IVA503 VERİ YÖNETİMİ VE ANALİZİ

FİNAL SINAVI RAPORU

HALE S. KIZILDUMAN

202167008

İçindekiler

[1. GİRİŞ 3](#_Toc105951548)

[1.1 VERİ TANIMA 3](#_Toc105951549)

[2. BELİRLEME 4](#_Toc105951550)

[3. TAHMİN 6](#_Toc105951551)

[4. KONTROL 6](#_Toc105951552)

[4.1 OVERFIT 6](#_Toc105951553)

[4.2 ARTIKLARDA BILGI VAR MI? 9](#_Toc105951554)

[5. YAPISAL KIRILMA 10](#_Toc105951555)

[6. ALTERNATİF MODEL 11](#_Toc105951556)

[7. TAHMİN SONUÇLARI 14](#_Toc105951557)

[8. R SCRIPT 17](#_Toc105951558)

## GİRİŞ

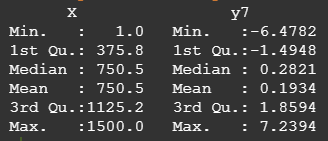
Kullanılan veri: y7

y7 verisine göre bir zaman serisi modeli tahmin edilmesi amaçlanmaktadır. Bu raporda model tahmin aşamaları, yapılan kabuller, testler ve sonuçları yeralmaktadır.

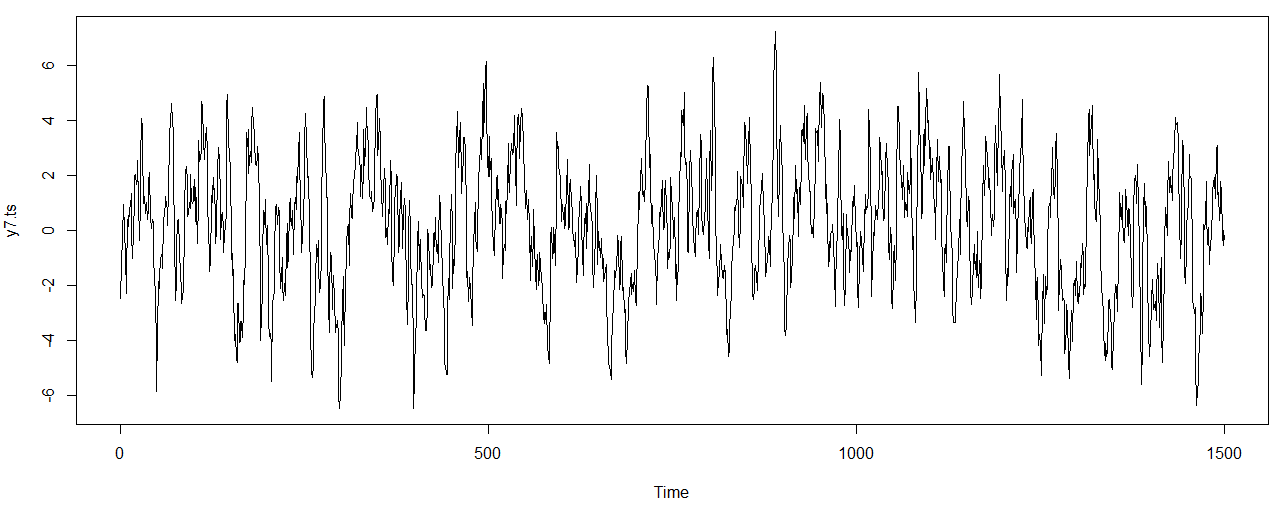
R kodu rapor sonuna kopyalanmıştır.

### VERİ TANIMA

Y7 verisinde 1500 adet gözlem bulunmaktadır. Bunlardan ilk 1450 si modeli geliştirmek için kullanılmış kalan 50’si ise test amaçlı kullanılmıştır. Aşağıda y7 verisine ait summary verilmiştir. Verile -6.4782 ile 7.2394 arasında değişmektedir ve 0 civarında oscillate etmektedir.



Aşağıda y7’nin zamana bağlı değişimi gösterilmiştir.



Zaman serisi yapısı olup olmadığı kontrol edilecek ve varsa zaman serisi modeli tahmin edilmeye çalışılacaktır.

## BELİRLEME

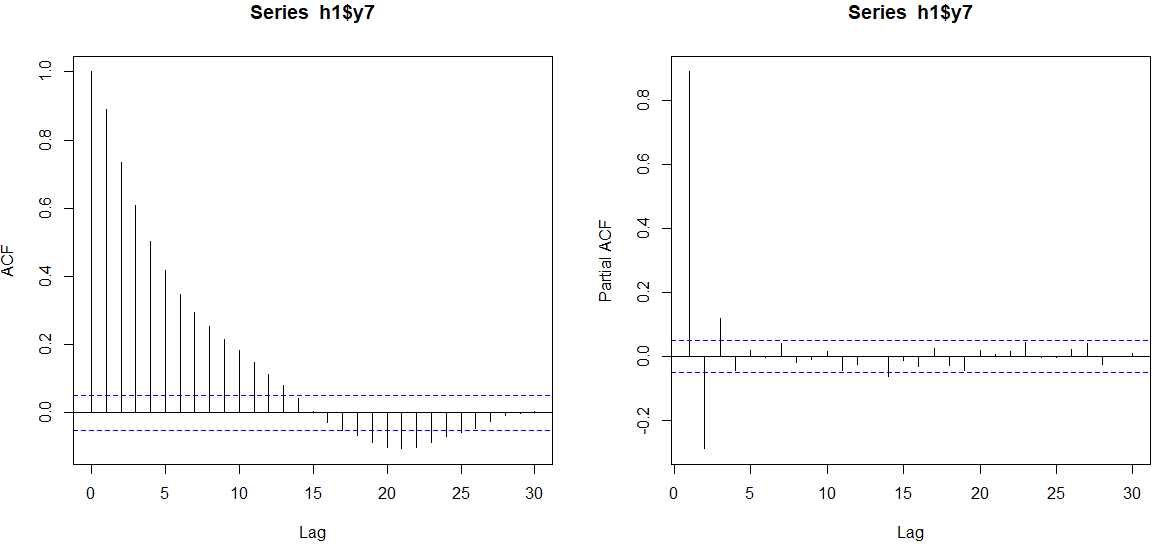
D: difference : stationary olup olmaması. D = 0 ise stationary.

Arima / arma : şöyle bir fark var. arma’da seri durağan arima’da durağan değil.

p, d, q değerleri belirlenirken correlogram ve bilgi kriterlerinden yararlanılacaktır.

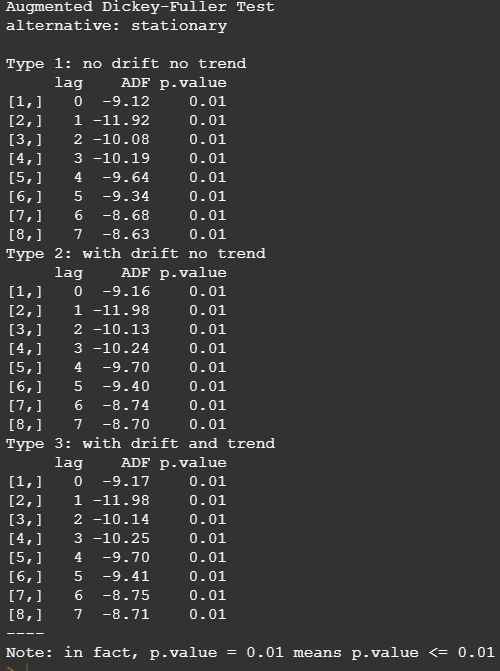
Otokorelasyon ve kısmı otokorelayon grafikleri aşağıda gösterilmiştir.

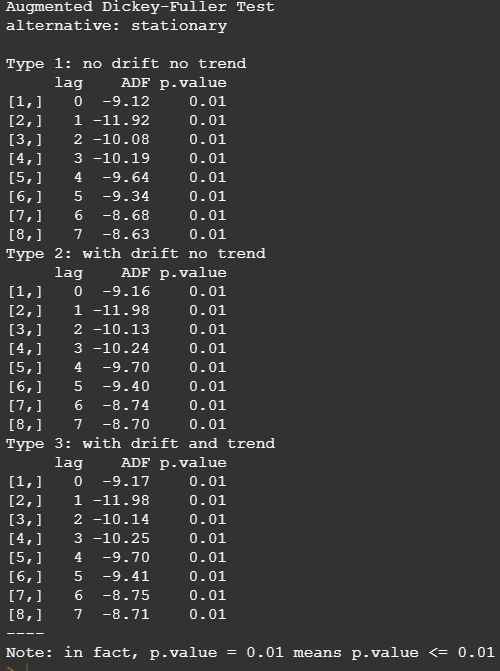
ACF, grafiği geometric olarak azalmakta, PACF cut off after lag 3. Bu durumda tahmin edilen model ARMA(3,0)



ACF grafiğine baktığımızda geçmiş şokların etkisinin azaldığı, geçmişin hızla unutulduğu görülmektedir. Buna göre durağan olduğu söylenebilir. ADF birim kök testi ile durağan olup olmadığı belirlenecektir.

ADF test sonuçları:



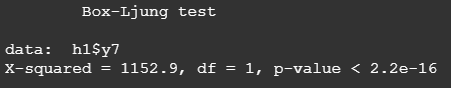


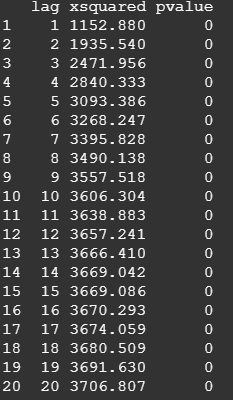
p value < 0. 01 for no drift no trend, with drift no trend, with drift and trend 🡺 Durağan

Bu durumda modeli durağan hale getirmeye gerek yok.

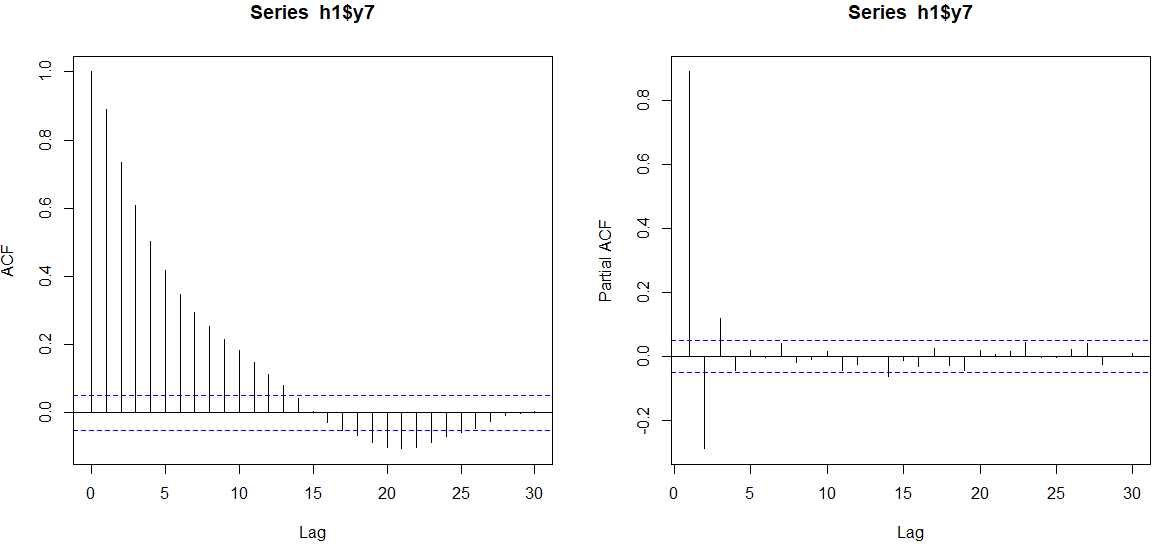
Zaman serisi olup olmadığı Ljung-Box Test ile kontrol edilmiştir.

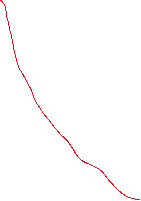
Ljung-Box Test: 20 lag için test sonuçları aşağıda verilmiştir.



 p\_value < 2.2e-16 🡺 alfa=0.05 anlamlılık düzeyine göre istatistiki olarak anlamlı, zaman serisi yapısı var diyebiliriz.

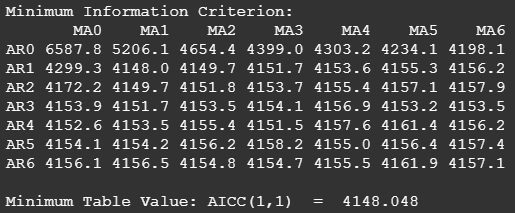
## TAHMİN





Grafiklere bakılarak ARMA(3,0) modeli tahmin edilmiştir.

AIC bilgi kriteri incelendiğinde ise ARMA(1,1) modeli’nin en düşük bilgi kriterine sahip olduğu görülmüştür. Bu durumda ARMA (3,0) ve ARMA(1,1) için diagnostic check yapılacaktır.

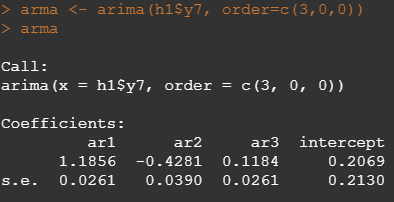


## KONTROL

### OVERFIT

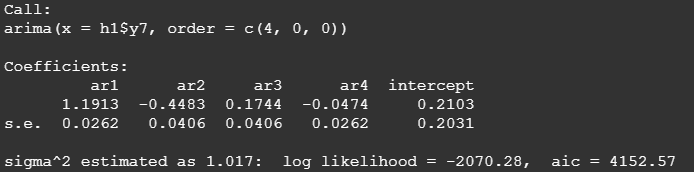
ARMA(3,0) modeline ARMA(4,0) ve ARMA(3,1) modelleriyle overfit edilerek ARMA(3,0) modelinin doğruluğu kontrol edilmiştir.

ARMA(3,0)model sonuçları:





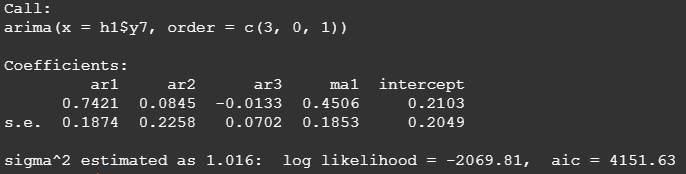
ARMA(4,0) model sonuçları:

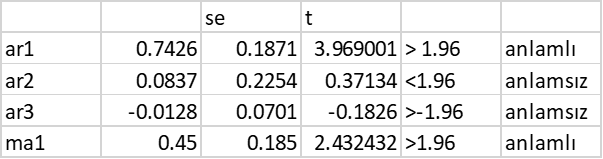




ar4 değişkeni non-rejection region’da kalmaktadır bu durumda H0: ar4=0 hipotezi reddedilemez. Yani istatistiki olarak 0 dan farksızdır. ar4 değişkeni istatistiki olarak anlamsız olduğu için ARMA(4,0) modeli reddedilmiştir.

ARMA(3,1) model sonuçları:

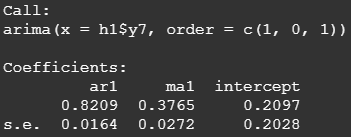


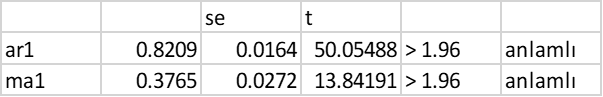


ar2, ar3 değişkenleri istatistiki olarak anlamsız çıkarken sonradan eklenen ma1 değişkeni istatistiki olarak anlamlı çıkmıştır. Bu durumda ARMA(3,1) modeli reddedilmiştir fakat ma1 değişkeni anlamlı çıkmıştır. Bu durumda AIC değeri daha düşük olan ARMA(1,1) modeli denenecektir.

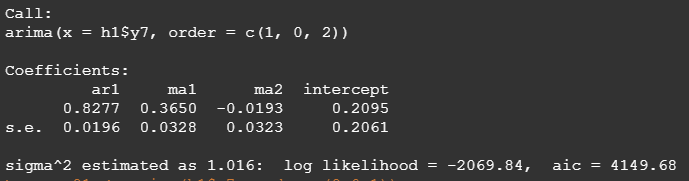
ARMA(1,1) modeline ARMA(1,2) ve ARMA(2,1) modelleriyle overfit edilerek ARMA(1,1) modelinin doğruluğu kontrol edilmiştir.

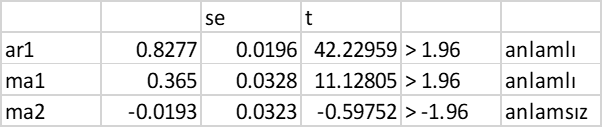
ARMA(1,1) model sonuçları:





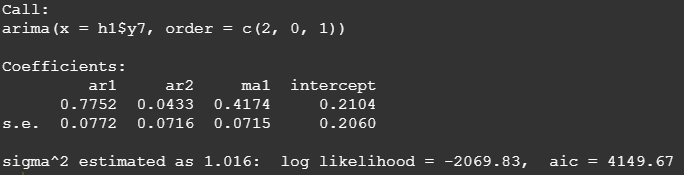
ARMA(1,2) model sonuçları:

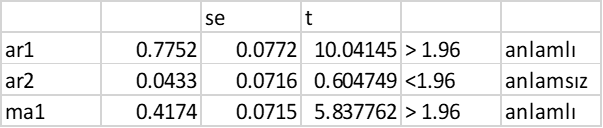




ma2 değişkeni istatistiki oalrak anlamsızdır, ARMA(1,2) modeli reddedilebilir.

ARMA(2,1) model sonuçları:





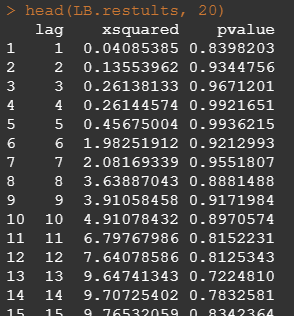
ar2 değişkeni istatistiki oalrak anlamsızdır, ARMA(2,1) modeli reddedilebilir.

ARMA(1,1) modleine overfit edilmşi ve over fit edilen değerler istatsitiki olarak anlamsız çıkmıştır. Bu durumda AIC değeri daha küçük olduğu için ARMA(1,1) modelinin kullanılabileceği düşünülmüşütür. Tahmin aşamasının ikinci aşamasına geçilebilir.

### ARTIKLARDA BILGI VAR MI?

Artıklarda bilgi olup olmadığı Q-stat testleri ile doğrulanabilir. Örneklem sayısı az olduğundan dolayı Ljung-Box testi kullanılmıştır.

Ljung-Box testi sonuçları:

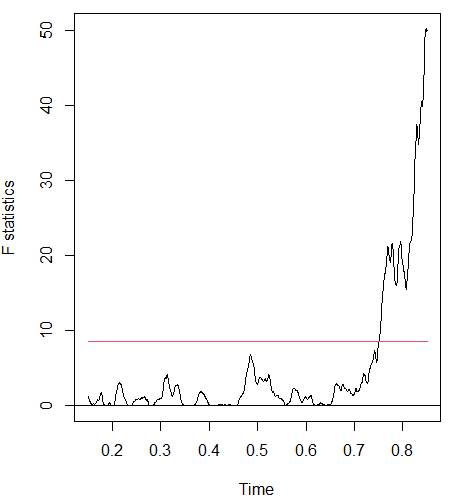
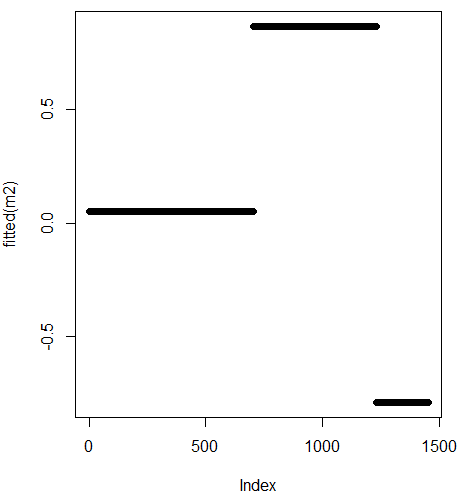


Zaman serisi yapısı yok. Artıklarda bilgi yok.

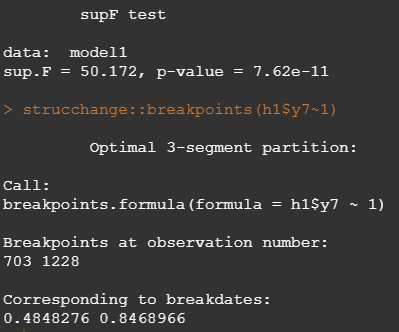
Yapılan testler sonucunda ARMA(1,1) modeli uygun görülmüştür. Fakat modelde structural break olup olmadığı control edilmelidir.

## YAPISAL KIRILMA

Modelde yapısal kırılmalar olup olmadığı chow test ile control edilmiştir. Aşağıdaki grafikten structural break olduğu görülebilir. 703. ve 1228. Noktalarda yapısal kırılmalar tespit edilmiştir.



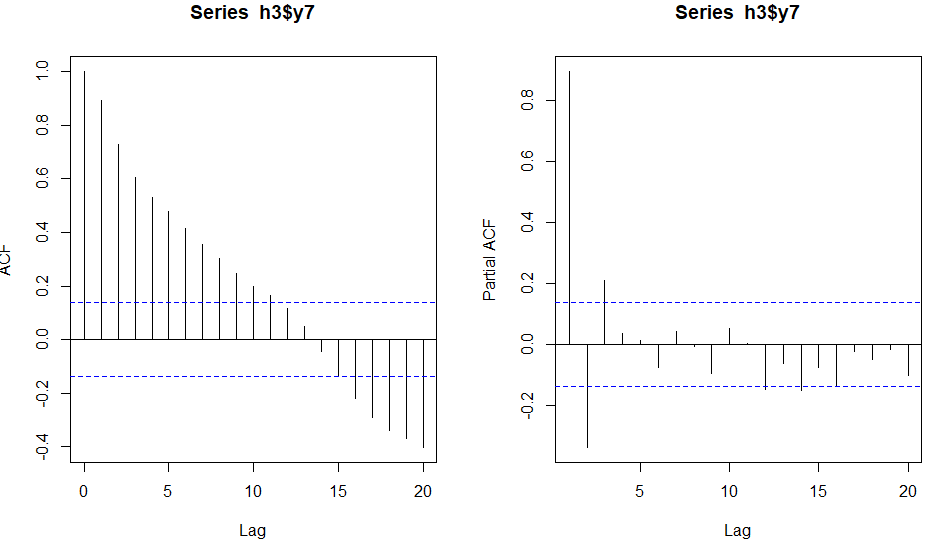
Chow test sonuçları:



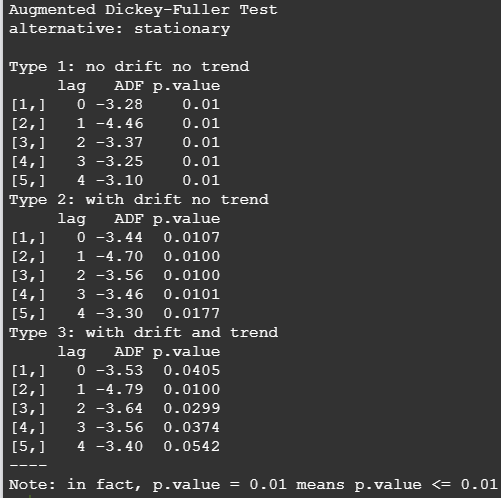
## ALTERNATİF MODEL

Son yapısal kırılma sonrası için yeni bir model oluşturularak predictionda alternatif model olarak kullanılacatır. structural break (gözlem 1228) sonrası [1250:1450] satırları modele dahil edilmiştir. 1250 seçilme sebebi tek bir structural break bakıldığında 1240 civarında yapısal kırılma göstermesidir. Yapısal kırılmadan sonraki kısımda kalındığından emin oolmak için 1250-1450 aralığına bakılmıştır.

Correlogram:

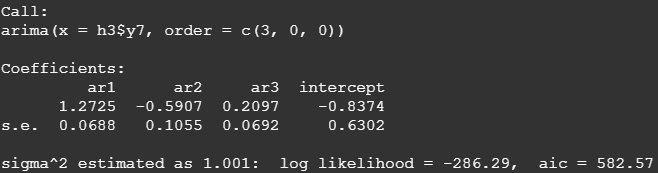


ADF Test sonuçları: Stationary



p\_value < 0.05 sıfır hipotezini reddediyorum yani durağan diyorum.

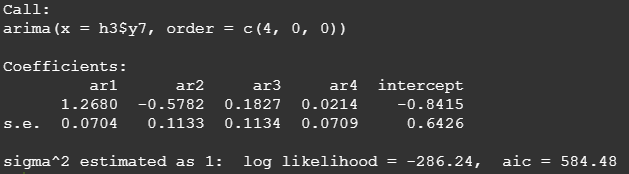
ARMA(3,0) modeli:





Overfit:

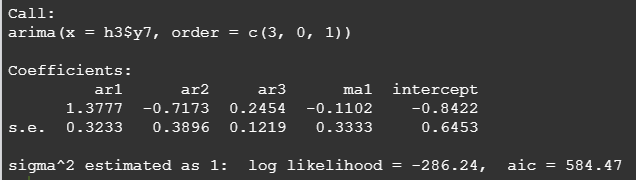
ARMA(4,0)





ar4 istatistiki olarak anlamsız. ARMA (3,0) modeli daha uygun görünüyor.

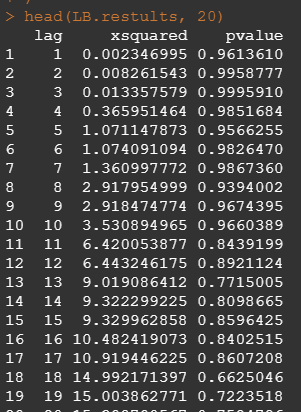
ARMA(3,1)





ma1 istatistiki olarak anlamsız, ARMA(3,0) modeli daha uygun görünüyor.

Artıklarda bilgi olup olmadığı control edildiğinde,



Zaman serisi yapısı yok. Artıklarda bilgi yok.

## TAHMİN SONUÇLARI

ARMA(1,1) ve ARMA(3,0) modeleri için ayrılan test datası tahmin edilmeye çalışılmış ve çıkan sonuçlar test datasıyla karşılaştırılmıştır. Ayrıca iki modelin hangisinin daha iyi sonuçlar verdiği MSE ve RMSE metrikleri kullanılarak da karşılştırılmıştır.

ARMA(1,1) MES değeri = 6,57 çıkarken ARMA(3,0) MSE değeri = 6,05 çıkmıştır buna göre hata oranı ARMA(3,0) modelinde daha düşüktür. Bu durumda ARMA(36,0) modelinin daha iyi tahmin yaptığını söyleyebiliriz.





Artış ve azalışları doğru tahmin etme oranına bakarak modelleri yazı tura atmakla karşılaştırırsak:

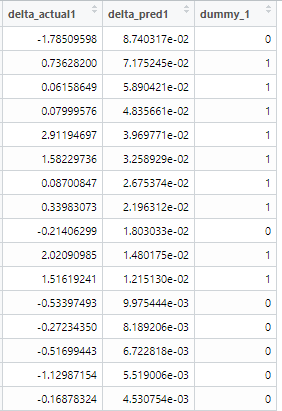
delta\_actual= y7t - y7t+1

delta pred = y7\_pred t – y7\_pred t+1

trend tahmini için dummy variable oluşturulmuştur. delta actual ve delta pred’in işarelerinin aynı oldu durumlarda dummy variable 1 e eşit olurken farklı olduğu durumlarda 0 a eşitlenmiştir. Bu sayede 1 olduğu durumların toplamı actual ve prediction değerlerinin aynı trended olduğunu gösterecektir.

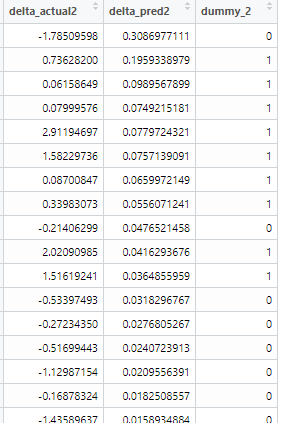
Elde edilen sonuçların bir kısmı aşağıda gösterilmiştir.

ARMA(1,1)



Dummy variabledaki toplam 1 çıkan veri sayısı 23 adettir. Aradaki farka bakıldığı için 50 yerine 49 veri kalmıştır. Doğruluk oranına bakıldığında 23/49 = 0.469 < 0.5 🡺 yazı tura atmaktan daha kötü bir model tahmin edilmiştir.

ARMA(3,0)



Dummy variabledaki toplam 1 çıkan veri sayısı 23 adettir. Aradaki farka bakıldığı için 50 yerine 49 veri kalmıştır. Doğruluk oranına bakıldığında 23/49 = 0.469 < 0.5 🡺 yazı tura atmaktan daha kötü bir model tahmin edilmiştir. Bu modelleri kullanmak yerine yazı tura atılması varlık fiyatı tahmininde daha iyi sonuç verir.

Her iki modelin de yazı tura karşısında aynı oranda başarısız olmakla beraber RMSE ölçütüne bakıldığında ARMA (3,0) ‘ın daha iyi sonuç verdiği söylenebilir.

## R SCRIPT

a <- read.csv("D:/IVA/IVA503 Nihat/Final/data\_final.csv")

View(a)

summary(a)

#Hale'nin verisi y7:

h <- a[, c(1, 8)]

View(h)

summary(h) #NA veri yok

h1 <- h[1:1450,]

h2 <- h[1451:1500,]

#grafik

par(mfrow=c(1,2))

plot(h)

plot(h1)

plot(h2)

y7.ts <- ts(h$y7)

plot.ts(y7.ts)

#Zaman serisine çevirelim

par(mfrow=c(1,1))

h1.ts <- ts(h1$y7)

plot.ts(h1.ts)

#ry7 <- 100\*diff(h1.ts, lag=1, diff=1)/lag(h1.ts, -1) #getirisi

#plot.ts(ry7)

#Correlogram

par(mfrow=c(1,2))

acf(h1$y7, lag.max=30) #geometrik azalıyor. geçmişi hızla unutuyor. geçmiş şoklar etkisin kaybediyor. durağan.

acf(h1$y7, lag.max=30, plot=F)#otokorelasyon fonk. değerlerini görelim

pacf(h1$y7, lag.max=30)#kismi otokorelason fonk.

pacf(h1$y7, lag.max=30, plot=F)#kismi otokorelason fonk. değerlerini görelim

#buraya ADF test ekleyip durağanlığını onayla!

library(aTSA)

adf.test(h1$y7, output=T) #birim kök yok, durağan.

#Ljung-Box Testi (Q-Stat)

Box.test(h1$y7, lag=1, "Ljung-Box") #zaman serisi yapısı var istatistiki olarak anlamlı p-value<2.2e-16

#20 lag için değerlere bakarsam

LB.test <- lapply(1:30, function(i) Box.test(h1$y7, type="Ljung-Box", i))

LB.restults <- data.frame(

lag=1:30,

xsquared=sapply(LB.test, "[[", "statistic"),

pvalue=sapply(LB.test, "[[", "p.value")

)

head(LB.restults, 30)

# sıfır hipotezini reddet. ilk 30 korelasyon katsayıları 0'a eşit.

#Estimate: ARMA(3,0)

arma30 <- arima(h1$y7, order=c(3,0,0))

arma30

#Diagnostic check

#1. Overfit

arma\_40 <- arima(h1$y7, order=c(4,0,0))

arma\_40 #Ho: ar4=0 ?

t = -0.0475/0.0263 #t= -1.806 > -1.96 , alfa=0.05 anlamlılık düzeyine göre non-rejection alanında kaldım yani istatistiki olarak ar4=0, ar(4) olamaz.

#modelim ar(3) geçerliliğini koruyor.

arma\_31 <- arima(h1$y7, order=c(3,0,1))

arma\_31

t= 0.45/0.185 # t= 2.43>1.96, rejection regionda, istatistiki olarak ma(1) 0 dan farklı

#istatistiki olarak anlamlı fakat ar2 ve ar3 anlamsız

#arma(1,1)'e göre tekrar overfit edeceğim.

arma\_11 <- arima(h1$y7, order=c(1,0,1))

arma\_11 #ar1 ve ma1 istatistiki olarak anlamlı

arma\_12 <- arima(h1$y7, order=c(1,0,2))

arma\_12 #Ho : ma2=0 ? ma2 istatistiki olarak 0 a eşit.

arma\_21 <- arima(h1$y7, order=c(2,0,1))

arma\_21 #Ho : ar2=0 ? ar2 istatistiki olarak 0 a eşit.

#arma(1,1) modeli geçerliliğini koruyor.

# 2. Artıklarda bilgi var mı?

arma <- arima(h1$y7, order=c(1,0,1))

arma

res <- residuals(arma)

plot.ts(res)

par(mfrow=c(1,2))

acf(res, lag.max=30)

pacf(res, lag.max=30)

Box.test(res, lag=1, "Ljung-Box") #zaman serisi yapisi yok diyor,

LB.test <- lapply(1:20, function(i) Box.test(res, type="Ljung-Box", i))

LB.restults <- data.frame(

lag=1:20,

xsquared=sapply(LB.test, "[[", "statistic"),

pvalue=sapply(LB.test, "[[", "p.value")

)

head(LB.restults, 20)

#Artıklarımın içinde zamandan kaynaklanan bir bilgi yok. O zaman modelim ARMA(1,1)

#ARMA(1,1) modelimi tahminlerde kullanabilirim.

#oto identification ile ARMA(1,1) AIC değeri en küçük mü kontrol edelim.

library(aTSA)

identify(h1$y7, p=6, q=6, nlag=6 , intercept = T, stat.test = F, output = T)

#ARMA (1,1) 'in doğru model olduğunu onaylamış oldum.

#fakat correlogram ile çıkan model örtüşmüyor. içimiz rahat etmedi structural break var mı bakalım

#Chow Test for structural break

library(strucchange)

model1=Fstats(h1$y7~1)

plot(model1)

sctest(model1) #yapısal kırılma var!

strucchange::breakpoints(h1$y7~1)

#breakpoints graph

bp\_ts <- breakpoints(h1$y7~1, breaks=2)

m1 <- lm(h1$y7~1)

coef(m1)

plot(fitted(m1), type="l")

m1.fac <- breakfactor(bp\_ts)

m2 <- lm(h1$y7~m1.fac-1)

coef(m2)

plot(fitted(m2))

#Bu durumda ARMA modelimi son breakpoint için yenilemek istiyorum.

#son breakpoint : 1228, 1230 den sonraki veriyi kullanacağım

h3 <- h1[1250:1450,]

View(h3)

#Correlogram

par(mfrow=c(1,2))

acf(h3$y7, lag.max=20) #geometrik azalıyor. geçmişi hızla unutuyor. geçmiş şoklar etkisin kaybediyor. durağan.

pacf(h3$y7, lag.max=20)#kismi otokorelason fonk.

#buraya ADF test ekleyip durağanlığını onayla!

adf.test(h3$y7, output=T) #birim kök yok, durağan.

#Ljung-Box Testi (Q-Stat)

Box.test(h3$y7, lag=1, "Ljung-Box") #zaman serisi yapısı var istatistiki olarak anlamlı p-value<2.2e-16

#20 lag için değerlere bakarsam

LB.test <- lapply(1:20, function(i) Box.test(h3$y7, type="Ljung-Box", i))

LB.restults <- data.frame(

lag=1:20,

xsquared=sapply(LB.test, "[[", "statistic"),

pvalue=sapply(LB.test, "[[", "p.value")

)

head(LB.restults, 20)

# sıfır hipotezini reddet. ilk 20 korelasyon katsayıları 0'a eşit.

#Estimate: ARMA(3,0)

arma2 <- arima(h3$y7, order=c(3,0,0))

arma2

#Diagnostic check

#1. Overfit

arma\_40 <- arima(h3$y7, order=c(4,0,0))

arma\_40 #Ho: ar4=0 ?

t = 0.0001/0.0671 #t= 0.0014 < 1.96 , alfa=0.05 anlamlılık düzeyine göre non-rejection alanında kaldım yani istatistiki olarak ar4=0, ar(4) olamaz.

#modelim ar(3) geçerliliğini koruyor.

arma\_31 <- arima(h3$y7, order=c(3,0,1))

arma\_31

t= -0.0989/0.3281 # t= -0.30>-1.96, non-rejection regionda, istatistiki olarak ma(1) 0 dan farksız, anlamsız

# 2. Artıklarda bilgi var mı?

arma2 <- arima(h3$y7, order=c(3,0,0))

arma2

res <- residuals(arma2)

plot.ts(res)

par(mfrow=c(1,2))

acf(res, lag.max=30)

pacf(res, lag.max=30)

Box.test(res, lag=1, "Ljung-Box") #zaman serisi yapisi yok diyor,

LB.test <- lapply(1:20, function(i) Box.test(res, type="Ljung-Box", i))

LB.restults <- data.frame(

lag=1:20,

xsquared=sapply(LB.test, "[[", "statistic"),

pvalue=sapply(LB.test, "[[", "p.value")

)

head(LB.restults, 20)

#Artıklarımın içinde zamandan kaynaklanan bir bilgi yok. O zaman modelim ARMA(3,0)

#Yine de iki model için de tahmin deneyeceğim.

arma

arma2

y7\_pred <- predict(arma, n.ahead = 50)

y7\_pred

h2\_new <- cbind(h2, y7\_pred)

View(h2\_new)

#fark fonsiyonu yt-yt+1

d\_y7 <- function(v){

b <- c()

i <- 1

while (i < length(v)){

b[i]<- v[i+1]-v[i]

i<- i+1

}

return(b)

}

delta\_actual1 <- d\_y7(h2\_new$y7)

data.frame(delta\_actual1)

delta\_pred1 <- d\_y7(h2\_new$pred)

data.frame(delta\_pred1)

dummyfortest <- function(v,k){

d<- c()

i <- 1

while (i < length(v)+1){

if((v[i]<0&&k[i]<0)||(v[i]>0&&k[i]>0)){

d[i] = 1

i<- i+1

}else{

d[i] = 0

i<- i+1

}

}

return(d)

}

dummy\_1 <- dummyfortest(delta\_actual1,delta\_pred1)

data.frame(dummy\_1)

result11 <-cbind(delta\_actual1, delta\_pred1, dummy\_1)

View(result11)

colSums(result11)

prob11= 23/49

MSE <- mean((h2\_new$y7 - h2\_new$pred)^2)

RMSE <- MSE^0.5

#MSPE <- mean((h2\_new$pred / h2\_new$y7)^2)

#MAPE <- mean(abs(h2\_new$y7 - h2\_new$pred)/abs(h2\_new$y7)) \* 100

y7\_pred2 <- predict(arma2, n.ahead = 50)

y7\_pred2

h2\_new2 <- cbind(h2, y7\_pred2)

View(h2\_new2)

MSE2 <- mean((h2\_new2$y7 - h2\_new2$pred)^2)

RMSE2 <- MSE2^0.5

delta\_actual2 <- d\_y7(h2\_new2$y7)

data.frame(delta\_actual2)

delta\_pred2 <- d\_y7(h2\_new2$pred)

data.frame(delta\_pred2)

dummy\_2 <- dummyfortest(delta\_actual2,delta\_pred2)

data.frame(dummy\_2)

result30 <-cbind(delta\_actual2, delta\_pred2, dummy\_2)

View(result30)

colSums(result30)

prob30 = 23/49