개요

- George Box, Gwilym Jenkins 제안한 시계열 모형
- 시계열 데이터는 (Trend + Cycle + Seasonality + Irregular) 성분이 있에 (1)설명변수 설정이 용이하지 못하거나 (2) $\{Y_t\}$에 대한 예측을 위하여(시계열 데이터 분석의 주요 목적) 설명변수에 대한 예측치(X_t)가 있어야 하는 문제가 있고 (3)독립성 가정을 만족하지 못해 이 문제를 해결하는 어려움이 있어 회귀모형에 의한 분석보다는 관측치의 이전 관측치를 활용하는 방법이 제안
- ARIMA(Auto-Regressive Integrated Moving-Average) 모형은 시계열 데이터 $\{Y_t\}$ 의 과거치(previous observation) $\{Y_{t-1}, Y_{t-2}, \cdots\}$ 가 설명변수인 AR과 과거 관측치가 설명하지 못하는 부분에 해당되는 오차항 ($e_{t-1}, e_{t-2}, \cdots$)들이 설명변수인 MA, 차분을 나타내는 integrate의 합성어이다.

AR 모형은 아래 가설에 의해 제안되었다.

- 과거의 패턴이 지속된다면 시계열 데이터 관측치 Y_t 는 과거 관측치 $Y_{t-1}, Y_{t-2}, Y_{t-p}, \cdots$ 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?

Backshift Notation

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

상관함수 Correlation Function

자기상관함수 Auto Correlation Function (ACF)

- $\rho(j) = \frac{\gamma(j)}{\gamma(0)} = \frac{Cov(Y_t, Y_{t-j})}{VAR(Y_t)}$, $\gamma(j)$ 는 주기 j의 자기 공분산 (auto-covariance with lag j)
- 현재 관측값과 j기 이전 관측값들의 상관계수이다.
- $\rho(0) = 1$ 주기 0인 상관계수는 현재 관측치 간 상관계수이므로 1이다.

부분자기상관함수 Partial Auto Correlation Function (PACF)

- 두 변수 (X, Y)의 상관관계를 시간의 효과를 제거한 후 구한 순수 상관관계
- $\rho_{XY.Z} = \frac{E(X - E(X|Z))E(Y - E(Y|Z))}{\sqrt{E(X - E(X|Z))^2 E(Y - E(Y|Z))^2}}$ ⇔ Z->X 회귀분석 잔차와 Z->Y 잔차의 상관계수

- 시계열 분석 : (Y_{t-1}, Y_{t-k+1})의 효과 제외한 (Y_t, Y_{t-k})의 순수 상관계수 ϕ_k 을 부분자기상관계수, 즉 $\phi_k = Corr(Y_t^z, Y_{t-k}^z)$
- $\phi_1 = \rho(1), \phi_2 = \frac{\rho(1) - \rho(1)^2}{1 - \rho(1)^2}$, ..., $\phi_k = \frac{\rho(k+1) - \sum \phi_{k,j} \rho(k+1-j)}{1 - \sum \phi_{k,j} \rho(j)}$

역자기상관함수 Inverse Auto Correlation Function (IACF)

- ARMA(p, q) 모형의 IACF는 ARMA(q, p)의 ACF이다.
- 그러므로 AR(p)의 IACF는 MA(p)의 ACF와 같고 MA(q)의 IACF는 AR(q)의 ACF와 같다.

상관함수 활용

ARMA 모형 인식에 활용

AR(p) 모형

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

- Markov process : $|\rho| < 1$ ⇔ stationary 프로세스
- $\rho = 0$: 서로 독립이고 유한인 평균과 분산을 갖는 동일 분포를 따르는(iid) white noise(백색 잡음)
- 만약 평균이 0, 분산이 σ^2 인 정규분포를 따른다면 이를 Guassian white noise라 한다. $\{Y_t\}$ 대신 $\{Z_t\} = \{Y_t - \mu\}$를 사용하기도 하는데 이는 평균을 0으로 하기 위함이다.
- μ 는 시계열 데이터의 총 평균(grand mean)에 해당된다.
- 평균 $E(Y_t) = a + \rho E(Y_{t-1})$ => $\mu = \frac{a}{1-\rho}$
- 분산 $V(Y_t) = \rho^2 \gamma(0) + \sigma^2$ => $V(Y_t) = \gamma(0) = \frac{\sigma^2}{1-\rho^2}$

자기상관함수

AR(1) 모형을 이를 다시 쓰면 다음과 같다. 즉 AR(1) 모형이더라도 과거의 흔적을 모두 모함하고 있다.

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

그리고 $|\rho|<1$ (stationary)이면 가중치는 지수적 감소

MA(∞) 모형: $$Y_t = \mu + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \beta_3 e_{t-3} + \ldots = \mu + \sum_{j=0}^{\infty} \beta_j e_{t-j}$$

$\gamma(j) = COV(Y_t, Y_{t-j}) = \rho^j \sigma^2 / (1-\rho^2)$ => (ACF) $\rho(k) = \rho^k$ 지수적으로 감소

부분자기상관함수

- 1차 이후 회귀계수가 0이므로 1차 PACF는 $\phi_1 = \rho$ 이고, 2차부터 이후는 0이다.

Unit-Root 검정

AR(1) 모형을 갖는 시계열 데이터의 경우 UNIT root 문제는 ($Y_t = \mu + \alpha Y_{t-1}, \alpha = 1$)임을 의미한다. Unit-root 갖는 데이터(Random Walk Model)는 안정적이지 못하므로 모형 설정의 의미가 없다.

test 방법 : augmented Dickey-Fuller 검정 방법, Phillips-Perron 검정 방법 등이 있음

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

- 설명변수의 개수 p개
- AR(p)도 MA(∞) 모형으로 쓸 수 있으므로 정상적인 AR(p)의 자기상관함수는 지수적으로 감소하며, 부분자기상관함수는 p차 이후부터 0이다.

자기상관함수

Stationary 시계열 데이터의 AR(p)의 ACF는 AR(1)과 동일하게 지수적으로 감소한다. 자기상관함수 $\rho(k)$는 Yule-Walker 방정식에 의해 구한다. (complicated)

부분자기상관함수

- $\phi_{k=}\alpha_k$ for $k \le p$
- p차부터 이후는 0이다.

MA(q) 모형

MA(1) 모형: $Y_t = e_t - \beta_1 e_{t-1}$, $e_t \sim iid\ N(0,\sigma^2)$

- 평균은 0이다.
- $\gamma(0) = V(Y_t) = (1+\beta_1^2)\sigma^2$, $\gamma(1) = COV(Y_t, Y_{t-1}) = -\beta_1\sigma^2$,
- $\gamma(2) = \gamma(3) = \gamma(4) = \ldots = 0$

자기상관함수

- $\gamma(0) = \sigma^2(1+\beta_1^2)$, $\gamma(1) = -\dfrac{\beta_1}{1+\beta_1^2}$
- 1차 이후 0이다.

부분자기상관함수

- invertibility에 의해 AR(∞)로 변환가능하다.
- $\phi_k = \dfrac{-\beta_1^k(1-\beta_1^2)}{1-\beta_1^{2(k+1)}}$

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 1보다 클 경우 MA 모형은 Invertibility하다. 이 말은 AR(∞)모형으로 변환할 수 있다는 것이다.

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

자기상관함수

- MA(1) with $\beta_1 = 0.7$: $V(Y_t) = (1+\beta_1^2)\sigma^2$, $\gamma(1) = \dfrac{Cov(Y_t, Y_{t-1})}{V(Y_t)} = \dfrac{0.7}{1+0.49} = 0.47$
- 2차부터 0이다. $\gamma(2) = \gamma(3) = \ldots = 0$

부분자기상관함수

- MA(1)가 invertibility 하면, $\phi_1 = 0.47$

MA(q) 모형: $Y_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \ldots - \beta_q e_{t-q}$, $e_t \sim iid\ N(0, \sigma^2)$

- 과거 오차항 $e_{t-1}, e_{t-2}, \cdots$ 의미 : 이전 관측치 $Y_{t-1}, Y_{t-2}, \cdots$ 에 포함되어 있지 않은 정보
- 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 q인 Moving-Average 이동평균) 모형을 따른다고 한다.
- MA(∞) 모형은 언제나 정상적(stationary)이다.

자기상관함수

- $\rho(k) = \dfrac{-\beta_k + \beta_1\beta_{k+1} + \ldots + \beta_{q-k}\beta_q}{1 + \beta_1^2 + \beta_2^2 + \ldots + \beta_k^2}$, $k \le q$
- $\gamma(q+1) = \gamma(q+2) = \ldots = 0$, q차 이후 0이다.

부분자기상관함수

invertibility에 의해 AR(∞)로 변환가능하다. 그러므로 MA(q) 모형의 PACF는 Invertibility 조건 하에서 지수적으로 감소한다.

ARMA(p, q) 모형

ARMA(p,q) 모형: $Y_t = e_t - \beta_1 e_{t-1} - \ldots - \beta_q e_{t-q} + \alpha_1 Y_{t-1} + \ldots + + \alpha_p Y_{t-p}$, $e_t \sim iid\ N(0, \sigma^2)$

- AR 모형과 MA 모형의 결합이다. 그러므로 AR(∞), MA(∞)로 표현될 수 있음.
- 일반적으로 (2, 2)가 최대

자기상관함수(acf) 부분자기상관함수(pacf)

지수적으로 감소

차분 Difference (계절성 및 추세 성분)

- ARMA 모형은 시계열 데이터 중 사이클 (cycle) 성분에 대한 패턴을 표현하게 된다.
- 물론 불규칙 irregularity 성분은 오차항으로 커버한다.
- 그럼 추세 trend, 계절성 seasonality 성분은 어떻게 하지? 차분이 답이다.
- 차분은 추세나 계절성 성분을 제외시키는 효과가 있다.
- 차분에 의해 추세나 계절성 성분을 제외하면 주기와 불규칙 성분만 남아 수평 상태, 사이클만 존재하게 된다.

정의

- 1차 차분 : $Y_t^* = \nabla Y_t = Y_t - Y_{t-1}$ => 직선 추세성분 해결
- 2차 차분 : $\nabla Y_t = Y_t^* - Y_{t-1}^*$ => 이차형식 추세성분 해결
- d차 차분 : $(Y_t - Y_{t-d})$ => 주기 d 계절성 성분

차분 필요성 진단

PACF에서 차분이 필요한 주기에서 Peak가 발생하며, ACF는 지수적으로 감소

ARMA 모형 진단 표

	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 지수적으로 감소
*) D(p): Drop off after p 차수 p 이후 0의 값

ARMA 모형 적합 절차

순서1) 시간도표

(1) 주기, 계절성 확인 (실제 진단은 상관함수 이용) => plot() 함수

(2) 안정성 stationary process

(a) 평균의 이동 => 평균이 이동하는 경우에는 시계열 데이터 분리하여 모형 적합

(b) 분산의 크기 변동 ⇔ 주기의 폭이 변함 => 분산 안정화, LN 혹은 제곱근(SQRT) 변환

순서2) 모형 적합 가능성 진단

(1) white noise 데이터는 모형 적합 불가

MN 검정통계량 test value 유의확률

(2) 또 다른 백색 잡음 검정 수정 Ljung Box-Pierce Q 통계량 $n(n+2)\sum_{j=1}^{k}\frac{\gamma(j)}{(n-j)} \sim \chi^2(k)$. Q-

=> Box.test(type="Ljung-Box") 검정

(3) unit root (단일근) 검정 => pp.test() 함수 ⇔ 시계열 데이터의 안정성 stationary

순서3) 모형 진단

ACF, PACF 활용하여 (p, q, d) 결정

순서4) 모형추정

- 회귀계수 추정 => arima() 함수
- method = c("CSS-ML", "ML", "CSS")
- maximum likelihood / minimize conditional sum-of-squares.

순서5) 모형 적합성

(1) 회귀계수의 유의성 검정 : 추정된 회귀계수는 모두 유의해야 함

(2) 잔차의 백색 잡음 Ljung Box-Pierce Q 통계량 : 오차의 분산 추정량인 잔차 $r_t = Y_t - \hat{Y}_t$ - 모형이 시계열 데이터를 완전하게 모형화 했으면 잔차는 백색잡음이어야 함

순서6) 예측모형 활용

(1) 여러 모형 중 가장 적합한 모형 : AIC, SBC 작은 값의 모형이 더 적합

- AIC (Akaike Information Criterion) $AIC = -\log\hat{\sigma}_e^2 + 2(p+q)$
- SBC (Schwartz Bayesian Criterion) $SBC = n\log\hat{\sigma}_e^2 + (p+q)\log(n)$
- $\hat{\sigma}_e^2$ 은 오차의 분산 σ^2 의 추정치로 MSE이다.

(2) 향후 필요한 주기까지 최종 모형을 활용하여 관심 변수 예측값 추정

ARMA in R

(1) AR(p)

```
y1=c(10);y2=c(10);y3=c(10);y4=c(10)
for (i in 2:100){
  y1[i]=0.7*y1[i-1]+rnorm(1)   #AR(1)
  y2[i]=y2[i-1]+rnorm(1)    #Random Walk
  y3[i]=0.3*y3[i-1]+0.4*y4[i-1]+rnorm(1) #AR(2)
  y4[i]=y3[i-1]
}

#(1) time plot
plot(y1,type="l",ylim=c(-4,20))
lines(y2,col="red"); lines(y3,col="blue")

#자기상관계수 출력
acf(y1,plot=F,type=c("correlation")); acf(y2,plot=F,type=c("correlation"))
acf(y3,plot=F,type=c("correlation"))

#(2) 모형적합성 검정
library(normwhn.test)
whitenoise.test(y1); whitenoise.test(y2); whitenoise.test(y3)

library(tseries)
pp.test(y1,alternative=c("explosive"))
pp.test(y2,alternative=c("explosive"))
pp.test(y3,alternative=c("explosive"))

#(3) 모형인식 acf, pacf
par(mfrow=c(2,1)); acf(y1, main="ACF of y1"); pacf(y1, main="PACF of y1")
par(mfrow=c(2,1)); acf(y2, main="ACF of y2"); pacf(y2, main="PACF of y2")
par(mfrow=c(2,1)); acf(y3, main="ACF of y3"); pacf(y3, main="PACF of y3")

#(4) 모형 추정
fit1=arima(y1,order=c(1,0,0)); fit2=arima(y2,order=c(1,0,0)); fit3=arima(y3,order=c(2,0,0))

#(5) 계수 유의성
ts1=fit1$coef[1:1]/sqrt(fit1$var.coef[1:1])
p1=1-pt(ts1,length(y1)-2); cat("ts=",ts1,"p-value",p1)

ts3_1=fit3$coef[1:1]/sqrt(fit3$var.coef[1:1])
p3_1=1-pt(ts3_1,length(y3)-2); cat("ar(1) ts=",ts3_1,"p-value",p3_1)
ts3_2=fit3$coef[2:2]/sqrt(fit3$var.coef[2,2])
p3_2=1-pt(ts3_2,length(y3)-2); cat("ar(2) ts=",ts3_2,"p-value",p3_2)

#(5) 잔차 유의성 검정
tsdiag(fit1,gof.lag=24); tsdiag(fit3,gof.lag=24)

#(6) 예측치
predict(fit1,n.ahead=12)$pred; predict(fit3,n.ahead=12)$pred
```

모형

$$1) Y_t = 0.7 * Y_{t-1} + e_t,\ e_t \sim N(0,1)$$
$$2) Y_t = Y_{t-1} + e_t,\ e_t \sim N(0,1) \quad , \text{초기값 } Y_0 = 10$$
$$3) Y_t = 0.3 * Y_{t-1} + 0.4 * Y_{t-2} + e_t,\ e_t \sim N(0,1)$$

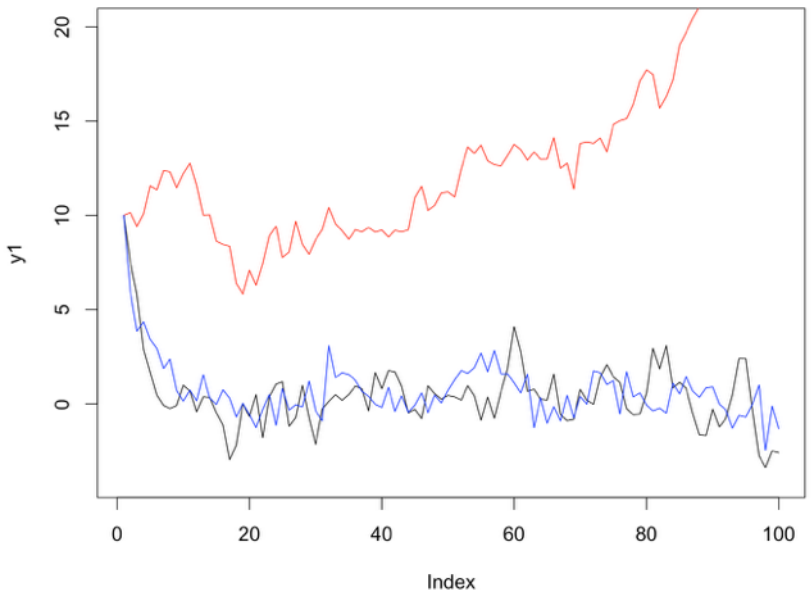
순서1) Time plot

모형1) AR(1)~검은색

모형2) R.Walk ~붉은색

모형3) AR(2)~파랑색

도표 그림만으로는 모형을 인식하는 것은 불가능하다.



순서2) 모형 적합 가능성

(1) White Noise 검정

- 귀무가설 : 시계열은 백색잡음이다.
- 대립가설 : 백색잡음이 아니다. 〈=〉 ARMA 모형 적합이 가능하다.
- tMN=검정통계량 값, test value=유의확률
- 유의확률이 모두 5%보다 작으므로 귀무가설이 기각되어 백색잡음 아니다.

```
[1] "tMN"
[1] 5.774125
[1] "test value"
[1] 0.003888567

[1] "tMN"
[1] 5.896913
[1] "test value"
[1] 0.003193649

[1] "tMN"
[1] 4.170712
[1] "test value"
[1] 0.03703699
```

(2) random walk – stationary

- 귀무가설 : 시계열이 Stationary
- 대립가설 : explosive (발산) – Random Walk
- y2 : 유의확률이 0.01이므로 귀무가설이 기각되어 explosive 시계열이므로 ARMA 모형 적합은 가능하지 않음

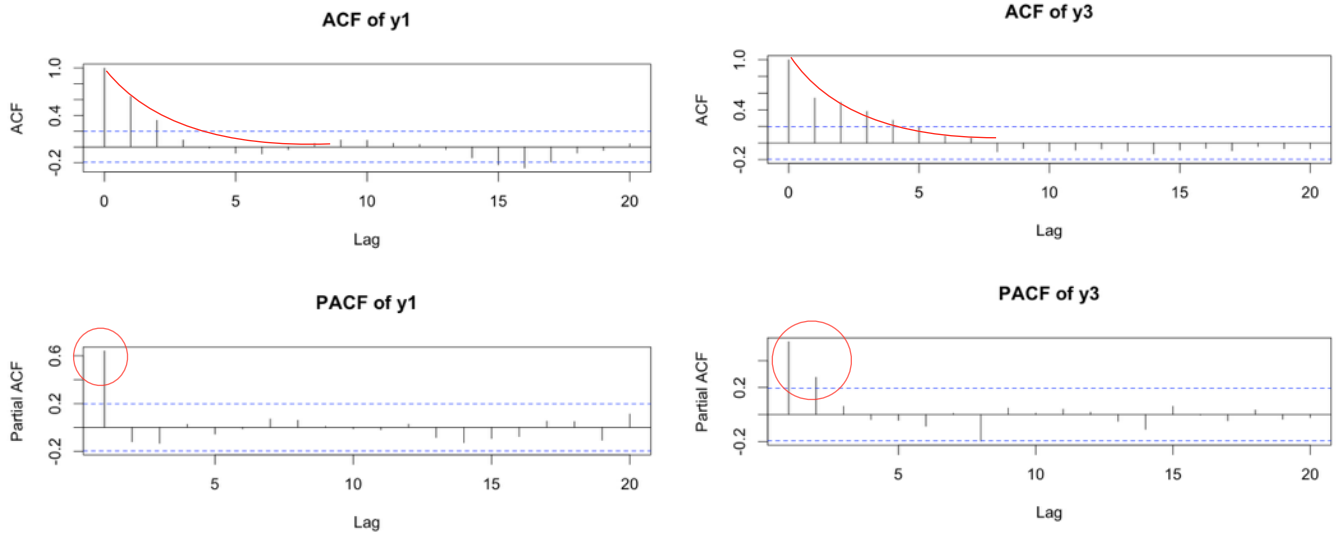
```
> pp.test(y2,alternative=c("explosive"))

        Phillips-Perron Unit Root Test

data:  y2
Dickey-Fuller Z(alpha) = 1.0161, Truncation lag parameter = 3,
p-value = 0.01
alternative hypothesis: explosive
```

순서3) 모형인식

- ARMA 모형 인식에 필요한 acf, pacf 그래프를 그린다. nlag=24가 디폴트이다.
- par(mfrow=c(2,1)) : 그래프를 2행 1열 (2개) 그래프를 그린다.

(ACF) 지수적으로 감소하므로 AR() 모형을 적합

(PACF) Y1은 주기 1에서 peak이므로 AR(1), Y2 는 주기 1, 2에서 peak이므로 AR(2) 적합



순서4) 모형추정

(모형1) $\hat{Y}_t = 0.857 + 0.874 * Y_{t-1}$

(모형2) non-stationary 하여 모형 추정 불가

(모형3) $\hat{Y}_t = 2.109 + 0.481 * Y_{t-1} + 0.483 * Y_{t-1}$

순서5) 모형 유의성

(회귀계수 유의성)

- 검정통계량 : $TS = \frac{\hat{\beta}}{SE} \sim t(n-2)$
- fit1$coef[1:1]/sqrt(fit1$var.coef[1:1])
  - 1행 1열 값 가져오기

```
> arima(y1,order=c(1,0,0))

Call:
arima(x = y1, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.8741     0.8573
s.e.  0.0706     0.9367
```

```
> arima(y3,order=c(2,0,0))

Call:
arima(x = y3, order = c(2, 0, 0))

Coefficients:
         ar1     ar2  intercept
      0.4809  0.4825      2.109
s.e.  0.0951  0.0967      2.331
```

```
> arima(y2,order=c(1,0,0))
다음에 오류가 있습니다arima(y2, order = c(1, 0, 0)) : non-stationary
```

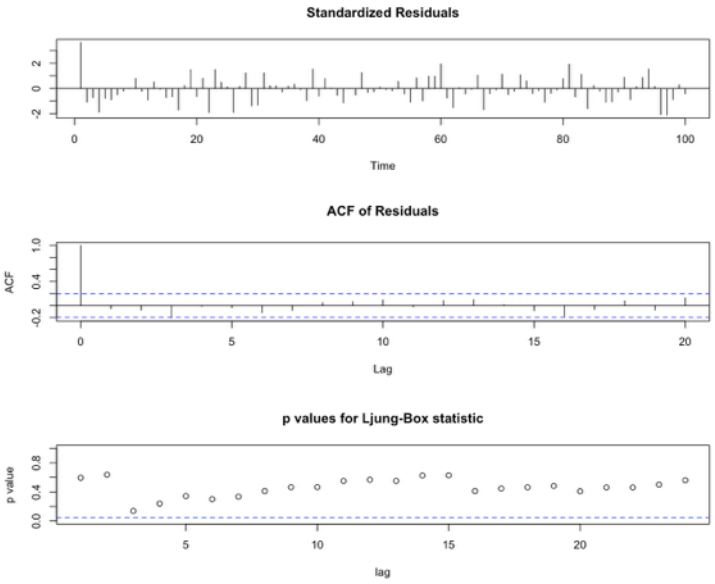
```
> cat("ts=",ts1,"p-value",p1)
ts= 12.38615 p-value 0
```

```
> cat("ar(2) ts=",ts3_2,"p-value",p3_2)
ar(2) ts= 4.98952 p-value 1.312348e-06
```

```
> cat("ar(1) ts=",ts3_1,"p-value",p3_1)
ar(1) ts= 5.05808 p-value 9.880445e-07
```

(잔차 유의성)

- 잔차 백색잡음 검정은 마지막 3행의 산점도를 보면 됨
- 모든 주기의 잔차가 유의확률이 모두 5% 이상이므로 백색잡음이다



순서6) 예측치

```
> predict(fit1,n.ahead=12)$pred
Time Series:
Start = 101
End = 112
Frequency = 1
 [1] -2.15391589 -1.77476213 -1.44334962 -1.15366701 -0.90045987
 [6] -0.67913538 -0.48567903 -0.31658178 -0.16877646 -0.03958209
[11]  0.07334473  0.17205234
> predict(fit3,n.ahead=12)$pred
Time Series:
Start = 101
End = 112
Frequency = 1
 [1] -0.6057238 -0.8453677 -0.6214891 -0.6294440 -0.5252577 -0.4789925
 [7] -0.4064782 -0.3492852 -0.2867962 -0.2291521 -0.1712829 -0.1156429
```