Day 1

MAS 640

1/17/2018

Reading

- Chapter 1 - Introduction
- Chapter 2 - Fundamental Concepts

Time Series

A **time series** is a sequence of ordered data. The “ordering” refers generally to time, but other orderings could be envisioned (e.g., over space, etc.).

In this class, we will be concerned exclusively with time series that are

- measured on a single continuous random variable Y
- equally spaced in discrete time; that is, we will have a single realization of Y at each second, hour, day, month, year, etc.

Time Series data are everywhere!

- Business: daily stock prices, weekly interest rates, quarterly sales, monthly supply figures, annual earnings
- Medicine: EKG measurements, drug concentrations, blood pressure readings
- Public Health: Flu cases per day, health-care clinic visits per week, annual disease incidence
- Agriculture: annual yields, daily crop prices
- Social Sciences: annual birth and death rates, accident frequencies, crime rates, school enrollment
- Meteorology: daily high temperatures, annual rainfall, hourly wind speeds, earthquake frequency

Time Series Notation

A time series is denoted as Y_t, where

$$Y_t = \text{Value of } Y \text{ at time } t, \text{ for } t = 1, 2, \cdots, n$$

The subscript t tells which time point the measurement Y_t corresponds.

Note that in the sequence $Y_1, Y_2, \cdots Y_n$ the subscripts are very important because they correspond to a particular ordering of the data.

This is perhaps a change in mind set from other courses where time is ignored and the subscripts rarely matter.

Time Series Plot

A **time series plot** is the most basic graphical display in the analysis of time series data. Always start here!

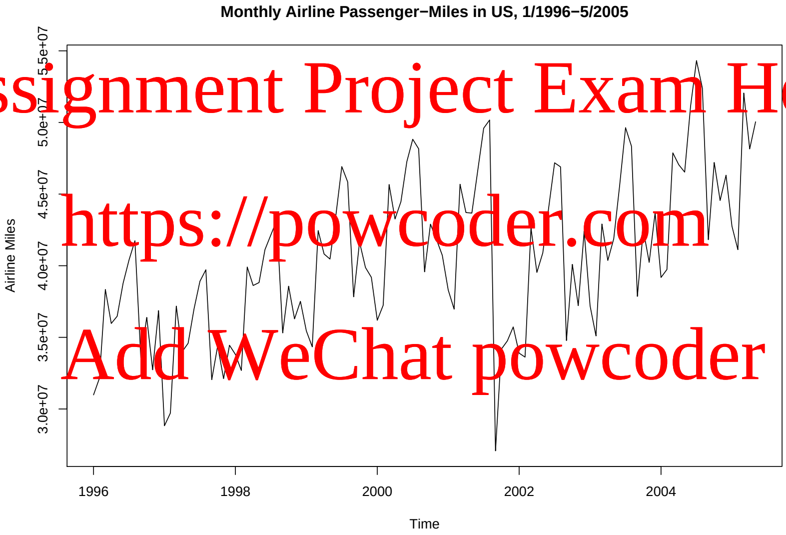
The plot is a scatterplot Y_t versus t, with straight lines connecting the points

Time Series Plots

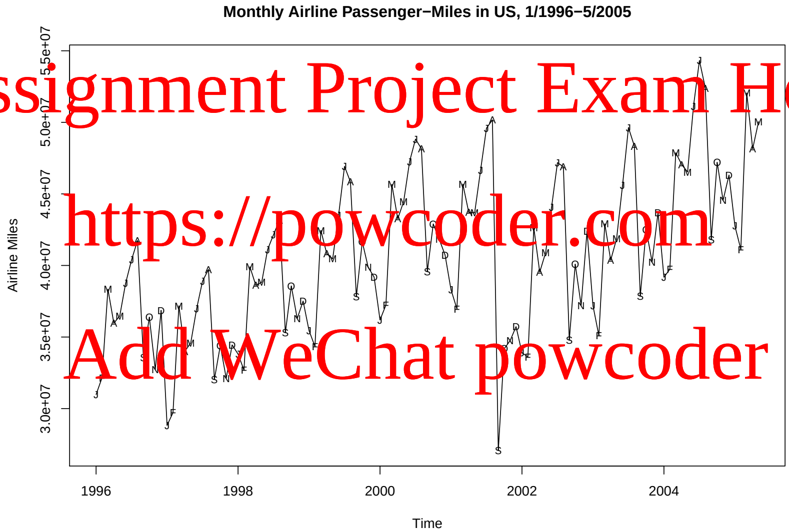
When looking at a time series plot. . .

- Is there a **trend**? On average, increasing or decreasing over time
- Is there **seasonality**? Regularly repeating patters corresponding to calendar time (seasons, quarters, months, weekday, etc. . .)
- Are there any **outliers**?
- Is there **constant variance** over time?
- Are there any **abrupt changes** to either the level or variance?

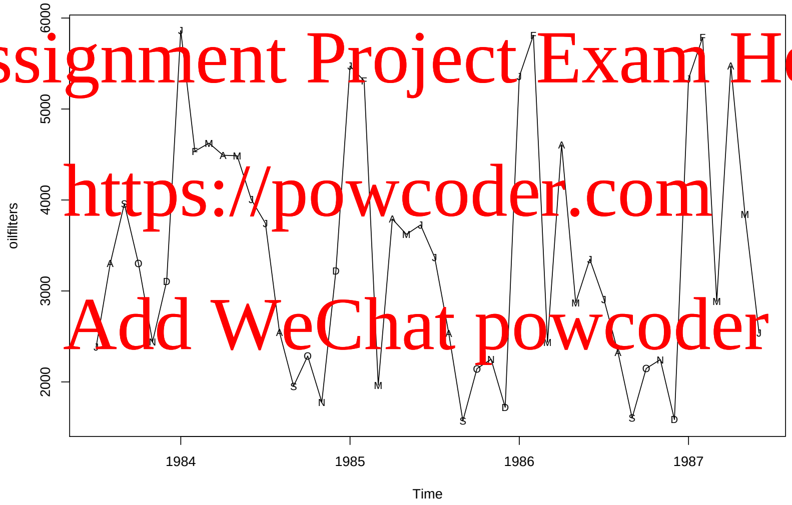
Airline Miles



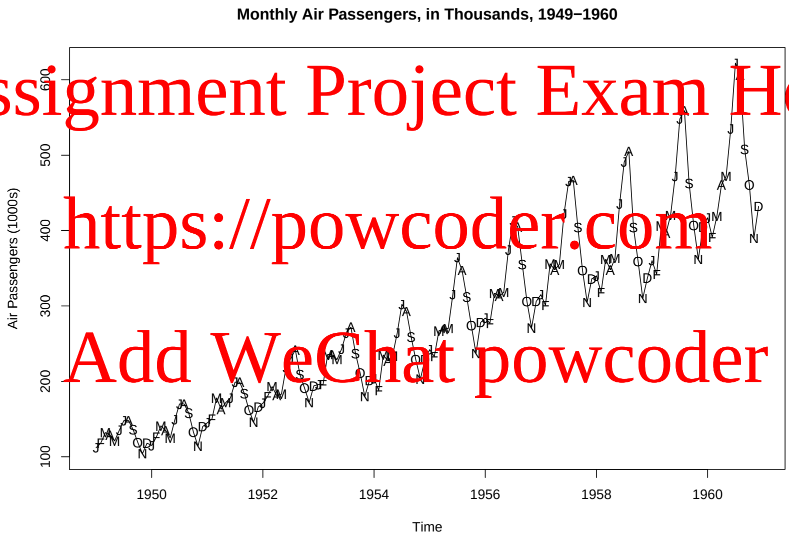
Airline Miles



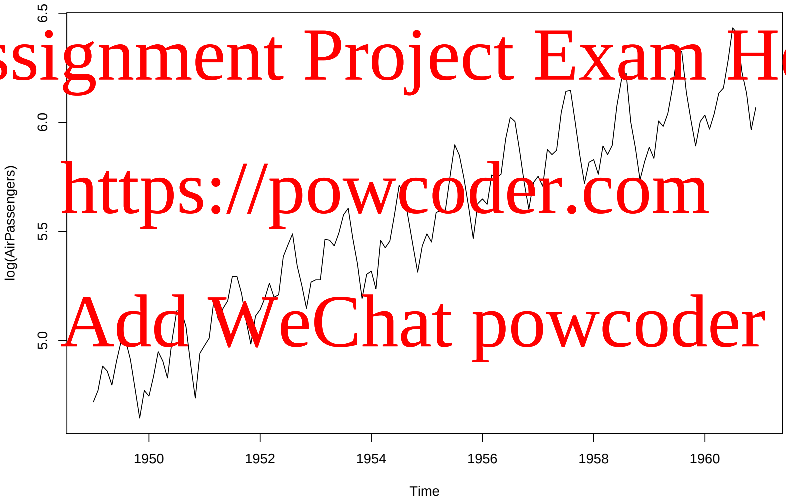
Monthly Oil Filter Sales



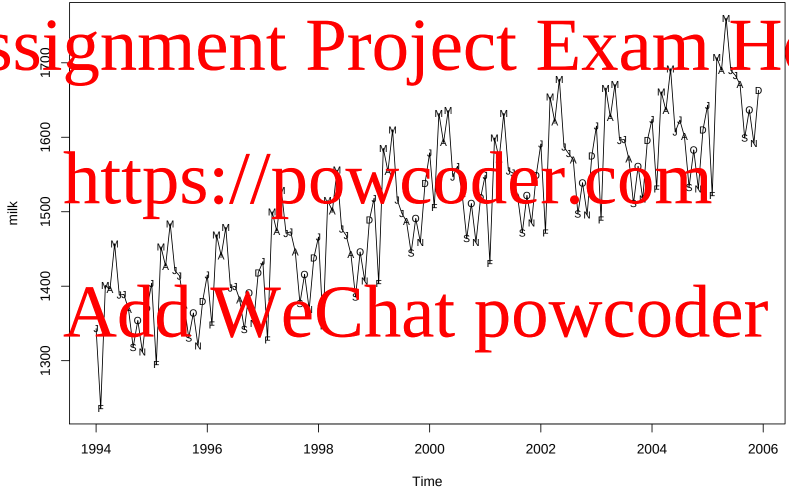
Airline Passengers



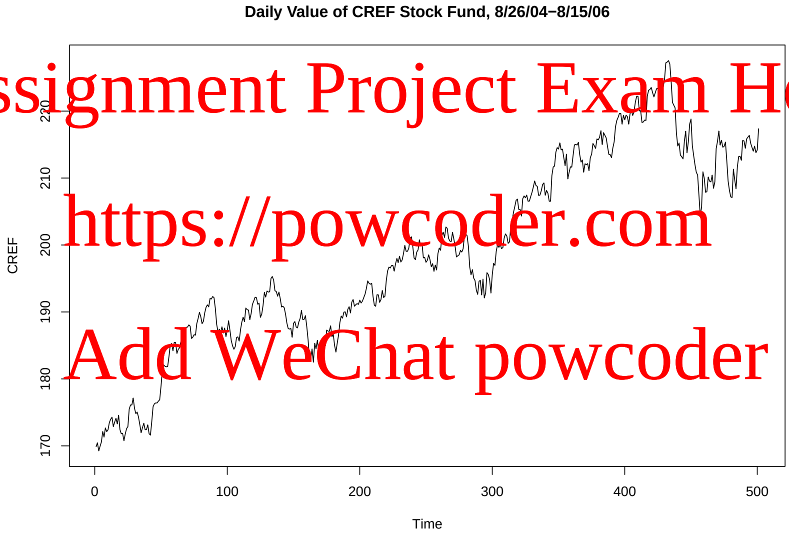
Airline Passengers - After Log Transformation. . .



Milk Production



CREF Stock Fund



Time Series Plots before R

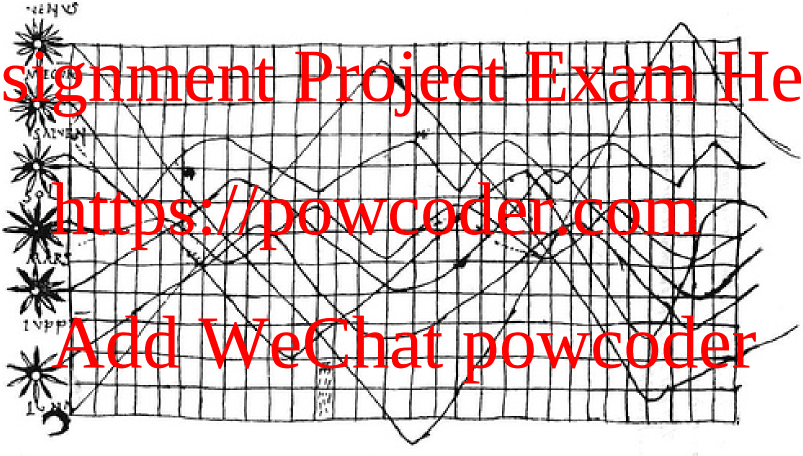


Figure 1: First known time series plot. From 10th or 11th century showing the inclinations of the planetary orbits.

Try one yourself!

- Using the birth dataset located in the astsa package:
  - Plot the time series
  - What noticeable patterns are present?
  - Which month tends to have the most births? The least?


```r
library(astsa)
data() # See what's available
data(birth, package='astsa')
```

Why Visualize the Time Series?

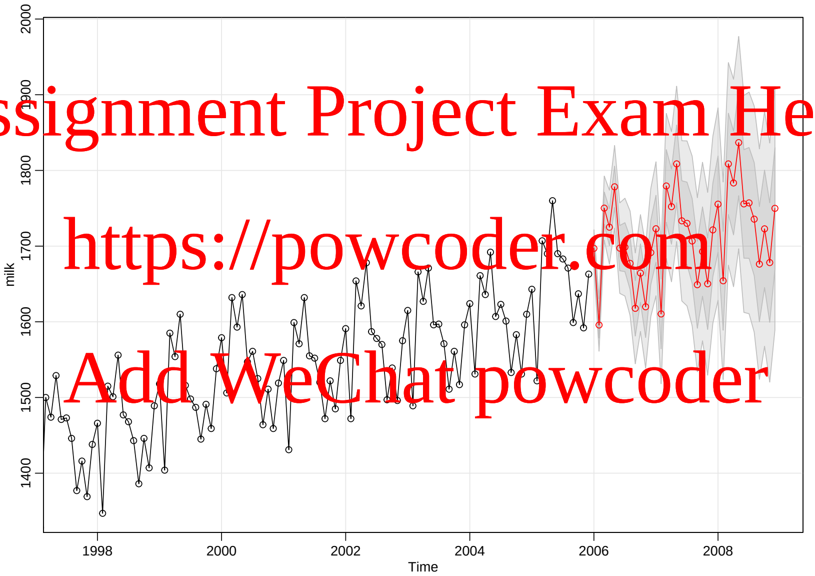
- It's important to know what patters are present, as they will guide the modeling process.

- If there's a trend, seasonality, non-constant variance, abrupt changes, etc, we need to account for that when modeling.
- How else can you detect these patterns if not via visualization?

Goals in Time Series

1. **Model** the stochastic (random) mechanism that gives rise to the data
2. Predict or **forecast** the future based on the past

Model and Predict



What's Different About Time Series?

The big thing about time series data is that they are **not independent**! Instead, observations are correlated through time.

- Correlated data are generally more difficult to analyze
- Statistical theory without independence is markedly more difficult

Classical Statistics

- Most classical methods (Regression, ANOVA, GLM) assume that observations are independent. Consider the linear model:

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$$

- We typically assume that ϵ_i are independent and identically distributed, normal with mean 0 and constant variance.

Course Goals

At the end of this course, I hope that you have an understanding of how to build and use time series models.

1. Model specification
   - Consider different classes of time series models
   - Use descriptive statistics, graphs, subject matter knowledge to propose sensible candidate models
   - Abide by **Principle of Parsimony**

2. Model fitting
   - After choosing a model, estimate it!
   - Least squares / MLE via software, understand output

3. Model diagnostics
   - Inference and graphics to determine how well the model fits the data
   - Might suggest your model is inappropriate, or point toward a more appropriate model

Expected Value

$$\mu = E(Y) = \int_R yf(y)dy$$

- For any real number a, we have...
  - $E(a) = a$
  - $E(aY) = aE(Y)$
  - $E(\sum_{j=1}^{k} Y_j) = \sum_{j=1}^{k} E(Y_j)$

Variance

$$\sigma_y^2 = var(Y) = E\left[(Y - \mu)^2\right]$$

- Typically easier to work with...

$$var(Y) = E(Y^2) - E(Y)^2$$

- Can you show these are equivalent?

Covariance

$$cov(X, Y) = E\left[(X - \mu_x)(Y - \mu_y)\right] = E(XY) - E(X)E(Y)$$

- If $cov(X, Y)$...
  - $> 0 \rightarrow$ X and Y are positively linearly related
  - $< 0 \rightarrow$ X and Y are negatively linearly related
  - $= 0 \rightarrow$ X and Y are not linearly related

- If X and Y are independent...
  - $E(XY) = E(X)E(Y) \rightarrow cov(X, Y) = 0$

Correlation

$$\rho = corr(X, Y) = \frac{cov(X, Y)}{\sigma_x \sigma_y}$$

- $-1 \leq \rho \leq 1$
  - $\rho = -1 \rightarrow$ perfectly negatively related
  - $\rho = 1 \rightarrow$ perfectly positively related
  - $\rho = 0 \rightarrow$ not linearly related

Time Series and Stochastic Processes

The sequence of random variables $Y_t : t = 0, 1, 2, \cdots, n$, or simply Y_t, is called a **stochastic process**. It is a collection of random variables indexed by time t, so . . .

$$
\begin{aligned}
Y_0 &= \text{value of the process at time } t = 0 \\
Y_1 &= \text{value of the process at time } t = 1 \\
&\vdots \\
Y_n &= \text{value of the process at time } t = n
\end{aligned}
$$

In most time series processes, most of what we need is captured with only $E(Y_t)$ and $E(Y_t Y_{t-k})$

For the stochastic process Y_t, define the **mean** function as

- Mean function

$$\mu_t = E(Y_t)$$

Note that the mean might depending on the time t (it can change through time)

Time Series and Stochastic Processes

For the stochastic process Y_t, define the **autocovariance** as

$$\gamma_{t,s} = cov(Y_t, Y_s)$$

For the stochastic process Y_t, define the **autocorrelation** as

$$\rho_{t,s} = \frac{cov(Y_t, Y_s)}{\sqrt{var(Y_t)var(Y_s)}}$$

- $\rho_{t,s}$ near ± 1 indicates strong linear dependence of Y_t and Y_s
- $\rho_{t,s}$ near 0 indicates weak linear dependence
- $\rho_{t,s} = 0$ indicates Y_t and Y_s are uncorrelated

Stationarity

- Stationarity is a very important concept in time series and one that you will often hear. Broadly speaking, a time series is called **stationary** if...

  - No systematic change in the mean (no trend),
  - No systematic change in the variance,
  - No noteable seasonal patterns exist

In other words, the properties of one section of the data are the same as any other section.

Stationarity

- Notationally, the stochastic process Y_t is said to be stationary if
  1. The mean function $\mu = E(Y_t)$ is constant through time
     - μ_t is free of t
  2. The covariance between any two observations depends only on the time lag between them
     - $cov(Y_t, Y_s) = cov(Y_{t-k}, Y_{s-k})$

Stationarity

- Why do we care if a time series is stationary or not?
  - Because almost all of the theory and models used in time series is applicable for stationary time series.
  - If you don't like that, enroll in a PhD and develop more general methods!

- Lucky for us, it's typically straightforward to transform a non-stationary time series to a stationary one.

Examples of Common Stochastic Processes

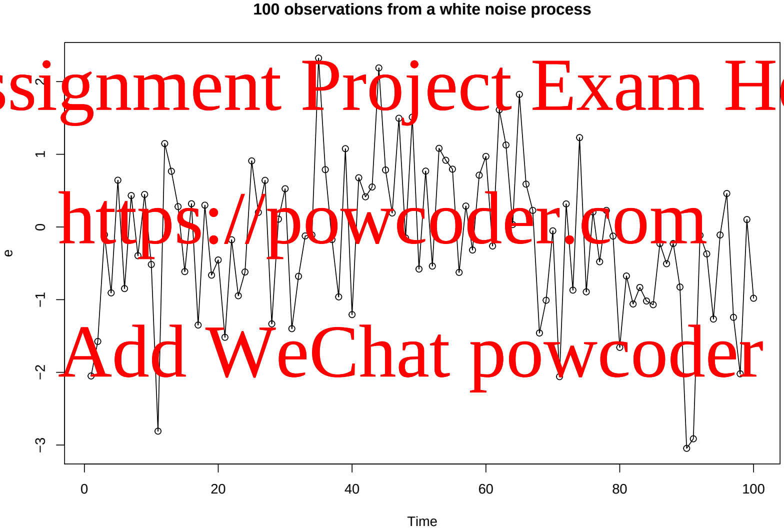
White Noise

- The process $e_t : t = 1, 2, \ldots$ is called a **white noise** process if...

$$E(e_t) = \mu$$
$$var(e_t) = \sigma^2$$

- Both mean and variance are constant through time
- Is this process stationary?

White Noise


100 observations from a white noise process
e
Time
0
20
40
60
80
100
−3
−2
−1
0
1
2

Random Walk

- Suppose e_t is a zero mean white noise process. Define

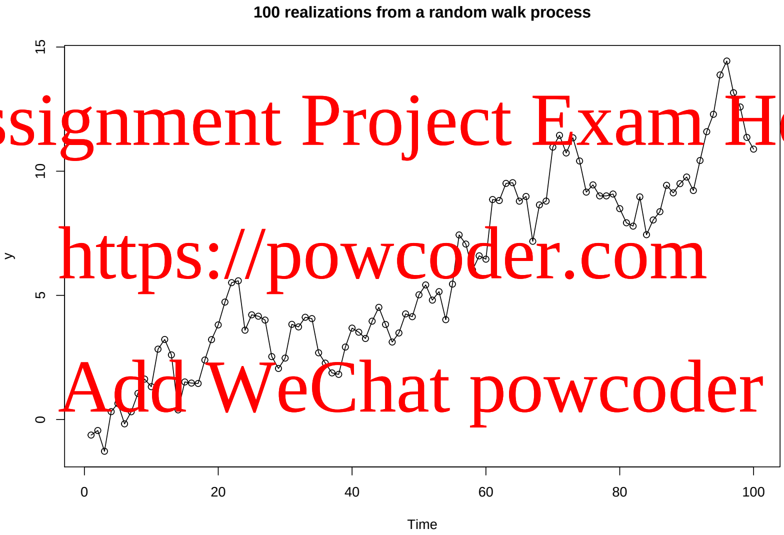
$$
\begin{aligned}
Y_1 &= e_1 \\
Y_2 &= e_1 + e_2 \\
&\vdots \\
Y_n &= e_1 + e_2 + \cdots + e_n
\end{aligned}
$$

or

$$Y_t = Y_{t-1} + e_t$$

- The process Y_t is referred to as a **random walk**. These are very frequently used in finance for modeling stock prices, among other things.

Random Walk



- Does the random walk appear stationary?
- How could we transform it to stationarity?

Differencing

- One thing we'll do a lot in this course is "difference" the time series in an effort to transform to a stationary process.
- Rather than look at Y_t, define

$$\Delta Y_t = Y_t - Y_{t-1}$$

- Try this for the random walk data and see how it looks...

Model and Subtract the Trend (Detrending)

- An alternative to differencing is to build a model for the trend and then subtract it from the original time series.

- This typically involves fitting a regression model of Y_t against time (or perhaps a more involved regression with quadratics, splines, etc. . .)

Next Class

- **Trends**
  - Modeling trends in time series data
    - Differencing or Detrending - goal of obtaining a stationary time series

  - Residual analysis
  - Chapter 3 of text