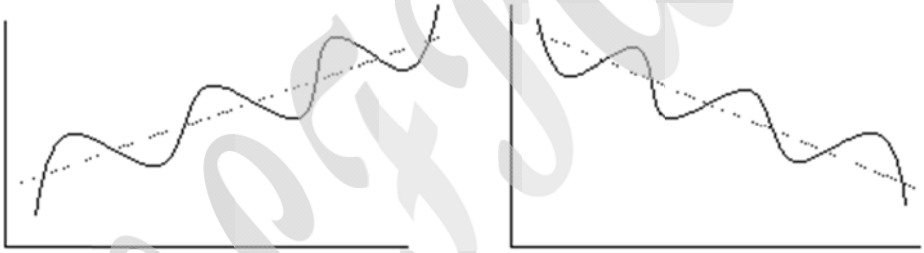
시계열 데이터

시계열(time series) 데이터는 관측치가 시간적 순서를 가지게 된다. 일정 시점에 조사된 데이터는 횡단(cross-sectional) 자료라 한다. ○○전자 주가, △△기업 월별 매출액, 소매물가지수, 실업률, 환율 등이 시계열 자료이다. $\{Y_t; t = 1,2,...,T\}$

시계열 데이터 4 가지 component $\{Y_t; t = 1,2,..,T\}$

- 경향(Trend): 데이터가 증가(감소)하는 경향이 있는지 혹은 안정적인지 알 수 있다. 직선의 기울기가 있는가?
- 주기(cycle): 일정한 주기(진폭)마다 유사한 변동이 반복된다. (sine, cosine 곡선)
- 계절성(seasonality): 주별, 월별, 분기별, 년별 유사 패턴이 반복된다.
- 불규칙성(irregular): 일정한 패턴을 따르지 않는다.
-

$$Y_t = Trend + Cycle + Seasonality + Irregular$$



시계열 형태

①white noise process

평균이 0 이고 분산이 σ^2 인 동일분포로부터 독립적으로(iid) 얻어진 시계열 데이터 $\{Y_t\}$ 을 백색 잡음(white noise) process 라 한다. 백색 잡음 데이터의 평균 수준을 μ 라 하면 이 시계열 데이터의 모형은 $Y_t = \mu + e_t$ 라 쓸 수 있다.

②stationary process

$F(y_{t_1}, y_{t_2}, ..., y_{t_n}) = F(y_{t_1+k}, y_{t_2+k}, ..., y_{t_n+k})$ 이면 시계열 데이터 $\{Y_t\}$ 를 strongly stationary process(강한 정상성)이라 한다. 일정한 기간의 종속변수 결합밀도함수는 동일한 분포를 가진다는 것을 의미한다.

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

다음 조건을 만족하는 시계열 데이터 $\{Y_t\}$는 weakly stationary process(약한 정상성)라 정의한다.

(1)평균이 일정하다. $E(Y_t) = \mu$

(2)분산이 존재하며 일정하다. $V(Y_t) = \gamma(0) < \infty$

(3)두 시점 사이의 자기 공분산(auto-correlation)은 시간의 차이에 의존한다.

$COV(Y_t, Y_{t-j}) = COV(Y_s, Y_{s-j}) = \gamma(j), for j \neq s$

가정 파괴 진단도구

가정의 만족은 모형의 타당성과 연결되므로 이를 검정해야 한다. 이에 대한 4 개의 그래프 기법은 다음과 같다.

> 순서 그림 Sequence plot, Y_i vs i

> 시차 그림 Lag plot, Y_i vs Y_{i-1}

> *히스토그램 Histogram => 분포의 형태, 봉우리 개수*

> *확률그림 Normal probability plot => 데이터 순서통계량 vs. 이론적 순서통계량*

Non-randomness

만약 무작위성이 (randomness) 무너진다면 : 확률표본 (random sample) , *iid*

> 모든 통계적 분석이 유효하지 않는다.

> 불확실성에 대한 측정 유효하지 않는다.

> 데이터에 근거한 판단 (표본 크기, 허용 오차) 유효하지 않는다.

> y = constant + error 유효하지 않는다.

> 모수 추정의 의미가 없다.

그래프 도구 : 시간플롯 time plot

o (그리기) X-축을 시간, Y-축을 관심 시계열 데이터 관측치

o (활용 1) 주기, 추세, 계절성

o (활용 2) 위치모수 (주기의 중심) 변화, 분산 (주기의 폭) 변화, 이상치 진단

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

데이터 IC.xls

Date: Time period (1-30) of the study (from 3/18/51 to 7/11/53)

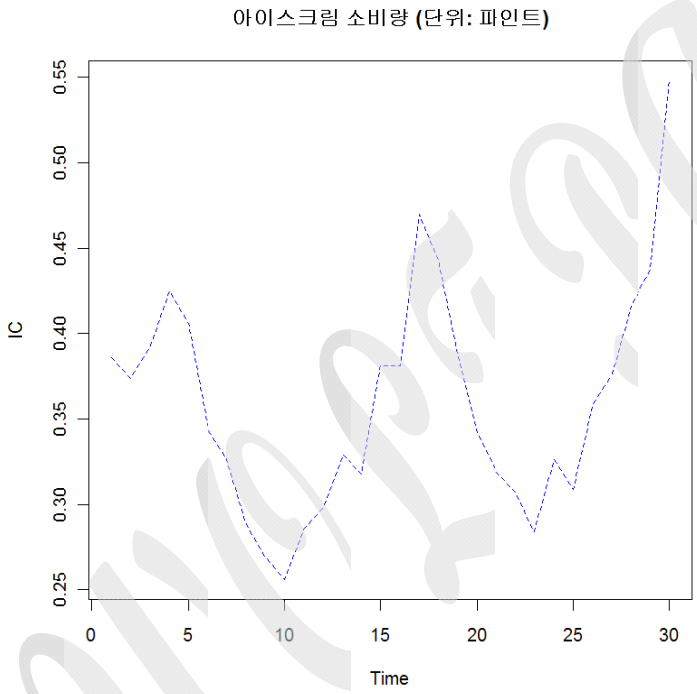
IC: Ice cream consumption in pints per capita

Price: Price of ice cream per pint in dollars

Income: Weekly family income in dollars

Temp: Mean temperature in degrees F.

```
ds=read.csv("ic.csv")
ts.ic=ts(ds[2:2], start=c(1,1), frequency=1)
plot(ts.ic, lty="dashed", col="blue",
 main="아이스크림 소비량 (단위: 파인트)")
```



o 추세 : 직선적 증가

o 계절성 : 주기 13

o 위치 모수 일정 : 위치 모수가 변하는 것은 주기의 점프가 있을 때

o 분산 (주기의 폭) 일정 => 시계열 데이터는 안정적 stationary

Professor Kwon, Sehyug | Dept. of Statistics, HANNAM Univ. 010.6365.7622 http://wolfpack.hnu.ac.kr

시계열 데이터 랜덤

(1) white noise 검정

시계열 데이터가 백색잡음이면 패턴이 존재하지 않아 시계열 모형 분석 적합하지 않음

귀무가설 : 시계열 데이터는 백색잡음이다.

대립가설 : 백색잡음이 아니다.

```
library(normwhn.test)
whitenoise.test(ts.ic)
```

```
[1] "tMN"
[1] 16.80531
[1] "test value"
[1] 4.364978e-17
```

귀무가설이 기각되어 백색잡음이 아니다.

(2) RUN 검정 (Dichotomous 개념)

관측치의 크기 부호가 바뀐 회수를 RUN 이라 정의한다.

+++++----- : Runs=2, 패턴존재

+-+-+-+-+- : Runs=10, 패턴 존재

귀무가설 : 관측치는 랜덤이다.

대립가설 : 랜덤이 아니다.

검정통계량 : RUNs 개수

```
x=as.matrix(subset(as.data.frame(sign(ts.ic-lag(ts.ic,1))), ts.ic!=0))
x.f=factor(x)
library(tseries)
runs.test(x.f)
```

```
        Runs Test

data:  x.f
Standard Normal = -1.9084, p-value = 0.05634
alternative hypothesis: two.sided
```

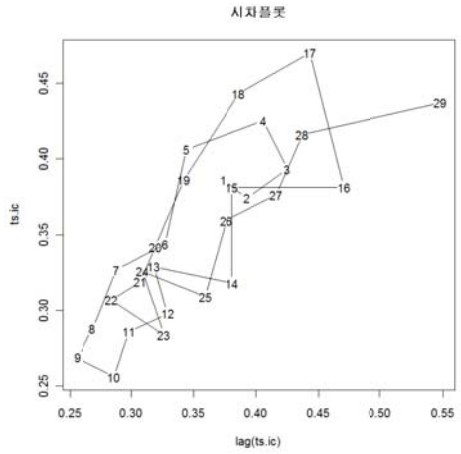
귀무가설이 채택되어 랜덤이다.

그래프 도구 : 시차 플롯 lag plot

o (그리기) X-축을 $Y_{(t-1)}$, Y-축을 $Y_{(t)}$

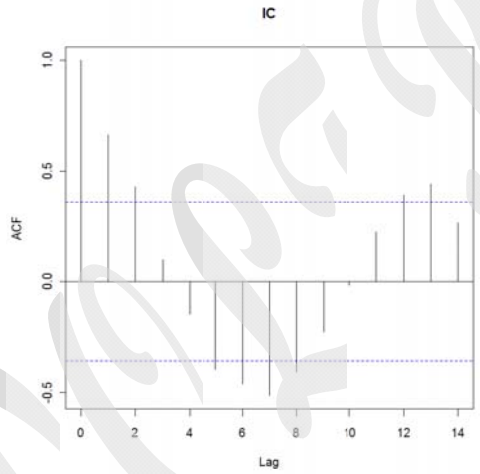
o (활용) 관측치의 랜덤 확인, 일정한 패턴을 가지면 랜덤 아님

```
plot(ts.ic~lag(ts.ic), main="시차플롯")
```

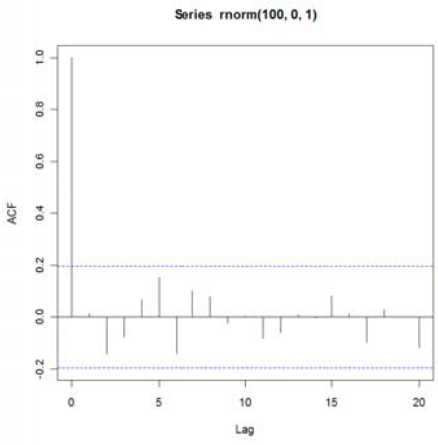


자기상관 autocorrelation.

자기상관은 관측치 Y_i와 Y_{i-k}의 상관계수이다. k는 시차이다.




```
acf(ts.ic)
```




```
acf(rnorm(100,0,1))
```

자기상관으로 인한 "무작위성"이 무너진다면,

> 근접 데이터는 상관되어 있다.
> n개의 독립적 특성을 가지지 못한다.
> 발견되지 못하는 "junk" 이상치 outlier 있음.
> 발견되지 않은 정보량 많은 관측치 있음.

이동평균법

자신의 m 개 관측치 평균으로 시계열 자료 $\{Y_t\}$의 패턴 인식
가중치는 1/m 으로 동일하다.
이를 이용하여 미래 값 $\{Y_{t+1}\}$ 예측한다.

$$MA = \frac{\sum \text{최근 } m \text{개자료}}{m}$$

$$\hat{Y}_{t+1} = MA_{t,m} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-m+1}}{m}$$ (다음 1 기만 예측 가능)

M 의 결정

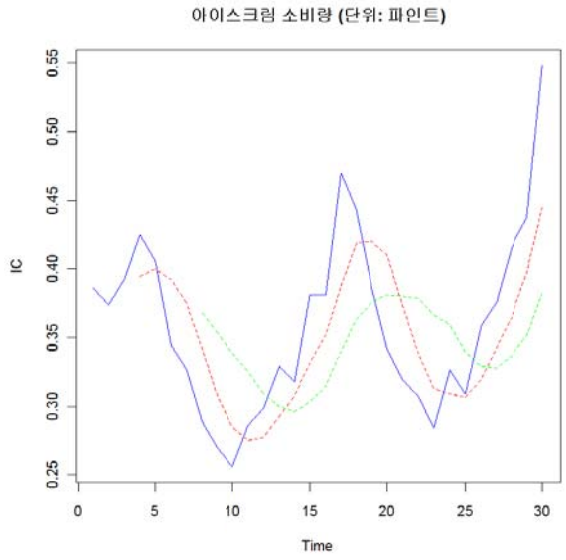
일반적으로 주기를 m 으로 놓는다.
주가의 경우 5 일, 20 일, 60 일, 120 일, … 이동평균을

이동평균법 특징

m 이 클수록 주기의 영향은 없어지고 직선에 가까워짐, Trend(경향)을 보는데 활용
작은 m 은 단기 예측, 큰 주기 m 은 장기 예측에 사용
주가 예측에 가장 많이 이용, 그러나 예측보다는 (실제 예측 가능은 다음 1 기) 추세분석에 가까움

예제 데이터

```
library(TTR)
fit.ic=SMA(ts.ic,n=4)
fit2.ic=SMA(ts.ic,n=8)
plot(ts.ic, col="blue", main="아이스크림 소비량 (단위: 파인트)")
lines(fit.ic, col="red", lty="dashed")
lines(fit2.ic, col="green", lty="dashed")
```



o 주가 4 는 한달, 주기 8 은 2 달 => 이제 감소할 것이다.

지수평활법 Simple Exponentially Smoothing

o 단순지수평활법은 경향이나 계절성이 없을 때 사용한다. (가중평균 weighted mean)
o 평활 가중치 값의 설정이 다소 주관적이나, 계산이 간편함

상수모형 $Y_t = \beta_0 + \varepsilon_t$ **: 사인 곡선, 시간 추세 없음**

시간 변동 모형 $Y_t = \beta_{0,t} + \varepsilon_t$

Locally 동일한 평균을 가지나 globally 평균 차이 보임

단순지수 평활 통계량 S_t **활용**

1) Y_{t+1} 예측치
2) $\beta_{0,t}$ 의 추정치
3) Y_{t+1} 예측치인 S_t 의 신뢰구간은 가중최소제곱법의 특수한 경우가 지수평활법 예측이므로

초기치 평활값 선택 초기평활 값 $S_0 = \frac{\sum_{i=1}^{T} y_i}{T}$ 이고 일반적으로 $T = 6$ 혹은 $T = n/2$

가중치 결정

지수평활법은 현재에 가까운 관측치에 높은 가중치를 주기 위하여 0.05 에서 0.3 사이의 값을 준다. (다른 측면에서 보면 μ 가 시간에 따른 변화가 느리기 때문이다)

-클수록 최근 관측치 영향이 크다.

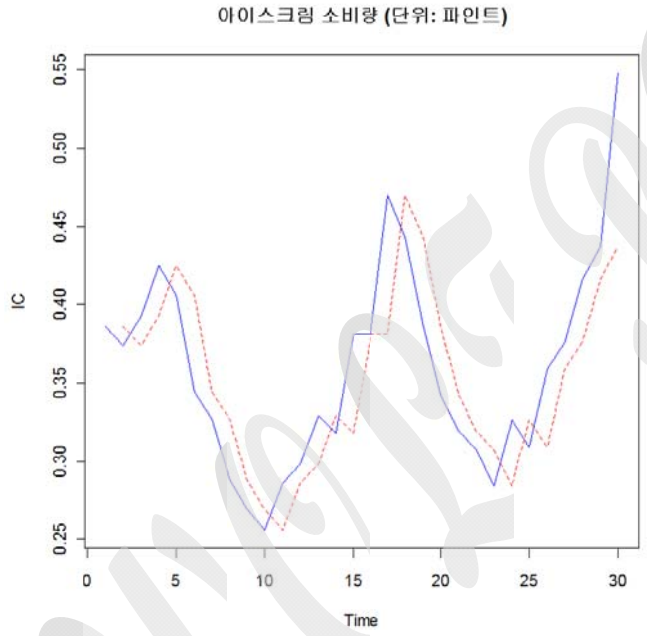
-일반적으로 0.05 와 0.3 사이의 값

-가중치 선택 : 모형 적합 정도를 나타내는 통계량을 이용하여 trial and error 방법

```
fit.s=HoltWinters(ts.ic,gamma=F,beta=F)
plot(ts.ic, col="blue", main="아이스크림 소비량 (단위: 파인트)")
lines(fitted(fit.s)[,1], col="red", lty="dashed")
```

```
> fit.s$SSE
[1] 0.0484111
```

o beat=F 제외하면 이중지수평활법 =>




```
> predict(fit.s, n.ahead=4)
Time Series:
Start = 31
End = 34
Frequency = 1
          fit
[1,] 0.5479927
[2,] 0.5479927
[3,] 0.5479927
```

주가 예측

임의의 기업 주가 (각자 찾기) 2010 년 1 월 2 일부터 2012 년 5 월 18 일까지 데이터를 이용하자.

(1) 이동평균법(m=6, 20, 60, 120)에 의해 장, 단기 주가 예측하시오.

(2) 지수평활법 이용하여 다음 **일주일** 예측하시오.