Ch. 1 : Autocorrelation

Contents

1 Random Sample vs Autocorrelation **2**
- 1.1 Forecasting Random Sample data 3
- 1.2 Examples of time series 9
- 1.3 Autocorrelation 13
- 1.4 Summary 22

2 ACF plot and the Assumption of Stationarity **24**
- 2.1 Sample ACF under no autocorrelation 28
- 2.2 Stationarity Assumption 33
- 2.3 Summary 41

3 Testing Mean **42**
- 3.1 Random Sample Case 43
- 3.2 Non-constant Trend and MA filter 56
- 3.3 Summary 62

1

Random Sample vs Autocorrelation

[ToC]

1.1 Forecasting Random Sample data

[ToC]

- idependent and identically distributed random variable (Random Sample from Normal)

$$e_t \sim N(\mu, \sigma^2) \qquad t = 1, \ldots, 100$$

```
X = rnorm(100, 2, 2)      #- 100 random num from N(mu=2,sd=2)
X
hist(X)

plot(X)
plot(X,type="l")
plot(X,type="o", ylim=c(-6,8) )

abline(h=2)
```

Forecasting iid

- Suppose you know the data $\{X_1, \dots, X_{100}\}$ is Random Sample from Normal distribution.

$$X_t = e_t \sim N(\mu, \sigma^2) \qquad t = 1, \dots, 100$$

- What can we say about the future X_t?
  - Assume the data comes from Normal distribution with unknown mean and variance.

CI and PI

- 95% Confidence Interval for μ

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

- 95% Prediction Interval for X_{101}

$$\bar{X} \pm 1.96\, S \sqrt{\frac{1}{n} + 1}$$

- 95% Prediction Interval for X_{102}

Forecasting iid

```
X = rnorm(100, 2, 2)
plot(X,type="o", ylim=c(-6,8), xlim=c(1,200) )

mean(X)
sd(X)

#- compute CI
upper.CI = mean(X) + 1.96 * sd(X) * sqrt(1/100)
lower.CI = mean(X) - 1.96 * sd(X) * sqrt(1/100)


#- compute PI
upper.PI = mean(X) + 1.96 * sd(X) * sqrt(1/100 + 1)
lower.PI = mean(X) - 1.96 * sd(X) * sqrt(1/100 + 1)

abline(h=upper.CI, col="red", lty=2)
abline(h=lower.CI, col="red", lty=2)

abline(h=upper.PI)
abline(h=lower.PI)
```

Forecasting iid 2

```
X2 = rnorm(100, 2, 2)

lines((101:200), X2,type="o", col="red" )   #- Ovrelay to existing plot
```

White Noise vs Random Sample (iid)

- Random Sample = iid = idependent and identically distributed random variable

$$\text{e.g.} \quad e_t \sim N(0,1) \quad \text{or} \quad e_t \sim U(0,1)$$

- WN = idependent, but may not be identically distributed random variable

$$e_t \sim WN(0,1)$$

- WN still have same mean and same variance.

- WN is like Random Sample without knoledge of the underlying distribution.

1.2 Examples of time series

[ToC]

Yearly Rainfall in LA (Cryer p2)

```
install.packages("TSA")   #- required for the first time on PC
library(TSA)              #- required every time you restart R

data(larain)              #- loads the dataset from TSA package
larain
is.ts(larain)             #- larain is in time series format

plot(larain, type="o")
```

Ex. 2

Chemical Coloring Process (Cryer p3)

```
data(color)    #- loads the dataset

plot(color,ylab="Color Property",xlab="Batch",type="o")
```

Ex. 3

Canadian Hare (Cryer p5)

```
data(hare)
plot(hare,ylab="Abundance",xlab="Year",type="o")
```

Ex. 4

Monthly Oil Filter Sales (Cryer p7)

```
data(oilfilters)
plot(oilfilters,type="o",ylab="Sales")

is.ts(oilfilters)  # is data in TS format?

plot(oilfilters,type="l",ylab="Sales")
points(y=oilfilters,  x=time(oilfilters),
                    pch=as.vector(season(oilfilters)))
```

1.3 Autocorrelation

[ToC]

In Canadian Hare Abundance Data

- May be this year's number of rabbit is correlated with last year's?

```
length(hare)

thisYear <- hare[2:31]     #- This year vs Last Year
lastYear <- hare[1:30]

plot(lastYear, thisYear,ylab="This Year",xlab="Previous Year",
                              xlim=c(-10,100), ylim=c(-10,100))
```

- There's autocorrelation of lag 1.
- This correlation will be useful in prediction for future values.
- How about lag 2?

```
cor(thisYear,lastYear)

thisYear <- hare[3:31]   #-- This year vs 2yrs ago
TwoYearsAgo <- hare[1:29]

cor(thisYear,TwoYearsAgo)

plot(TwoYearsAgo, thisYear,ylab="This Year",xlab="Previous Year",
                              xlim=c(-10,100), ylim=c(-10,100))
```

Autocorrelation in Other Examples

```
#-  Chemical Process
plot(color,ylab="Color Property",xlab="Batch",type="o")

cor( color[1:34], color[2:35])
plot(color[1:34], color[2:35], ylab="Color Property",
                               xlab="Previous Batch Color Property")


#- LA rain
cor(larain[1:114], larain[2:115])
plot(larain[1:114], larain[2:115], ylab="This year",xlab="Last year")
```

Formula for Correlation

- When you have measurement on two variables

$$(Y_1, X_1), \ldots, (Y_n, X_n)$$

Sample Correlation r

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{S_X} \frac{(Y_i - \bar{Y})}{S_Y}$$

What it measures is the sample correlation between

$$(Y_1, Y_2, Y_3, \ldots, Y_n)$$

vs

$$(X_1, X_2, X_3, \ldots, X_n)$$

Formula for Autocorrelation

- We need to measure correlation between

$$\begin{array}{ll} \text{[This Year]:} & (X_2, X_3, X_4, \dots, X_n) \\ & \text{vs} \\ \text{[Last Year]:} & (X_1, X_2, X_3, \dots, X_{n-1}) \end{array}$$

- So the formula for autocorrelation is Sample Correlation r

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{S_X} \frac{(X_{i-1} - \bar{X})}{S_X}$$

- This is called **Autocorrelation at Lag 1**.

ACF and ACVF

Given sequence of random variable $\{X_1, \ldots, X_n\}$,

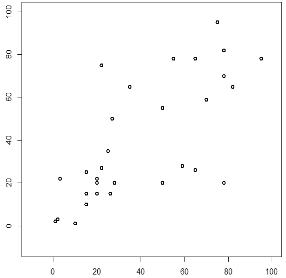
- ACF : AutoCorrelation Function (at lag h)

$$\rho(h) = \mathrm{COR}(X_t, X_{t-h})$$

- ACVF : AutoCoVariance Function (at lag h)

$$\Gamma(h) = \mathrm{COV}(X_t, X_{t-h})$$

From Hare Example



cor(hare[2:31], hare[1:30]) = 0.703

Cov and ACVF

- Sample Covariance

$$\hat{\rho} = \frac{COV(X_t, Y_t)}{\sqrt{V(X_t)V(Y_t)}}$$

- Sample Autocovariance Function

$$\hat{\Gamma}(h) = \frac{1}{n}\sum_{i=1}^{n-|h|}(X_t - \bar{X})(X_{t+|h|} - \bar{X})$$

$$\hat{\Gamma}(0) = V(X_t)$$

- ACF and ACVF is related as:

$$\hat{\rho}(h) = \frac{\hat{\Gamma}(h)}{\hat{\Gamma}(0)}$$

Properties

- ACF and ACVF is symmetric in h. (e.g. $\Gamma(h) = \Gamma(-h)$)
- $\rho(0) = 1$ and $\hat{\rho}(0) = 1$.
- Don't plot for h that is too big relative to n. ($n \geq 50$ and $h \leq n/4$)

1.4 Summary

[ToC]

---

1. Some Time Series data exibits **Autocorrelation**.

2. **Autocorrelation** at lag 1 is a correlation between this year's data points against last year's data points.

3. **Autocorrelation** can be detected by plotting autocorrelation function (ACF) at many lags.

4. If data is random sample (iid), then there should be no ACF, except at lag 0.

5. Autocovariance function ACVF and ACF are related as

$$\hat{\rho}(h) = \frac{\hat{\Gamma}(h)}{\hat{\Gamma}(0)}$$

because $\hat{\Gamma}(0) =$ Sample Variance of $X = S_X^2$.

2

ACF Plot and the Assumption of Stationarity

[ToC]

From Hare Example


```
library(TSA)
data(hare)

cor(hare[2:31], hare[1:30]) = 0.703
```

From Hare Example

```
cor(hare[2:31], hare[1:30])   #- lag 1
cor(hare[3:31], hare[1:29])   #- lag 2
cor(hare[4:31], hare[1:28])   #- lag 3

acf(hare)
?acf  #- look up help page

acf(hare, type = "covariance")

#-- warning:  TSA package overrites acf() function.
```

ACF and ACVF

```
acf(hare)                        # ACF

Rho <- acf(hare)                 # keep numbers from ACF
Rho

acf(hare, type="covariance")   # ACVF

Gam <- acf(hare, type="covariance")   # take numbers from ACVF
Gam

Gam[0]

var(hare)                        # not same as Gam[0]

var(hare) * 30 / 31              # same as Gam[0]
```

2.1 Sample ACF under no autocorrelation

[ToC]

- If your data was iid, X_t and X_{t+h} should be uncorrelated.
- Theoretical ACVF and ACF:

$$\begin{aligned} \Gamma(0) &= V(X_t) \quad &\text{and} \quad &\Gamma(h) = 0 \text{ for } h \neq 0. \\ \rho(0) &= 1 \quad &\text{and} \quad &\rho(h) = 0 \text{ for } h \neq 0. \end{aligned}$$

- Sample ACVF and ACF, $\hat{\Gamma}(h)$ and $\hat{\rho}(h)$ are estimating 0.
- Distribution of $\hat{\rho}(h)$ when $\rho(h) = 0$

$$\hat{\rho}(h) \sim N\left(0, \frac{1}{\sqrt{n}}\right) \qquad h \neq 0, \text{ under iid.}$$

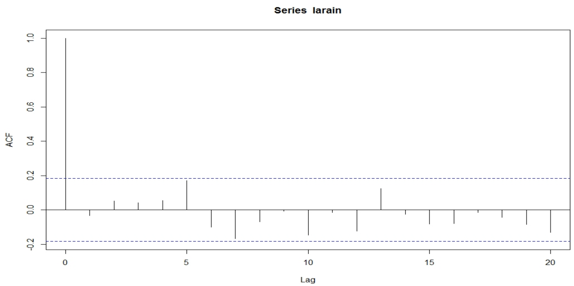
Monte Carlo Simulation

```
n = 40
X <- rnorm(n, 2, 2)
plot(X)

Rh <- acf(X)

Th.Rh <- c(1, rep(0,15))

plot(Rh, type="h", xlim=c(0,15), ylim=c(-0.4,1))
par(new=T)
plot(0:15, Th.Rh, type="p", col="red", xlim=c(0,15), ylim=c(-0.4,1))
```

```
#--- Put above in a loop ---

for (i in 1:100) {
  X <- rnorm(n, 2, 2)
  Rh <- acf(X, plot=FALSE)

  plot(Rh, type="p", xlim=c(0,15), ylim=c(-0.4,1))
  par(new=T)
  plot(0:15, Th.Rh, type="p", col="red", xlim=c(0,15), ylim=c(-0.4,1), xlab="",ylab="")
  par(new=T)
}
```

Diagnosis for LArain data

- If data is Random Sample, then plot of ACF should show almost all the bars within 95% CI under iid ($1.96/\sqrt{n}$).



- Blue dotted line in `acf()` $= \pm 1.96/\sqrt{n}$

```
data(larain)
plot(larain)

length(larain)     # this is n

acf(larain)

1.96/sqrt(115)     # size of the blue line
```

- More than 95% of acf plots are within the blue dotted line $\Rightarrow$ LArain data is white noise.

2.2 Stationarity Assumption

[ToC]

Condition needed to talk about ACF, ACVF

- ACVF of lag 1

```
plot( hare[2:31], hare[1:30], ylab="This Year", xlab="Lat Year" )
cor( hare[2:31], hare[1:30] )   #- lag 1
```

- We assumed $\text{cor}(X_2, X_1)$ is same as $\text{cor}(X_{31}, X_{30})$.
- That's a big assumption!

Weak Stationarity

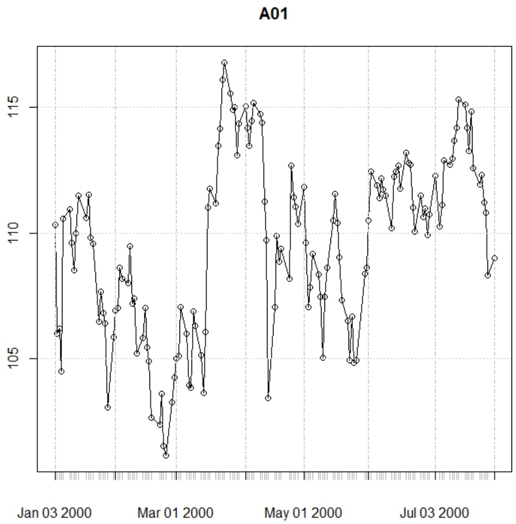
Series of r.v. $\{X_1, \ldots, X_n\}$ is called weakly stationary if

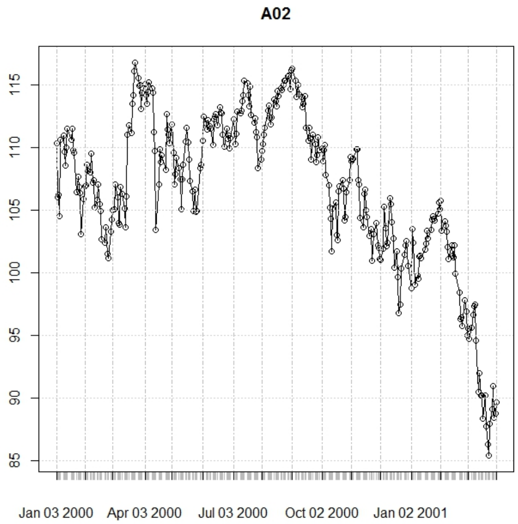
- $E(X_t)$ does not depend on t.
- $V(X_t)$ does not depend on t.
- $\text{cor}(X_t, X_{t+h})$ does not depend on t.

Stationary?

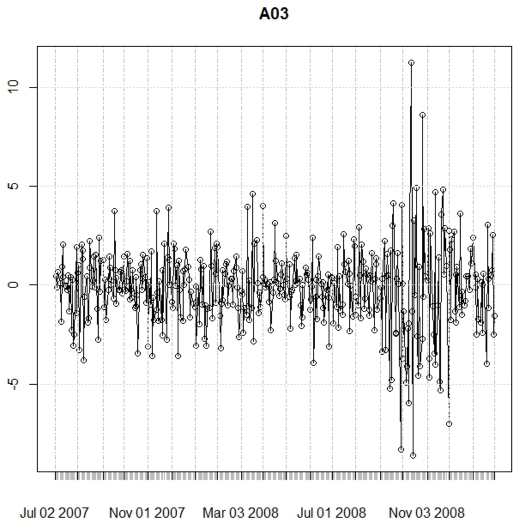
```
plot(hare,type="o")

plot(color,type="o")

plot(larain,type="o")

plot(oilfilters,type="o",ylab="Sales")
```

Stationary?

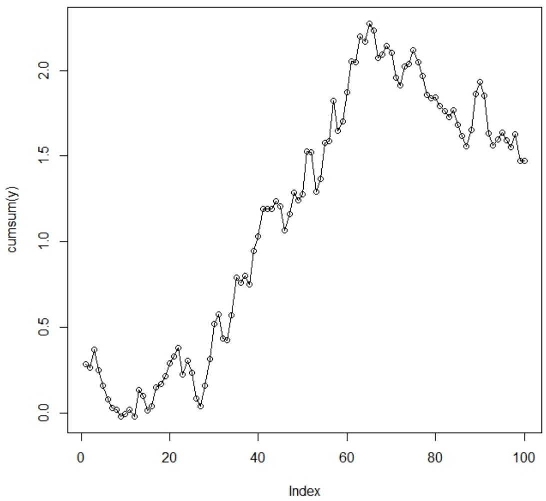

A01
115
110
105
Jan 03 2000
Mar 01 2000
May 01 2000
Jul 03 2000


A02
115
110
105
100
95
90
85
Jan 03 2000
Apr 03 2000
Jul 03 2000
Oct 02 2000
Jan 02 2001

Stationary?



A03
10
5
0
-5
Jul 02 2007
Nov 01 2007
Mar 03 2008
Jul 01 2008
Nov 03 2008



cumsum(y)
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80
100
Index

Warning:

- How 'stationary' it looks depend on scale of the plot.

```
plot(hare,type="o")
```

```
plot(hare,type="o", xlim=c(1917,1922))
```

Strong Stationarity

Series of r.v. $\{X_1, \ldots, X_n\}$ is called strongly stationary if

- Joint pdf of $\{X_1, \ldots, X_n\}$ are identical to joint pdf of $\{X_{t+1}, \ldots, X_{t+n}\}$ for all t.
- This is pretty strong assumption.
- Stationary = (weakly) stationary

Example of Non-sationariy process

- Series with Trend
- Series with non-constant variance
- Random Walk

2.3 Summary

[ToC]

---

1. To check if a time series is Random Sample (White Noise), then plot its ACF, see if 95% of them are between the blue dashed line.

2. The blue dashed line in `acf()` is $\pm 1.96/\sqrt{n}$.

3. For ACF and ACVF to be plotted and analyzed, the series must be **Weak Stationary**.

4. **Weak Stationarity** means
    - $E(X_t)$ is constant over time.
    - $V(X_t)$ is constant over time.
    - $\Gamma = cov(X_t, X_{t-|h|})$ does not depend on time.

Testing Mean

[ToC]

3.1 Random Sample Case

[ToC]

Suppose we observe Y_t, and model it as:

$$Y_t = \mu + X_t \qquad \begin{cases} \mu : & \text{Constant Trend (deterministic)} \\ X_t : & \text{Random Noise with mean 0, variance } \sigma^2 \end{cases}$$

How can we test if $\mu = 0$?

We will use $\bar{Y}$ to estimate μ, as usual.

```
mu = sample(c(0,2,1),1)
Y  = mu + rnorm(100,0,5)
plot(Y, type="o")

mean(Y)

abline(h=0)
abline(h=mean(Y), col="blue", lty=2)
```

CI for mean μ

- By CLT, we know how 'well' $\bar{Y}$ estimates μ.

$$\bar{Y} \sim N\Big(\mu, \frac{\sigma^2}{n}\Big)$$

- 95% CI for μ (large sample)

$$\bar{Y} \pm 1.96\frac{S}{\sqrt{n}}$$

- If 0 is outside of the CI, then we reject $H_0 : \mu = 0$ with 5% confidence level.

```
CI.u = mean(Y) + 1.96*sd(Y)/sqrt(100)
CI.u
CI.l = mean(Y) - 1.96*sd(Y)/sqrt(100)
CI.l

abline(h=c(CI.u,CI.l), col="red", lty=2)
```

Time Series Case

[top]

Suppose we observe Y_t, and model it as:

$$Y_t \quad = \quad \mu + X_t$$

$$\begin{aligned} \mu &: \quad \text{Constant Trend (deterministic)} \\ X_t &: \quad \text{Staionary Time Series} \end{aligned}$$

We assume that $EX_t = 0, \quad V(X_t) = \sigma^2$.

Note that if X_t has autocorrelation, then Y_t is also autocorrelated.

We will use $\bar{Y}$ to estimate μ, as usual.

When Y_t is time series,

Does property of $\bar{Y}$ changes? Expectation is

$$E(\bar{Y}) \;=\; E\Big(\frac{1}{n}\sum_{t=1}^{n} Y_t\Big) \;=\; \frac{1}{n}\sum_{t=1}^{n} E(Y_t)$$

We have

$$E(Y_t) \;=\; E(\mu + X_t) \;=\; \mu + E(X_t) \;=\; \mu,$$

therefore

$$E(\bar{Y}) = \mu.$$

i.e. $\bar{Y}$ is still unbiased estimator of μ.

Variance of sample mean

Variance of $\bar{Y}$ actually changes when Y has autocorrelation.

Since adding a constant does not change the variance,

$$\begin{aligned} V(\bar{Y}) &= V(\bar{Y} - \mu) \quad = \quad V\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)\right) \\ &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= V(\bar{X}) \end{aligned}$$

$$
\begin{aligned}
V(\bar{X}) &= V\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) \\
&= \frac{1}{n^2}\,\mathrm{Cov}\Big(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\Big) \\
&= \frac{1}{n^2}\Big[\text{ sum of Cov of all pairs } (X_i, X_j)\Big]
\end{aligned}
$$

$$
\frac{1}{n^2}\Big[\text{ sum of Cov of all pairs } (X_i, X_j)\Big]
$$

$$
\begin{aligned}
&= \frac{1}{n^2}\ \text{sum of Cov of pairs} \begin{array}{c|c}
 & \begin{matrix} X_1 & X_2 & \cdots & X_n \end{matrix} \\ \hline
\begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{matrix} & \begin{bmatrix} \ddots & & \cdots & \\ & \ddots & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \cdots & \end{bmatrix}
\end{array} \\
&= \frac{1}{n^2}\ \text{sum of} \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(-1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix} \\
&= \frac{1}{n^2}\sum_{h=-n}^{n}(n-|h|)\gamma(h)
\end{aligned}
$$

$$= \frac{1}{n^2}\sum_{h=-n}^{n}(n-|h|)\gamma(h) \quad = \quad \frac{1}{n}\sum_{h=-n}^{n}\left(1-\frac{|h|}{n}\right)\gamma(h).$$

So the variance of sample mean is different under the presence of autocorrelation.

If X_t were iid, then $\gamma(0)=1$, and $\gamma(h)=0$ for all $h \neq 0$. Then above reduces to

$$V(\bar{X}) \quad = \quad \frac{1}{n}\left(1-\frac{|0|}{n}\right)\gamma(0) \quad = \quad \frac{\sigma^2}{n}$$

Variance of Sample Mean

So we have

$$V(\bar{Y}) = \frac{\nu^2}{n} \qquad \text{where} \qquad \nu^2 = \sum_{h=-n}^{n} \left(1 - \frac{|h|}{n}\right)\gamma(h).$$

That means approximately,

$$\bar{Y} \sim \mathcal{N}\left(\mu, \nu^2\right).$$

Then the confidence interval for μ is

$$\bar{Y} \pm 1.96\sqrt{\frac{\nu^2}{n}}$$

- In practice, we don't know the true value of $\gamma(h)$, so we need to use the sample version $\hat{\gamma}(h)$.

$$\hat{\nu}^2 = \sum_{h=-\sqrt{n}}^{\sqrt{n}} \left(1 - \frac{|h|}{n}\right)\hat{\gamma}(h).$$

- Also, we can't really use $\hat{\gamma}(h)$ with h close to n. So sum goes from $-\sqrt{n}$ to $\sqrt{n}$ instead.

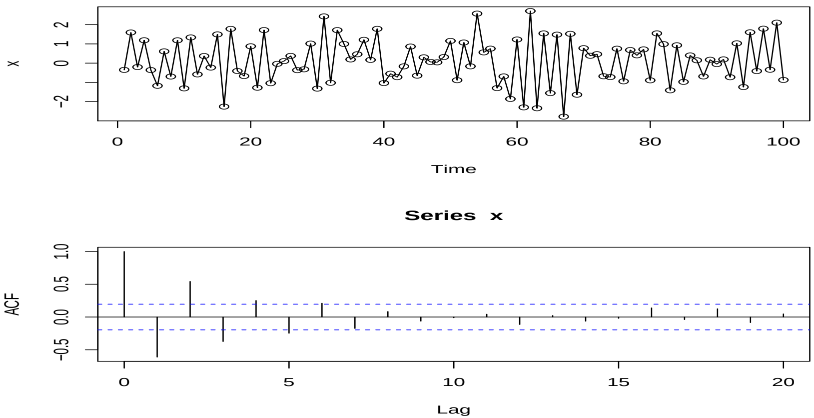
- Finally, modify the sum to:

$$\hat{\nu}^2 = \sum_{h=-\sqrt{n}}^{\sqrt{n}} \left(1 - \frac{|h|}{n}\right)\hat{\gamma}(h) = \gamma(0) + 2\sum_{h=1}^{\sqrt{n}} \left(1 - \frac{|h|}{n}\right)\hat{\gamma}(h).$$

In R:

```
data(color)
plot(color, type="o")
abline(h=mean(color), col="blue")
n = length(color)
n               # 35
sqrt(n)         # let's say 6

Ga <- acf(color, type="covariance")   #- extract ACVF values
str(Ga)
Ga.hat = Ga$acf

nu.sq <- Ga.hat[1] + 2*sum( (1-(1:6)/n)*Ga.hat[2:7] )   #- sqrt n is 6, which is 7th element in Ga.hat

mean(Y)
1.96*sqrt(nu.sq/n)

1.96*sqrt(var(color)/n)

CI.u <- mean(color) + 1.96*sqrt(nu.sq/n)
CI.l <- mean(color) - 1.96*sqrt(nu.sq/n)

plot(color, type="o")
abline(h=mean(color), col="blue")
abline(h=c(CI.u, CI.l), col="red", lty=2)
```

Example: Is the mean significantly different from zero? ($\bar{Y} = 0.117$)



Is the mean of this time series zero?

$$\bar{X} = 0.117 \qquad \hat{\nu}^2 = 0.822 \qquad 1.96 * \sqrt{\frac{\hat{\nu}^2}{n}} = 0.178.$$

Yes, zero mean is plausible. ($\bar{Y}$ not significant)

3.2 Non-constant Trend and MA filter

[ToC]

Suppose Y_t is the observation. The Model says:

$$Y_t = m_t + X_t$$

$$\begin{array}{rcl} m_t & : & \text{Trend Component (deterministic)} \\ X_t & : & \text{Stationary Time Series} \end{array}$$

We assume that $EX_t = 0$. If it is not, we can always absorb it in m_t.

Note that Y_t is not stationary.

Linear MA filter

Let

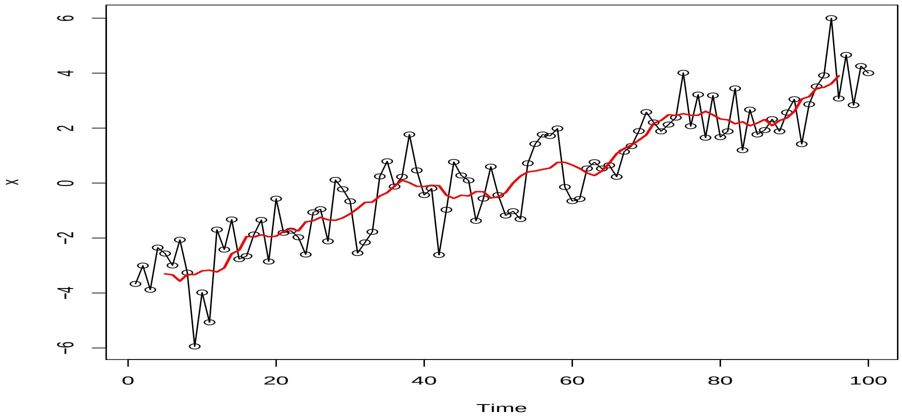
$$\begin{aligned} W_t &= \frac{1}{(2q+1)} \sum_{j=-q}^{q} Y_{t-j} \\ &= \frac{1}{(2q+1)} \sum_{j=-q}^{q} (m_{t-j} + X_{t-j}) \end{aligned}$$

Suppose the trend is linear over $(t-q, t+q)$.
i.e. $m_t = a + bt$

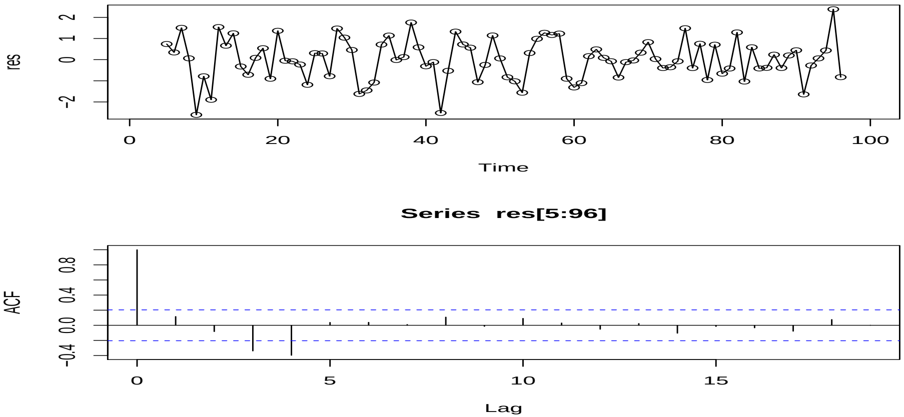
$$
\begin{aligned}
W_t &= \frac{1}{(2q+1)} \sum_{j=-q}^{q} (m_{t-j} + X_{t-j}) \\
&= \frac{1}{(2q+1)} \sum_{j=-q}^{q} (a + b(t-j) + X_{t-j}) \\
&= a + bt + \underbrace{\frac{1}{(2q+1)} \sum_{j=-q}^{q} (-bj)}_{0} + \underbrace{\frac{1}{(2q+1)} \sum_{j=-q}^{q} X_{t-j}}_{\bar{X}_q}
\end{aligned}
$$

The last term $\bar{Y}_q$ should be small because $EY_t = 0$.

Example:



Example:



res
-2
0
1
2
0
20
40
60
80
100
Time
Series res[5:96]
ACF
-0.4
0.0
0.4
0.8
0
5
10
15
Lag

```
t <- 1:100
Y <- -3 +.07*t + arima.sim(n = 100, list(ma = c(.4, .2) ) )

m <- filter(Y, rep(1/9, 9))
res <- Y-m

plot(Y, type="o")
lines(m, lwd=2, col="red")

layout(c(1,2))            #- change layout of the plot window
plot(res, type="o" )
acf(res[5:96])
layout(c(1,1))
```

3.3 Summary

[ToC]

- For random sample, 95% CI for the mean is

$$\bar{X} \pm 1.96 \frac{S}{\sqrt{n}}$$

- For Time Series Data with autocorrelation, 95% CI for the mean is

$$\bar{X} \pm 1.96 \sqrt{\frac{\hat{\nu}^2}{n}} \qquad \text{where} \qquad \hat{\nu}^2 = \hat{\gamma}(0) + 2 \sum_{h=1}^{\sqrt{n}} \Big(1 - \frac{|h|}{n}\Big) \hat{\gamma}(h).$$

$\hat{\gamma}(h)$ is ACVF at lag h.