w271: Homework 5 (Due: Week 6)

Grace Lin

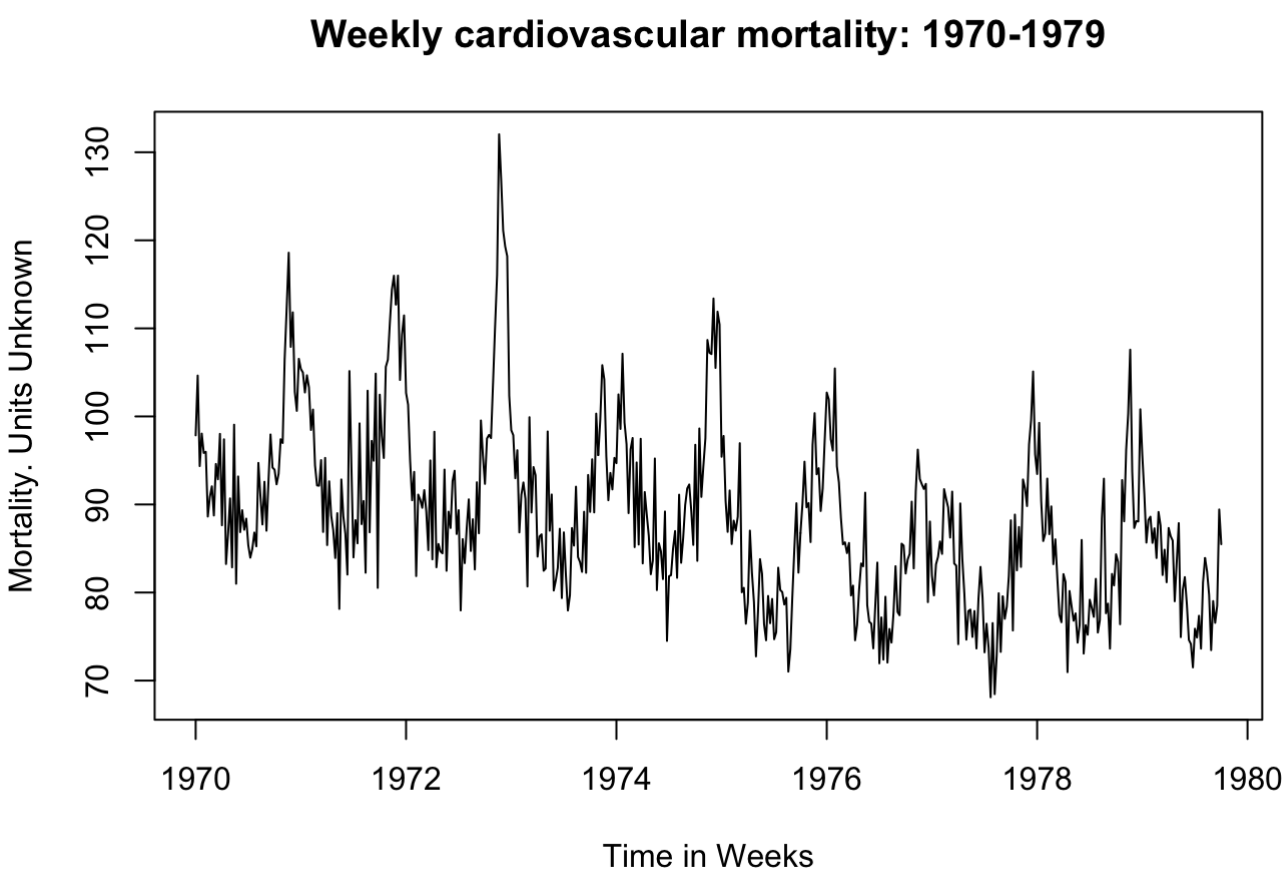
Due: Before the Live Session of Week 6

Instructions (Please Read it Carefully!):

- **Page limit of the pdf report: None, but please be reasonable**
- Page setup:
- Use the following font size, margin, and linespace:
    - fontsize=11pt
    - margin=1in
    - line_spacing=single
- Submission:
    - Homework needs to be completed individually; this is not a group project.
    - Each student submits his/her homework to the course github repo by the deadline; submission and revision made after the deadline will not be graded
    - Submit 2 files:
        1. A pdf file that details your answers. Include all the R codes used to produce the answers. *Please do not suppress the codes in your pdf file.*
        2. R markdown file used to produce the pdf file
    - Use the following file-naming convensation; fail to do so will receive 10% reduction in the grade:
        - StudentFirstNameLastName_HWNumber.fileExtension
        - For example, if the student's name is Kyle Cartman for homework 1, name your files as
            - KyleCartman_HW1.Rmd
            - KyleCartman_HW1.pdf
    - Although it sounds obvious, please write your name on page 1 of your pdf and Rmd files.
    - For statistical methods that we cover in this course, use only the R libraries and functions that are covered in this course. If you use libraries and functions for statistical modeling that we have not covered, you have to (1) provide an explanation of why such libraries and functions are used instead and (2) reference to the library documentation. **Lacking the explanation and reference to the documentation will result in a score of zero for the corresponding question.** For data wrangling and data visualization, you are free to use other libraries, such as dplyr, ggplot2, etc.
- For mathematical formulae, type them in your R markdown file. **Do not write them on a piece of paper, snap a photo, and either insert the image file or sumbit the image file separately. Doing so will receive a 0 for that whole question.**
- Students are expected to act with regards to UC Berkeley Academic Integrity.

```
rm(list = ls())
library(astsa)

plot(cmort, xlab= "Time in Weeks", ylab="Mortality. Units Unknown")
title(main="Weekly cardiovascular mortality: 1970-1979")
```



1. Conduct the EDA of the `weekly cardiovascular mortality` time series.

```
df<- cmort
summary(df)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.11   81.90   87.33   88.70   94.36  132.04
```

```
cat("Number of missing values:", sum(is.na(df)), "\n")
```

```
## Number of missing values: 0
```

2. What features do you notice of the `weekly cardiovascular mortality` time-series plot? There is seasonality and downward trend in this plot.
3. Do you think that it is stationary in the mean? In the variance? The mean is not stationary as of the downward trend. Variance is relatively stationary.
4. What pieces of information did you use from your EDA to arrive at your conclusion? Mean and standard deviation.
5. Do you find any evidence that there is a dependency structure in this time series data? Please explain. I have not found evidence that there is a dependency structure.

6. What is the difference between strict and weak stationarity? Strict, or strong, stationarity means that in the probability distribution of the random variable (RV) tossed in each time instant is exactly the same along time, and that the joint probability distribution of RVs in different time instants is invariant to time shifting (this joint probability is usually evaluated with correlation or covariance).

In a weak, or wide-sense, the mean and the correlation and covariance of the RV are invariant to time shift (e.g. the variance of the RV eventually changes with time).

7. What is the difference between an acf and pacf plot? The autocorrelation function (ACF) measures how a series is correlated with itself at different lags. If the data is strongly seasonal, the peaks will coincide with the seasonality period, so it can help to infer the seasonality (sometimes its obvious from the chart, but not always).

If stationary, the ACF can help guide your choice of moving average lags. Also it's a good way to confirm any trend, for a positive trend you'll see the ACF taking ages to die out.

The partial autocorrelation function can be interpreted as a regression of the series against its past lags. It helps to come up with a possible order for the auto regressive term. The terms can be interpreted the same way as a standard linear regression, that is the contribution of a change in that particular lag while holding others constant.

As a rule of thumb, the ACF is used to confirm trend and infer possible values of the moving average parameters, and the PACF for the auto regressive part.

8. (Open-ended question) Give two examples of questions people in your industry might ask that, based on what you learn in the async lecture, you think can be addressed using time-series analysis.