Introduction to Time Series Analysis

OC Data Driven Insights

Eric Weber and Sarah Nooravi

August 22, 2018

Announcements

- Thank you to MobilityWare for hosting us! We are always looking to hire talent, so look at our website for open reqs.
- Next 2 meetups (topics we are tossing around): data visualization, portfolio optimization, loss function
- No meetups planned for November/December
- Interested in starting your own study group? Let's talk..

Resources

- Forecasting: Principles and Practice by Rob J. Hyndman and George Athanasopoulos
- Time Series Analysis and Its Applications with R Examples by Robert H. Shumway and David S. Stoffer
- astsa
  - Includes datasets and scripts to accompany *Time Series Analysis and Its Applications with R Examples* by Robert H. Shumway and David S. Stoffer

Outline

- What is a time series? Time series model?
  - Time Series Examples
- Fundamental building blocks of time series
  - White noise
  - Moving average
  - Random walk
- Second order properties and stationarity
- When/why linear regression fails

What is a Time Series: Definition

A sequence of values of a variable at **equally spaced** intervals with a **natural temporal ordering**.

What is a Time Series: Johnson and Johnson

Let's look at a dataset provided in the astsa package which shows **quarterly** earnings per share for Johnson and Johnson from the first quarter of 1960 to the last quarter of 1980.

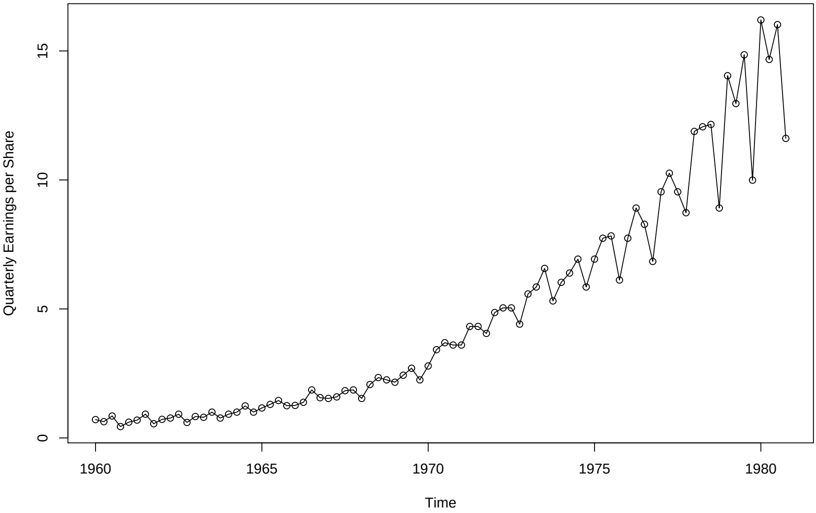
```
> require(astsa)
> window(jj, 1960, c(1965,4))
```

```
##      Qtr1 Qtr2 Qtr3 Qtr4
## 1960 0.71 0.63 0.85 0.44
## 1961 0.61 0.69 0.92 0.55
## 1962 0.72 0.77 0.92 0.60
## 1963 0.83 0.80 1.00 0.77
## 1964 0.92 1.00 1.24 1.00
## 1965 1.16 1.30 1.45 1.25
```

What is a Time Series: Johnson and Johnson

If we plot the time series, we get:

```
> plot(jj, type="o", ylab="Quarterly Earnings per Share")
```



What is a Time Series: Global Temperature

Let's look at another dataset which shows the **yearly** global mean land-ocean temperature deviations (from 1951-1980 average), measured in degrees centigrade, for the years 1880-2015.

```
> class(globtemp)
```

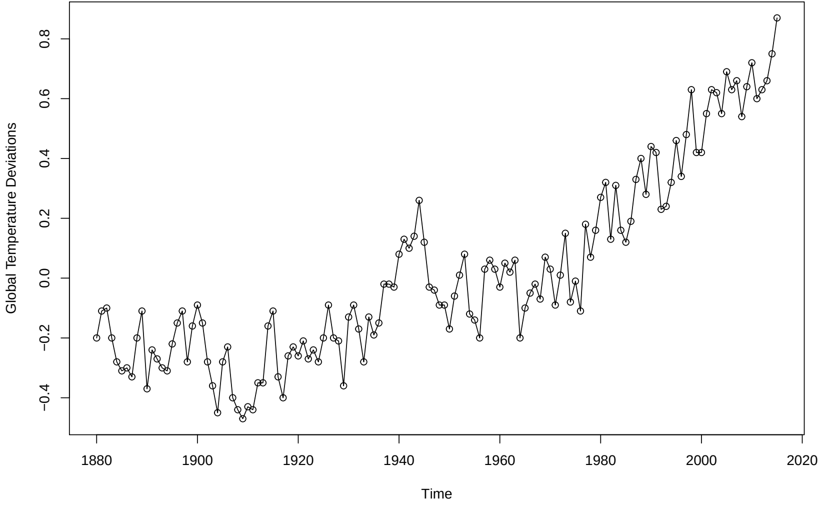
```
## [1] "ts"
```

```
> window(globtemp, 1880, 1885)
```

```
## Time Series:
## Start = 1880
## End = 1885
## Frequency = 1
## [1] -0.20 -0.11 -0.10 -0.20 -0.28 -0.31
```

What is a Time Series: Global Temperature

```r
plot(globtemp, type="o", ylab="Global Temperature Deviations")
```



Key Assumption

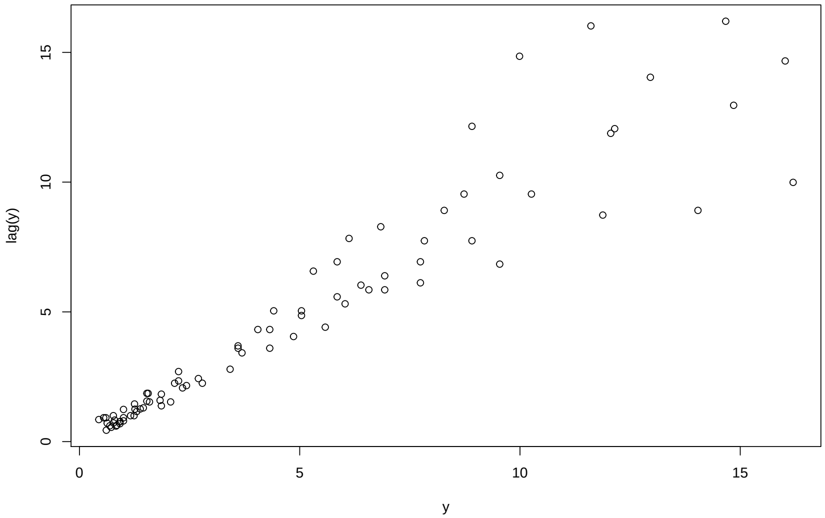
Can't assume consecutive observations are independent

```r
> jj_v <- unclass(jj)
> lag <- lag(jj_v, 1)
> lm <- lm(jj_v ~ lag)
> cor.test( ~ jj_v + lag,
+           method = "pearson",
+           conf.level = 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  jj_v and lag
## t = 25.987, df = 81, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9159239 0.9641248
## sample estimates:
##       cor
## 0.9449363
```

Key Assumption

```
> plot(jj_v, lag, ylab="lag(y)", xlab="y")
```



Why Care about Time Series?

- They are everywhere!
- Economics (stock market, unemployment, etc)
- Social sciences (population series like birth rates or school enrollments)
- Epidemiology (influenza outbreaks)
- Medicine (blood pressure measurements)

Working with Time Series

- Time series analysis
  - Analyzing observed data
  - Focus on characteristics of the data
  - Explanatory focus
- Time series forecasting
  - Generating a model
  - Predictive focus

Some Additional Terminology

- **Stochastic process**
  - A sequence of random variables
  - *Example*: flipping a coin
- **Sample path**
  - Sample path of a stochastic process
  - One sample from a stochastic process
  - *Example*: HTHHTT (six coin flips)
- A stochastic process can generate MANY sample paths (infinitely many)

Fundamental Stochastic Processes

- White noise
- Moving average
- Random walks

White Noise

- Fundamental building block of other stochastic processes
- Y_t is called white noise (process looks like white light of spectrometers)

Key: No correlation between observations

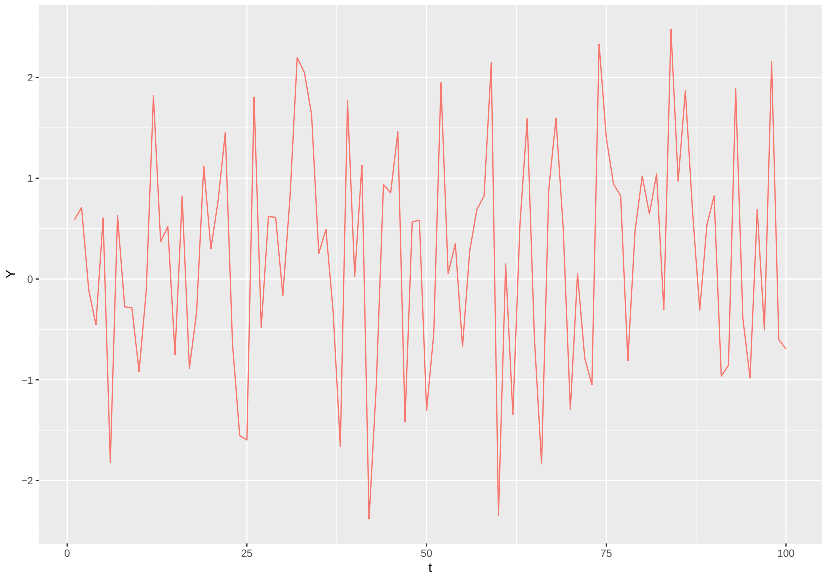
White Noise: Definition

- Assume $e_t \sim N(0, 1)$ is a collection of IID random variables all following a normal distribution
- Normal distribution? Anyone? Anyone at all?

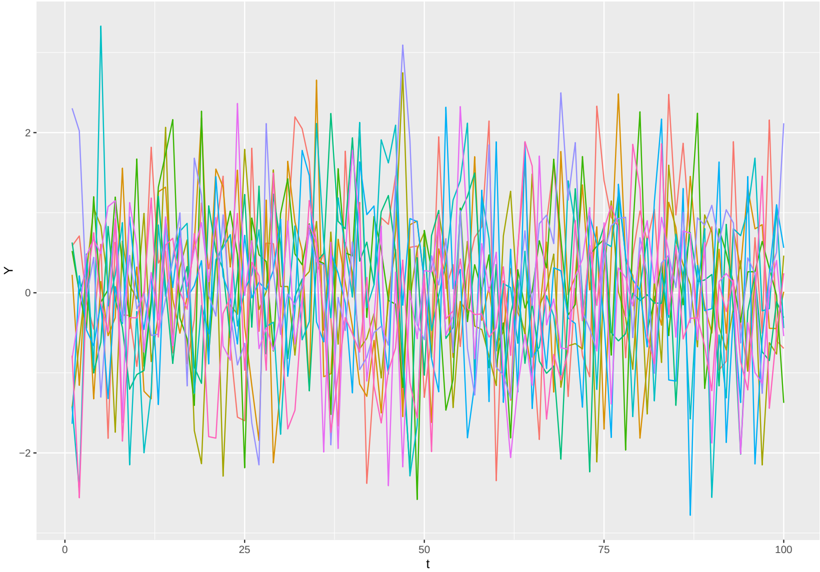
$$\text{Define } \boxed{Y_t = e_t} \text{ for all } t$$

- A white noise process satisfies:
  - $\text{Mean}(Y_t) = \text{constant}$
  - $\text{Var}(Y_t) = \text{constant}$
  - $\text{Auto-covariance} = 0$

White Noise: 1 Sample Path



White Noise: Many Sample Paths



White Noise: Implications

Since the data is random there is no point in modeling

Moving Average

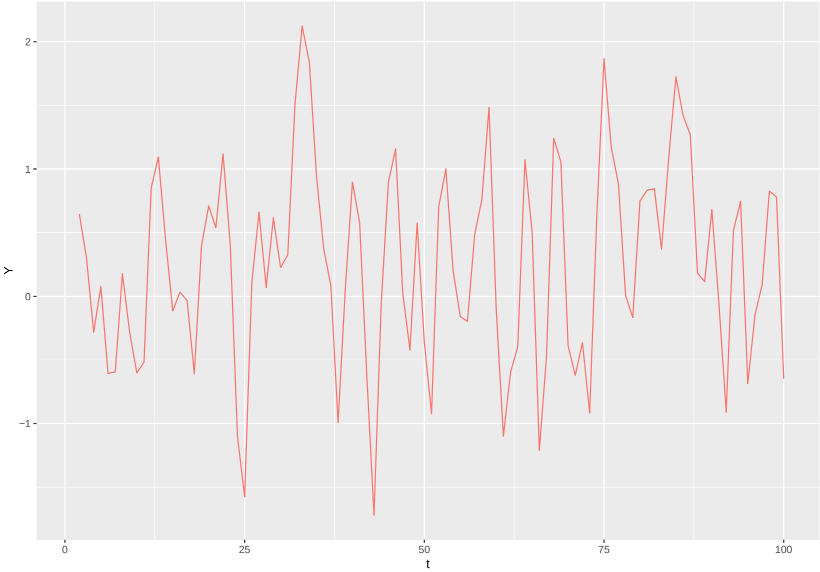
- Assume $e_t \sim N(0, 1)$ is a collection of IID random variables all following a normal distribution

$$\text{Define } \boxed{Y_t = \frac{e_t + e_{t+1}}{2}} \text{ for all } t$$

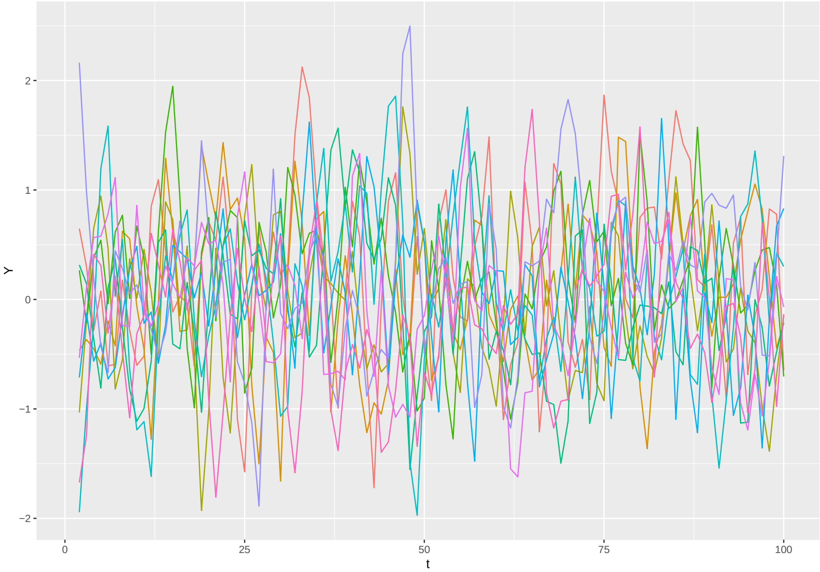
- Y_t is a type of moving average stochastic process

Big difference: there IS correlation between observations

Moving Average: 1 Sample Path



Moving Average: Many Sample Paths



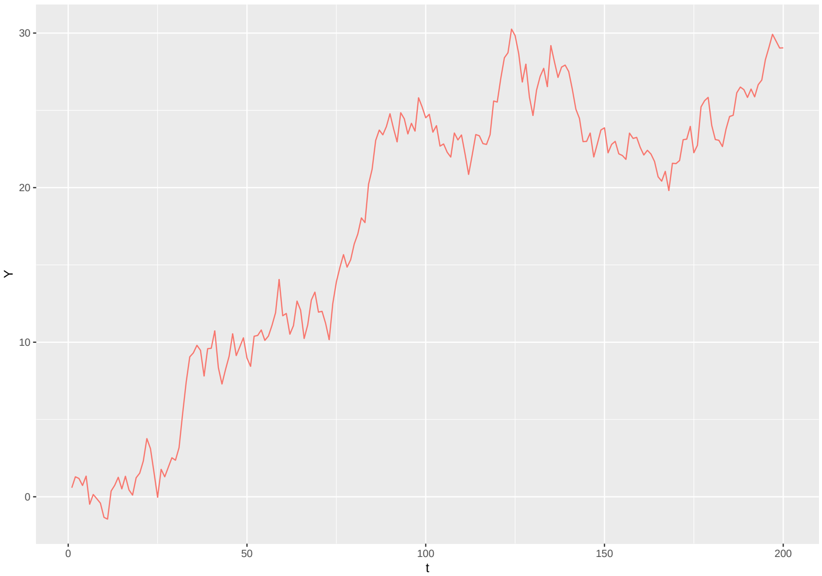
Random Walk

- Assume $e_t \sim N(0, 1)$ is a collection of IID random variables all following a normal distribution

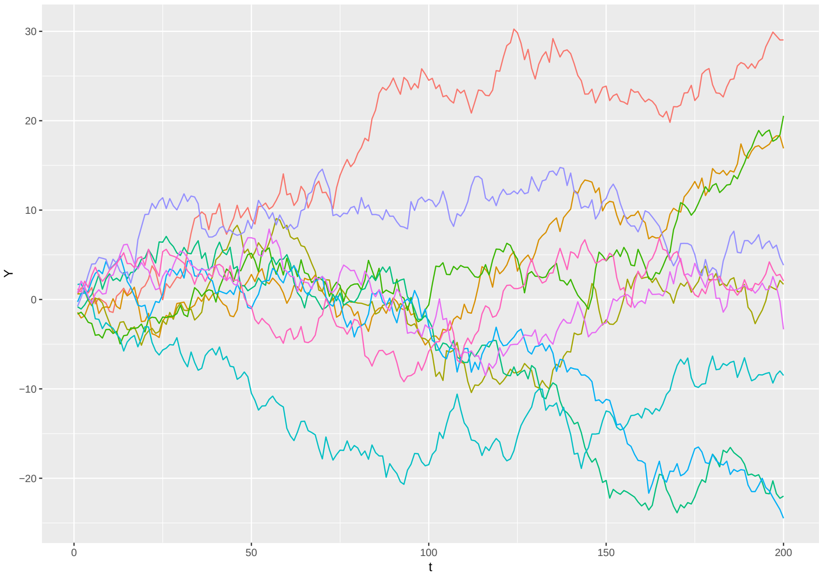
$$\text{Define } \boxed{Y_t = e_1 + e_2 + \ldots + e_t} \text{ for all } t$$

- Y_t is a random walk

Random Walk: 1 Sample Path



Random Walk: Many Sample Paths



Why Care about Stochastic Processes?

- We have ONE sample path of observations
- From this sample path we intend to infer the stochastic process that generated it
- We are NOT fitting lines to the data
- We are understanding the sample paths the process could take

Typical Process

- We observe a process to a specific point
- We determine what sample paths are likely as we look later in time

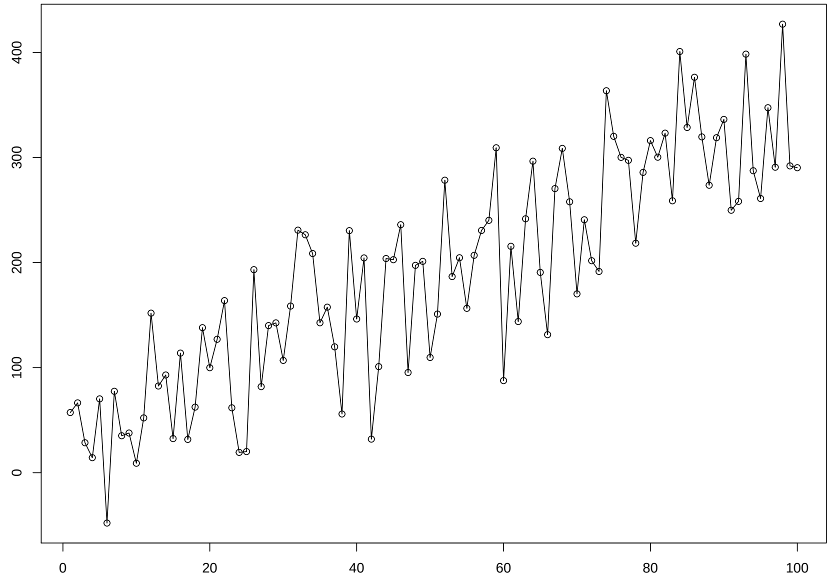
Stationarity

- Heavily mathematical definition
- Essence of it:
  - The mean (or expected value of the stochastic process) is constant
  - Covariance (or corelation between points only depends on distance between points and not on the value of t)
- Stationarity is an extremely common assumption in time series modeling!

Time Series: Linear Trend

- Assume stochastic process is $Y_t = \beta_0 + \beta_1 t + X_t$
- Assume $E[X_t] = 0$ and X_t is stationary
- This is just like linear regression so let's use OLS

Time Series: Linear Trend



Time Series

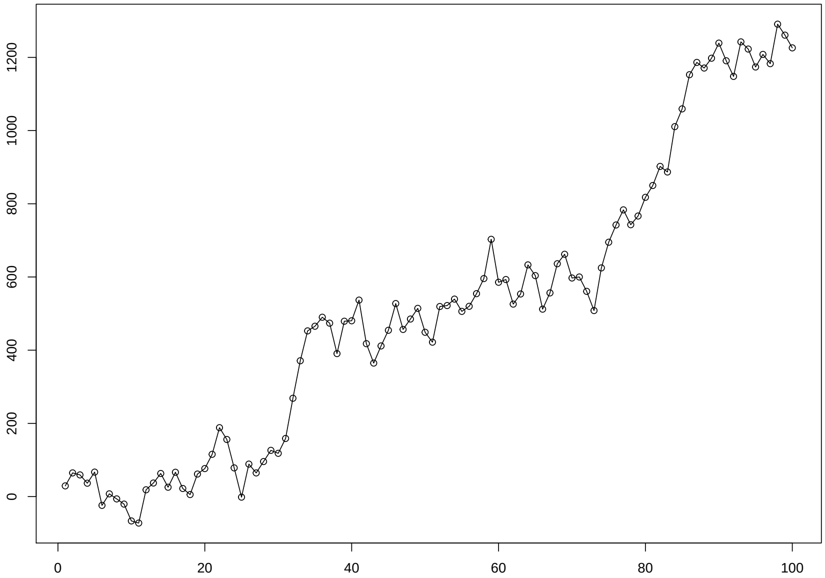
- Linear regression recovers the true parameters (close)

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    30.7     11.3        2.73 7.57e- 3
## 2 t               3.13     0.194     16.2  2.13e-29
```

Time Series: Linear Model for Random Walk

- What if we fit a random walk Y_t with a linear model?
- Put another way: what if we regress a random walk on time with a linear model?
- Good choice, bad choice?

Linear Model for a Random Walk



Linear Model for a Random Walk

- Linear regression fails. Hard
- Significant trend when there is NOT one
- Reminder: there is not a trend because the expected value of Y_t is 0

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -150.     22.9       -6.53 2.97e- 9
## 2 t               13.1     0.394     33.2  4.34e-55
```

What Happened?

- Spurious regression
- DO NOT regress non-stationary time series
- Our mistake
  - Coefficient is okay
  - Standard error is underestimated
- A random walk does not follow required linear regression assumptions!
  - Which key assumption does it not follow?