PART 3. 데이터 분석 - 4장. 통계분석

# 데이터 준전문가

## ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

강사 소개

# 류영표

## Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~4기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

### 강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 이젠 종로 아카데미(파이썬, ADSP 강사) / 강남 : ADSP
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원
- 보건 · 바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

### 주요 프로젝트 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 폐플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

# 시계열 분석

- **시계열 자료**

  : 시간의 흐름에 따라 관찰된 값들을 시계열 자료라고 함.

  : 시계열 데이터의 분석을 통해 미래의 값을 예측하고 경향, 주기, 계절성 등을 파악하여 활용한다.

- **시계열 예측이란?**

  a. 시계열 자료(Time Series, 시간의 흐름에 따라 기록된 자료)를 분석하는 것.

  · 일정 시점에 조사된 데이터는 횡단(cross-sectional) 자료라고 한다. ex) 소비자물가지수

  b. 시계열 예측의 목적

  · 미래의 값을 예측하는 것 ex) 향후 일주일간 주가예측

  · 시계열 데이터의 특성을 파악하는 것

  · 경향(trend), 주기(cycle), 계절성(seasonality), 불규칙성(irregular) 등

- **시계열 분석**

  : 어떤 현상에 대하여 과거에서부터 현재까지의 시간에 흐름에 따라 기록된 데이터를 바탕으로 미래의 변화에 대한 추세를 분석하는 방법

  : 시간의 흐름을 고려한다는 점이 일반분석과 다른 점.

# 데이터 관점에 따른 분류

| | 횡단면 데이터<br>(Cross Sectional) | 시계열 데이터<br>(Time Series) | 패널 데이터<br>(Panel) |
|---|---|---|---|
| 정의 | 특정시점 + 다수독립변수<br>(여러 변수의 관측치) | 다수시점+특정독립변수<br>(여러 시점에 대해 관측한 자료) | 다수독립변수+다수시점<br>(여러 시점에 대해 여러 변수 자료) |
| 예시 | 2000년의 각 기업의 매출액 | 기업 A의 연도별 매출액 | 기업 A,B,C,D의 2000년부터 2004년까지의 관측된 모든 매출액 자료 |
| 특징 | 값 독립적,<br>모집단 중 특정 시점 표본추출 | 값 Serial-correlation<br>/Trend/Seasonality 등 | 시점/변수 일치로 연구자들이 가장 선호 |

## Sales

Annual sales figures for each company in millions of KRW.

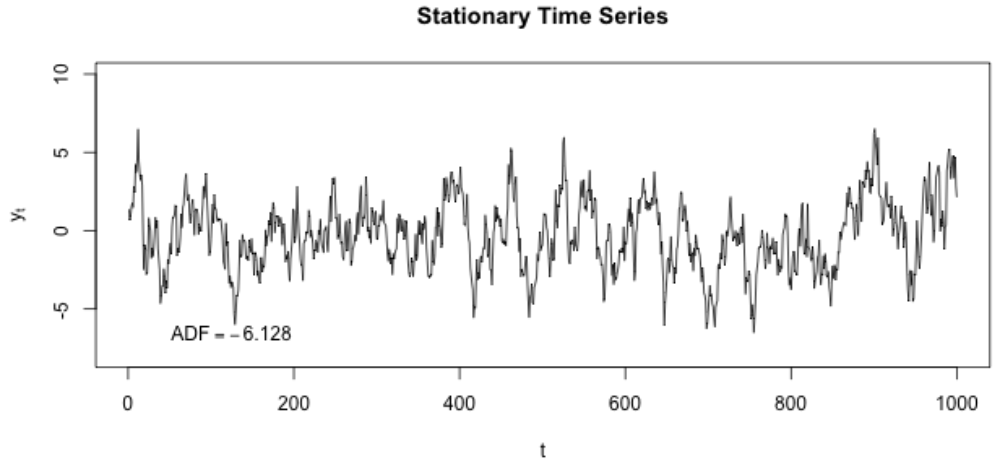
패널

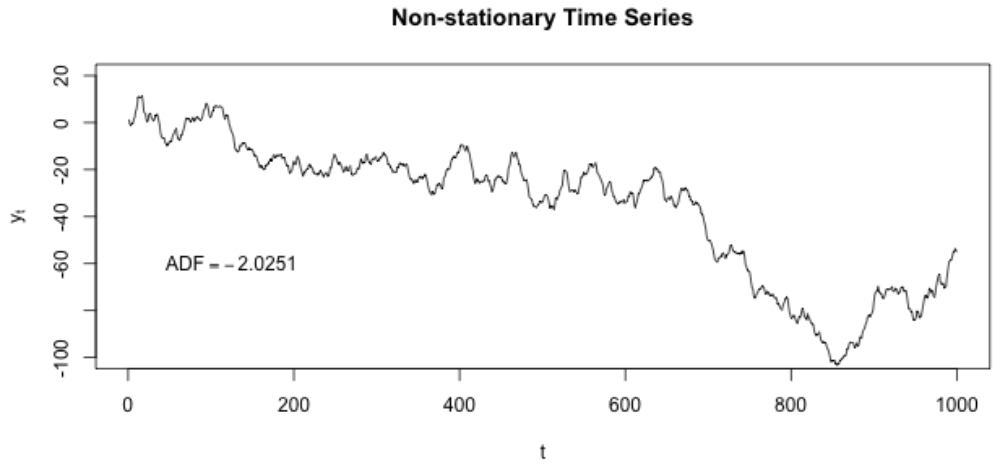
| Year | A | 시계열 B | C | D 횡단면 |
|---|---|---|---|---|
| 2000 | 1,881 | 11,296 | 24,855 | 6,929 |
| 2001 | 1,900 | 12,007 | 23,130 | 5,693 |
| 2002 | 1,994 | 12,659 | 23,519 | 6,145 |
| 2003 | 15,24 | 13,091 | 20,761 | 6,769 |
| 2004 | 2,107 | 13,636 | 22,505 | 7,902 |

출처 : http://triki.net/study/3343

# 시계열 데이터의 특성

- 시계열 데이터는 시간의 흐름에 따라 평균이나 분산 등의 통계적 특성이 변하지 않고, 일정한 추세가 없는 정상성(Stationary) 데이터와 통계적 특성이 변화하는 비정상성(Non-Stationary)데이터로 나눌 수 있음.




출처 https://jjeongil.tistory.com/664

# 시계열 구성 요인


## 시계열 구성 요인
## (Time Series Component Factors)

*시계열 분해*
*(de-composition)*

*시계열 구성*
*(composition)*

**시계열 가법 모형**
**(Timeseries additive model)**

**추세 요인**
**(trend factor)**

+

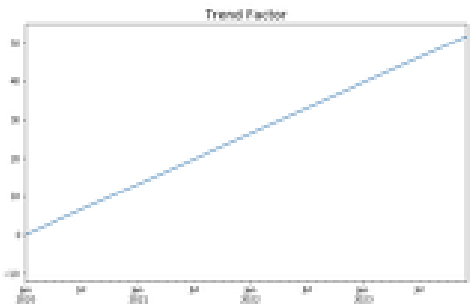
**순환 요인**
**(cycle factor)**

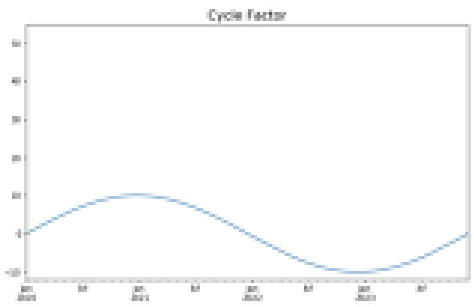
+

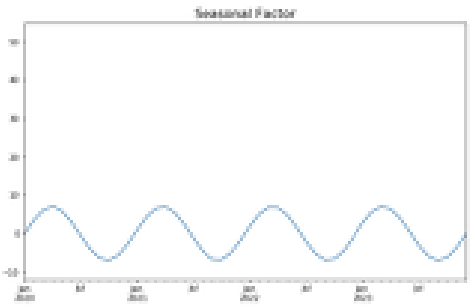
**계절 요인**
**(seasonal factor)**

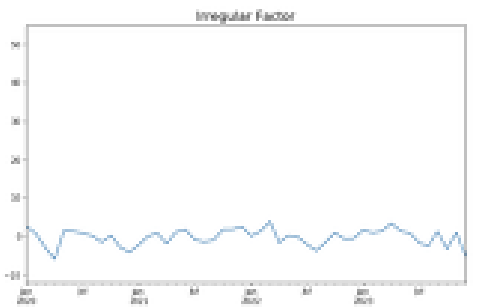
+

**불규칙 요인**
**(irregular factor)**









R, Python 데이터 분석과 프로그래밍의 친구 **http://rfriend.tistory.com**

# 정상성(Stationary)

- 시계열 분석을 위해서는 **정상성**을 만족해야 됨.
- **정상성(Stationary)**

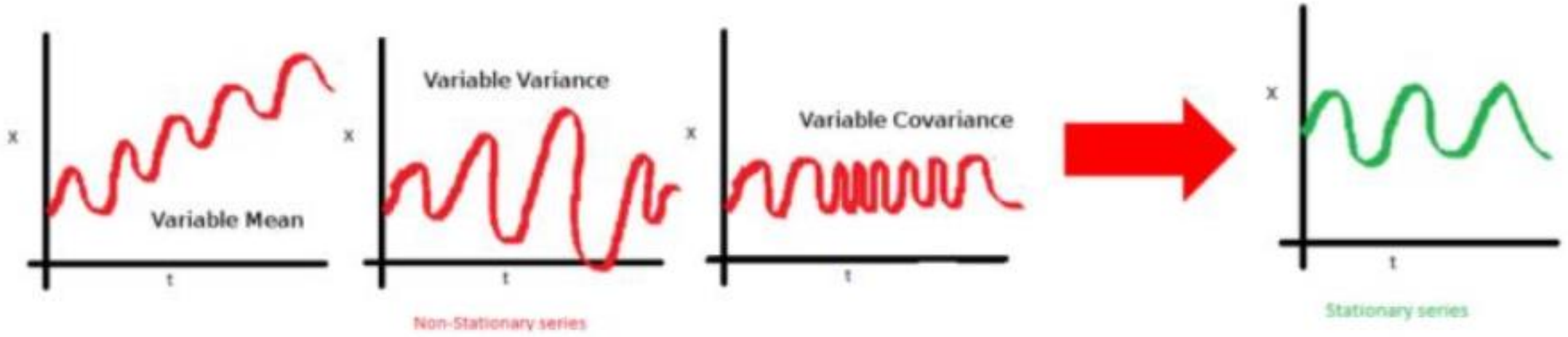
  : 시계열의 수준과 분산에 체계적인 변화가 없고, 주기적 변동이 없다는 것.

  : 미래는 확률적으로 과거와 동일하다는 것.
- **정상 시계열의 조건**

  : 평균은 모든 시점(시간 t)에 대해 일정하다.

  : 분산은 모든 시점(시간 t)에 대해 일정하다.

  : 공분산은 시점(시간 t)에 의존하지 않고, 단지 **시차에만 의존**한다.




출처 : https://towardsdatascience.com/achieving-stationarity-with-time-series-data-abd59fd8d5a0

# 정상성을 확보하는 목적

- 비정상 데이터를 정상화하여 예측하고 다시 정상성 데이터로 환원하는 것이 정상성 데이터를 활용하는 목적
- 정상성을 확보하는 이유는 예측값이 무한대로 가지 않고 / 값이 튀지 않고, 특정한 범위 내에서 안정적이게 예측되도록 하는 것.
- 정상성 확보는 바로 매출을 예측하는 것이 아닌, 점유율을 예측하고 이를 매출로 바꾸는 것과 비슷하다.
- 즉, 넓은 범위의 값을 좁은 범위의 값으로 바꾸어서 예측의 정확성을 높이기 위해 정상성을 확보하는 것 이다.

1. 시계열 모형은 데이터가 Stationary라 가정한다. -〉 Stationary여야 분석효과가 높다, 추정해야하는 파라미터가 적어지고 알고리즘이 단순해 질 수 있음.(과적합 방지)
2. 백색잡음 또한 Stationary 이다. -〉 잔차검증 역시 Stationary 가정을 전재로 함.

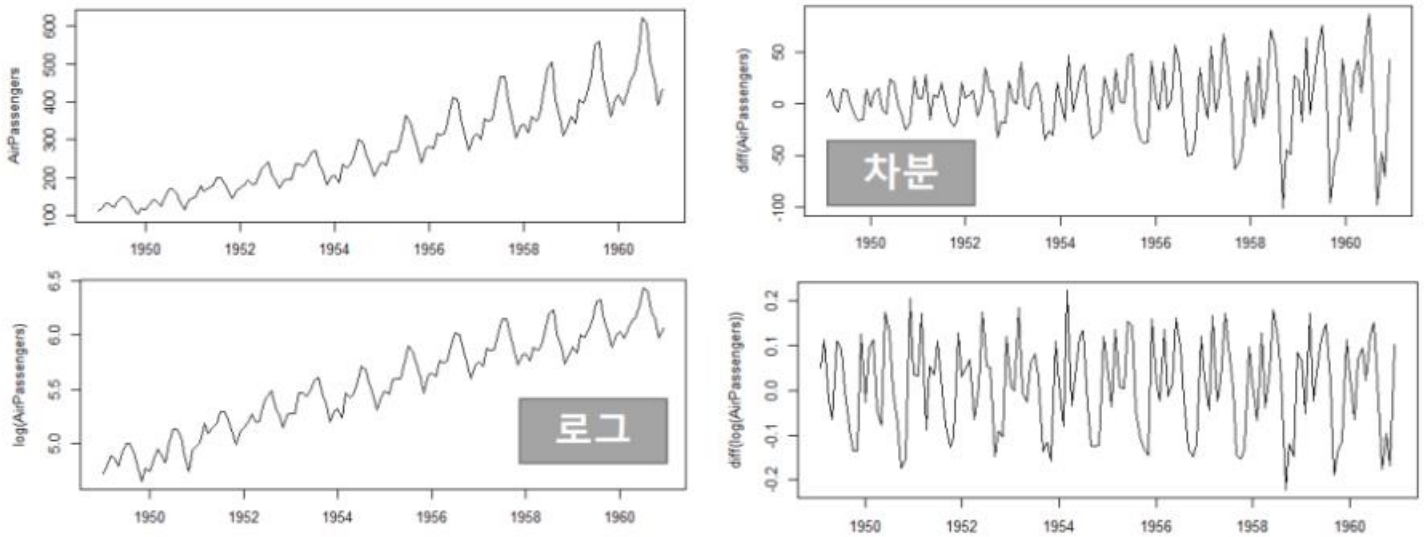
출처 : https://ysyblog.tistory.com/232

# 정상 시계열 전환하는 방법

- 비정상시계열 자료는 정상성을 만족하도록 데이터를 정상성을 만족하도록 데이터를 정상 시계열로 만든 후 시계열 분석을 수행함.
- 평균이 일정하지 않은 경우 : 원계열에 차분 사용
- 계절성을 갖는 비정상 시계열 : 계절 차분 사용
- 분산이 일정하지 않는 경우 : 원계열에 자연 로그(변환) 사용

- **차분**

  : **현 시점의 자료 값에서 전 시점의 자료 값을 빼 주는 것 의미함**. / 현재시점($t_i$)의 자료에서 인접시점($t_{i-1}$)의 자료를 차감하는 것.

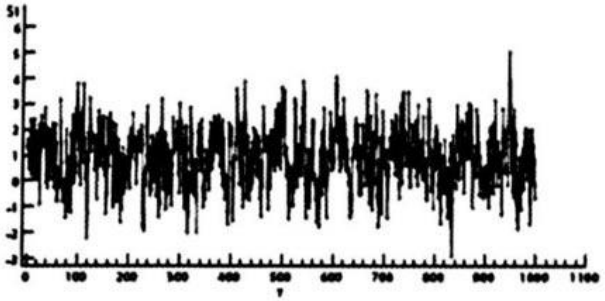



출처 : https://starrykss.tistory.com/1959

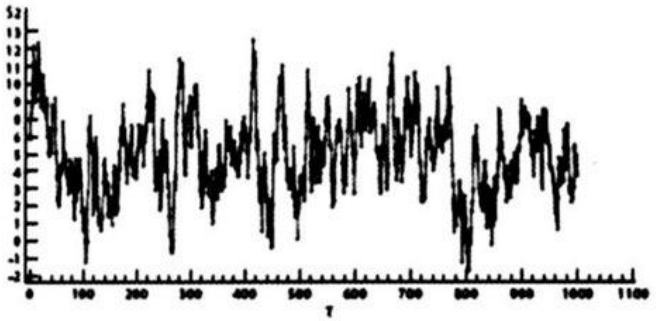
# 정상 시계열 전환하는 방법

정상 시계열의 모습

$y_t = 0.5 + 0.5\, y_{t-1} + e_t,\ e_t \sim iid\ N(0,1)$



$y_t = 0.5 + 0.9\, y_{t-1} + e_t$



$E(y_t) = \mu$ [일정한 평균]

$var(y_t) = \sigma^2$ [일정한 분산]

$cov(y_t, y_{t+s}) = cov(y_t, y_{t-s}) = \gamma_s$ [공분산은 t가 아닌 s에 의존함]

- 정상 시계열은 어떤 시점에서 평균과 분산 그리고 특정한 시차의 길이를 갖는 자기공분산을 측정하더라도, 동일한 값을 가진다.
- 정상 시계열은 항상 그 평균값으로 회귀하려는 경향이 있으며, 그 평균값 주변에서의 변동은 대체로 일정한 폭을 갖는다.
- 정상 시계열이 아닌 경우 특정 기간의 시계열 자료로부터 얻은 정보를 다른 시기로 일반화 할 수 없다.

# 시계열자료 분석방법

## 1) 분석방법

- 수학적 이론모형 : 회귀분석(계량경제)방법, Box-Jenkins 방법
- 직관적 방법 : 지수평활법, 시계열 분해법으로 시간에 따른 변동이 느린 데이터 분석에 활용
- 장기예측 : 회귀분석방법
- 단기예측 : Box-Jenkins 방법, 지수평활법, 시계열 분해법

## 2) 자료 형태에 따른 분석 방법

1. 일변량 시계열분석 : Box-Jenkins 방법(ARMA), 지수평활법, 시계열 분해법
   - 시간(t)을 설명변수로 한 회귀모형주가, 소매물가지수 등 하나의 변수에 관심을 갖는 경우
2. 다중 시계열 분석 : 계량경제 모형, 전이함수 모형, 개입분석, 상태공간 분석, 다변량 ARIMA 등
   - 여러 개의 시간(t)에 따른 변수들을 활용하는 시계열 분석

# 시계열자료 분석방법

### 3) 이동평균법

: 과거로부터 현재까지의 시계열 자료를 대상으로 일정기간별 이동평균을 계산하고, 이들의 추세를 파악하여 다음 기간을 예측

: 시계열 자료에서 계절변동과 불규칙변동을 제거하여 추세변동과 순환변동만 가진 시계열로 변환하는 방법으로 사용.

$$F_{n+1} = \frac{1}{n}\left(Z_n + Z_{n-1} + \dots + Z_{n-m-1}\right) = \frac{1}{m}\sum_{t}^{n} Z_t, \quad t = n-m+1$$

- m은 이동평균한 특정기간, $Z_n$은 가장 최근 시점의 데이터

  -〉 n개의 시계열 데이터를 m기간으로 이동평균하면 n-m+1개의 이동평균 데이터가 생성된다.

**[특징]**

- 간단하고 쉽게 미래를 예측할 수 있으며 자료의 수가 많고 안정된 패턴을 보이는 경우 예측의 품질이 높음.
- 특정 기간 안에 속하는 시계열에 대해서는 동일한 가중치를 부여함
- 일반적으로 시계열 자료가 뚜렷한 추세가 있거나 불규칙변동이 심하지 않은 경우에는 짧은 기간(m개의 개수가 적음)의 평균을 사용, 반대로 불규칙변동이 심한 경우 긴 기간(m개의 개수가 많음)의 평균을 사용.
- 가장 중요한 것은 적절한 기간을 사용하는 것, 즉 적절한 n의 개수를 결정하는 것임.

# 시계열자료 분석방법

## 4) 지수평활법

: 일정기간의 평균을 이용하는 이동평균법과 달리 모든 시계열 자료를 사용하여 평균을 구하며,

시간의 흐름에 따라 최근 시계열에 더 많은 가중치를 부여하여 미래를 예측

$$
\begin{aligned}
F_{n+1} &= \alpha Z_n + (1-\alpha)F_n \\
&= \alpha Z_n + (1-\alpha)[\alpha Z_{n-1} + (1-\alpha)F_{n-1}] \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2 F_{n-1} \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + (1-\alpha)^2[\alpha Z_{n-2} + (1-\alpha)F_{n-2}] \\
&\vdots \\
&= \alpha Z_n + \alpha(1-\alpha)Z_{n-1} + \alpha(1-\alpha)^2 Z_{n-2} + \alpha(1-\alpha)^3 Z_{n-3} + \cdots
\end{aligned}
$$

$F_{n+1}$은 n시점 다음의 예측값

$\alpha$는 지수평활계수

$Z_n$은 n시점의 관측값

->지수평활계수가 과거로 갈수록 지수형태로 감소하는 형태, 최근에 더 많은 가중치

**[특징]**

- 단기간에 발생하는 불규칙변동을 평활하는 방법
- 자료의 수가 많고 안정된 패턴을 보이는 경우일수록 예측 품질이 높음
- 지수평활계수는 과거로 갈수록 지속적으로 감소함
- 지수평활법은 불규칙변동의 영향을 제거하는 효과가 있으며, 중기 예측 이상에 주로 사용됨.

( 단, 단순지수 평활법의 경우, 장기추세나 계절변동이 포함되는 시계열의 예측에는 적합하지 않음)

# 시계열모형

## 1) 자기회귀 모형(AR 모형, autoregressive model)

: 과거와 현재의 자신과의 관계를 정의 한 것.

: 이전 관측값(과거)이 이후 관측값(현재)에 영향을 주는 원리를 사용하기 때문에 Z를 활용함.

t를 현재 시점, p를 과거 시점이라고 할 때,

$Z$ = 시계열 자료, $\Phi$ = 모수, $\alpha$ = 오차항

$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p} + \alpha_t$$

$Z_t$: 시계열 자료 현재 시점

$\Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \cdots + \Phi_p Z_{t-p}$: 과거가 현재에 미치는 영향을 나타내는 모수 × 시계열 자료 과거 시점

$\alpha_t$: 오차항 (백색 잡음 과정)

-> p에 특정한 숫자 n을 대입하여 n 시점부터 회귀를 하고 싶으면 AR(n)으로 쓴다.

-> 평균이 0, 분산이 $\sigma^2$, 자기공분산이 0인 경우, 시계열간 확률적 독립인 경우 강(Strictly)백색잡음과정이라고 한다.

-> 백색잡음 과정이 정규분포를 따를 경우 이를 가우시안(Gaussian) 백색잡음과정이라고 한다.

출처 : https://blog.naver.com/tjgml1343/222077619748

# 시계열모형

## 1) 자기회귀 모형(AR 모형, autoregressive model)

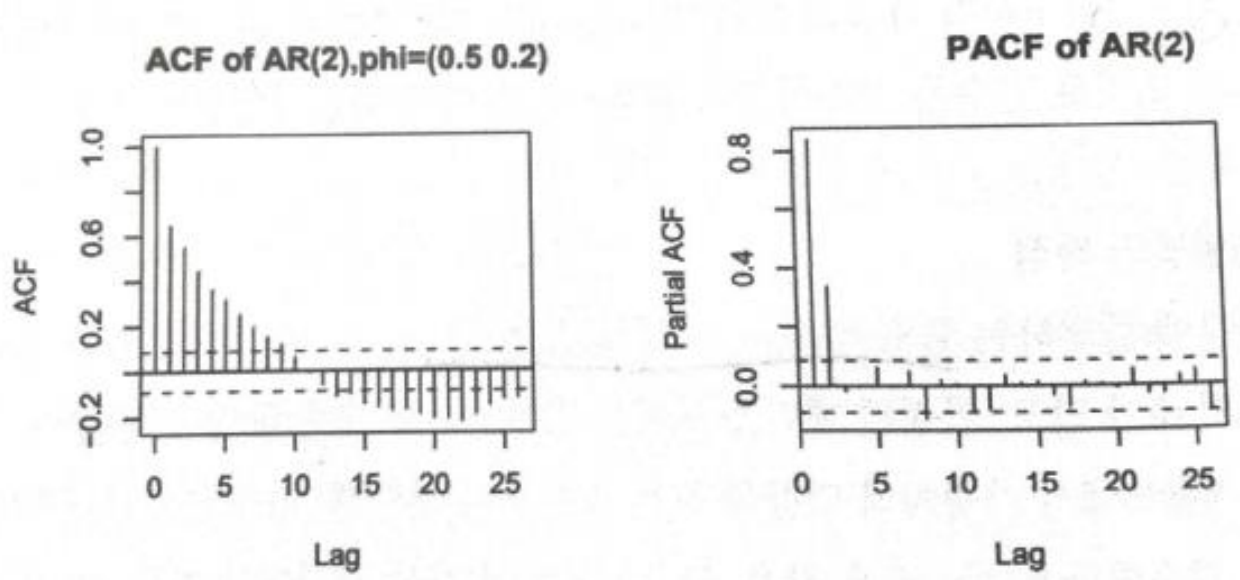
: p 시점 전의 자료가 현재 자료에 영향을 주는 모형

[AR(1) 모형] $Y_t = \Phi_1 Y_{t-1} + \epsilon$, 직전 시점 데이터로만 분석

[AR(2) 모형] $Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \epsilon$, 연속된 3시점 정도의 데이터로 분석

-〉 자기상관함수(ACF)는 빠르게 감소, PACF(부분자기함수)는 어느 시점에서 절단점을 가진다.

-〉 ACF가 빠르게 감소하고 PACF가 3시점에서 절단점을 갖는 그래프가 있다면, 2시점 전의 자료까지가 현재에 영향을 미치는 AR(2) 모형이라 볼 수 있다.

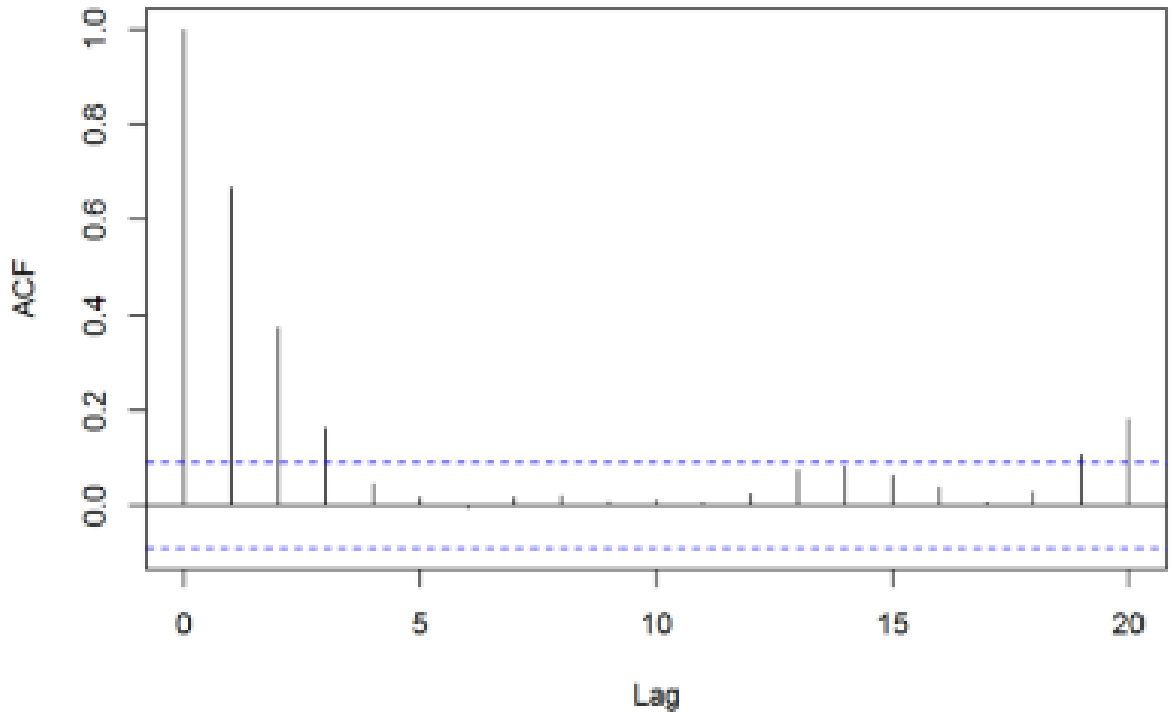



- lag : 시차
- 상관 : 두 변수들간의 관계, 관계정도는 피어슨의 상관관계수 값을 가지고 평가
- 다중공선성 : y가 하나가 있고 x가 여러 개 있을 경우 x들간의 상관관계 - 측정하는 방법 VIF(분산팽창인수)
- 자기상관 : 데이터 한컬럼을 두고 시차를 두어 쌍을 이뤘을때 관계

  -〉 시차가 많이 날수록 자기상관이 떨어짐.

# 자기 상관 함수(ACF, AutoCorrelation Function)

- 시간 단위로 구분된 시계열의 관측치 간 상관 관계 함수이다.

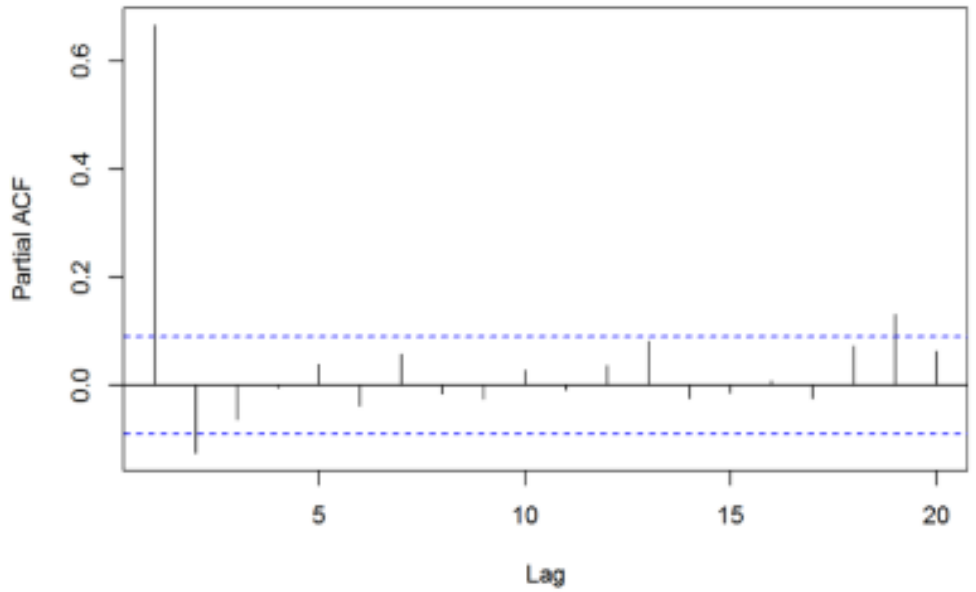



- 점선으로 유의미한 상관과 유의미하지 않은 상관을 확인할 수 있는데, 선의 위쪽이 유의미한 값이다.

출처 : https://blog.naver.com/tjgml1343/222077619748

# 부분(편) 상관 함수(PACF, Partial AutoCorrelation Function)

- 부분 상관이란 두 확률 변수 X와 Y에 의해 다른 모든 변수들에 나타난 상관관계를 설명하고 난 이후에도 여전히 남아 있는 상관 관계
- 편 자기 상관 함수(PACF)는 자기 상관 함수와 마찬가지로 시계열 관측치 간 상관 관계 함수이고, 시차 k에서의 k단계만큼 떨어져 있는 모든 데이터간의 상관 관계를 말함.




- PACF도 역시 선 위쪽이 유의미한 값이다.

출처 : https://blog.naver.com/tjgml1343/222077619748

# 시계열모형

## 2) 이동평균 모형(MA 모형, Moving Average model)

: 유한한 개수의 백색잡음의 결합이므로 언제나 정상성을 만족

$$Y_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2\, \alpha_{2-1} - \cdots - \theta_p \alpha_{t-p}$$

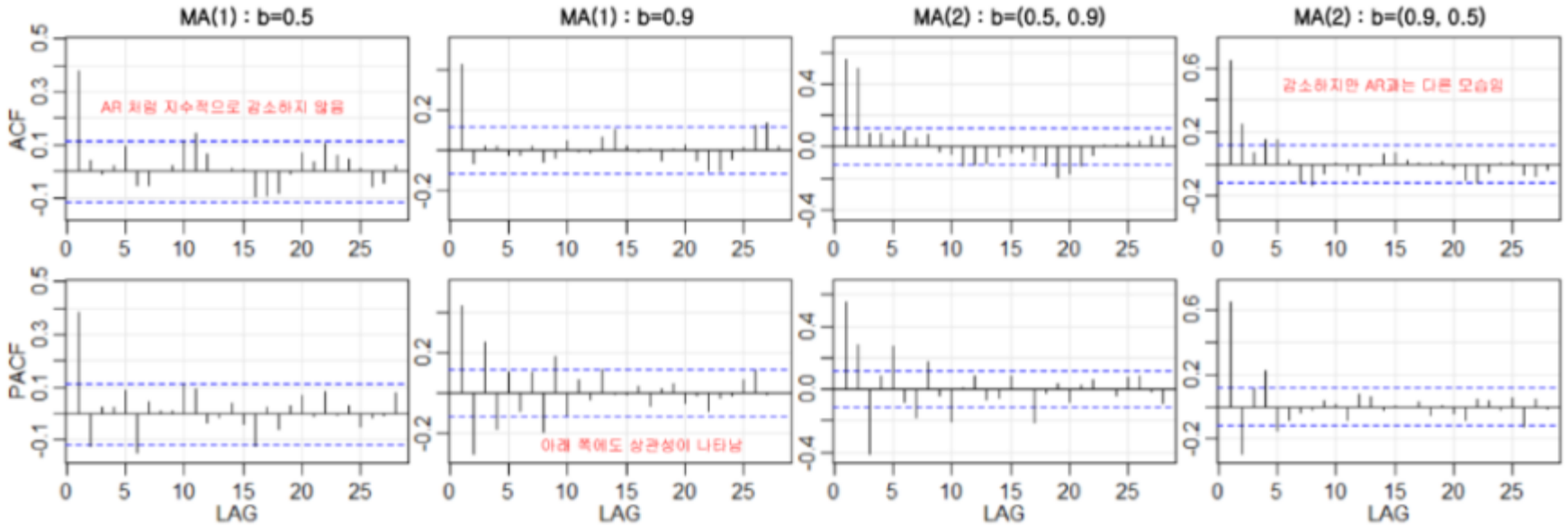
[MA1 모형] 1차 이동평균모형 : 이동평균모형 중에서 가장 간단한 모형으로 시계열이 같은 시점의 백색잡음과 바로 전 시점의 백색잡음의 결합으로 이뤄진 모형

$$Y_t = \alpha_t - \theta_1 \alpha_{t-1}$$

[MA2 모형] 2차 이동평균모형 : 바로 전 시점의 백색잡음과 시차가 2인 백색잡음의 결합으로 이뤄진 모형

$$Y_t = \alpha_t - \theta_1 \alpha_{t-1} - \theta_2\, \alpha_{t-2}$$

-〉 AR모형과 반대로 ACF에서 절단점을 갖고, PACF가 빠르게 감소



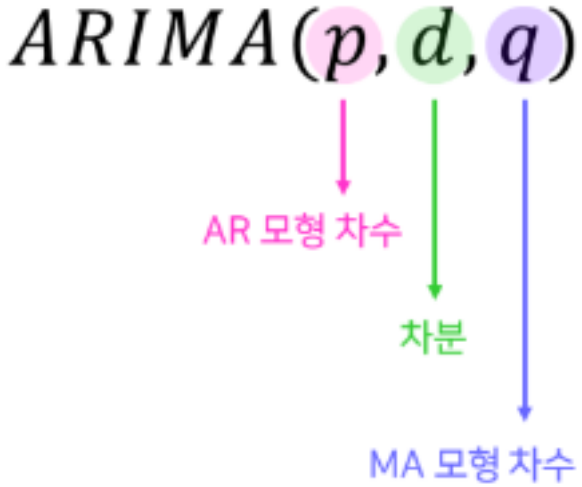
# 시계열모형

3) **자기회귀누적이동평균 모형(ARIMA(p, d, q) 모형**, autoregressive integrated moving average model)

: ARIMA 모형은 비정상시계열 모형이다.

: ARIMA 모형은 차분이나 변환을 통해 AR모형이나 MA모형, 이 둘을 합친 ARMA 모형으로 정상화 할 수 있다.

- P는 AR모형, q는 MA모형과 관련이 있는 차수이다.
- 시계열 $\{Z_t\}$의 d번 차분한 시계열이 ARMA 모형이면, 시계열 $\{Z_t\}$는 차수가 p, d, q인 ARIMA 모형, 즉 ARIMA(p,d,q)모형을 갖는다고 한다.




ARIMA는 차분, 변환을 통해
AR, MA, ARMA로 정상화

- p=0이면 IMA(d,q) -> d번 차분하면 MA(q)
- d=0이면 ARMA(p,q) -> 정상성 만족
- q=0이면 ARI(p,d) -> d번 차분하면 AR(p)

- ARIMA(1,1,2)의 경우에는 1차분 후 AR(1) MA(2) ARMA(1,2) 선택 활용

  -〉 이런 경우 가장 간단한 모형을 선택하거나 AIC를 적용하여 점수가 가장 낮은 모형을 선택한다.

출처 : https://blog.naver.com/tjgml1343/222077619748

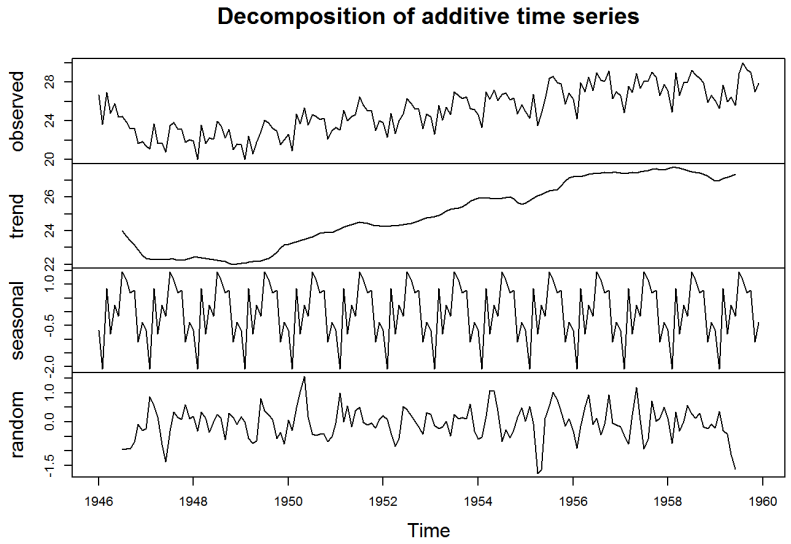
# 시계열모형

## 4) 분해시계열

: 시계열에 영향을 주로 일반적인 요인을 시계열에서 분리해 분석하는 방법. 회귀분석적인 방법을 주로 사용

$$Z_t = f(T_t, S_t, C_t, I_t)$$

- $T_t$ : 경향(추세)요인 : 자료가 오르거나 내리는 추세, 선형, 이차식 형태, 지수적 형태 등
- $S_t$ : 계절요인 : 요일, 월, 사계절 각 분기에 의한 변화 등 고정된 주기에 따라 자료가 변하는 경우
- $C_t$ : 순환요인 : 경제적이나 자연적인 이유 없이 알려지지 않은 주기를 가지고 변화하는 자료
- $I_t$ : 불규칙 요인 : 위의 세가지 요인으로 설명할 수 없는 오차에 해당하는 요인




시계열 분석 참고 : R 시계열 분석 Time Series ARIMA(woosa7.github.io/R-시계열분석-Time Series-ARMIA/)

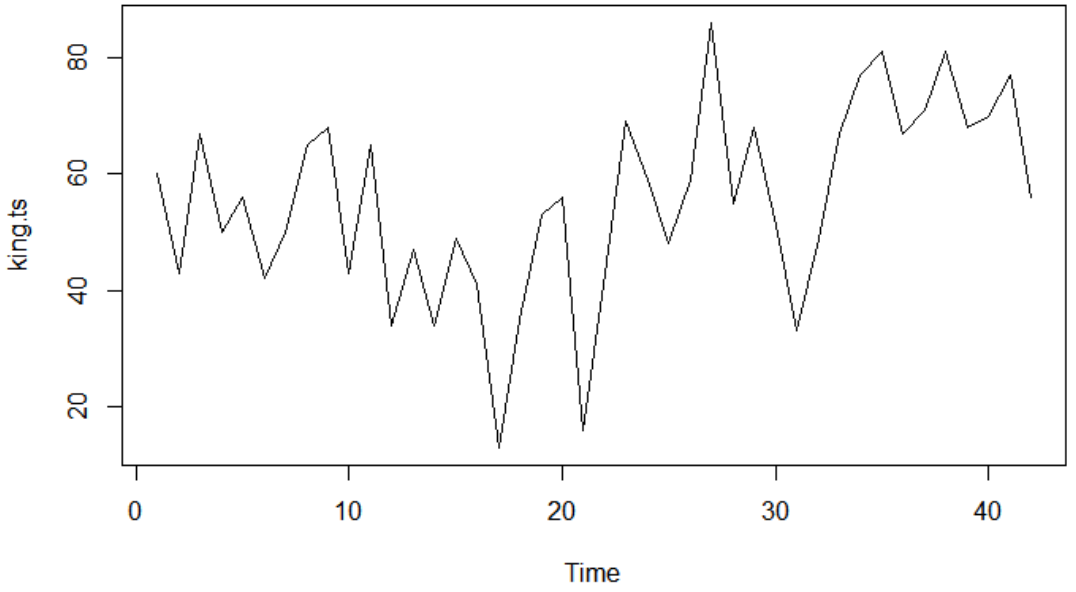
출처 : https://woosa7.github.io/R-%EC%8B%9C%EA%B3%84%EC%97%B4%EB%B6%84%EC%84%9D-Time-Series-ARIMA/

# R을 이용한 시계열분석

- 영국 왕들의 사망 시 나이 데이터를 이용한 시계열 분석

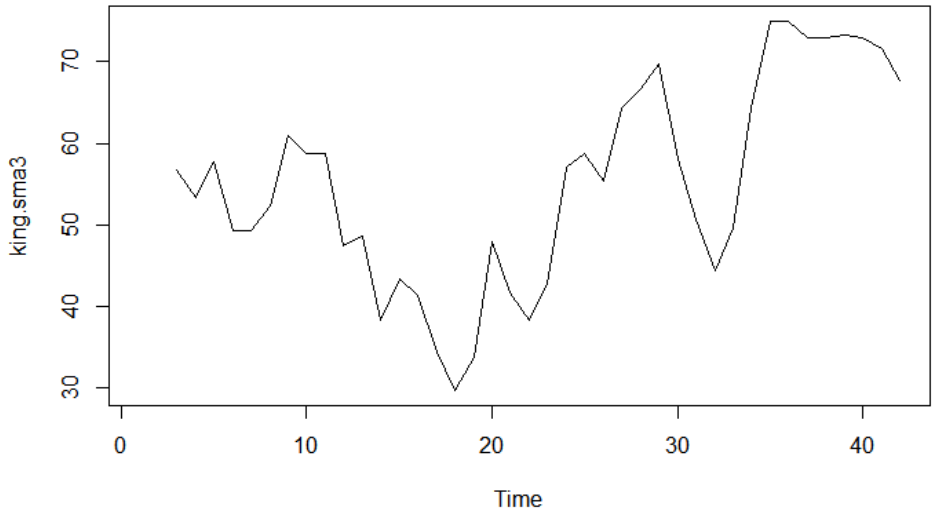
  - 영국 왕 42명의 사망 시 나이 예제는 비 계절성을 띄는 시계열 자료
  - 비계절성을 띄는 시계열 자료는 트렌드 요소, 불규칙 요소로 구성
  - 20번째 왕까지는 38세에서 55까지 수명을 유지하고, 그 이후부터는 수명이 늘어서 40번째 왕은 73세까지 생존

```
library(tseries)
library(forecast)
library(TTR)
king <- scan("http://robjhyndman.com/tsdldata/misc/kings.dat", skip = 3)
king.ts <- ts(king)
plot.ts(king.ts)
```

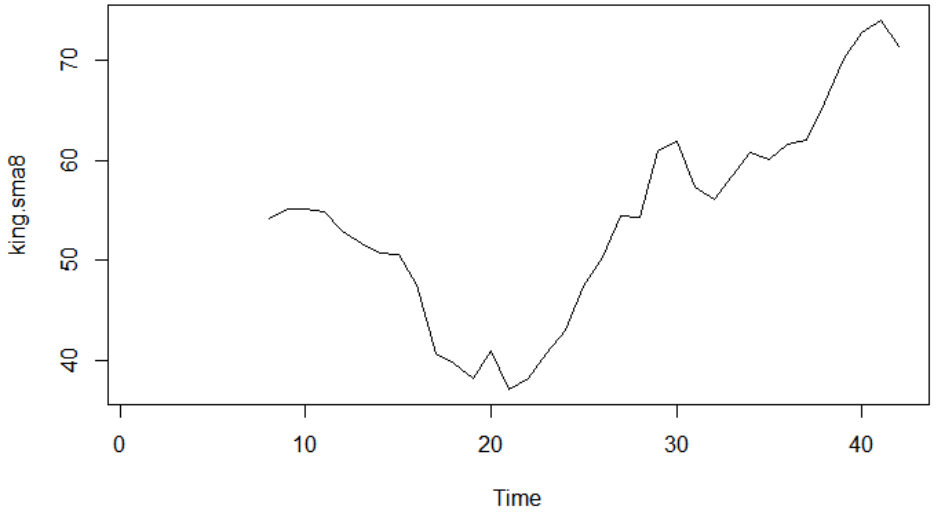


# R을 이용한 시계열분석

```
#3년마다 평균을 내서 그래프를 부드럽게 표현
king.sma3 <- SMA(king.ts, n=3)
plot.ts(king.sma3)
```



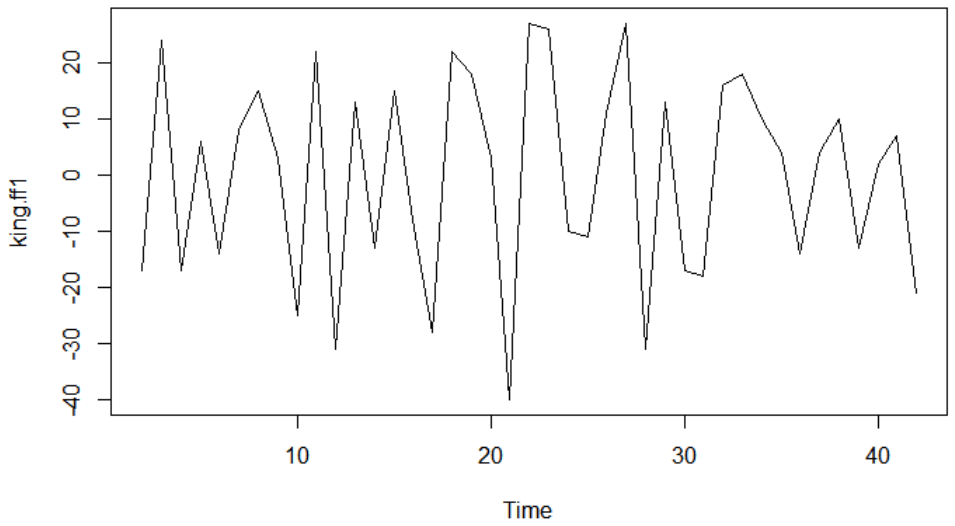

```
#8년마다 평균을 내서 그래프를 부드럽게 표현
king.sma8 <- SMA(king.ts, n=8)
plot.ts(king.sma8)
```



# ARIMA 모델

- ARIMA모델은 정상성 시계열에 한해 사용한다
- 비정상 시계열 자료는 차분해 정상성으로 만족하는 조건의 시계열로 바꿔준다
- 분해시계열 그래프에서 평균이 시간에 따라 일정치 않은 모습을 보이므로 비정상 시계열이다
- 1차 차분 결과에서 평균과 분산이 시간에 따라 의존하지 않음을 확인한다
- ARIMA(p,1,q)모델이며 차분을 1번 해야 정상성을 만족한다

```
#ARIMA적용->1차 차분
king.ff1 <- diff(king.ts, difference=1)
plot.ts(king.ff1)
```

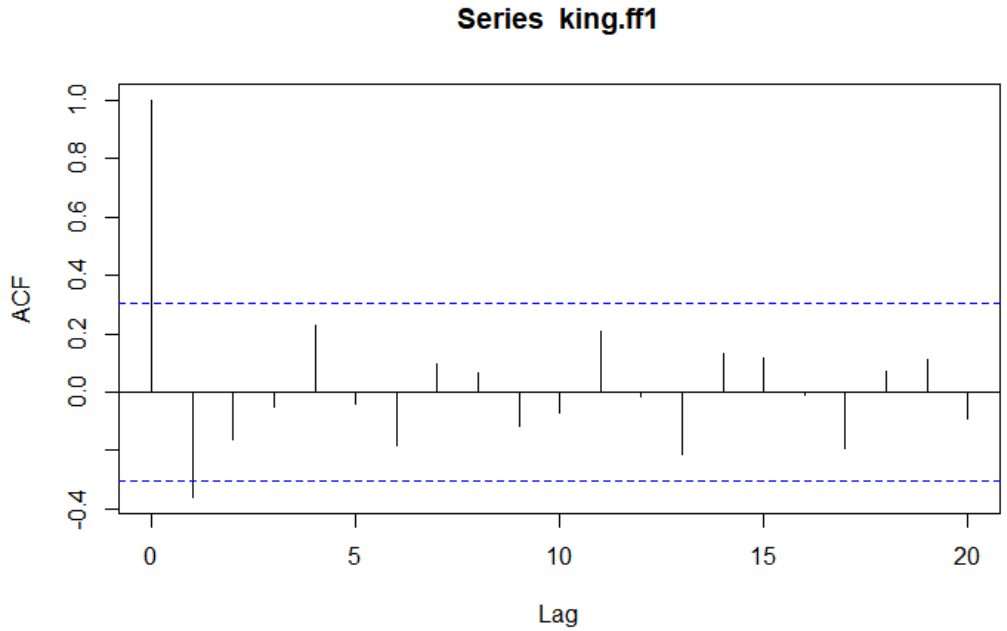


# ACF와 PACF를 통한 적합한 ARIMA 모델 결정

① ACF

- lag는 0부터 값을 갖는데, 너무 많은 구간을 설정하면 그래프를 보고 판단하기 어렵다.
- ACF값이 lag 1인 지점 빼고는 모두 점선 구간 안에 있고, 나머지는 구간 안에 있다.

```
#acf를 통해 ARIMA모델 결정
acf(king.ff1, lag.max=20)
acf(king.ff1, lag.max=20, plot=false)
```

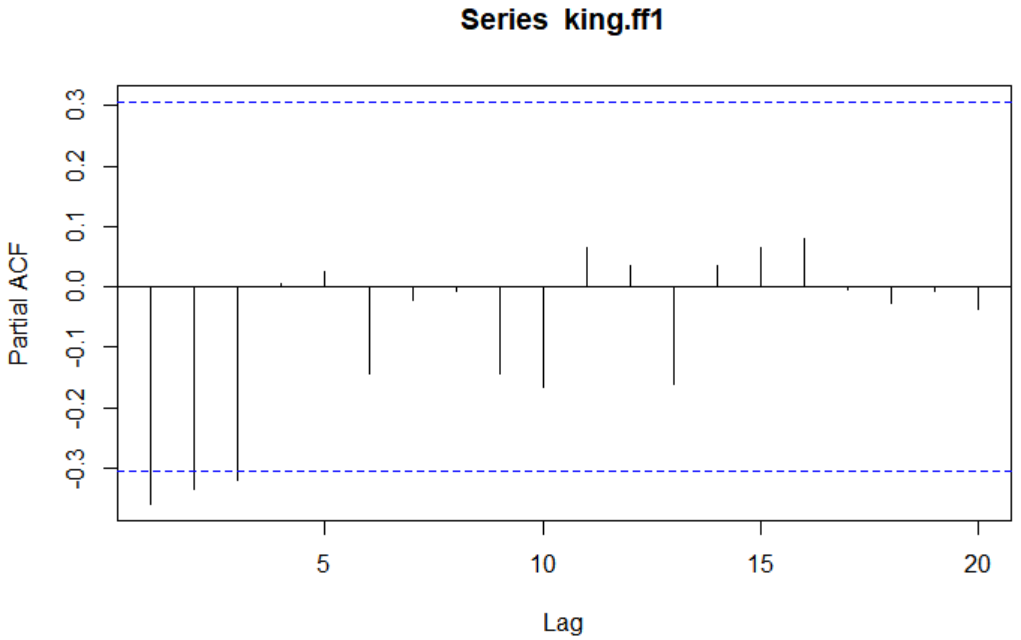


# ACF와 PACF를 통한 적합한 ARIMA 모델 결정

② PACF

- PACF 값이 lag 1, 2, 3에서 점선 구간을 초과하고 음의 값을 가지며 절단점이 lag 4이다.

```
#PACF를 통해 lag 절단점 결정
pacf(king.ff1, lag.max=20)
pacf(king.ff1, lag.max=20, plot=FALSE)
```



# 종합

- 아래와 같이 ARMA 후보들이 생성

  - ARMA(3,0) 모델 : PACF값이 lag4에서 절단점을 가짐, AR(3)모형
  - ARMA(0,1) 모델 : ACF값이 lag2에서 절단점을 가짐, MA(1)모형
  - ARMA(p,q) 모델 : 그래서 AR모형과 MA을 혼합

```
> auto.arima(king)
Series: king
ARIMA(0,1,1)

Coefficients:
          ma1
      -0.7218
s.e.   0.1208

sigma^2 = 236.2:  log likelihood = -170.06
AIC=344.13   AICc=344.44   BIC=347.56
```

```
> #ARIMA 적용
> king.arima <- arima(king, order=c(0,1,1))
> king.forecasts <- forecast(king.arima)
> king.forecasts
   Point Forecast    Lo 80    Hi 80    Lo 95     Hi 95
43       67.75063 48.29647 87.20479 37.99806  97.50319
44       67.75063 47.55748 87.94377 36.86788  98.63338
45       67.75063 46.84460 88.65665 35.77762  99.72363
46       67.75063 46.15524 89.34601 34.72333 100.77792
47       67.75063 45.48722 90.01404 33.70168 101.79958
48       67.75063 44.83866 90.66260 32.70979 102.79146
49       67.75063 44.20796 91.29330 31.74523 103.75603
50       67.75063 43.59372 91.90753 30.80583 104.69543
51       67.75063 42.99472 92.50653 29.88974 105.61152
52       67.75063 42.40988 93.09138 28.99529 106.50596
```

# 다차원 척도법(MultiDimensional Scaling)

- MDS(MultiDimensional Scaling)은 군집분석과 같이 개체들을 대상으로 변수들을 측정한 후, 개체들 사이의 유사성/비유사성을 측정하여 개체들을 2차원 또는 3차원 공간상에 점으로 표현하는 분석 방법

## 1. 다차원척도법

: 객체간 근접성(Proximity)을 시각화하는 통계기법

: 군집분석과 같이 개체들을 대상으로 변수들을 측정한 후에 개체들 사이의 유사성/비유사성을 측정하여 개체들 사이의 집단화를 2차원 또는 3차원 공간상에 점으로 표현.

## 2. 목적

: 데이터 속에 잠재해 있는 패턴, 구조를 찾아내어 소수 차원의 공간에 기하학적으로 표현한다.

: 데이터 축소(Data Reduction)의 목적으로 다차원척도법을 이용. 즉, 데이터에 포함되는 정보를 끄집어내기 위해 탐색수단으로써 사용

: 다차원축소법에 의하여 얻은 결과를 데이터가 만들어진 현상이나 과정에 고유의 구조로서 의미를 부여

# 다차원 척도법(MultiDimensional Scaling)

3. 방법

: 개체들의 거리 계산에는 '유클리드 거리행렬'을 이용한다.

$$d_{ii} = \sqrt{\left(x_{iI} - x_{iI}\right)^2 + \ldots + \left(x_{iR} - x_{iR}\right)^2}$$

〈유클리드 거리 Euclidean distance〉

: 다차원 공간에서 두 점 간의 거리, 자로 측정한 거리의 일종

: 관측대상들의 상대적 거리의 정확도를 높이기 위해 적합 정도를 스트레스 값(Stress Value)으로 나타냄.

$$S = \sqrt{\frac{\sum_{i=0j=1}^{n} \left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i=0j=1}^{n} d_{ij}^2}}$$ , $d_{ij}$ = 관측대상 $i$부터 $j$까지의 실제거리, $\hat{d}_{ij}$ = 프로그램에의해 추정된 거리

| STRESS | 적합도 수준 |
|---|---|
| 0 | perfect |
| 0.05 이내 | excellent |
| 0.05~0.1 | satisfactory |
| 0.1~0.15 | acceptable, but doubt |
| 0.15 이상 | poor |

# 다차원 척도법(MultiDimensional Scaling)

## 4. 종류

1) 계량적 MDS(Metric MDS)

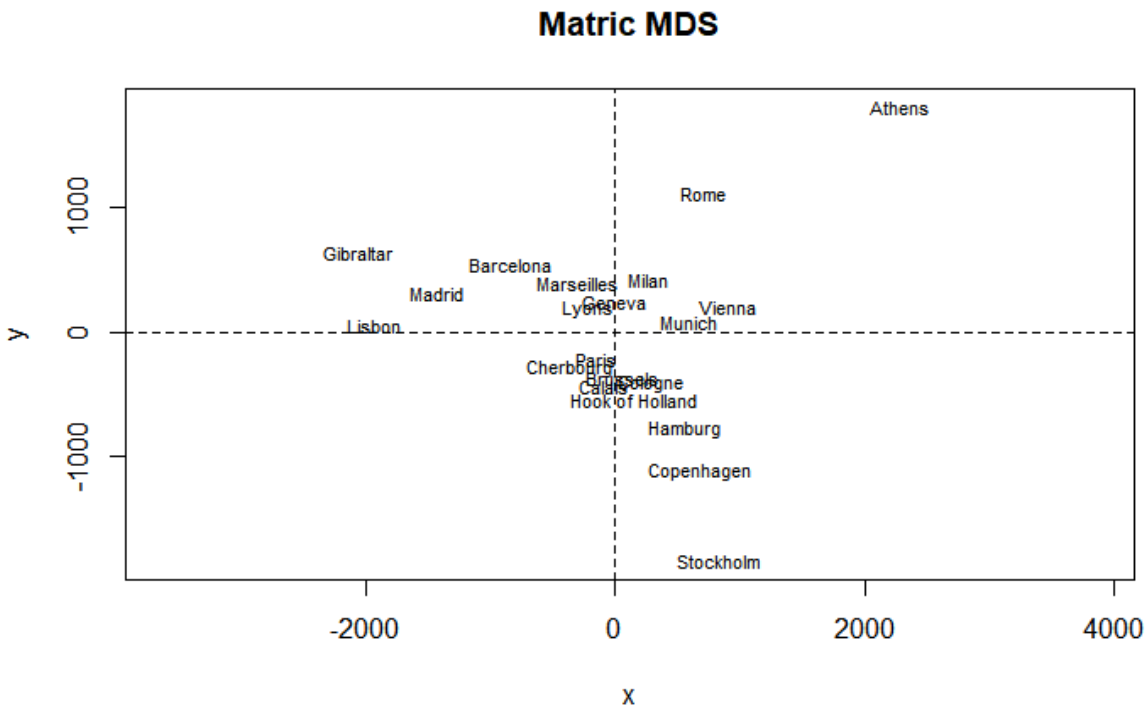
: 데이터가 구간척도나 비율척도인 경우 활용, 전통적인 다차원척도법

: N개의 케이스에 대해서 p개의 특성변수가 있는 경우, 각 개체들간의 유클리드 거리행렬을 계산하고 개체들간의 비유사성 S (거리제곱 행렬의 선형함수)를 공간상에 표현

- cmdscale 사례
- MASS package의 eurodist 자료를 이용한다.

```
library(MASS)
loc <- cmdscale(eurodist) #2차원으로 21개 도시들을 매핑

x <- loc[,1]
y <- loc[,2]
plot(x, y, type='n', asp=1, main="Matric MDS")
text(x,y,rownames(loc), cex=0.7)#종축은 북쪽 도시를 상단에 표시하기 위해 부호 변경
abline(v=0,h=0,lty=2,lwd=0.5)
```



# 다차원 척도법(MultiDimensional Scaling)

## 4. 종류

2) 비계량적 MDS(Nonmetric MDS)

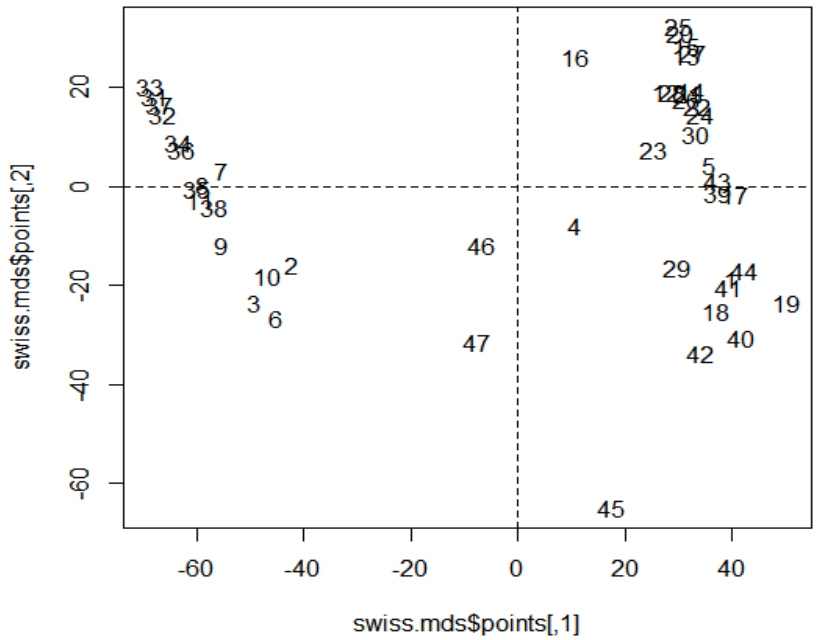
: 데이터가 순서척도인 경우 활용

: 개체들간의 거리가 순서로 주어진 경우에는 순서척도를 거리의 속성과 같도록 변환(monotone transformation)하여 거리를 생성한 후 적용

- isoMDS 사례
- MASS package의 Swiss 자료를 이용하여 2차원으로 도시들을 매핑.
  1888년경 스위스연방 중 47개의 불어권 주의 토양의 비옥도 지수와 여러 사회경제적 지표를 측정한 자료이다.

```
library(MASS)

data(swiss)
swiss.x <- as.matrix(swiss[,-1])
swiss.dist <- dist(swiss.x)
swiss.mds <- isoMDS(swiss.dist)
plot(swiss.mds$points, type="n")
text(swiss.mds$points, labels=as.character(1:nrow(swiss.x)))
abline(v=0, h=0, lty=2, lwd=0.5)
```



# 주성분 분석(Principal Componet Analysis)

## 1. 주성분분석

- 여러 변수들의 변량을 '주성분(Principal Component)이라는 서로 상관성이 높은 변수들의 선형결합으로 만들어 기존의 상관성이 높은 변수들을 요약, 축소하는 기법
- 첫번째 주성분으로 전체 변동을 가장 많이 설명할 수 있도록 하고, 두 번째 주성분으로는 첫번째 주성분이 설명하지 못하는 나머지 변동(첫번째 주성분과는 상관성이 낮아서)을 정보의 손실 없이 가장 많이 설명할 수 있도록 변수들의 선형조합을 만듦.

## 2. 목적

- 여러 변수들 간에 내재하는 상관관계, 연관성을 이용해 소수의 주성분으로 차원을 축소함으로써 데이터를 이해하기 쉽고 관리하기 쉽게 함.
- 다중공선성이 존재하는 경우(회귀분석 또는 의사결정나무 등의 모형 개발 시) 상관성이 없는 주성분으로 변수들을 축소하여 모형개발에 활용
- 연관성이 높은 변수를 주성분분석을 통해 차원을 축소한 후에 군집분석을 수행하면 군집화 결과와 연산속도를 개선
- 기계에서 나오는 다수의 센서데이터를 주성분 분석으로 차원을 축소한 후에 시계열로 분포나 추세의 변화를 분석하면 기계의 고장징후를 사전에 파악하는데 활용.

# 주성분 분석(Principal Component Analysis)

- PCA는 먼저 데이터에 가장 가까운 초평면(hyperplane)을 구한 다음, 데이터를 초평면에 투영(Projection) 시킴.

  : 고차원의 원본 데이터를 저 차원의 부분 공간으로 투영하여 데이터를 축소하는 방법

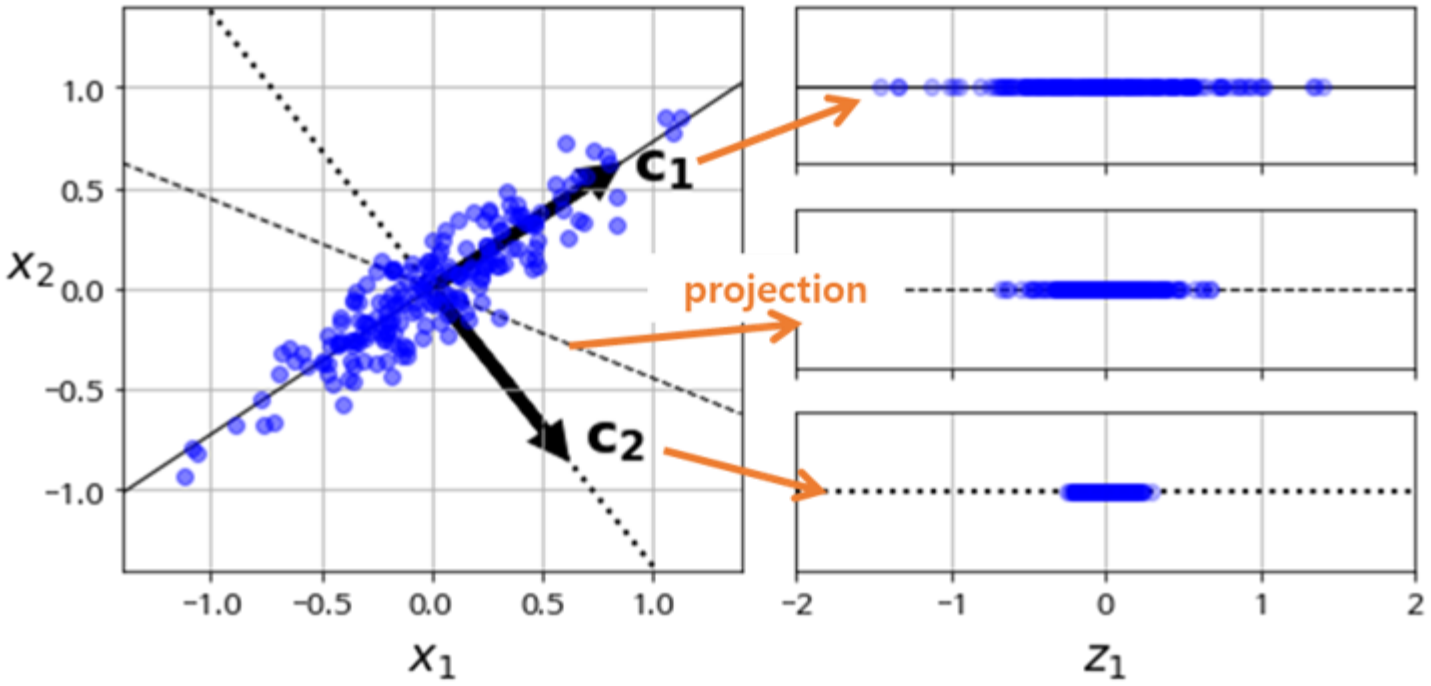
  : PCA는 원본 데이터가 가지는 데이터 변동성을 가장 중요한 정보로 간주하며, 이 변동성에 기반한 원본 데이터 투영으로 차원 축소를 수행.

  : PCA는 원본 데이터 변동성이 가장 큰 방향으로 순차적으로 축들을 생성하고, 이렇게 생성된 축으로 데이터를 투영하는 방식.

## 1) 분산 보존

: 저차원의 초평면에 데이터를 투영하기 전에 먼저 적절한 초평면을 선택해야 함.

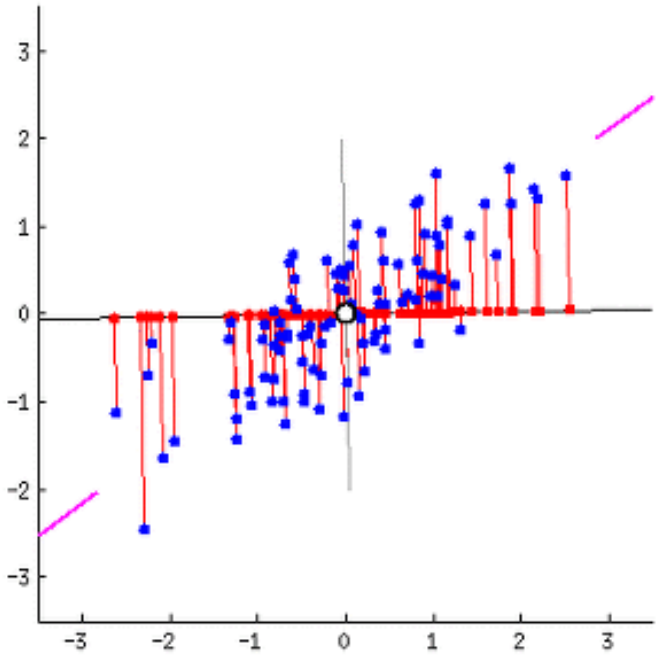
: PCA는 데이터의 분산이 최대가 되는 축을 찾음. 즉, 원본 데이터셋과 투영된 데이터셋 간의 평균제곱거리를 최소화하는 축을 찾는다.



# 주성분 분석(Principal Component Analysis)

## 2) 주성분(Principal Component)

1. 학습 데이터셋에서 분산이 최대인 축(axis)을 찾는다.
2. 이렇게 찾은 첫번째 축과 직교(Orthogonal)하면서 분산이 최대인 두 번째 축을 찾는다.
3. 첫번째 축과 두번째 축에 직교하고 분산을 최대한 보존하는 세번째 축을 찾는다.
4. 1~3과 같은 방법으로 데이터셋의 차원(특성 수)만큼의 축을 찾는다.



# 주성분 분석(Principal Componet Analysis)

3. 주성분 분석 VS 요인 분석

1) 요인분석(Factor Analysis) : 등간척도(혹은 비율척도)로 측정한 두개 이상의 변수들에 잠재되어있는 공통인자를 찾아내는 기법.

2) 공통점 : 모두 데이터를 축소하는데 활용, 원래 데이터를 활용해서 몇 개의 새로운 변수들을 만들 수 있음.

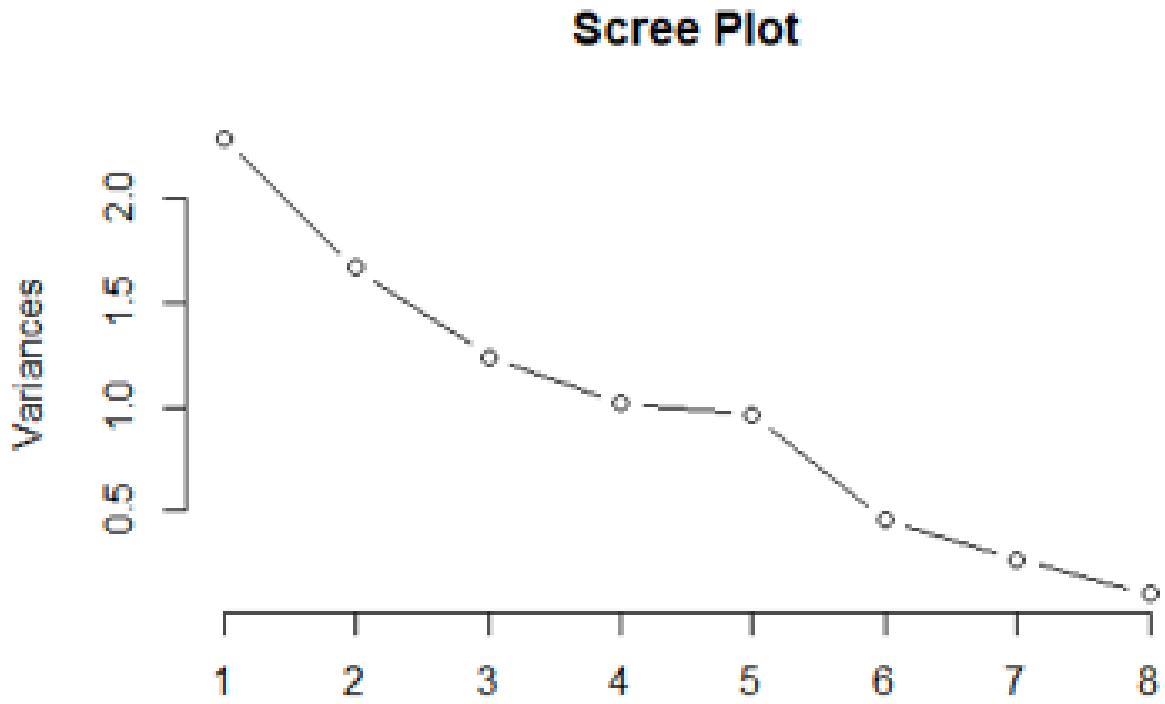
3) 차이점

| | 요인분석 | 주성분분석 |
|---|---|---|
| 생성된 변수의 수 | 몇 개라고 지정없이 만들 수 있음. | 제1, 제2, 제3주성분 정도로 활용<br>(4개 이상 넘지 않음) |
| 생성된 변수의 이름 | 분석자가 요인의 이름을 명명 | 제1, 제2주성분 등으로 표현 |
| 생성된 변수들 간의 관계 | 새 변수들을 기본적으로 대등한 관계를 갖고 '어떤 것이 더 중요하다'라는 의미가 없음,<br>단, 분류/예측에 그 다음 단계로 사용된다면 그 때 중요성의 의미가 부여 | 제1주성분이 가장 중요하고 그 다음 제2주성분이 중요하게 취급 |
| 분석 방법의 의미 | 목표변수를 고려하지 않고 그냥 데이터가 주어지면 변수들을 비슷한 성격들로 묶어서 새로운 [잠재] 변수들을 만듦. | 목표변수를 고려하여 목표변수를 잘 예측/분류하기 위하여 원래 변수들의 선형결합으로 이루어진 몇 개의 주성분(변수)들을 찾아냄 |

# 주성분 분석(Principal Componet Analysis)

4. 주성분 선택법

: 주성분 분석의 결과에서 누적기여율(Cumulative Proportion)이 85%이상이면 주성분의 수로 결정할 수 있다.




: Scree plot을 활용하여 고유값(eigenvalue)이 수평을 유지하기 전단계로 주성분의 수를 선택.

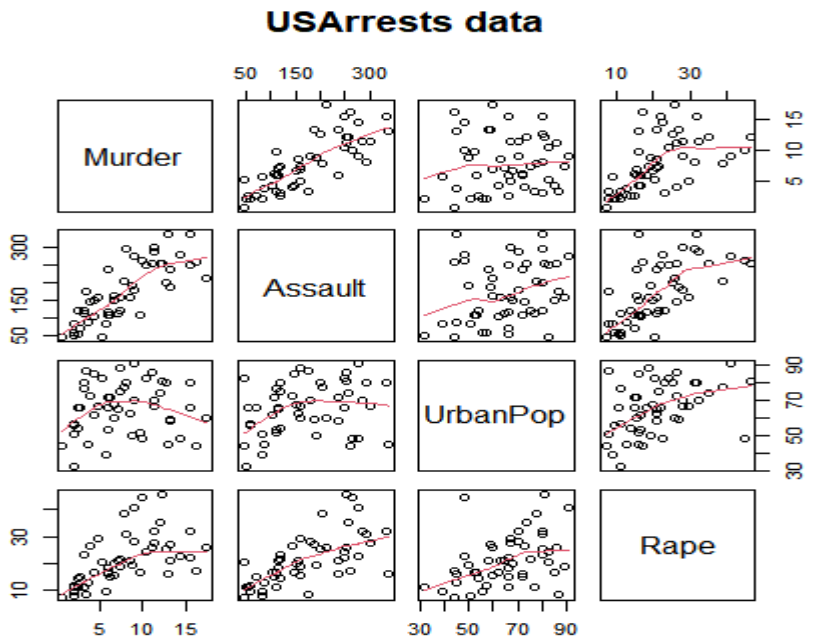
# 주성분 분석(Principal Componet Analysis)

5. 분석사례

- 1973년 미국 50개 주의 100,000명의 인구 당 체포된 세가지 강력범죄수(assault, murder, rape)와 각 주마다 도시에 거주하는 인구의 비율(%)
- 변수들 간의 척도의 차이가 상당히 크기 때문에 상관행렬을 사용하여 분석
- 특이치 분해를 사용하는 경우 자료 행렬의 각 변수의 평균과 제곱의 합이 1로 표준화되었다고 가정할 수 있다.

1) 4개의 변수들 간의 산점도

```
library(datasets)
data(USArrests)
pairs(USArrests, panel=panel.smooth, main="USArrests data")
```




Murder와 UrbanPop비율 간의 관련성이 작아 보인다.

# 주성분 분석(Principal Componet Analysis)

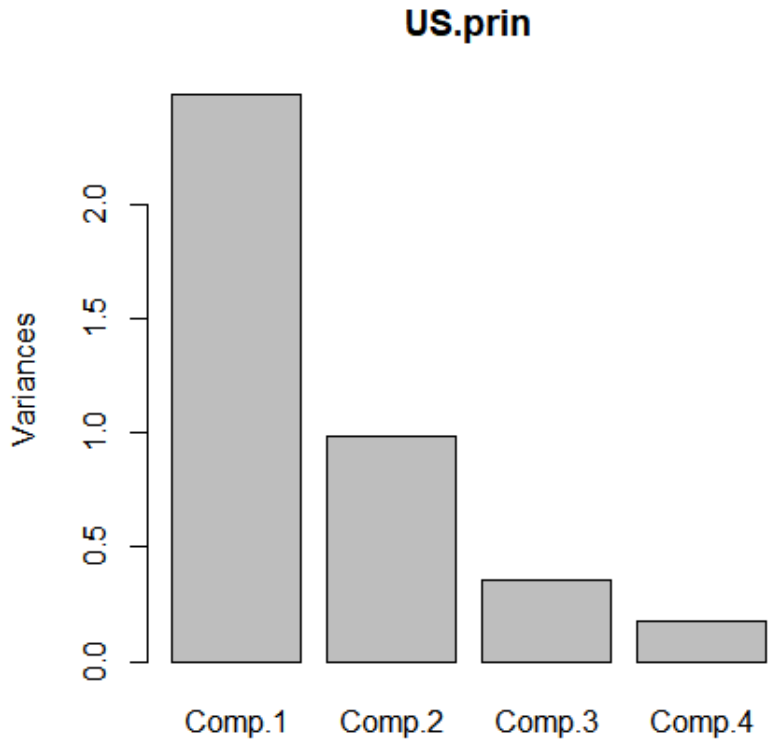
5. 분석사례

- 1973년 미국 50개 주의 100,000명의 인구 당 체포된 세가지 강력범죄수(assault, murder, rape)와 각 주마다 도시에 거주하는 인구의 비율(%)
- 변수들 간의 척도의 차이가 상당히 크기 때문에 상관행렬을 사용하여 분석
- 특이치 분해를 사용하는 경우 자료 행렬의 각 *변수의 평균과 제곱의* 합이 1로 표준화되었다고 가정할 수 있다.

1) Summary

```
US.prin <- princomp(USArrests, cor = TRUE)
summary(US.prin)
screeplot(US.prin, npcs=4, tpye="lines")
```

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4
Standard deviation     1.5748783 0.9948694 0.5971291 0.41644938
Proportion of Variance 0.6200604 0.2474413 0.0891408 0.04335752
Cumulative Proportion  0.6200604 0.8675017 0.9566425 1.00000000
```




제1성분과 제2성분까지의 누적 분산비율은 대략 86.8%로 2개의 주성분 변수를 활용하여 전체 데이터의 86.8%를 설명

# 주성분 분석(Principal Componet Analysis)

```
> loadings(US.prin)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4
Murder    0.536  0.418  0.341  0.649
Assault   0.583  0.188  0.268 -0.743
UrbanPop  0.278 -0.873  0.378  0.134
Rape      0.543 -0.167 -0.818

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
```

네 개의 변수가 각 주성분 Comp1 1~4까지 기여하는 가중치가 제시

제1성분에는 네 개의 변수가 평균적으로 기여

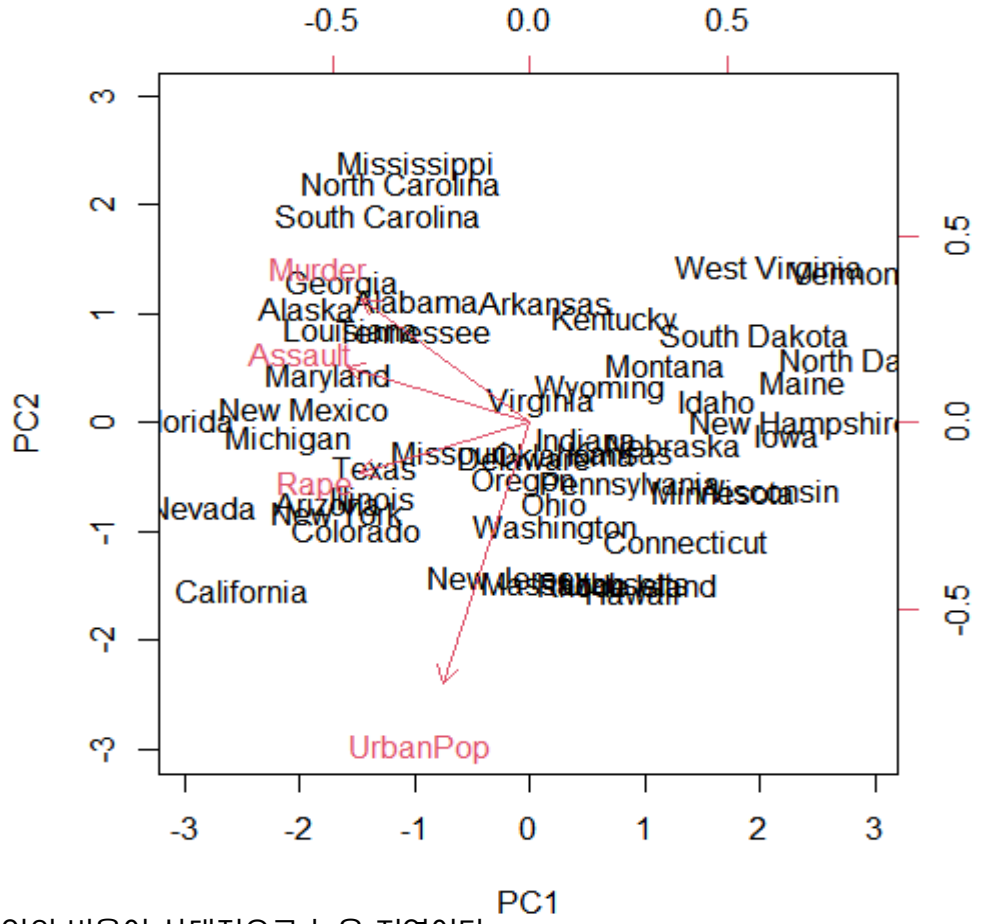
제2성분에는 (Murder, Assault)와 (Urbanpop, Rape)의 계수의 부호가 서로 다르다.

```
> US.prin$scores
                 Comp.1      Comp.2      Comp.3      Comp.4
Alabama      0.98556588  1.13339238  0.44426879  0.156267145
Alaska       1.95013775  1.07321326 -2.04000333 -0.438583440
Arizona      1.76316354 -0.74595678 -0.05478082 -0.834652924
Arkansas    -0.14142029  1.11979678 -0.11457369 -0.182810896
California   2.52398013 -1.54293399 -0.59855680 -0.341996478
Colorado     1.51456286 -0.98755509 -1.09500699  0.001464887
Connecticut -1.35864746 -1.08892789  0.64325757 -0.118469414
Delaware     0.04770931 -0.32535892  0.71863294 -0.881977637
Florida      3.01304227  0.03922851  0.57682949 -0.096284752
Georgia      1.63928304  1.27894240  0.34246008  1.076796812
Hawaii      -0.91265715 -1.57046001 -0.05078189  0.902806864
Idaho       -1.63979985  0.21097292 -0.25980134 -0.499104101
Illinois     1.37891072 -0.68184119  0.67749564 -0.122021292
Indiana     -0.50546136 -0.15156254 -0.22805484  0.424665700
Iowa        -2.25364607 -0.10405407 -0.16456432  0.017555916
Kansas      -0.79688112 -0.27016470 -0.02555331  0.206496428
Kentucky    -0.75085907  0.95844029  0.02836942  0.670556671
```

각 주성분 Comp 1~4의 선형식을 통해 각 지역(record)별로 얻은 결과를 계산.

# 주성분 분석(Principal Componet Analysis)

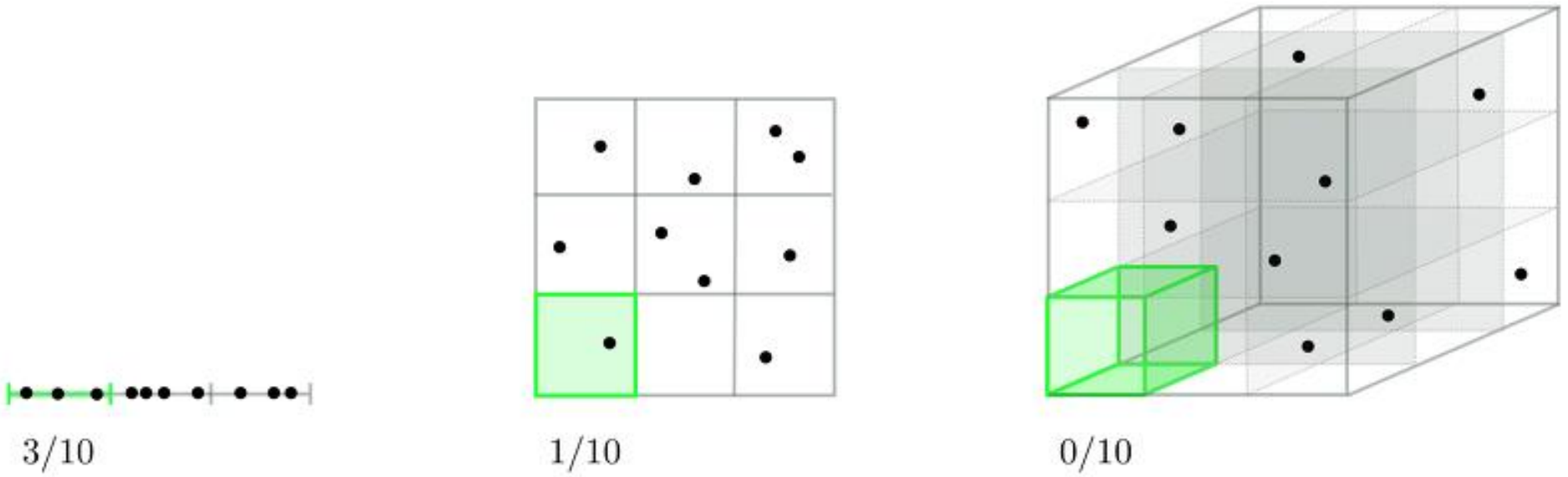
```
arrests.pca <- prcomp(USArrests,center = TRUE,scale. = TRUE)
biplot(arrests.pca, scale=0)
```




- 조지아, 메릴랜드, 뉴 멕시코 등은 폭행과 살인의 비율이 상대적으로 높은 지역이다.
- 미시간, 텍사스 등은 강간의 비율이 높은 지역이다.
- 콜로라도, 캘리포니아, 뉴저지 등은 도시에 거주하는 인구의 비율이 높은 지역이다.
- 아이다호, 뉴 햄프셔, 아이오와 등의 도시들은 도시에 거주하는 인구의 비율이 상대적으로 낮으면서 3대 강력범죄도 낮다.

# 차원의 저주

- 데이터 학습을 위해 차원이 증가하면서 학습데이터 수가 차원의 수보다 적어져 성능이 저하되는 현상
- 차원이 증가할 수록 개별 차원 내 학습할 데이터 수가 적어지는(Sparse) 현상 발생



# 연습문제

1. 다음 시계열분석에서 정상성의 특징이 아닌 것은?

① 평균이 일정하다. 즉, 모든 시점에 대해 일정한 평균을 가진다.

② 분산도 시점에 의존하지 않는다.

③ 자기회귀식에는 백색잡음이 없다.

④ 공분산은 단지 시차에만 의존하고 실제 어느 시점 t,s에는 의존하지 않는다.

# 연습문제

2. 시계열을 구성하는 4가지 요소에 해당되지 않은 것은?

① 계절요인

② 교호요인

③ 순환요인

④ 추세요인

# 연습문제

3. Data는 메이저리그에서 활약하는 263명의 선수에 대한 타자 기록으로 연봉(Salary)을 비롯한 17개 변수를 포함하고 있다. 아래는 17개의 변수들을 사용하여 주성분분석을 시행한 결과이다. 다음 설명 중 잘못된 것은?

```
> pca=princomp(data,cor=TRUE)
> summary(pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     2.7733967 2.0302601 1.3148557 0.9575410 0.84109683 0.72374220 0.69841796 0.50090065 0.42525940
Proportion of Variance 0.4524547 0.2424680 0.1016968 0.0539344 0.04161435 0.03081193 0.02869339 0.01475891 0.01063797
Cumulative Proportion  0.4524547 0.6949227 0.7966195 0.8505539 0.89216822 0.92298014 0.95167354 0.96643244 0.97707042
                          Comp.10     Comp.11     Comp.12     Comp.13     Comp.14      Comp.15       Comp.16
Standard deviation     0.363901982 0.312011679 0.243641510 0.232044829 0.163510472 0.1186398422 0.0693395039
Proportion of Variance 0.007789685 0.005726546 0.003491834 0.003167341 0.001572687 0.0008279654 0.0002828216
Cumulative Proportion  0.984860104 0.990586651 0.994078485 0.997245826 0.998818513 0.9996464785 0.9999293001
                          Comp.17
Standard deviation     3.466841e-02
Proportion of Variance 7.069994e-05
Cumulative Proportion  1.000000e+00
```

① 최소 80% 이상의 분산설명력을 갖기 위해서는 4개 이상의 주성분을 사용해야 한다.

② 가장 큰 분산설명력을 가지는 주성분은 전체 분산의 45.25%를 설명한다.

③ 공분산행렬을 사용하여 주성분분석을 시행한 것이다.

④ 17차원을 2차원으로 축소한다면 잃게 되는 정보량은 약 30.5%이다.

# 연습문제

4. 교차분석은 2개 이상의 변수를 결합하여 자료의 빈도를 살펴보는 기법이다. 다음 중 교차분석에 대한 설명으로 부적절한 것은 무엇인가?

① 범주의 관찰도수에 비교될 수 있는 기대도수를 기대한다.

② 교차분석은 두 문항 모두 범주형 변수가 아니어도 사용할 수 있으며, 두 변수 간 관계를 보기 위해 실시한다.

③ 교차분석은 교차표를 작성하여 교차빈도를 집계할 뿐 아니라 두 변수들 간의 독립성 검정을 할 수 있다.

④ 기대빈도가 5미만인 셀의 비율이 20%를 넘으면 카이제곱분포에 근사하지 않으며, 이런 경우 표본의 크기를 늘리거나 변수의 수준을 합쳐 셀의 수를 줄이는 방법 등을 사용한다.

# 연습문제

5. 시계열의 요소분해법은 시계열 자료가 몇 가지 변동들의 결합으로 이루어져 있다고 보고 변동요소별로 분해하여 쉽게 분석하기 위한 것이다. 다음 중 분해 요소에 대한 설명이 부적절한 것은?

① 추세 분석은 장기적으로 변해가는 큰 흐름을 나타내는 것으로 자료가 장기적으로 커지거나 작아지는 변화를 나타내는 요소이다.

② 계절변동은 일정한 주기를 가지고 반복적으로 같은 패턴을 보이는 변화를 나타내는 요소이다.

③ 순환변동은 경제 전반이나 특정 산업의 부침을 나타내 주는 것을 말한다.

④ 불규칙변동은 불규칙하게 변동하는 급격한 환경변화, 천재기변 같은 것으로 발생하는 변동을 말한다.

# 연습문제

6. 다음 중 아래 주성분 분석을 시행한 결과에 대한 설명으로 가장 부적합한 것은?

```
> college_s<-scale(college)
> summary(college_s)
    Outstate          Room.Board          Books             Personal          Grad.Rate
 Min.   :-2.0136   Min.   :-2.3503   Min.   :-2.7460   Min.   :-1.6108   Min.   :-3.22880
 1st Qu.:-0.7757   1st Qu.:-0.6935   1st Qu.:-0.4808   1st Qu.:-0.7247   1st Qu.:-0.72555
 Median :-0.1120   Median :-0.1436   Median :-0.2991   Median :-0.2077   Median :-0.02697
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 0.6175   3rd Qu.: 0.6314   3rd Qu.: 0.3066   3rd Qu.: 0.5308   3rd Qu.: 0.72982
 Max.   : 2.7987   Max.   : 3.4344   Max.   :10.8453   Max.   : 8.0632   Max.   : 3.05842
> fit<-princomp(college_s)
> fit$loadings

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Outstate    0.587         0.155  0.142  0.779
Room.Board  0.531  0.230  0.155  0.574 -0.557
Books              0.812 -0.561 -0.153
Personal   -0.329  0.532  0.776
Grad.Rate   0.514         0.187 -0.789 -0.279
```

① 두 번째 주성분은 0.230*Room.Board+0.812*Books+0.532*Personal로 계산된다.

② 두 번째 주성분에 가장 큰 영향을 미치는 원변수는 Books이다.

③ Personal 값이 클수록 첫 번째 주성분은 작아진다.

④ fit<-princomp(college, cor=T)의 결과는 위의 결과와 다르다.

# 연습문제

7. 아래 주성분분석의 결과에서 두 개의 주성분을 사용할 때 설명 가능한 전체 분산의 비율은?

```
> model<-princomp(Car)
> summary(model)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation        1.503     1.075     0.840     0.752     0.555
Proportion of Variance    0.453     0.231     0.141     0.113     0.061
Cumulative Proportion     0.453     0.684     0.825     0.938     1.000
```


Thank you.
ADSP / 류영표 강사
ryp1662@gmail.com