CS 660: Mathematical Foundations for Analytics

Dr. Francis Parisi

Pace University

Spring 2017

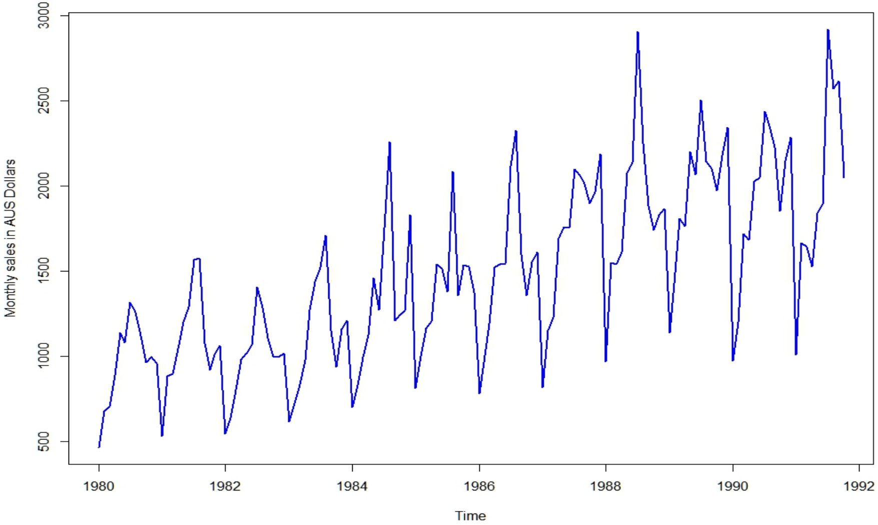
Course Overview

1. Part I: Data Science Fundamentals
   - Data Science Concepts and Process
   - The R Language
   - Exploratory Data Analysis
   - Cleaning & Manipulating Data
   - Presenting Results
2. Part II: Graphs & Statistical Methods
   - Basic Graphics
   - Advanced Graphics
   - Probability & Statistical Methods
3. Part III: Modeling Methods
   - Model Selection and Evaluation
   - Linear and Logistic Regression
   - Unsupervised Methods
   - Advanced Modeling Methods

Time Series

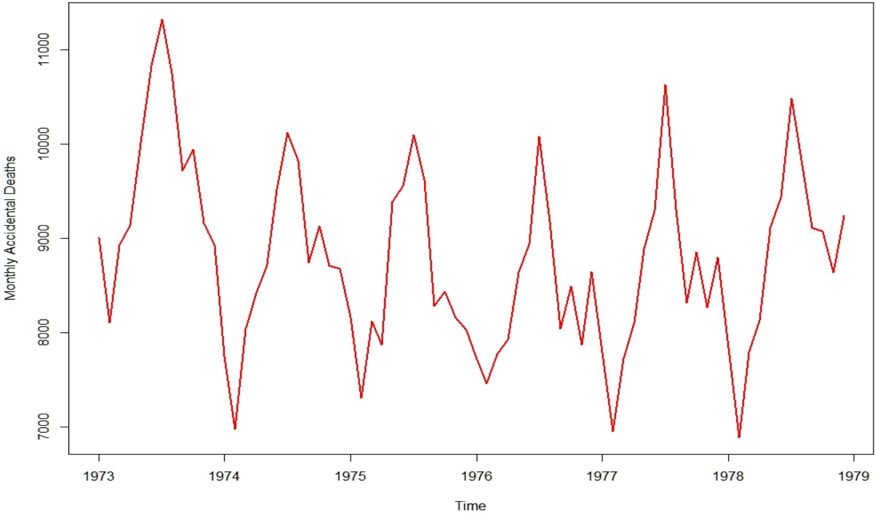
- Time series is a collection of observations X_t, each one being recorded at time t
- Time could be discrete, $t = 1, 2, 3, ,$ or continuous $t > 0$
- Time series data occur very often in reality and thus it is important to know how to deal with them
- Let's look at some examples of time series data

Time Series

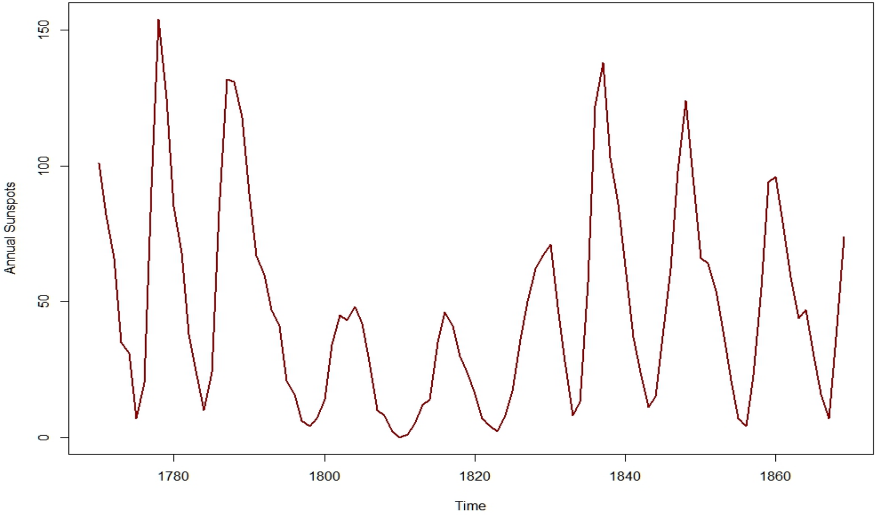


Monthly sales in AUS Dollars
500
1000
1500
2000
2500
3000
1980
1982
1984
1986
1988
1990
1992
Time

Time Series


Monthly Accidental Deaths
7000
8000
9000
10000
11000
1973
1974
1975
1976
1977
1978
1979
Time

Time Series



Time Series

- Time series show up in many fields including engineering, science, sociology, and economics
- We analyze time series to draw inferences from them
- A time series is a sequence of random variables $\{X_t\}$ measured over time
- We often denote a time series as

$$X_t = m_t + s_t + Y_t \tag{1.1}$$

  for an additive model and

$$X_t = m_t * s_t * Y_t \tag{1.2}$$

  for a multiplicative model
- This is known as classical decomposition

Time Series

- In equations (1.1) and (1.2) on the previous slide, m_t is the trend component, s_t is the seasonal component, and Y_t is the random or irregular component
- m_t is a slowly varying function and is often estimated using least squares
- That is, we fit a function for example

$$m_t = a_0 + a_1 t + a_2 t^2 \cdots + a_p t^p$$

  to the data $\{x_1, \ldots, x_n\}$ by finding parameters that minimize

$$\sum_{t=1}^{n} (x_t - m_t)^2$$

- For a linear trend we just have $m_t = a_0 + a_1 t$

Time Series – General Approach to Modeling

- Plot the time series and look for features do we have
  1. a trend
  2. a seasonal component
  3. any sudden change in the series
  4. any outliers
- Remove any trend and seasonal components to get *stationary* residuals
- Choose a model to fit the residuals
- Forecast based on the residuals and invert any transformations to get the original series
- An alternative approach is to express the series in terms of Fourier components (we won't cover this)

Time Series

A time series $\{X_t\}$ is (loosely speaking) *stationary* if it has similar statistical properties to those of the "time shifted" series $\{X_{t+h}\}$ for each integer h

Definition

Let $\{X_t\}$ be a time series with $\mathbb{E}\, X_t^2 < \infty$. The **mean function** of $\{X_t\}$ is

$$\mu_X(t) = \mathbb{E}(X_t).$$

The **covariance function** of $\{X_t\}$ is

$$\gamma_X(r,s) = \mathrm{Cov}(X_r, X_s) = \mathbb{E}[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for all integers $r,\ s$, and t.

Definition

$\{X_t\}$ is **(weakly) stationary** if

$$\text{(i) } \mu_X(t) \text{ is independent of } t,$$

and

$$\text{(ii) } \gamma_X(t+h, t) \text{ is independent of } t \text{ for each } h.$$

Definition

Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (ACVF) of $\{X_t\}$ is

$$\gamma_X(h) = \mathrm{Cov}(X_{t+h}, X_t).$$

The **autocorrelation function** (ACF) of $\{X_t\}$ is

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \mathrm{Cor}(X_{t+h}, X_t).$$

Time Series

Some examples...

- Creating time series
- Plotting to inform our analysis
- Seasonal decomposition to remove trend and seasonality
- Forecasting methods
  1. Exponential modeling
  2. Autoregressive integrated moving averages (ARIMA) models

Time Series

Function	Package	Use
`ts()`	`stats`	Creates a time-series object.
`plot()`	`graphics`	Plots a time series.
`start()`	`stats`	Returns the starting time of a time series.
`end()`	`stats`	Returns the ending time of a time series.
`frequency()`	`stats`	Returns the period of a time series.
`window()`	`stats`	Subsets a time-series object.
`ma()`	`forecast`	Fits a simple moving-average model.
`stl()`	`stats`	Decomposes a time series into seasonal, trend, and irregular components using loess.
`monthplot()`	`stats`	Plots the seasonal components of a time series.
`seasonplot()`	`forecast`	Generates a season plot.
`HoltWinters()`	`stats`	Fits an exponential smoothing model.
`forecast()`	`forecast`	Forecasts future values of a time series.
`accuracy()`	`forecast`	Reports fit measures for a time-series model.
`ets()`	`forecast`	Fits an exponential smoothing model. Includes the ability to automate the selection of a model.
`lag()`	`stats`	Returns a lagged version of a time series.
`Acf()`	`forecast`	Estimates the autocorrelation function.
`Pacf()`	`forecast`	Estimates the partial autocorrelation function.
`diff()`	`base`	Returns lagged and iterated differences.

Time Series

Function	Package	Use
`ndiffs()`	`forecast`	Determines the level of differencing needed to remove trends in a time series.
`adf.test()`	`tseries`	Computes an Augmented Dickey–Fuller test that a time series is stationary.
`arima()`	`stats`	Fits autoregressive integrated moving-average models.
`Box.test()`	`stats`	Computes a Ljung–Box test that the residuals of a time series are independent.
`bds.test()`	`tseries`	Computes the BDS test that a series consists of independent, identically distributed random variables.
`auto.arima()`	`forecast`	Automates the selection of an ARIMA model.

Smoothing with Simple Moving Averages

- Smoothing dampens fluctuations so we may see patterns in the data
- Simple moving averages is the simplest method to smooth time series
- We replace each data point with the mean of that observation and one or more points before and after
- This is a centered moving average

$$S_t = (Y_{t-q} + \cdots + Y_t + \cdots + Y_{t+q})/(2q+1)$$

- R code

Time Series

- Time series data often have trend and seasonal components
- We decompose the series into its trend, seasonal, and irregular or random components
- After we decompose the series we can model the random component
- R code

Exponential Smoothing Models

- Simple or single exponential models only a level with irregular variation, no trend or seasonal
- Double exponential aka Holt exponential model has level with irregular variation, and a trend
- Triple exponential aka Holt-Winters has level with irregular variation, and a trend and seasonal component
- R code

Time Series

- ARIMA models forecast values as a linear function of one or more recent values and recent errors
- Before modeling we need to understand key terms
  - Lag – shift the time series back by one or more periods
  - Autocorrelation – measures the association between observations in different time periods
  - Partial autocorrelation – measures the association between two observations, Y_t and Y_{t-k} after removing the effects of all the observations in between
  - Differencing – replace each value in the series with the difference between successive values, removing any trend
  - Stationarity – the statistical properties (mean, variance, autocorrelations for any lag k) of the series do not change over time the Augmented Dickey-Fuller (ADF) test is used to evaluate stationarity

Autoregressive Model of order p

$$AR(p): \; Y_t = \mu + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t$$

Moving Average Model of order q

$$MA(q): \; Y_t = \mu - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} + \epsilon_t$$

ARMA model of order (p, q)

$$ARMA(p,q): \; Y_t = \mu + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q} + \epsilon_t$$

Typically we subtract the mean so $\mu = 0$

Guidelines for model selection using ACF and PACF

Model	ACF	PACF
ARIMA(p, d, 0)	Trails off to zero	Zero after lag p
ARIMA(0, d, q)	Zero after lag q	Trails off to zero
ARIMA(p, d, q)	Trails off to zero	Trails off to zero

- $ARIMA(p, d, q)$ – Autoregressive Integrated Moving Average model with AR order p, MA order q series differenced d times
- R code for an $ARIMA$ example

Time Series Methods – Summary

- Time series data appear in every field of study
- Fundamental to time series analysis is to forecast future values
- There are many techniques for time series modeling and forecasting including the two we looked at: exponential smoothing and $ARIMA(p, d, q)$ models
- Be aware that we are forecasting beyond the observed data based on historical behavior and forecasts are prone to error if things change
- This is why forecast errors increase the further out you go
- This is less of a concern when working with say stable natural phenomena, or other phenomena where the historical behavior is less volatile

References

Brockwell, P. J. and Davis, R. A. (1991).
Time Series: Theory and Methods.
Springer-Verlag, New York, second edition.

Brockwell, P. J. and Davis, R. A. (1996).
Introduction to Time Series and Forecasting.
Springer-Verlag, New York.

James, G., Hastie, T., Witten, D. and Tibshirani, R. (2013).
An Introduction to Statistical Learning with Applications in R.
Springer, second edition.
6th Printing 2015.

Kabacoff, R. I. (2015).
R in Action.
Manning, Shelter Island, NY, second edition.

Lander, J. P. (2014).
R for Everyone.
Addison-Wesley, Upper Saddle River.

References (cont.)

Zumel, N. and Mount, J. (2014).
Practical Data Science with R.
Manning, Shelter Island, NY, second edition.