## ARMA 개념

ARIMA(Auto-Regressive Integrated Moving-Average) 모형은 시계열 데이터 $\{Y_t\}$ 의 과거치(previous observation $Y_{t-1}, Y_{t-2}, ...$)들이 설명 변수인 AR 과 과거의 오차항($e_{t-1}, e_{t-2}, ...$)들이 설명변수인 MA 모형의 합성어이다.

## AR(1) 모형

ARIMA 모형에 대한 개념 파악을 위하여 가장 간단한 AR(1) 모형을 먼저 살펴 보자.

### 용어와 기호

AR 모형은 아래 가설에 의해 제안되었다.

○과거의 패턴이 지속된다면 시계열 데이터 관측치 $Y_t$ 는 과거 관측치 $Y_{t-1}, Y_{t-2}, Y_{t-p}, ...$ 에 의해 예측할 수 있을 것이다.

○어느 정도의 멀리 있는 과거 관측치까지 이용할 것인가? 그리고 멀어질수록 영향력을 줄어들 것이다. 이런 상황을 고려할 수 있는 가중치를 사용해야 하지 않을까?

**AR(1) 모형:** $Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t$, $e_t \sim iid\ N(0, \sigma^2)$

만약 시계열 데이터가 서로 독립이고 유한인 평균과 분산을 갖는 동일 분포를 따르면(iid) 이 데이는 white noise(백색 잡음)이라 한다. 만약 평균이 0, 분산이 $\sigma^2$ 인 정규분포를 따른다면 이를 Guassian white noise 라 한다. $\{Y_t\}$ 대신 $\{Y_t - \mu\}$ 를 사용한 이유는 평균을 0 으로 하기 위함이다. $\mu$ 는 시계열 데이터의 총 평균(grand mean)에 해당된다.

만약 $\{Y_t\}$ 를 $\mu 0$ 가 되게 shift 하면 AR(1) 모형은 $Y_t = \rho Y_{t-1} + e_t$ 이고 개념 설명을 위하여 가장 많이 사용된다. 이를 일반화 하면 AR($p$) 모형은 다음과 같다.

**AR($p$) 모형:** $Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + e_t$, $e_t \sim iid\ N(0, \sigma^2)$

AR(1) 모형을 이를 다시 쓰면 다음과 같다.

$$Y_t = \mu + e_t + \rho e_{t-1} + \rho^2 e_{t-2} + \rho^3 e_{t-3} + ... + \rho^{t-1} e_1 + \rho^t (Y_o - \mu)$$

즉 AR(1) 모형이더라도 과거의 흔적을 모두 모함하고 있다. AR($p$)도 MA($\infty$) 모형으로 쓸 수 있다.

**MA($\infty$) 모형:** $Y_t = \mu + e_t + \beta e_{t-1} + \beta^2 e_{t-2} + \beta^3 e_{t-3} + \ldots = \mu + \sum_{j=0}^{\infty} \beta^j e_{t-j}$

**$\{Y_t\}$ 분산과 공분산**

**공분산** $\gamma(j) = \mathrm{cov}(Y_t, Y_{t-j})$, **분산** $\gamma(0) = \mathrm{Var}(Y_t)$

분산, 공분산 개념은 시계열 데이터에 적절한 AR, MA 모형을 찾는 함수인 ACF, PACF, IACF 에 이용된다. (다음 절에서 상세히 논한다.) 앞에서 우리는 AR(1)을 MA($\infty$)로 쓸 수 있음을 알았다, 이 사실을 이용하면 AR(1) 모형을 따르는 시계열 데이터 $\{Y_t\}$의 분산과 공분산을 구하면 다음과 같다.

공분산 $\gamma(j) = \mathrm{cov}(Y_t, Y_{t-j}) = \rho^j Var(Y_t)$, 분산 $\gamma(0) = \mathrm{Var}(Y_t) = \dfrac{\sigma^2}{1-\rho^2}$

그러므로 $\sigma^2$의 추정치는 $\hat{\sigma}^2 = \gamma(0)(1-\hat{\rho}^2)$ 이다.

**예측(Forecasting)**

AR(1)의 경우 $\rho$을 추정하면 $\{Y_{t-1}, Y_{t-2}, \ldots\}$ 예측치를 다음과 같이 구할 수 있다

(n+1) 시점 예측치: $\hat{Y}_{n+1} = \hat{\mu} + \hat{\rho}(Y_n - \hat{\mu})$ ($\because e_{t+1}$의 평균은 0 이기 때문이다) 즉,

$\mu = 100$ 이고 $\rho = 2/3$로 추정되었다면 $\hat{Y}_{n+1} = 100 + 2/3(Y_n - 100)$

예측 오차(forecasting error) $Y_{n+1} - \hat{Y}_{n+1} = e_{n+1}$

(n+2) 시점 예측치 $\hat{Y}_{n+2} = \hat{\mu} + \hat{\rho}^2(Y_n - \hat{\mu})$

(n+2) 시점 예측 오차 $Y_{n+2} - \hat{Y}_{n+2} = e_{n+2} + \hat{\rho} e_{n+1}$

**Backshift Notation**

$B(Y_t) = Y_{t-1}$, $B^2(Y_t) = Y_{t-2}$, …, $B^p(Y_t) = Y_{t-p}$

$Y_t - \mu = \rho(B(Y_t) - \mu) + e_t \Rightarrow (1-B)Y_t = \mu + \rho\mu + e_t$

만약 $\mu = 0$ 이면 $AR(1) \Rightarrow (1-B)Y_t = e_t$

# ARIMA 모형

## Process 정의

①white noise process

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

만약 $Y_0 = \mu$ 라 하면 $Y_t = Y_0 + e_1 + e_2 + ... + e_t$ 가 되며 $\{Y_t\}$ 을 random walk process 라 한다. $\{Y_t\}$ 는 동일한 분포를 가지며 서로 독립이라는 가정이다.

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.

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.

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

정상적 확률 모형(시계열 데이터 $\{Y_t\}$ 는 확률 변수)의 대표적인 것이 AR, MA, ARMA 모형이다.

### ARMA 모형

①AR(p) 모형

시계열 데이터 $\{Y_t\}$ 에서 시점 $t$ 의 관측치 $Y_t$ 가 과거 관측치 $Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ 들에 의해 설명될 때 AR(p) (차수가 p 인 Auto-Regressive, 자기회귀) 모형을 따른다고 한다.

$$Y_t \sim AR(p) \blacktriangleright Y_t = u + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + e_t$$

②MA(q) 모형

시계열 데이터 $\{Y_t\}$ 에서 시점 $t$ 의 관측치 $Y_t$ 가 과거 오차 $e_{t-1}, e_{t-2}, \ldots, e_{t-q}$ 들에 의해 설명될 때 MA(q) (차수가 q 인 Moving-Average 이동평균) 모형을 따른다고 한다.

$$Y_t \sim MA(q) \blacktriangleright Y_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \ldots - \beta_q e_{t-q}$$

③ARMA(p, q) 모형

시계열 데이터 $\{Y_t\}$ 에서 시점 $t$ 의 관측치 $Y_t$ 가 과거 관측치 $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ 들과 과거 오차 $e_{t-1}, e_{t-2}, \ldots, e_{t-q}$ 들에 의해 설명될 때 ARMA(p, p) (차수가 p, q 인 Auto-Regressive and Moving Average) 모형을 따른다고 한다.

$$Y_t = \mu + \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - \ldots - \alpha_p Y_{t-p} + e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} + \ldots - \beta_q e_{t-q} + e_t$$

**Stationarity and Invertibility**

MA(∞) 모형은 언제나 정상적(stationary)이다. why?

AR 모형 $Y_t = u + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + e_t$ 은

$1 - \alpha_1 M - \alpha_2 M^2 - \ldots - \alpha_p M^p = 0$ 의 방정식을 만족하는 근들의 절대값이 모두 1 보다 클 경우 stationary 하다. 정상적인 AR(p) 모형은 MA(∞) 모형으로 변환할 수 있음을 의미한다. 정상적인 process 인 경우

- $\{Y_t\}$ 는 $e_t, e_{t-1}, e_{t-2}, \ldots$ 으로 표현할 수 있으며,
- $\{Y_t\}$ 에 대한 $e_t, e_{t-1}, e_{t-2}, \ldots$ 들의 영향은 시점이 멀어질수록 줄어든다.
- 그러므로 $Y_{t+1}$ 에 대한 예측치를 구할 경우 $e_0 = 0$ 으로 사용해도 무방하다.

**Invertibility**

$Y_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \ldots - \beta_q e_{t-q}$ MA(q) 모형에서

$1 - \beta_1 M - \beta_2 M^2 - \ldots - \beta_q M^q = 0$ 의 방정식을 만족하는 근들의 절대값이 모두 1 보다 클 경우 MA 모형은 Invertibility 하다. 이 말은 AR(∞)모형으로 변환할 수 있다는 것이다.

- $\{Y_t\}$ 를 AR(∞)로 표현할 수 있으며, 즉 $Y_{t-1}, Y_{t-2}, \ldots$ 들로 표현되며
- $\{Y_t\}$ 에 대한 $Y_{t-1}, Y_{t-2}, \ldots$ 들의 영향은 시점이 멀어질수록 줄어든다.

## 상관 함수

시계열 자료 $\{Y_t\}$의 상관 함수는 acf, pacf, iacf 가 있는데 이는 ARMA 모형 진단에 사용된다.

### Auto Correlation Function (ACF)

자기상관함수(ACF)는 다음과 같이 정의한다.

$\rho(j) = \dfrac{\gamma(j)}{\gamma(0)} = \dfrac{Cov(Y_t, Y_{t-j})}{VAR(Y_t)}$ 그러므로 $\rho(0) = 1$, $\rho(j) = \rho(-j)$

**MA(1) 경우:** $Y_t = e_t - \beta_1 e_{t-1}$

$\gamma(0) = V(Y_t) = (1+\beta_1^2)\sigma^2$, $\gamma(1) = COV(Y_t, Y_{t-1}) = -\beta_1\sigma^2$, 그러나

$\gamma(2) = \gamma(3) = \gamma(4) = \ldots = 0$

그러므로 이를 요약하면 MA(q) 모형의 경우 $j > q$ 이면 ACF $\rho(j) = 0$ (drop off)이다.

**AR(1) 경우:** $Y_t = \alpha_1 Y_{t-1} + e_t$

정상적인(stationary) AR 모형은 MA(∞)로 바꾸어 쓸 수 있다. AR(1)인 경우

$Y_t = \mu + e_t + \alpha_1 e_{t-1} + \alpha_1{}^2 e_{t-2} + \alpha_1{}^3 e_{t-3} + \ldots + \alpha_1{}^{t\text{-}1} e_1 + \alpha_1{}^t (Y_o - \mu)$ 이다.

$\gamma(0) = V(Y_t) = \sigma^2 / (1-\alpha_1^2)$ 가정: $|\alpha_1| < 1$, 즉 정상성(stationary) 가정이 필요

$\gamma(j) = COV(Y_t, Y_{t-j}) = \alpha_i^j \sigma^2 / (1-\alpha_1^2)$

이를 정리하면 $\rho(j) = \alpha_1^j$ 이므로 ACF 는 지수적으로 감소한다.(exponentially decay)

이를 일반화 하면 AR(p) 모형의 경우 ACF 는 지수적으로 감소한다.

ARMA(p, q) 경우

$Y_t - \alpha_1 Y_{t-1} - \alpha_2 Y_{t-2} - \ldots - \alpha_p Y_{t-p} = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} + \ldots - \beta_q e_{t-q}$

AR(p) 모형처럼 지수적으로 감소한다. 그러나 MA(q) 모형의 drop off 효과가 있으므로 꼬리 부분이 갑자기 줄어들게 된다. 이를 exponentially tail off 라 한다.

### Partial Auto Correlation Function (PACF)

LAG 1 인 부분상관함수(PACF)는 $Y_t$ 를 종속변수, $Y_{t-1}$ 을 설명변수로 한 단순 회귀모형에서 $Y_{t-1}$ 의 회귀계수를 의미한다. LAG 2 인 부분상관함수(PACF)는 $Y_t$ 를 종속변수, $Y_{t-1}, Y_{t-2}$, 을 설명변수로 한 다중회귀모형에서 $Y_{t-2}$ 의 회귀계수를 의미한다. LAG 3 인 부분상관함수(PACF)는 $Y_t$ 를 종속변수, $Y_{t-1}, Y_{t-2}, Y_{t-3}$ 설명변수로 한 다중회귀모형에서 $Y_{t-3}$ 의 회귀계수를 의미한다.

- AR(p) 모형의 경우 PACF 는 LAG p 이후에는 0 이다.
- MA(q) 모형의 PACF 는 Invertibility 조건 하에서 지수적으로 감소한다.
- ARMA(p, q) 모형의 PACF 도 지수적으로 감소한다.

### Inverse Auto Correlation Function (IACF)

역상관함수(IACF) 다음과 같이 정의한다.
ARMA(p, q) 모형의 IACF 는 ARMA(q, p)의 ACF 이다.
그러므로 AR(p)의 IACF 는 MA(p)의 ACF 와 같고 MA(q)의 IACF 는 AR(q)의 ACF 와 같다. IACF 는 Drop off 와 Tail off 판단이 어려운 경우 사용한다.

### ARMA 모형 인식 방법

| | AR(p) | MA(q) | ARMA(p, q) |
|---|---|---|---|
| ACF | T | D(q) | T |
| PACF | D(p) | T | T |
| IACF | D(p) | T | T |

*) T: Tail off exponentially *) D(p): Drop off to 0 after lag p

계절성이 존재하는 경우 ACF 는 주기 k 마다 peak 가 생긴다. 왜냐하면 $Y_t$ 에 $Y_{t-k}$ 가 영향을 주기 때문이다.

## ARMA 모형 추정 순서

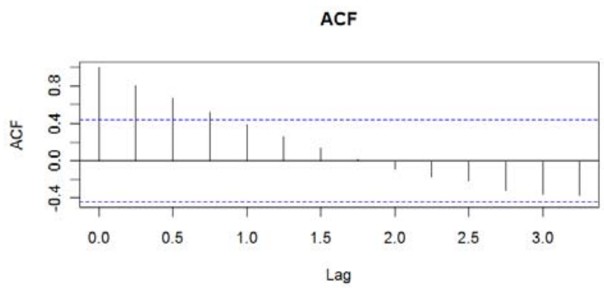
### (1) 시계열 데이터 white noise Test

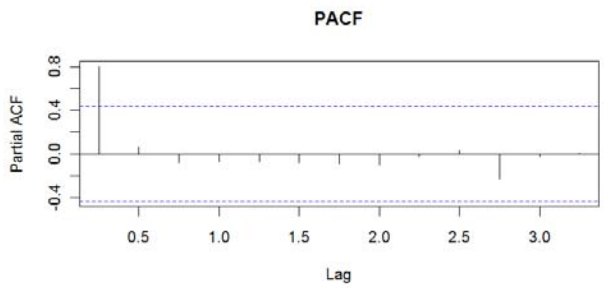
시계열 데이터기 백색 잡음(white noise)인 경우 자기상관계수는 Chi-square 분포에 근사한다. Ljung modified Box-Pierce Q 통계량 $n(n+2)\sum_{j=1}^{k}\frac{\gamma(j)}{(n-j)} \sim \chi^2(k)$. Q-통계량은 시계열 데이터의 백색 잡음 여부를 판단하는 것으로 원 시계열 자료는 백색 잡음이 아니어야 모형 설정이 가능하다. 또한 모형 설정 후 잔차는 백색 잡음이면 모형 설정이 올바로 된 경우이다.

```
> Box.test(ts.hj, type="Ljung-Box")

        Box-Ljung test

data:  ts.hj
X-squared = 14.9131, df = 1, p-value = 0.0001126
```

### (2) ACF, PACF 그래프 진단

```
par(mfrow=c(2,1))
acf(ts.hj, main="ACF")
pacf(ts.hj, main="PACF")
```




AR(1)이 적절해 보인다.

(Unit Root 문제)

AR(1) 모형을 갖는 시계열 데이터의 경우 UNIT root 문제는 $(Y_t = \mu + \alpha Y_{t-1}, \alpha = 1)$임을 의미한다. Unit-root 갖는 데이터는 안정적이지 못하므로 모형 설정의 의미가 없다. 이에 대한 test 방법으로 augmented Dickey-Fuller 검정 방법, Phillips-Perron 검정 방법 등이 있다.

```
library(tseries)
pp.test(ts.hj,alternative = c("explosive"))
```

```
        Phillips-Perron Unit Root Test

data:  ts.hj
Dickey-Fuller Z(alpha) = -14.6937, Trun
p-value = 0.8393
alternative hypothesis: explosive
```

단일근인 시계열 데이터는 1 차 차분 데이터에 MA(1)을 적용하여 미래 값을 예측한다. why? $Y_t = \mu + Y_{t-1} + e_t \Rightarrow Y_t - Y_{t-1} = \mu + e_t$, $e_t$ is white noise

### (3) ARMA 모형 추정, 회귀계수 유의성 검정

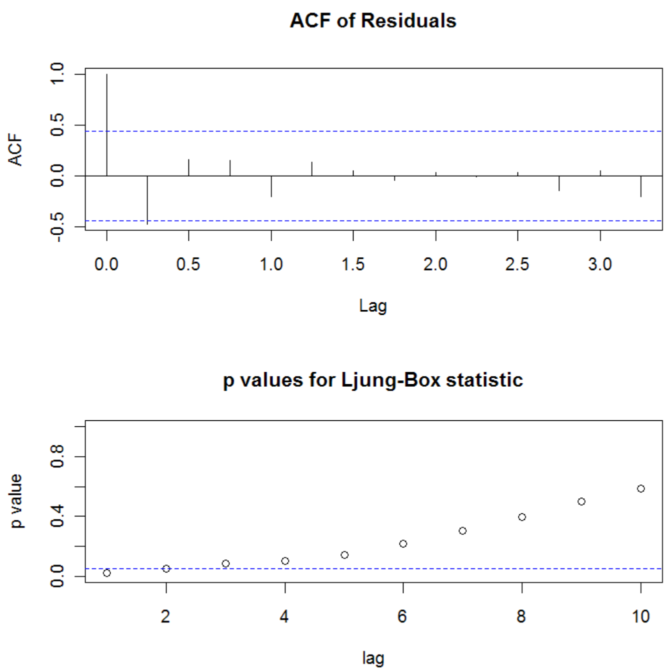
```
fit.hj=arima(ts.hj,order=c(1,0,0))
tsdiag(fit.hj)
```

```
Call:
arima(x = ts.hj, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.9675    81.6348
s.e.  0.0420     8.0148

sigma^2 estimated as 5.394:  log likelihood = -46.61,  aic = 99.21
```

회귀계수 검정통계량 $TS = \frac{0.9675}{0.042} = 23.02$, highly significant (±2 기준)

Akaike Information Criteria : 모형 적합성 통계량, 검정통계량이 아니므로 서로 다른 모형의 적합 정도를 비교 판단할 때 사용, 작을수록 적합성 높음

### (4) 모형 추정 잔차 white noise 검정




모형이 적합하다면 잔차는 white noise 여야 한다. ACF 는 지수적으로 감소해야 하며 (why? 백색잡음은 MA 모형) 유의확률이 0.05 미만, 즉 점선 위에 있어야 함.

### (5) 최종모형

높이뛰기 기록 시계열 데이터는 ACF, PACF 에 의해 AR(1)을 적용하였다. 그러나 단일근 문제가 있고 모형 추정 결과 회귀계수는 유의하나 잔차가 백색잡음을 따르지 않는다. 모형 적합 실패

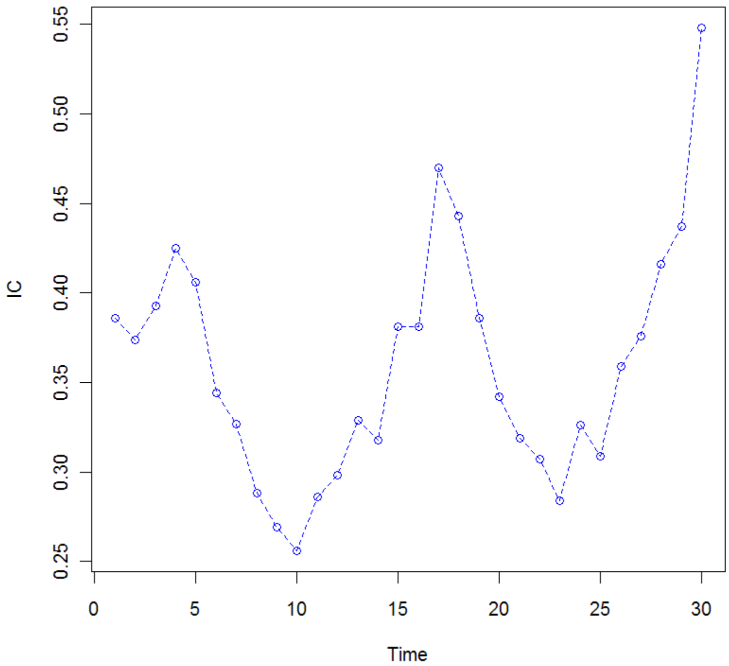
```
d.ts.hj=diff(ts.hj) #1차 차분 데이터 D(t)=Y(t)-Y(t-1)
fit2.hj=arima(d.ts.hj,order=c(0,0,1))
predict(fit2.hj, n.ahead=4)$pred
```

```
> predict(fit2.hj, n.ahead=4)$pred
          Qtr1      Qtr2      Qtr3      Qtr4
2006                     0.8754475 1.0967947
2007 1.0967947 1.0967947
```

## 아이스크림 예제

### 데이터 시간도표

```
ds.ic=read.table("icecream.csv", header=T, sep=",")
ts.ic=ts(ds.ic[2:2])
x11()
plot(ts.ic,type="o",col="blue",lty="dashed")
```




직선 trend, seasonality 주기 13 있는 것으로 보임

### 시계열 데이터 백색잡음 및 stationary 검정

```
> Box.test(ts.ic, type="Ljung-Box")

        Box-Ljung test

data:  ts.ic
X-squared = 14.5389, df = 1, p-value = 0.0001373
```

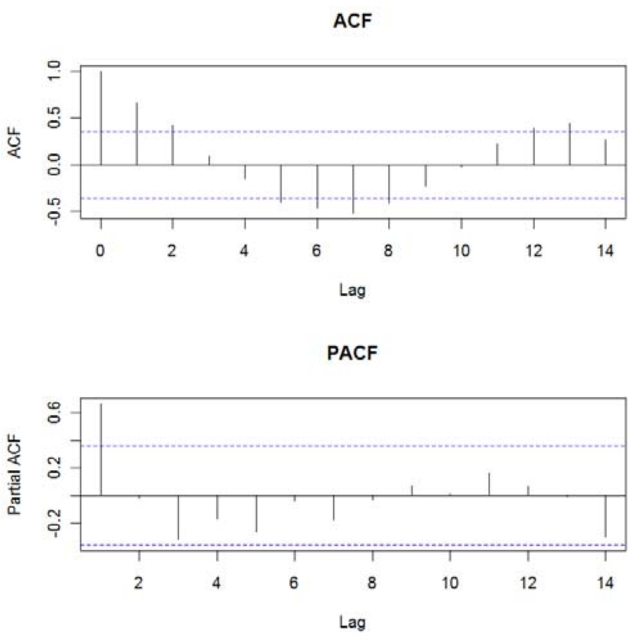
시계열 데이터는 백색잡음이 아니므로 ARMA 모형 추정 가능

```
> library(tseries)
> pp.test(ts.ic,alternative = c("explosive"))

        Phillips-Perron Unit Root Test

data:  ts.ic
Dickey-Fuller Z(alpha) = -5.0509, Truncation lag parameter = 2,
p-value = 0.1912
alternative hypothesis: explosive
```

시계열 데이터는 stationary 하므로 ARMA 모형 추정 가능

## ACF, PACF 이용한 모형진단

```
par(mfrow=c(2,1))
acf(ts.ic, main="ACF")
pacf(ts.ic, main="PACF")
```



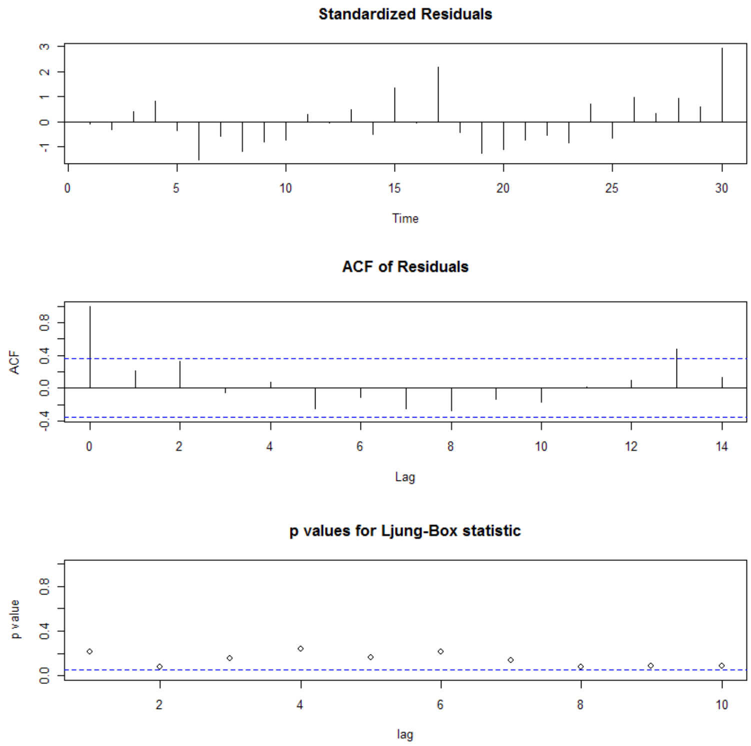

ACF 는 지수적 감소, PACF 는 Lag=1 에서 유의 => AR(1) 진단

## 모형 추정 및 잔차 진단

```
fit.ic=arima(ts.ic,order=c(1,0,0))
tsdiag(fit.ic)
```

회귀계수 검정통계량 $TS = \dfrac{0.8679}{0.1034} = 8.4$ high significant

```
> arima(ts.ic,order=c(1,0,0))

Call:
arima(x = ts.ic, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.8679     0.3922
s.e.  0.1034     0.0503

sigma^2 estimated as 0.001585:  log likelihood = 53.44,  aic = -100.88
```




잔차의 ACF 는 13 에서 peak 가 있고 (계절성 문제) 잔차가 백색잡음을 따르지 않으므로 문제가 있음

## 최종 진단 모형

```
fit2.ic=arima(ts.ic,order=c(1,13,0))
tsdiag(fit2.ic)
```