

ts.plot(data3, xlim = c(length(data3)-12, length(data3)+12), ylim= c(0, max(u.pred2)))

points((length(data3)+1):(length(data)+12), pred2, col = "red")

lines((length(data3)+1):(length(data)+12), u.pred2, col = "blue")

lines((length(data3)+1):(length(data)+12), l.pred2, col = "blue")

# TODO

# Test forecasting against holdout data

SARAH - UPDATED MODELING 11/29/18

```{r}

# fit and estimation based on MLE method

# model 1: SARIMA (1,1,1,1,1,0)

sarima1 <- arima(diff2\_data, order = c(1,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

sarima1

plot.roots(NULL, polyroot(c(1, -0.3499)), main = "roots for ar")

plot.roots(NULL, polyroot(c(1, -0.9999)), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6788)), main = "roots for sar")

# test diagnostics, looks good

resid1 <- residuals(sarima1)

hist(resid1)

qqnorm(resid1)

qqline(resid1)

shap1 <- matrix(c(shapiro.test(resid1)$statistic, shapiro.test(resid1)$p.value))

shap1

acf(resid1)

pacf(resid1)

#model 2: sarima(3,1,1,1,1,0), gives same forecast as 1

sarima2 <- arima(diff2\_data, order = c(3,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

sarima2

plot.roots(NULL, polyroot(c(1, -0.3634, -0.0921, -0.2173)), main = "roots for ar")

plot.roots(NULL, polyroot(c(1, -0.9703 )), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6818)), main = "roots for sar")

# test diagnostics for model2

resid2 <- residuals(sarima2)

hist(resid2)

qqnorm(resid2)

qqline(resid2)

shap2 <- matrix(c(shapiro.test(resid2)$statistic, shapiro.test(resid2)$p.value))

shap2

Box.test(resid2, type = "Box-Pierce")

Box.test(resid2, type = "Ljung-Box")

shapiro.test(resid2)

# all tests pass because p > 0.05

# model 3:SARIMA (0,1,2,1,1,0)

sarima3 <- arima(diff2\_data, order = c(0,1,2), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

sarima3

plot.roots(NULL, polyroot(c(1, -1.3714, 0.3856)), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6827)), main = "roots for sar")

resid3 <- residuals(sarima3)

hist(resid3)

qqnorm(resid3)

qqline(resid3)

shap3 <- matrix(c(shapiro.test(resid3)$statistic, shapiro.test(resid3)$p.value))

shap3

acf(resid3)

pacf(resid3)

Box.test(resid3, type = "Ljung-Box")