Lab 4

Time Series in R[1]

!!!Please read C6 and C7 before.

Building a model

1. The quarterly U.S. Gross National Product time series

```
library(astsa)   #install the package before
par(mfrow=c(2,1))
plot(gnp)
acf(gnp, lag.max=30)
```

The slowly decaying ACF indicates that differencing is necessary. #see C6-1

```
X=diff(gnp)
```

Plot the sample ACF and the PACF for X (acf2 function from library(astsa)).

The ACF suggests a possible MA(3); the PACF suggests a possible AR(2), AR(9), AR(16).

```
ar2<-arima(X ,order=c(2,0,0))

ar9<-arima(X ,order=c(9,0,0))

ar16<-arima(X ,order=c(16,0,0))


ma3<-arima(X , order=c(0,0,3))

#the goodness of fit for ar2

tsdiag(ar2)   #see III. Diagnostics in C6

#it's ok if the ACF (except for ACF(0) which is 1) in within the limits (the blue dotted lines)

#it's ok if the p-values for Ljung-Box statistics are above the blue dotted line -- we'll talk about p-values in a future course

rs <- ar2$residuals
stdres <- rs/sqrt(ar2$sigma2)  #standardized residuals

qqnorm(stdres, main = "Normal Q-Q Plot of Std Residuals")  #it's ok if the plot is approximately linear – see C6-1

ar2$aic   # or AIC(ar2)   The Akaike's criterion

        # BIC(ar2)     and/or use Bayesian Information Criterion
```

[1] Robert H. Shumway, David S. Stoffer. Time Series Analysis and Its Applications – with R examples, Springer 2017

d=1 is the order of differencing of the gnp time series => a possible model for gnp is ARIMA(2,1,0)

```
fore = predict(arima(gnp ,order=c(2,1,0)), n.ahead=15)   #we forecast 15 values

ts.plot(gnp, fore$pred, col=1:2)
```

Repeat the diagnostics steps for ar9, ar16 and ma3. Which model (ar2, ar9, ar16 or ma3) do you choose for gnp? (based on AIC)

2. The international airline passengers time series

```
par(mfrow=c(2,1))
plot(AirPassengers)
acf(AirPassengers, lag.max=30)
```

The increasing variability of the time series indicates towards the log transformation and the slowly decaying ACF indicates that differencing is necessary.

```
X=diff(log(AirPassengers))
plot(X)
acf2(X)
```

The ACF suggests a possible MA(12), MA(24), MA(36), …; the PACF suggests a possible AR(12).

Do the diagnostics for all these models for X and choose the best one according to the AIC.

As the order of differencing *d*=1 => possible models for log(AirPassengers) are ARIMA(*p*,1,*q*), where *p* and *q* are determined in the step above.

```
fore = predict(arima(log(AirPassengers) ,order=c(?,1,?)), n.ahead=15)

ts.plot(AirPassengers, exp(fore$pred), col=1:2)   #don't forget to inverse log, so
                    #that the forecasts are in the initial domain of the time series
```

Further comments

The ACF of X=diff(log(AirPassengers)) shows strong correlations at lags that are multiple of 12 – that's an indication of the presence of a seasonal component.

We deal with seasonality of period *d* by applying the lag-*d* difference operator (see Brockwell&Davis, p. 24):

$$\nabla_d X_t = X_t - X_{t-d} = (1-B^d)X_t$$

```
Y=diff(X,12)
acf2(Y)
```

Inspection of the ACF and the PACF of Y indicates as possible models for Y AR(12), MA(12), MA(23).

This analysis suggests that log(AirPassengers) can be modeled using SARIMA processes. Follow the steps in C6 slides 22-24 to determine the parameters of the best SARIMA model.

SARIMA models are implemented in R using arima function
arima(xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S)

[More on SARIMA models, in Brockwell&Davis, pages 310-316 and Shumway&Stoffer pages 145-154]

Density estimation, parameter estimation[2]

#see C6 slides 27-30

```
d<-read.csv("C:\\Users\\Marina\\Desktop\\curs Statistics for Data Science\\pima-indians-diabetes.csv", header=FALSE)

a<-d[,6]
a


a=a[a>0]
hist(a,breaks=20,freq=FALSE)  #freq=FALSE  - the plot has area=1 to have the same scale with the density

lines(density(a,kernel="epanechnikov"))  #try different kernels

#looking at the density, we shall assume that the data come from a Gamma
#distribution
```

#apply MM to estimate the parameters of Gamma -- C6 slide 33

```
m1=mean(a)
m2=var(a)+m1^2

alpha=m1^2/(m2-m1^2)
teta=(m2-m1^2)/m1

curve(dgamma(x,alpha,1/teta),0,70,add=TRUE)
```

[2] Norman Matloff, Probability and Statistics for Data Science, CRC Press 2019

```
#MLE for the parameters of the Negative Binomial distribution
#see C7 slides 8-9

X=rnbinom(100,6,0.5) #X is the number of failures until we get 6 successes
library(stats4)

f_log<-function(a)       #we have only one parameter to estimate k; p=0.5
     sum(-log(dnbinom(X,a,0.5)))

z<-mle(minuslogl=f_log,start=list(a=1))   #a=1 is an initialization in the
                                            optimization algorithm
z
```

Compare the real value of K (=6) with MLE for K.

Bootstrapping[3]

```
d<- read.csv(file.path('C:\\Users\\Marina\\Desktop\\curs Statistics for Data
Science\\loan_data.csv'))

d<- d[,3]
```

From the population d, take a random sample of size n (n=100; n=1000; …). Write a function that implements the bootstrap resampling of the mean (C7 slide 13) for a sample. Create the plots in C7 slide 16 for samples with different sizes *n* and observe how the standard errors for the sample mean is changing.

Confidence interval

```
 conf_int<-function(data,x){              #x is the level of confidence 90; 95

 q1<-((100-x)/2)/100

 q2<-1-q1

 c(quantile(data,q1), quantile(data,q2))

 }

a<-conf_int(boot_res,90)          #boot_res is returned by the function that
                                  #implements the bootstrap resampling of the mean
hist(boot_res,breaks=15)
abline(v=a[1],col="red")
abline(v=a[2],col="red")
```

[3] Peter Bruce, Andrew Bruce. Practical Statistics for Data Scientists, O'Reilly Media, 2017

Permutation tests[3]

We have two groups, denoted by groupA and groupB. The average in groupB is greater than the average in groupA. We want to know whether this difference is due to chance or is statistically significant.

```
web_times<- read.csv(file.path('C:\\Users\\Marina\\Desktop\\curs Statistics for
Data Science\\web_page_data.csv'))


groupA<- web_times[web_times['Page']=='Page A', 'Time']
groupB<- web_times[web_times['Page']=='Page B', 'Time']
initial_difference<- mean(groupB) - mean(groupA)


# Permutation test
perm_fun <- function(x, sizeA, sizeB)
{#x is data from both groups
    n <- sizeA+sizeB
    idx_b <- sample(1:n, sizeB)
    idx_a <- setdiff(1:n, idx_b)
    mean(x[idx_b]) - mean(x[idx_a])
    #return the difference of the means of the shuffled groups
}


perm_diffs <- rep(0, 1000)
x<-c(groupA,groupB)
for(i in 1:1000)
    perm_diffs[i] = perm_fun(x, length(groupA), length(groupB))

hist(perm_diffs, main='')
abline(v = initial_difference,col='red')
```

The red line lies well within the permuted values. This suggests that the observed difference bet ween groupA and groupB is due to chance, thus is not statistically significant.
Draw in blue the margins of the 95% confidence interval for perm_diffs (as in C7 slide 24).

What is the conclusion if instead of groupB we have groupB+0.5?

[3] Peter Bruce, Andrew Bruce. Practical Statistics for Data Scientists, O'Reilly Media, 2017