Statistical Methods for Discrete Response, Time Series, and Panel Data: Live Session 6

Devesh Tiwari

June 20, 2017

Main Topics Covered in Lecture 6:

- Introduction to time series analysis
- Basic terminology of time series analysis
- Steps to analyze time series data
- Common empirical time series patterns
- Examples of simple time series models
- Notion and measure of dependency
- Examining time series correlation - autocorrelation function (ACF)
- Notion of stationarity

Required Readings:

CM2009: Paul S.P. Cowpertwait and Andrew V. Metcalfe. *Introductory Time Series with R.* Springer. 2009.

SS2016: Robert H. Shumway and David S. Stoffer. *Time Series Analysis and Applications.* EZ Edition with R Examples

- CM2009: Ch. 1, 2.1.1, 2.2.4, 2.2.5, 2.3, 4.2
  - Skip Ch. 1.5.4, 1.5.5
- SS2016: Ch.1

Agenda for the Live Session

1. Discussion of what time-series data is and why it is different.
2. What are the types of questions we can ask with time-series data?
3. An overview of the time-series portion of this course.
4. Exericse 1 (Estimated Time: Total - 30 minutes (Breakout 15 minutes, Classwide disscussion: 15 minutes)
5. Exericse 2 (Estimated Time: Total - 30 minutes (Breakout 15 minutes, Classwide disscussion: 15 minutes)
6. Exericse 3 (Estimated Time: Total - 15 minutes (Breakout 10 minutes, Classwide disscussion: 5 minutes)

1 Why is time series different? (Breakout Session: 10 mins, group discussion: 10 mins)

Question for class: What is the difference between cross-sectional data, time-series data, and panel data? What challenges do we face when we have to analyze time-series data? For simpliclity, assume that your dependent variable is continuous.

2 Time-series data and two different questions we can ask (group discssion: 10 mins)

Consider the following example:

Suppose you were interested in examining the relationship between the Presidential Approval rate and how well the stock market is performing. You suspect that voters like their President more when the stock market is doing well. In order to test this hypothesis, you collect 20 years worth of data. To be more precise, you collect the weekly (or monthly) Presidential approval rate over a 20 years span along with the weekly (or monthly) value of the Dow Jones Industrial Average (DJIA), and now you wish to regress Presidential Approval on DJIA to see if there is indeed a positive relationship between the two.

We can actually ask two different questions with these data:

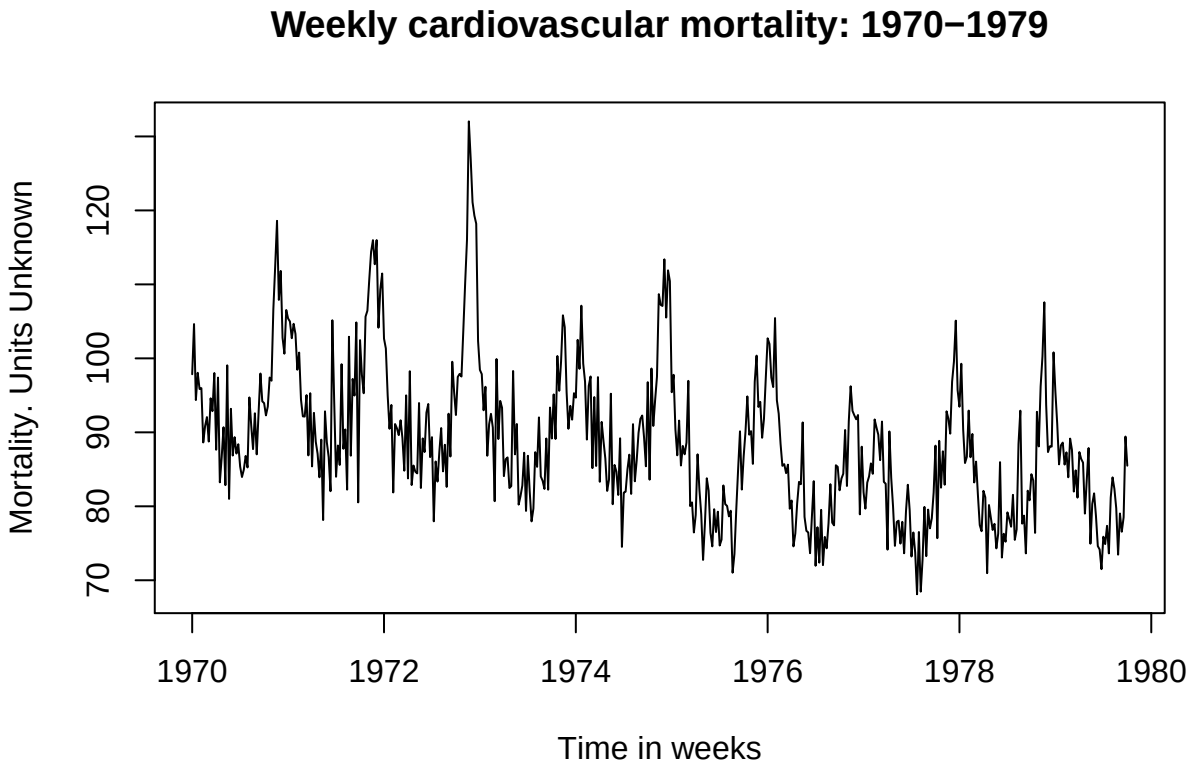
1. Is there a relationship between the Presidential approval rate and stock-market performance? Is there a causal relationship and if so, what causes what?
2. What will the President's approval rating (or DJIA) be over the next X weeks?

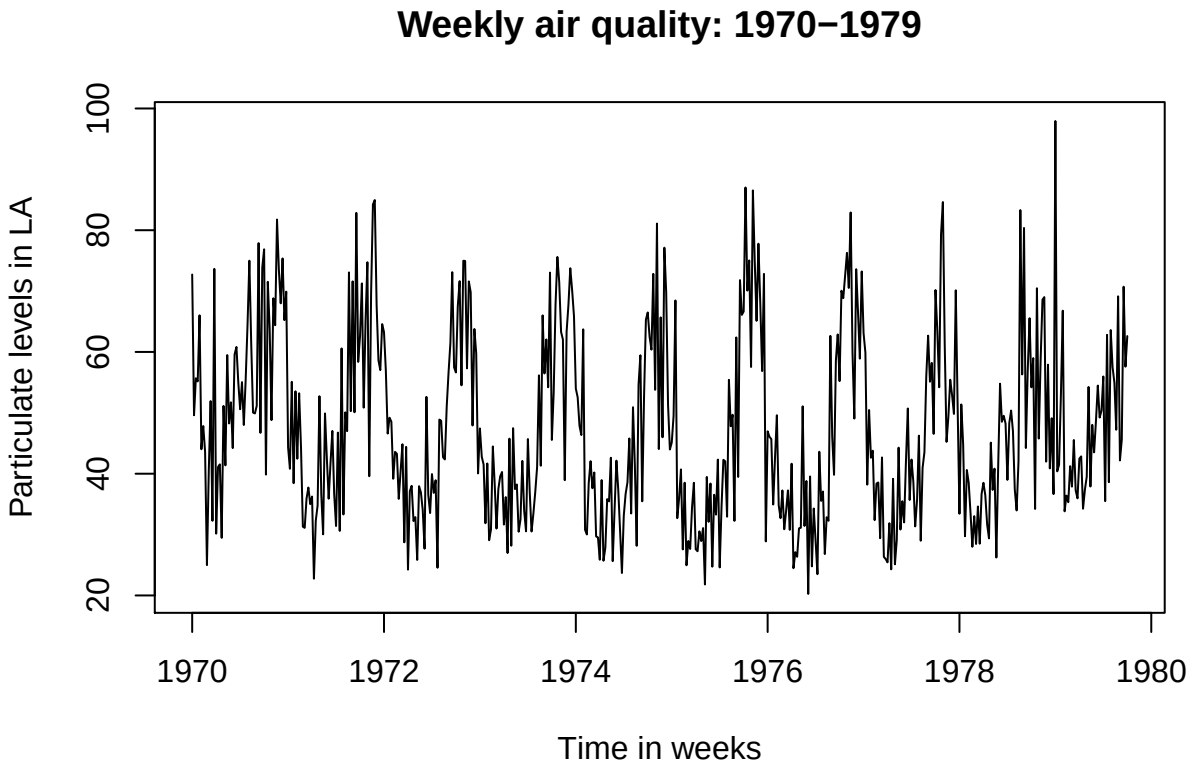
Think about these two questions for a moment. Think about what equation you would be estimating and the challenges you face answering these questions. Ask questions if this is not clear to you!

In order to answer any of these questions, we need to deal with the temporal dimension of the data, and we have many different methods at our disposal to answer each question (which is why question formation is so important!). With respect to forecasting, it turns out that we can generate forecasts by analyzing each variable in isolation (univariate time-series analysis) and possibly improve these forecasts by including both variables within the same analysis (multi-variate time-series analysis).

Similarily, suppose you were interested in understanding the relationship between air pollution and deaths related to cardio-vascular issues. You could use time-series data to answer questions about the relationship between these two variabes and you could use these data to generate forecasts.

Over the next few weeks, we will cover all of the steps you will need to go through in order to generate forecasts for cardiovascular mortality rate.


Weekly cardiovascular mortality: 1970−1979
Mortality. Units Unknown
70
80
90
100
120
1970
1972
1974
1976
1978
1980
Time in weeks


Weekly air quality: 1970−1979
Particulate levels in LA
20
40
60
80
100
1970
1972
1974
1976
1978
1980
Time in weeks

3 Overview of the next few weeks

Week 7: Smoothing, EDA, and intro to modeling We will learn how to smooth time-series data in order to remove noise and observe it's underlying trend, how to conduct an EDA and an overview on the modelling process.

Week 8: Auto-Regressive (AR), Moving-Average (MA), and Auto-Regressive and Moving Average (ARMA) models.

We will learn how to model time-series data where the current observation can be expressed as a function of prior observations. These models are applied on data that are *stationary in the mean.* In doing so, we will have to answer the following questions:

1. Which model is appropriate, AR, MA, or ARMA?
2. How many lagged values do we need to include in the model?
3. How do we know that we have a "good" model?

Week 9: Integrated and Seasonal models (ARIMA and SARIMA)

We will then learn how to model data that are *not* stationary in the mean. We will answer the same questions as above, but only this time we will also answer:

1. Do I need to difference (or transform) the data in order to make it stationary?
2. Is there a seasonal component to the data and if so, how do I incorporate it in the model?

Week 10: Vector-Auto Regression Models (VAR) In the last week of the time-series portion of class, we will cover how to include more than one time-series variable in the model. We will answer the same questions as above, and add:

1. What is the relationship between the multiple variables?
2. How many lagged terms to we inlcude in the VAR model?

4. Exercise 1: Basic Concepts

- *Estimated Time: Total 30 minutes (Breakout 15 minutes, Classwide disscussion: 15 minutes)*

What do we mean by the term "stochastic process" and what is the analalgous concept in generalized linear models?

Suppose you were interested in modeling the cardiovascular mortality data for forecasting purposes. You have a series of models to choose from (we will talk about why this is the case in great length in later sessions). Based on your prior experiences in linear regression and discrete response models, what are the characteristics of a good model?

5. Exercise 2: Walking through an EDA

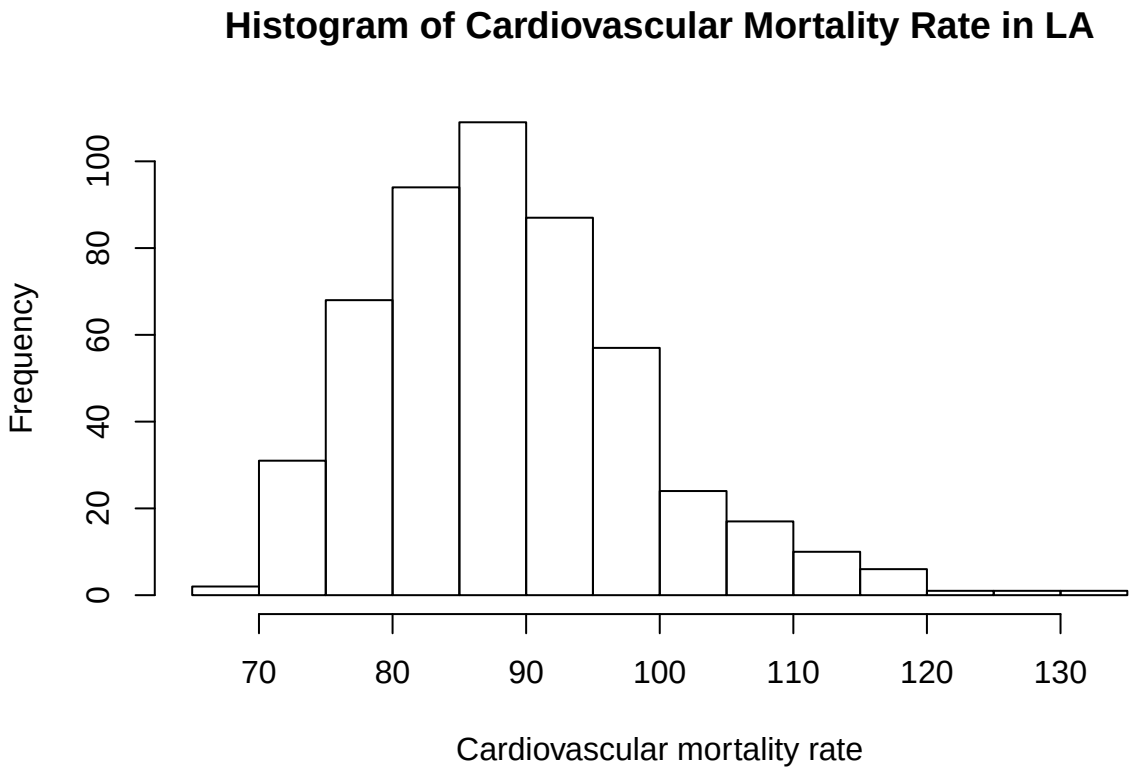
- *Estimated Time: Total 10 minutes (Classwide disscussion: 10 minutes)*

Examine the EDA below. Answer the following questions:

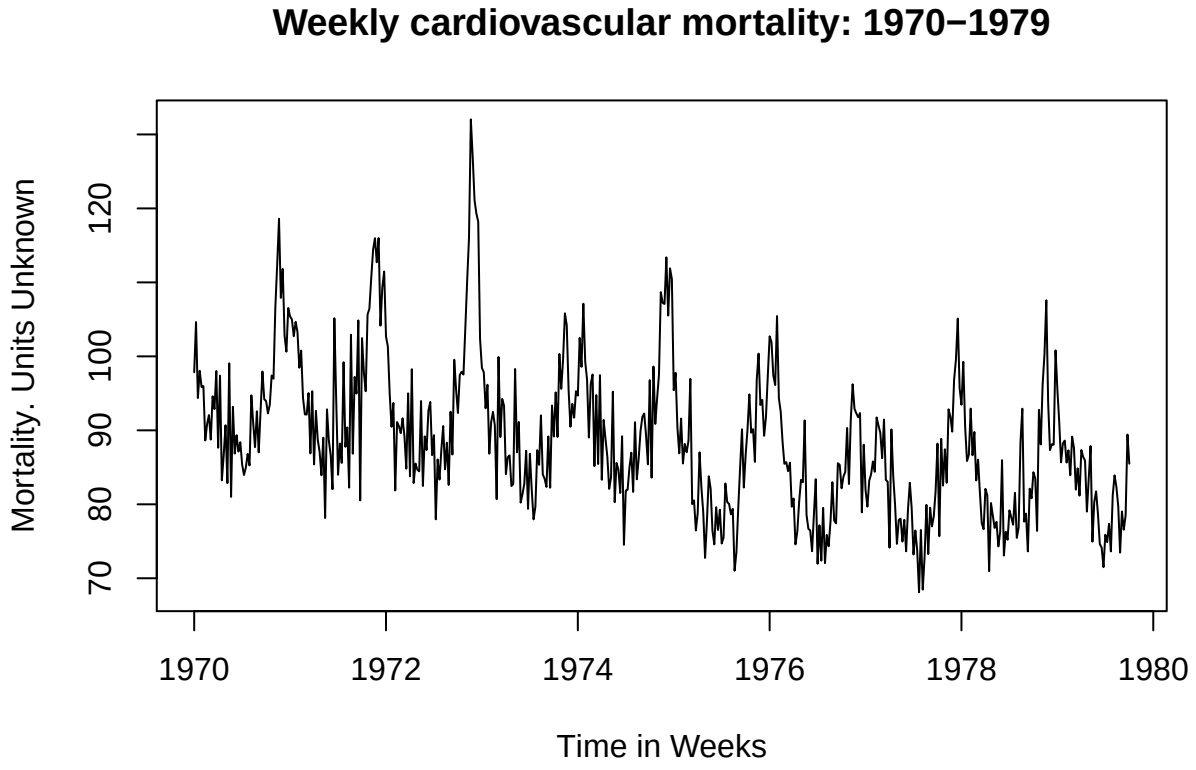
1. What features do you notice of the time-series plot?
2. Do you think that it is stationary in the mean? In the variance?
3. What pieces of information did you use from the EDA below to come to that conclusion?

4. Do you find any evidence that there is a dependency structure in this time series data?
5. What is the difference between strict and weak stationarity?
6. What is the difference between an acf and pacf plot?

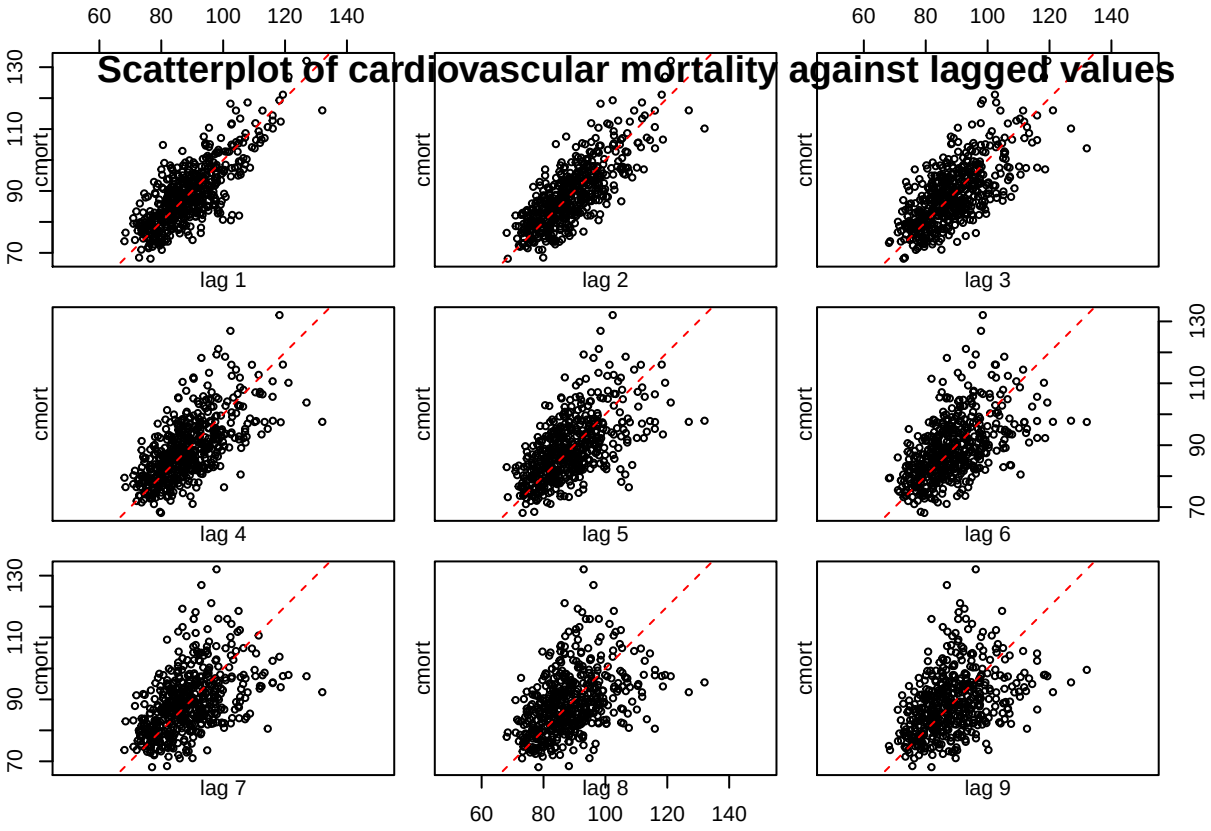
```
hist(cmort, main = "Histogram of Cardiovascular Mortality Rate in LA",
     xlab = "Cardiovascular mortality rate")
```



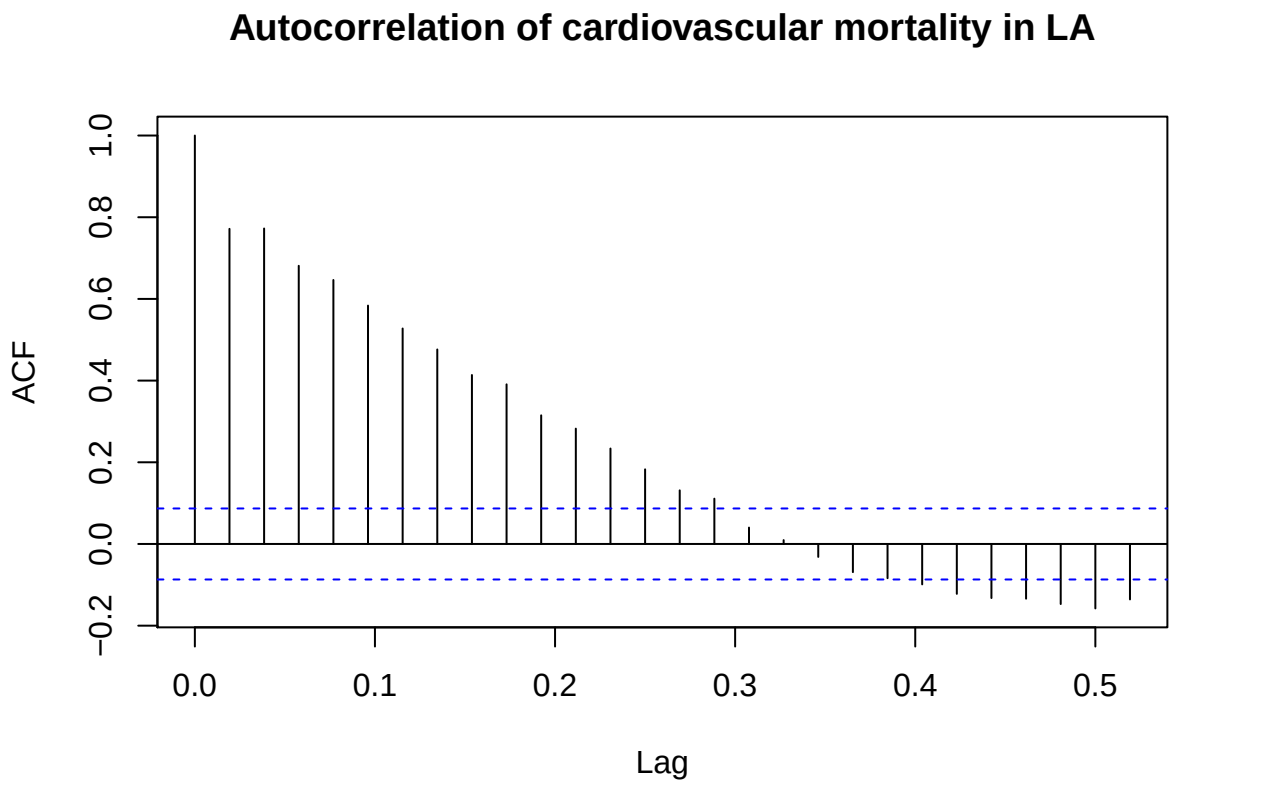

```
plot(cmort, xlab= "Time in Weeks", ylab="Mortality. Units Unknown")
title(main="Weekly cardiovascular mortality: 1970-1979")
```



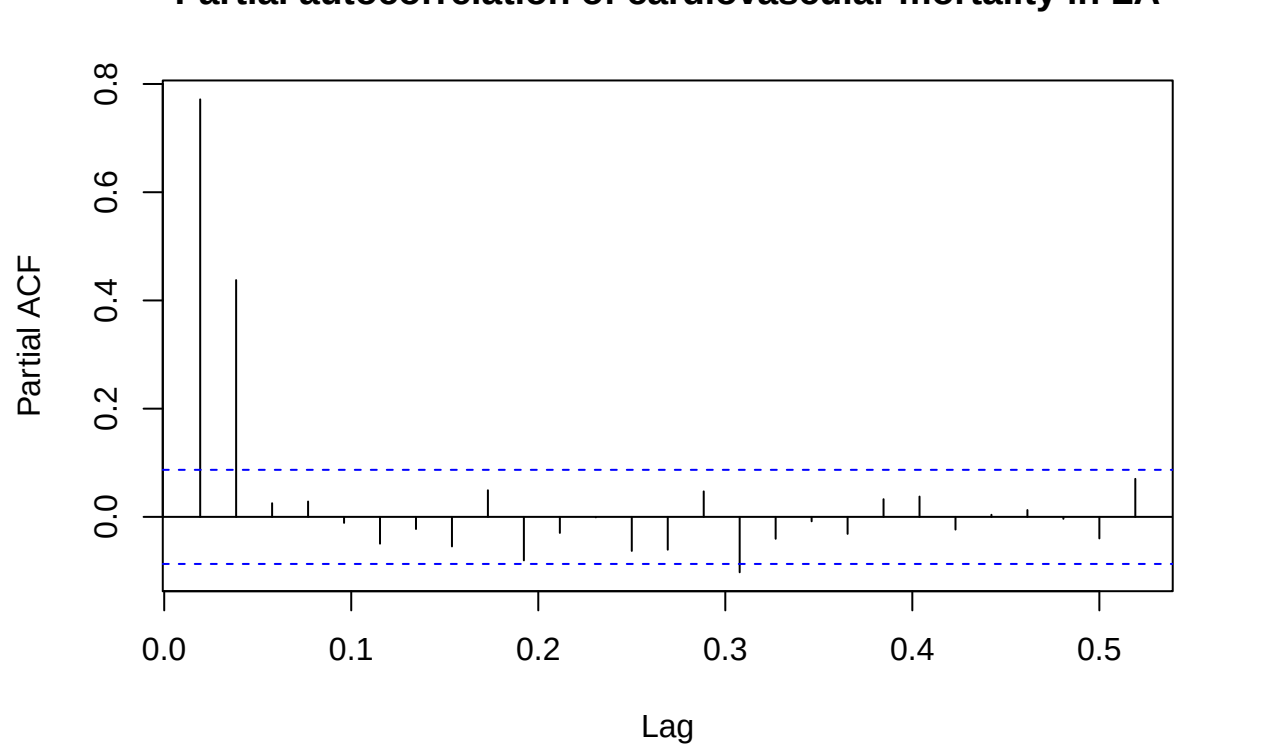

```
lag.plot(cmort, lags = 9, layout = c(3,3),
         diag = TRUE, diag.col = "red")
title(main = "Scatterplot of cardiovascular mortality against lagged values")
```



```r
acf(cmort, main = "Autocorrelation of cardiovascular mortality in LA")
```




```r
pacf(cmort, main = "Partial autocorrelation of cardiovascular mortality in LA")
```



6. Exercise 3: Time-Series Data and You

- *Estimated Time: Total 15 minutes (Breakout 10 minutes, Classwide disscussion: 5 minutes)*

In your line of work or industry, brainstorm how you might encounter time-series data and which of the aforementioned types is the most common. Give some examples of the type of question people in your industry might ask and how they could use time-series analysis to answer them.