Task 1

Oleg Chaban

# 1 этап

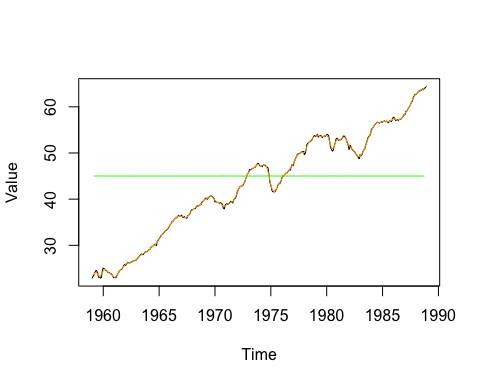
Считываем данных из training.csv.

dates\_rows<-read.csv("training.csv", header = TRUE, row.names = 1)  
dates\_rows.ts<-ts(dates\_rows, start=c(1959,1), frequency=12)

Построим скользящие статистики - скользящее среднее и стандартное отклонение.

dates\_rows\_rl<-rollmean(dates\_rows.ts, 5)  
dates\_rows\_d<-((dates\_rows\_rl-dates\_rows.ts)^2/359)^(1/2)

Строим график, на котором отрисованны сам ряд и его скользящие статистики. Скользящее среднее окрашено оранжевым цветом, стандартное отклонение - зеленым.



Изучив данные графики, можно сказать, что что ряд является нестационарным.

Проведем тест Дики-Фуллера.

adf.test(dates\_rows.ts)

##   
## Augmented Dickey-Fuller Test  
##   
## data: dates\_rows.ts  
## Dickey-Fuller = -3.2505, Lag order = 7, p-value = 0.07962  
## alternative hypothesis: stationary

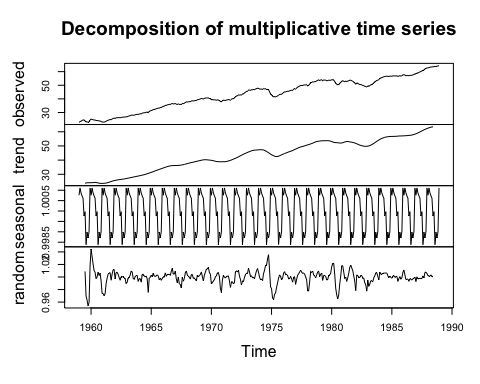
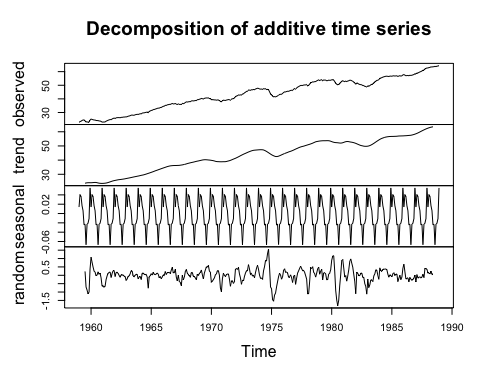
Так как корень один, то можно считать, что ряд нестационарен. В результате, после проведения двух тестов, можно сказать, что исходный ряд является нестационарным.

# 2 этап

Расскладываем временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями.

dates\_rows\_add<-decompose(dates\_rows.ts,type = "additive")  
dates\_rows\_mul<-decompose(dates\_rows.ts,type = "multiplicative")

Построим графики.



В обеих моделях видно, что тренд не стационарен, но сезональность и остаток - стационарны в широком смысле, т.к. статистические характеристики не изменяются с течением времени. Чтобы удостовериться в этом, проведем тест Дики-Фуллера.

adf.test(na.remove(dates\_rows\_add$seasonal), k = 0)

##   
## Augmented Dickey-Fuller Test  
##   
## data: na.remove(dates\_rows\_add$seasonal)  
## Dickey-Fuller = -10.169, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

adf.test(na.remove(dates\_rows\_add$trend), k = 0)

##   
## Augmented Dickey-Fuller Test  
##   
## data: na.remove(dates\_rows\_add$trend)  
## Dickey-Fuller = -0.6874, Lag order = 0, p-value = 0.9711  
## alternative hypothesis: stationary

adf.test(na.remove(dates\_rows\_add$random), k = 0)

##   
## Augmented Dickey-Fuller Test  
##   
## data: na.remove(dates\_rows\_add$random)  
## Dickey-Fuller = -7.3949, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

adf.test(na.remove(dates\_rows\_mul$seasonal), k = 0)

##   
## Augmented Dickey-Fuller Test  
##   
## data: na.remove(dates\_rows\_mul$seasonal)  
## Dickey-Fuller = -9.192, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

adf.test(na.remove(dates\_rows\_mul$trend), k = 0)

##   
## Augmented Dickey-Fuller Test  
##   
## data: na.remove(dates\_rows\_mul$trend)  
## Dickey-Fuller = -0.6874, Lag order = 0, p-value = 0.9711  
## alternative hypothesis: stationary

adf.test(na.remove(dates\_rows\_mul$random), k = 0)

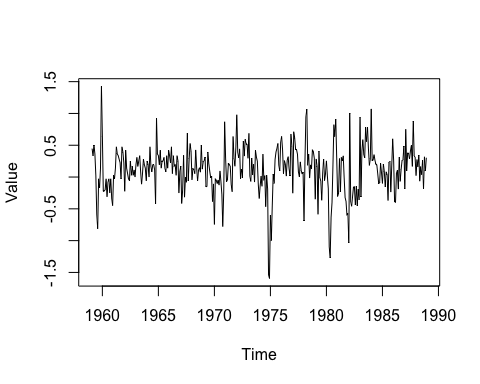
##   
## Augmented Dickey-Fuller Test  
##   
## data: na.remove(dates\_rows\_mul$random)  
## Dickey-Fuller = -7.5842, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

Наше предположение верно. В обеих моделях значение параметра p-value у тренда большое, следовательно, ряд не является стационарным. А для остатка и сезональности он мал, что подтверждает их стационарность.

# 3 этап

Проанализируем последовательно разности временного ряда, для того, чтобы определить порядок интегрированности этого ряда.

diff\_n<-diff(dates\_rows.ts, differences=1)  
plot.ts(diff\_n)



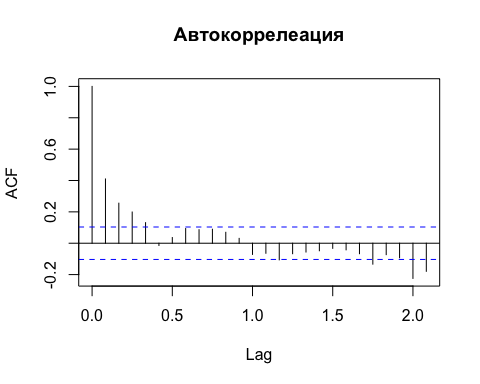
adf.test(diff\_n, k = 0)

##   
## Augmented Dickey-Fuller Test  
##   
## data: diff\_n  
## Dickey-Fuller = -12.2, Lag order = 0, p-value = 0.01  
## alternative hypothesis: stationary

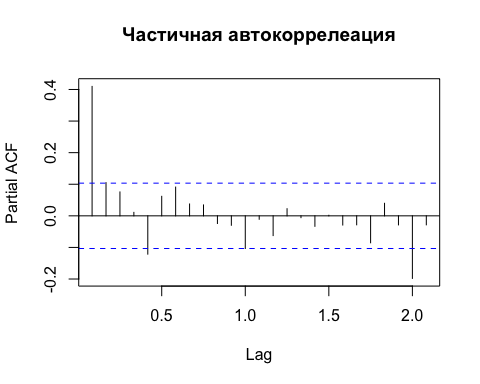
Из результатов теста Дики-Фуллера можно сделать вывод, что порядок интегрированности равен 1, т.к. первая разность исходного ряда - стационарна.

Построим функции автокоррелеации и частичной автокоррелеации

acf(diff\_n,main="Автокоррелеация")



pacf(diff\_n,main="Частичная автокоррелеация")



С помощью pacf и acf находим параметры p и q для построения моделей ARIMA. Параметр d равен порядку интегрированного ряда. Потроим несколько моделей ARIMA с параметрами, выбранными таким образом, что: 0<=p<=2, d=1, 0<=q<=4.

model01<-Arima(dates\_rows.ts,order=c(2,1,0))  
model02<-Arima(dates\_rows.ts,order=c(0,1,4))  
model03<-Arima(dates\_rows.ts,order=c(2,1,4))  
model04<-Arima(dates\_rows.ts,order=c(1,1,1))

Считываем данные из testing.csv.

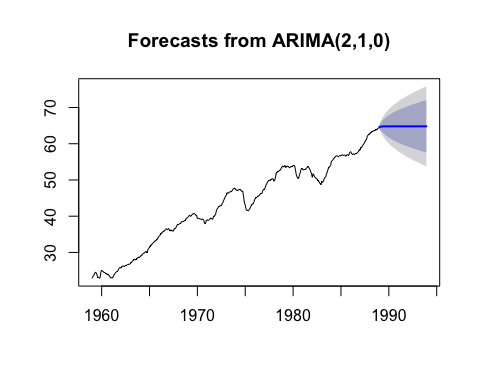
test<-read.csv("testing.csv",header=TRUE, row.names = 1)  
test.ts<-ts(test, start=c(1989,1), frequency=12)

Вычисляем для каждой модели прогноз с помощью forecast.Arima.Так же проверяем, что для каждого прогноза значение r2\_score достаточно близко к нулю. Визуализируем forecast.Arima для всех моделей.

frcst\_1<-forecast.Arima(model01,h=60)  
R2\_Score(frcst\_1$mean,test.ts)

## [1] -0.08292902

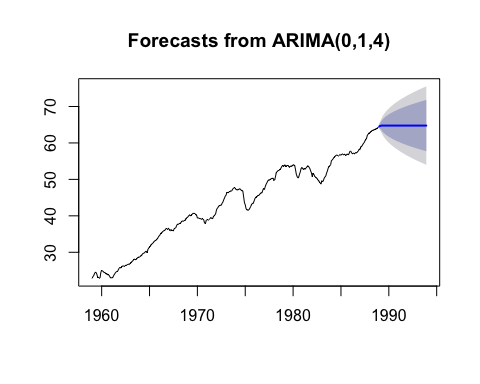
plot.forecast(frcst\_1)



frcst\_2<-forecast.Arima(model02,h=60)  
R2\_Score(frcst\_2$mean,test.ts)

## [1] -0.09951206

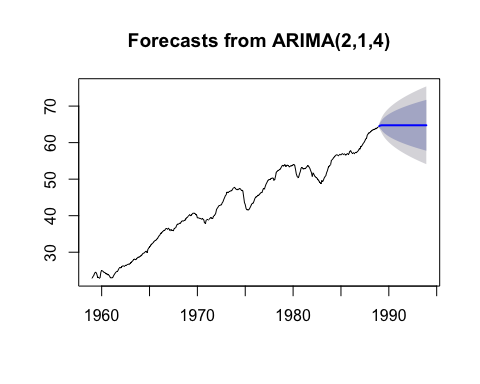
plot.forecast(frcst\_2)



frcst\_3<-forecast.Arima(model03,h=60)  
R2\_Score(frcst\_3$mean,test.ts)

## [1] -0.104477

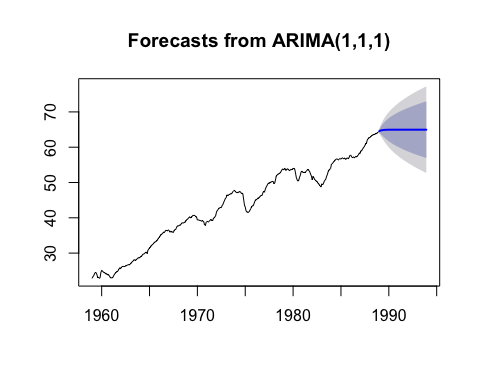
plot.forecast(frcst\_3)



frcst\_4<-forecast.Arima(model04,h=60)  
R2\_Score(frcst\_4$mean,test.ts)

## [1] -0.02447902

plot.forecast(frcst\_4)



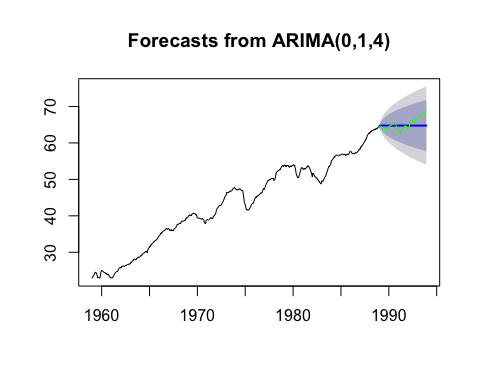
С помощью информационного критерия Акаике(AIC) из всех этих моделей находим наилучшую: выбираем модель с минимальным значением критерия AIC.

AIC(model01,model02,model03,model04)

## df AIC  
## model01 3 258.5099  
## model02 5 253.3399  
## model03 7 257.3009  
## model04 3 255.6413

Визуализируем forecast.Arima выбранной модели и получаем область прогноза дальнейших значений временного ряда. Строим график тестовой выборки.

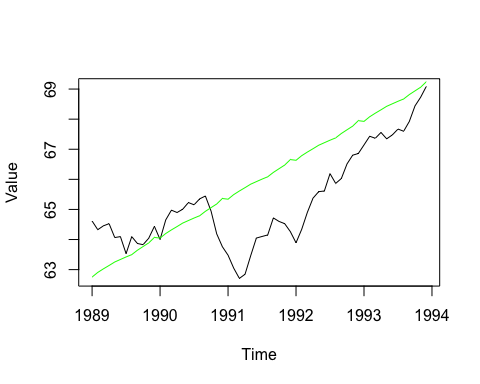
plot.forecast(frcst\_2)  
lines(test.ts,col='green')



Мы убедиись, что график лежит внутри спрогнозируемой области. Однако, r2-score для наших моделей - отрицательный. Отсюда можно сделать вывод, что прогнозирование на 60 месяцев с помощью модели ARIMA не подходит. Вопрос в том, что же нам подойдет?

Давайте рассмотрим линейную модель, разбив ряд на тренд и сезональность:

dates\_rows.lin <-tslm(dates\_rows.ts ~ trend + season)  
frcst\_lin <- forecast.lm(dates\_rows.lin, h = 60)  
plot(test.ts)  
lines(frcst\_lin$mean, col = 'green')



R2\_Score(frcst\_lin$mean,test.ts)

## [1] 0.1968743

В линейной модели получили, что r2\_score у нас больше нуля, отсюда можно сделать вывод, что данная модель является применимой для тестовых данных. Так же, хочется заметить, что 60 месяцев, предоставленных к прогнозу по заданию, являются неподходящим сроком для прогнозирования моделью ARIMA.