Pstat274_hw5_sandy

2023-11-09

1. Create a glossary of R-commands for time series. It should contain all commands that you learned so far in the labs, doing homework, and reviewing posted lecture slides. At the minimum, the glossary should include commands that allow

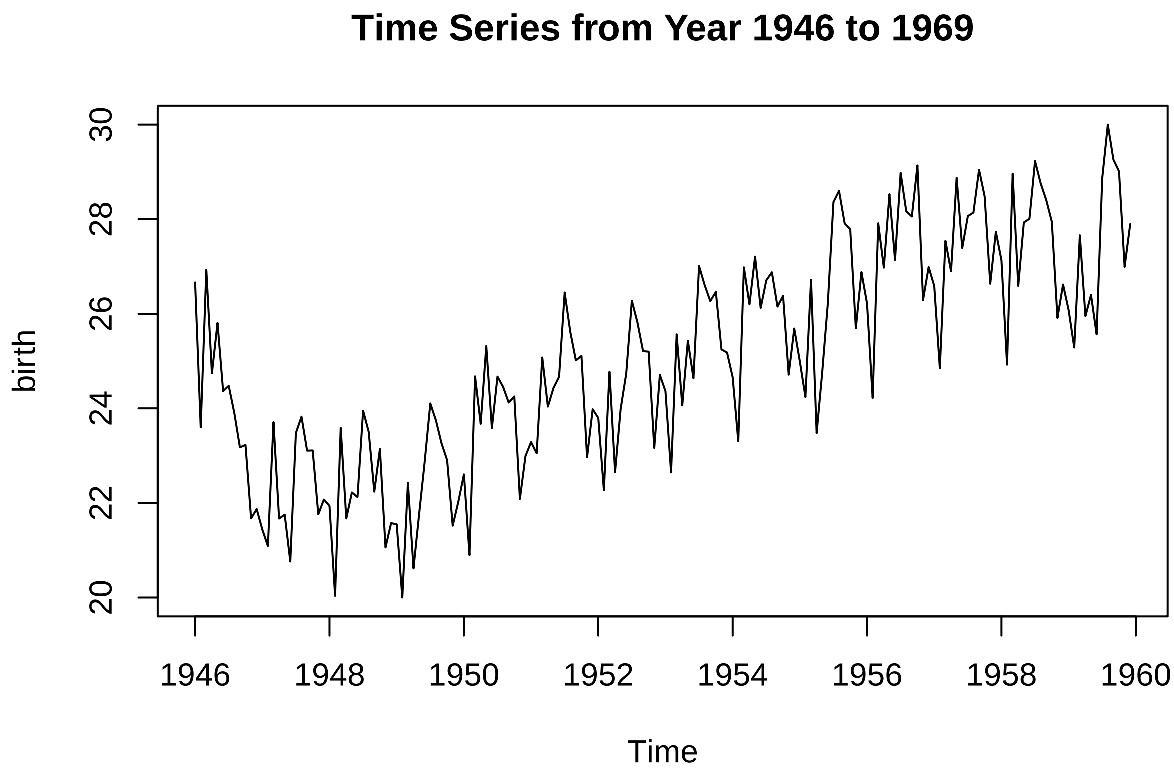
– define working directory;

```
getwd()
```

```
## [1] "/Users/zejiegao/Desktop/PSTAT 274/hw5"
```

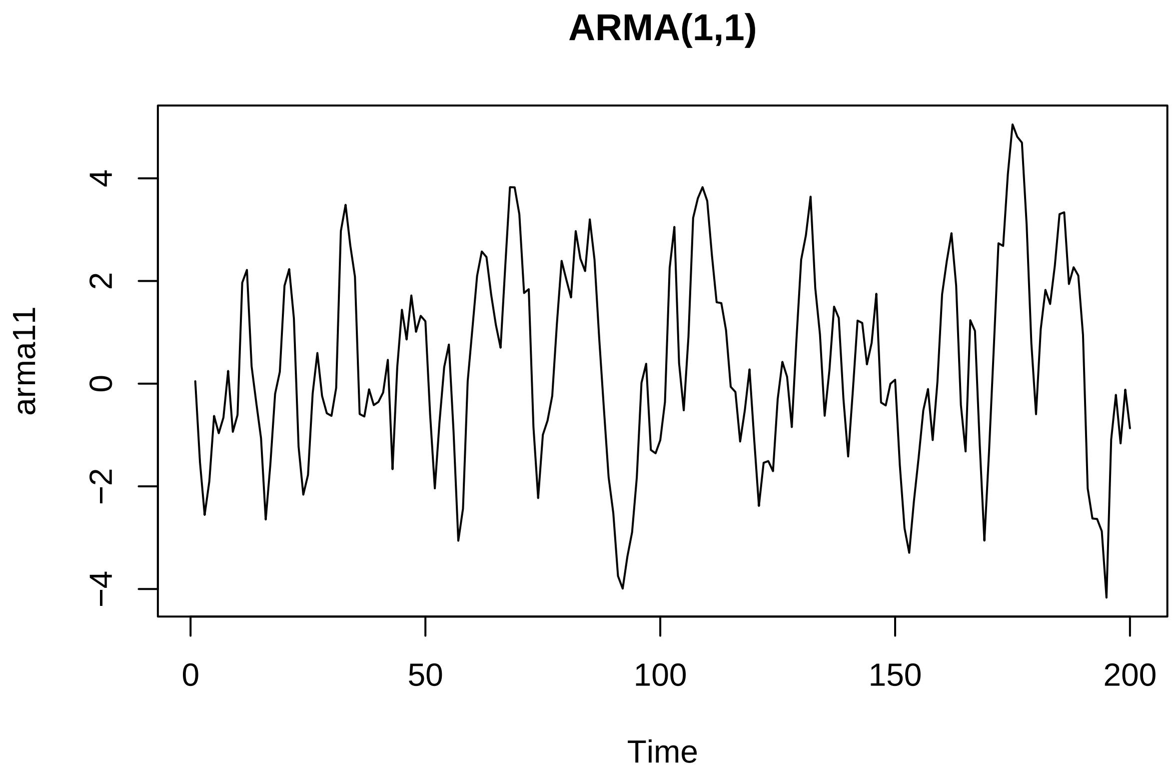
– read and plot data;

```
# read data
births <- scan("http://robjhyndman.com/tsdldata/data/nybirths.dat")
# convert into time series
births_ts = ts(births, start = c(1946,1), frequency = 12)
# plot data
plot(births_ts, main = "Time Series from Year 1946 to 1969", ylab = "birth")
```



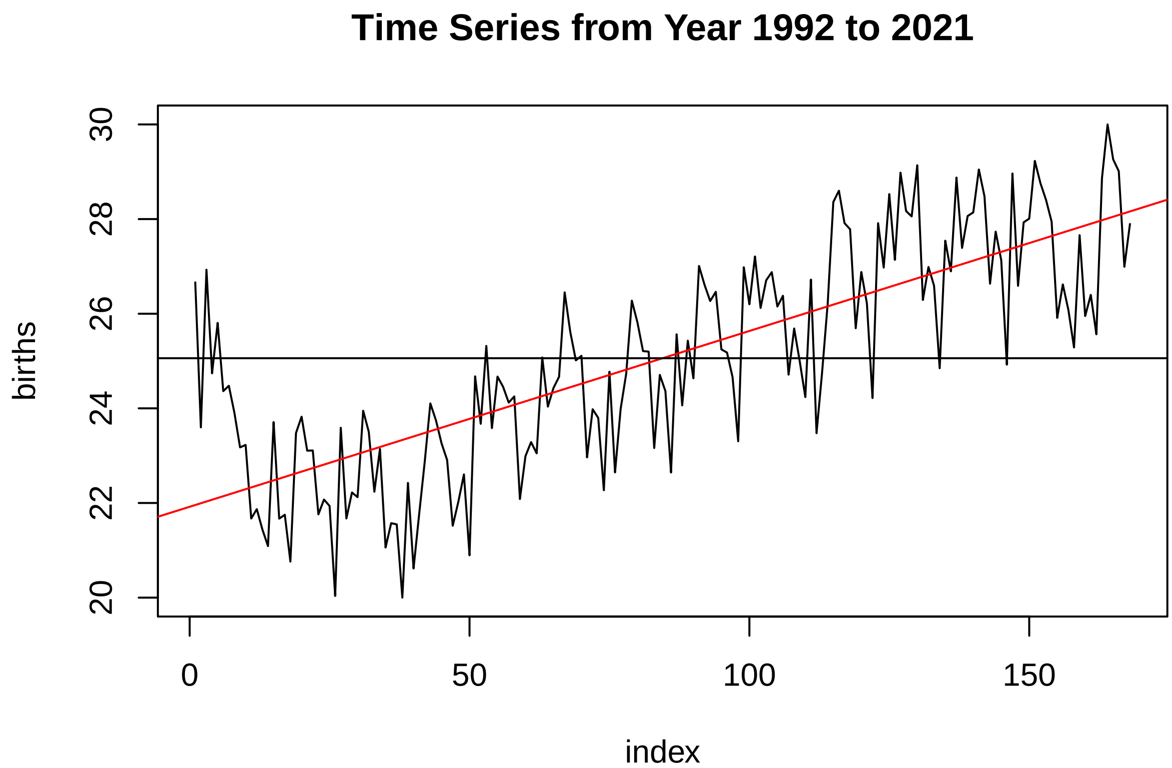
– simulate and plot ARMA models;

```
# Simulate model
set.seed(1014)
arma11 <- arima.sim(model = list(ar = c(0.6),ma = 0.6,sd = 1),n = 200)
# Plot ARMA model
plot(arma11, main = "ARMA(1,1)")
```



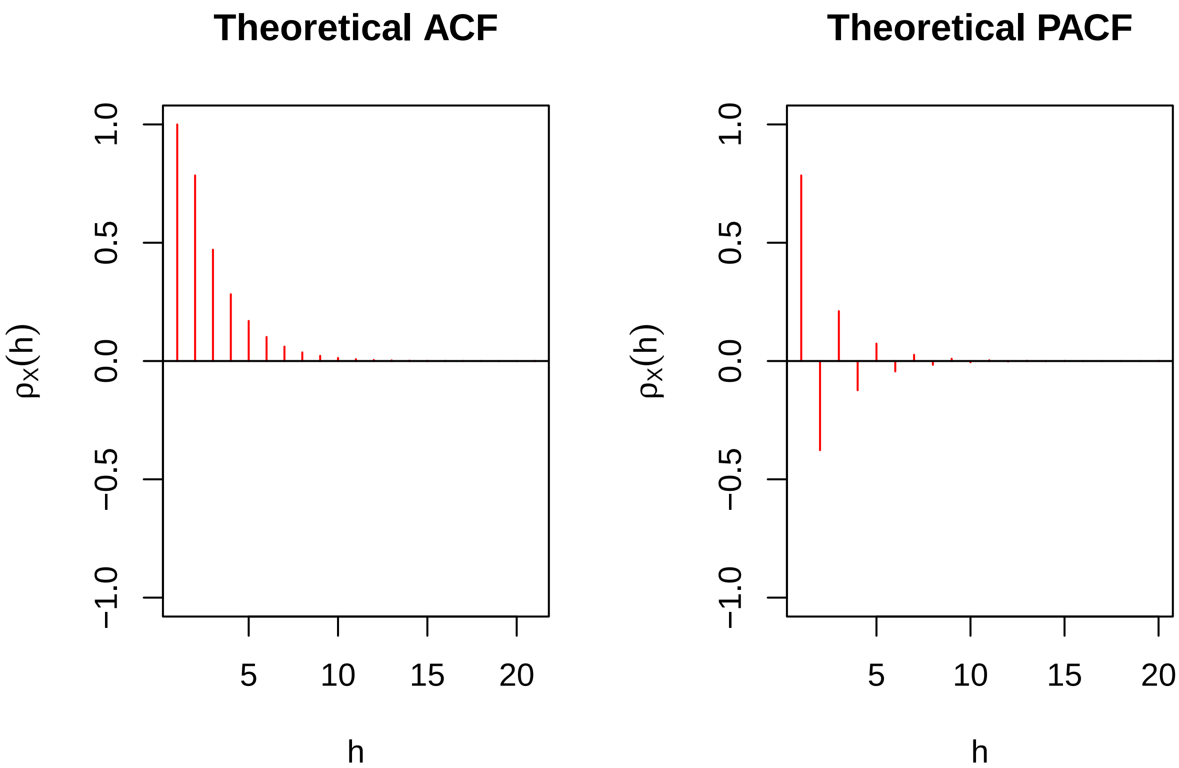
– add trend and mean line to the original time series plot;

```
plot(1:length(births),births, main =
"Time Series from Year 1992 to 2021", type = 'l',xlab='index')
index = 1: length(births)
trend <- lm(births ~ index)
# add trend to the original time series plot
abline(trend, col="red")
# add mean line to the original time series plot
abline(h=mean(births , col='green'))
```



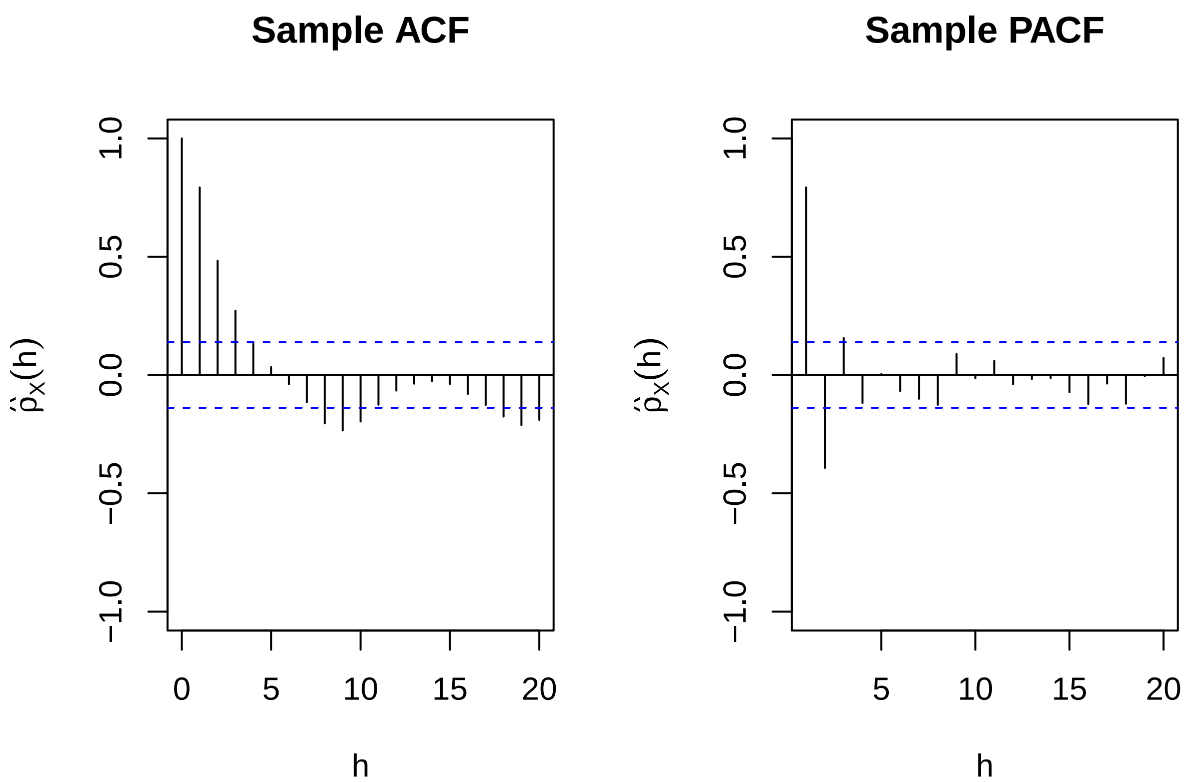
– calculate and plot theoretical acf/pacf for ARMA models;

```
# Theoretical ACF
theo_acf <- ARMAacf(ar = 0.6, ma = 0.6,lag.max = 20, pacf = FALSE) # Plot
op <- par(mfrow = c(1,2))
plot(theo_acf,type = "h",ylim = c(-1,1),
     main = "Theoretical ACF",
     col = "red",
     ylab = expression(rho[X](h)), xlab = "h")
abline(h = 0) # Add horizontal line

# Theoretical PACF
theo_pacf <- ARMAacf(ar = 0.6, ma = 0.6,lag.max = 20, pacf = TRUE) # Plot
plot(theo_pacf,type = "h",ylim = c(-1,1),
main = "Theoretical PACF",
col = "red",
ylab = expression(rho[X](h)), xlab = "h")
abline(h = 0) # Add horizontal line
```



– calculate and plot sample acf/pacf;

```
# Sample ACF
op <- par(mfrow = c(1,2))
acf(arma11,lag.max = 20, main = "Sample ACF",
ylim = c(-1,1),
xlab = "h",
ylab = expression(hat(rho)[X](h)))

# Sample PACF
pacf(arma11,lag.max = 20,
main = "Sample PACF",
ylim = c(-1,1),
xlab = "h",
ylab = expression(hat(rho)[X](h)))
```




```
par(op)
```

– check whether a particular model is causal/invertible (R commands to find and plot roots of polynomials)

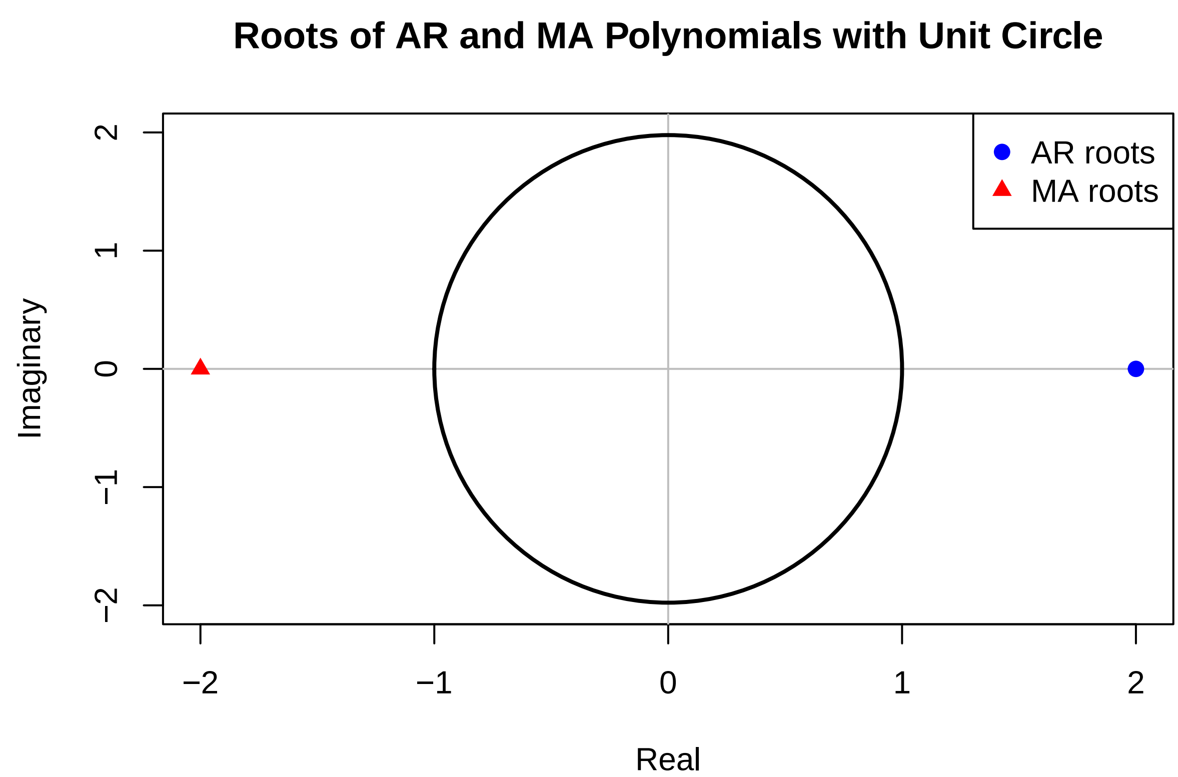
```
# Define AR and MA coefficients (including the constant term for the polynomials)
ar_coefficients <- c(1, -0.5)  # AR(1) coefficient (phi)
ma_coefficients <- c(1, 0.5)   # MA(1) coefficient (theta)

# Find the roots of the AR and MA polynomials
ar_roots <- polyroot(ar_coefficients)
ma_roots <- polyroot(ma_coefficients)

# Check causality (AR roots)
ar_causal <- all(Mod(ar_roots) > 1)
cat("The AR part is causal:", ar_causal, "\n")
```

```
## The AR part is causal: TRUE
```

```
# Check invertibility (MA roots)
ma_invertible <- all(Mod(ma_roots) > 1)
cat("The MA part is invertible:", ma_invertible, "\n")
```

```
## The MA part is invertible: TRUE
```

```r
# Plot the roots along with the unit circle
plot(0, 0, xlim = c(-2, 2), ylim = c(-2, 2), type = "n", xlab = "Real", ylab = "Imaginary", main = "Roo
abline(h = 0, v = 0, col = "gray")
# Draw the unit circle
symbols(0, 0, circles = 1, add = TRUE, inches = FALSE, lwd = 2)
# Plot AR roots
points(Re(ar_roots), Im(ar_roots), pch = 19, col = "blue", xlab = "Real Part", ylab = "Imaginary Part")
# Plot MA roots
points(Re(ma_roots), Im(ma_roots), pch = 17, col = "red", xlab = "Real Part", ylab = "Imaginary Part")
legend("topright", legend = c("AR roots", "MA roots"), col = c("blue", "red"), pch = c(19, 17))
```



– perform Box-Cox transforms;

```r
t = 1:length(births)
fit = lm(births~ t)
# bcTransform = boxcox(births ~ t,plotit = TRUE)
```

```r
# lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))];lambda
```

– perform differencing data at lags 1 and 12;

```r
x = births_ts
# Differencing data at lag 1
y_1 = diff(x,lag = 1)
# Differencing data at lag 12
y_12 = diff(y_1,lag = 12)
```

– perform Yule-Walker estimation and find standard deviations of the estimates.

```
# Perform Yule-Walker estimation to estimate parameters of an AR model;
ar_model_yw <- ar.yw(births_ts, order.max = NULL)
ar_model_yw$var.pred
```

```
## [1] 0.9162396
```

```
stv <- sqrt(ar_model_yw$var.pred)
```

– perform MLE and check AICC associated with the model.

```
library(forecast) # for ARIMA modeling
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# library(qpcR)     # for AICc function
mle_fit <- arima(births_ts, order = c(1, 0, 1), method = "ML")

# Calculate the AICc for the fitted model using the qpcR package
# aicc <- AICc(mle_fit)

# Print the AICc value
# print(aicc)
```

2. Choose a dataset that you will be interested to analyze for your class final project. URLs of time series libraries are posted on Canvas. Provide the following information about the project:

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(forecast)

# Read in the data from a CSV file
energy <- read.csv("/Users/zejiegao/Desktop/PSTAT 274/hw5/World Energy Overview.csv")
```

```r
names(energy)[names(energy) == "Total.Fossil.Fuels.Consumption"] <- "tffc"

# Now select the Date and tffc columns
fossil_c <- energy[, c("Date", "tffc")]

# Filter the data to include only rows after 1999-12-31
fossil_c <- fossil_c[fossil_c$Date > as.Date("1999-12-31"),]

# write to subset data to a new CVS file

write.csv(fossil_c, file = "/Users/zejiegao/Desktop/PSTAT 274/hw5/TFFC.csv", row.names = FALSE)

head(fossil_c)
```

```
##           Date     tffc
## 325 2000-01-31 7.692062
## 326 2000-02-29 7.222425
## 327 2000-03-31 7.071465
## 328 2000-04-30 6.486421
## 329 2000-05-31 6.789661
## 330 2000-06-30 6.783197
```

(a) Data set description: briefly describe the data set you plan to use in your project.

I plan to work on the world energy, focusing on the total fossil fuels consumption(TTFC),started from 2000 to 2022.

(b) Motivation and objectives: briefly explain why this data set is interesting or important. Provide a clear description of the problem you plan to address using this dataset (for example to forecast).

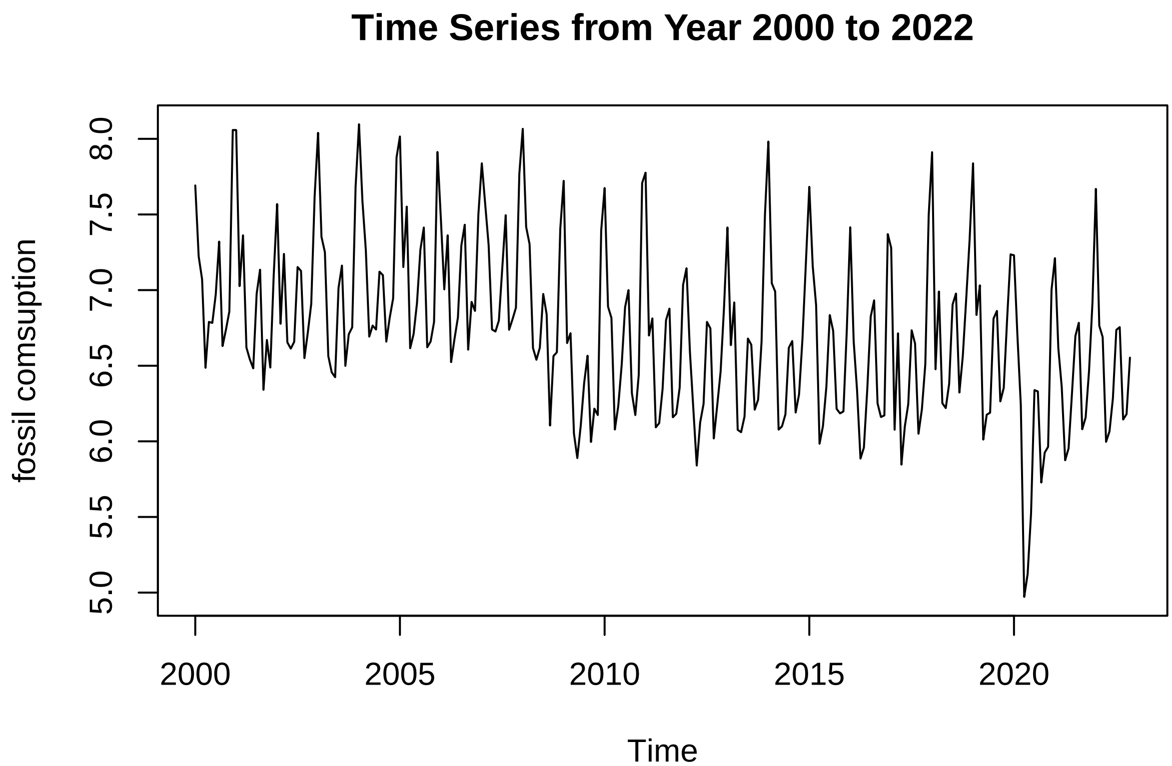
This dataset is particularly interesting and of great importance due to the central role of fossil fuels in the global energy landscape. Despite the growing emphasis on renewable energy, fossil fuels continue to be the predominant source of energy worldwide, deeply intertwined with economic growth, environmental policy, and geopolitical dynamics.

The primary objective of analyzing this dataset is to understand the trends in fossil fuel consumption over the last two decades, which can illuminate patterns related to economic cycles, the impact of energy policies, and the response to climate change initiatives.

By employing time series analysis techniques, I plan to forecast future fossil fuel consumption based on historical data. This forecast can help in assessing the trajectory of global energy consumption and the potential implications for both the economy and the environment.

(c) Plot and examine the main features of the graph, checking in particular whether there is (i) a trend; (ii) a seasonal component, (iii) any apparent sharp changes in behavior. Explain in detail.

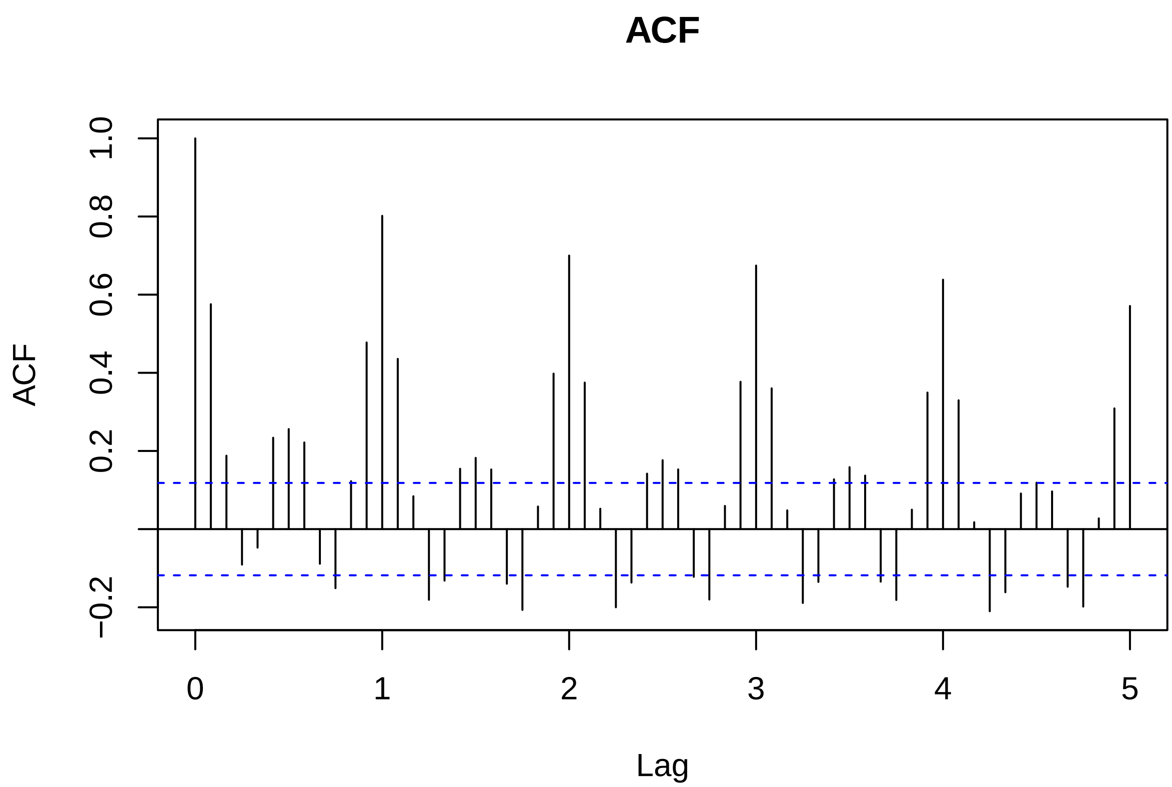
```r
fossil_ts = ts(fossil_c[,2], start=c(2000,1), frequency=12)
# plot data
plot(fossil_ts, main = "Time Series from Year 2000 to 2022", ylab = "fossil comsuption")
```



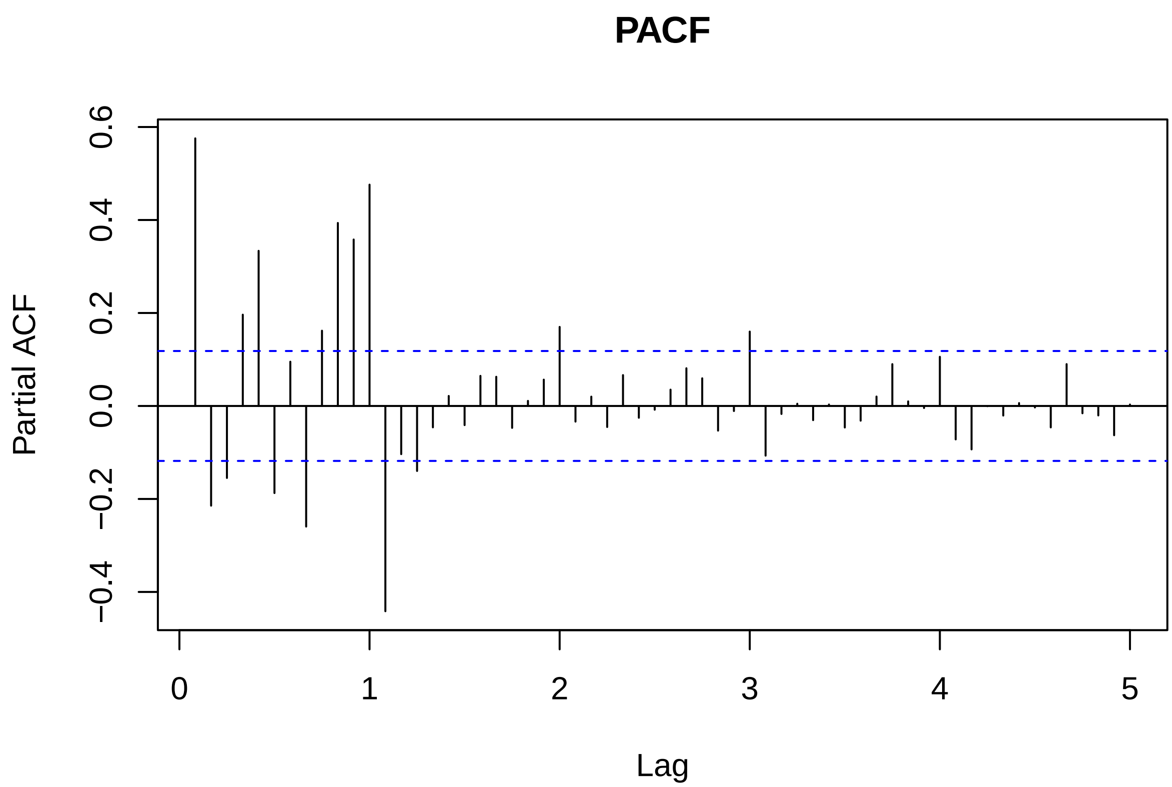
There is a slight decreasing trend in the data. There is a pattern that seems to repeat itself regularly each year, which suggests the presence of a seasonal component. There are several points in the plot where there are sharp spikes and drops in fossil fuel consumption.

(d) Use any necessary transformations to get stationary series. Give a detailed explanation to justify your choice of a particular procedure. If you have used transformation, justify why. If you have used differencing, what lag did you use? Why? Is your series stationary now?

```
# Plot acf and pacf of original data
acf(fossil_ts,lag.max = 60, main = "ACF")
```



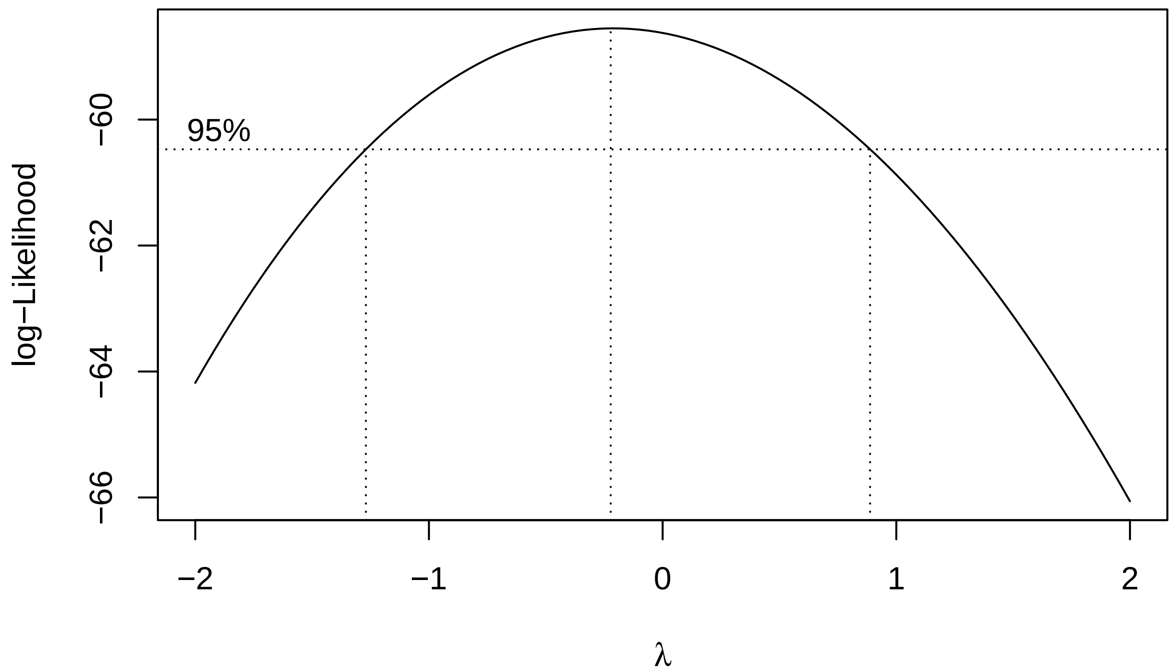

```
pacf(fossil_ts,lag.max = 60, main = "PACF")
```




```r
# Using Box-Cox transformation to stabalize variance
library(MASS)
```

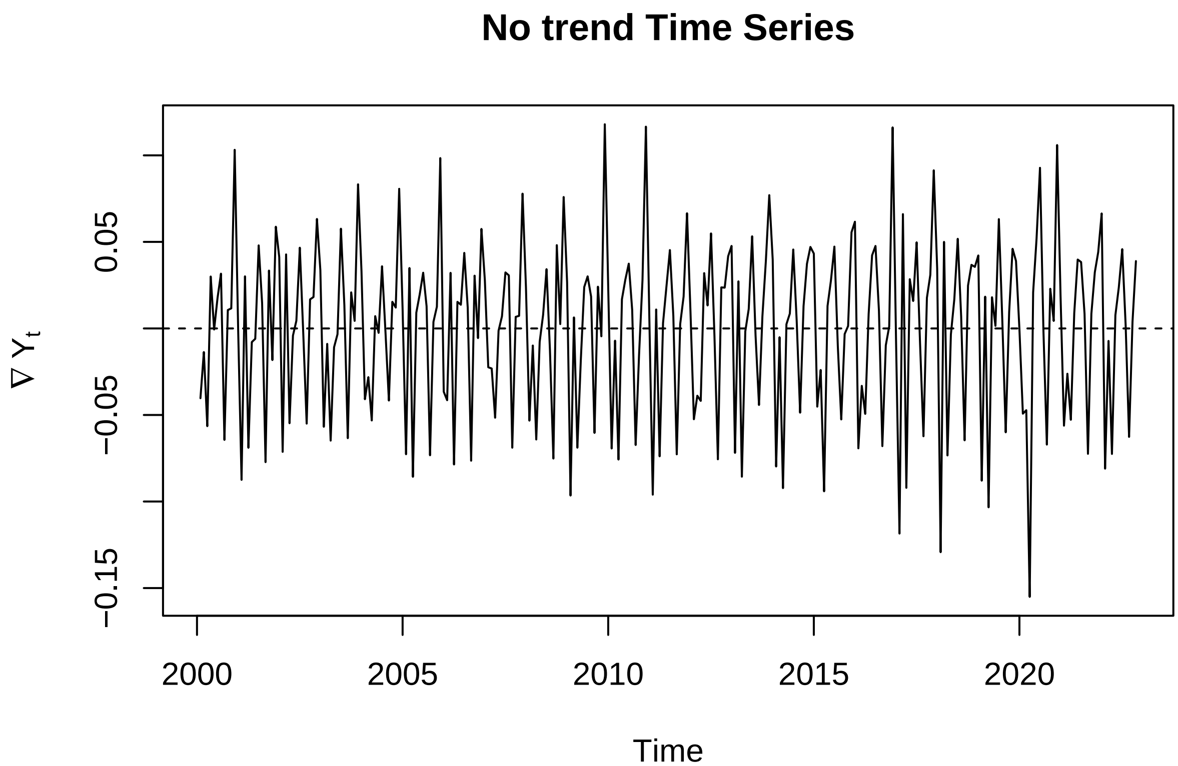
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
t = 1:length(fossil_ts)
fit = lm(fossil_ts ~ t)
bcTransform = boxcox(fossil_ts ~ t,plotit = TRUE)
```



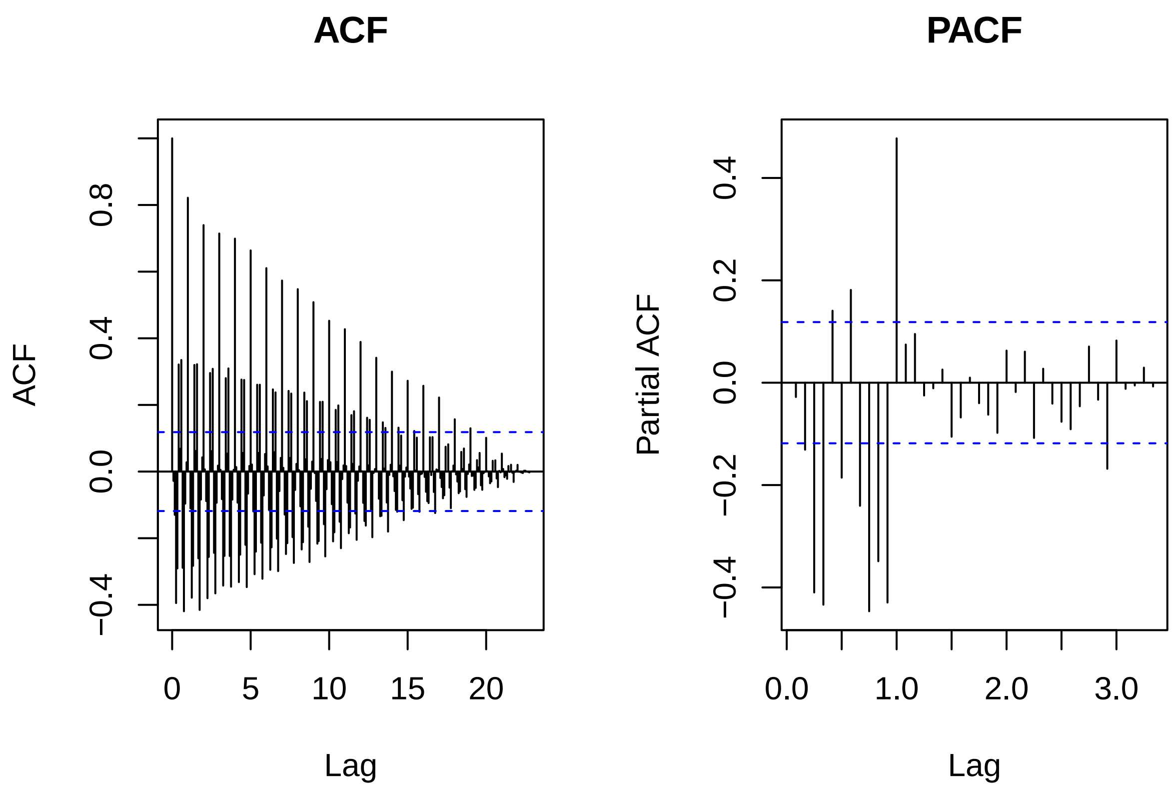

```
lambda = bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
fossil_ts = (1/lambda)*(fossil_ts^lambda-1)
```

```
# Differencing at lag 1 to eliminate trend
y1 = diff(fossil_ts, 1)
plot(y1,main = "No trend Time Series",ylab = expression(nabla~Y[t]))
abline(h = 0,lty = 2)
```



(e) Plot and analyze the ACF and PACF to preliminary identify your model(s): Plot ACF/PACF. What model(s) do they suggest? Explain your choice of p and q here.

```
op <- par(mfrow = c(1,2))
# Plot acf and pacf of new data
acf(y1,lag.max = 300, main = "ACF")
pacf(y1,lag.max = 40, main = "PACF")
```



The ACF plot shows a gradual decay, which suggests an autoregressive process, while the PACF plot has a significant spike at lag 1 and then cuts off. These patterns suggest an AR(1) model, so a preliminary model could be an ARIMA(1,0,0), indicating one autoregressive term. Consider we do have seasonal that could affect the model, this could be seasonal AR(1) with s = 12.