**시계열 분석**

**1. 시계열 자료 (time series data)**

**1.1. 시계열 자료 정의**

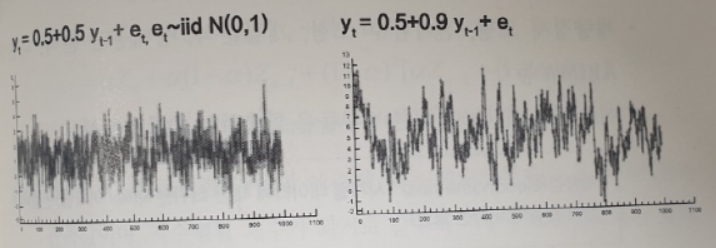
* 시간의 흐름(경과)에 따라 순서대로(ordered in time) 관측되는 자료

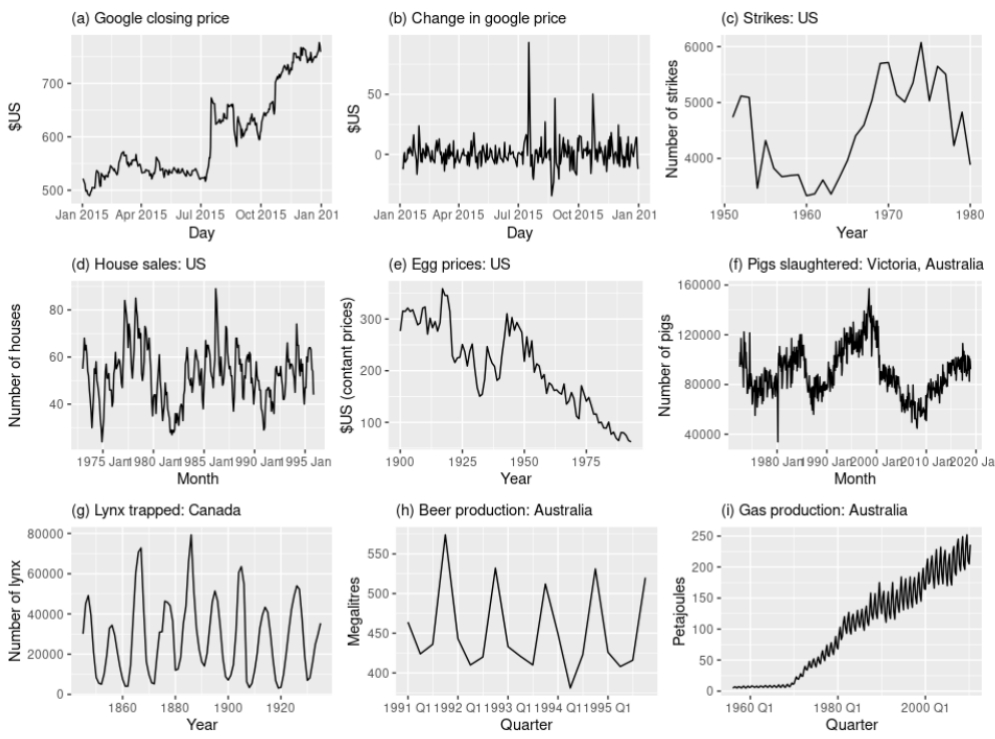
예) 국내총생산(GDP), 물가지수, 판매량, 종합주가지수, 강우량, 실험 및 관측자료 등

* 시계열 자료 표기

**1.2. 시계열 자료 특징**

* 시간의 흐름에 따라 관측된 자료이므로 시간에 영향을 받음
* 따라서 시계열 자료를 분석할 때 관측 시점들 간의 시차(time lag)가 중요한 역할을 함. 예를 들어 오늘의 주가가 한달 전, 일주일 전의 주가보다는 어제의 주가에 더 많은 영향을 받는 것과 마찬가지로 가까운 관측시점일수록 관측자료들 간에 상관관계가 커짐
* 실제로 많은 시계열 자료들이 연속적으로 생성되고 있지만 일정한 시차를 두고 관측되므로 이산 시계열 자료 형태를 지니는 경우가 많음
* 정상성(stationarity)
  + 시계열의 특성이 시간에 상관없이 일정; 시계열의 특성이 시간에 상관없이 평균, 분산이 일정 & 공분산도 단지 시차에만 의존하고 실제 특정 시점에는 의존하지 않을 때
* 평균이 일정한 경우
* 모든 시점에 대해 일정한 평균을 가짐
* 평균이 일정하지 않은 시계열은 차분(difference)을 통해 정상성을 회복할 수 있음
* 차분(difference)은 현시점 자료에서 전 시점 자료를 빼는 것; 일반차분(regular difference)은 바로 전 시점의 자료를 빼는 방법, 계절차분(seasonal difference)은 여러 시점 전의 자료를 빼는 방법
* 1차 차분 데이터를 대상으로 시계열 분석 & 예측을 진행했을때 예측값은 아래와 같이 구한다
* 분산이 일정한 경우
  + 모든 시점에 일정한 분산을 가짐
  + 분산이 일정하지 않은 시계열은 변환(transformation)을 통해 정상성을 회복할 수 있음
  + 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 일반화할 수 없다





* 변동성(variation)
  + 불규칙 변동(irregular variation) 또는 확률적 변동(random variation)
    - 시계열 자료에서 시간에 따른 규칙적인 움직임과는 달리 어떤 규칙성이 없이 예측이 불가능하게 우연적으로 발생하는 변동
    - 예, 전쟁, 홍수, 화재, 지진, 파업 등

라인, 텍스트, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* + 체계적 변동(systematic variation)
    - 장기간에 걸쳐 어떤 추세로 나타나는 추세 변동(trend variation), 추세선을 따라 주기적으로 오르고 내림을 반복하는 순환 변동(cyclical variation), 그리고 계절적 요인이 작용하여 1년 주기로 나타나는 계절 변동(seasonal variation) 이 있음
    - 추세변동
      * 시계열 자료가 갖는 장기적인 변화 추세임
      * 추세란 장기간(일반적으로 1년 초과)에 걸쳐 지속적으로 증가 또는 감소하거나 또는 일정산 상태(stationary)를 유지하려는 성향을 의미함
      * 그러므로 시계열 자료에서 짧은 기간 동안에는 추세변동을 찾기 어려움. 따라서 추세변동은 짧은 기간 동안 급격하게 변동하는 것이 아니라 장기적인 추세경향이 나타나는 것으로 직선이나 부드러운 곡선의 연장선으로 표시함. 이러한 추세는 직선뿐만 아니라 곡선, S자 형태의 추세를 가질 수도 있음
      * 추세 파악은 무작위(random)적인 잡음을 제거하여 큰 흐름을 파악
      * 예, 국내총생산, 인구증가율, 기술변화 등

라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

* + - 순환변동
      * 대체로 2~3년 정도의 일정한 기간을 주기로 순환적으로 나타남. 즉 1년 이내의 주기로 곡선을 그리며 추세변동에 따라 변동하는 것을 말함
      * 시간의 흐름에 따라 상하로 반복되는 변동으로 추세선을 따라 변화하는 것임
      * 예, 경기변동곡선은 불황과 경기회복, 호황과 경기후퇴로 인하여 수년을 주기로 나타나고 있는 순환변동의 예

라인, 그래프, 도표, 경사이(가) 표시된 사진

자동 생성된 설명

* + - 계절변동
      * 시계열 자료에서 보통 계절적 영향과 사회적 관습에 따라 1년 주기로 발생하는 변동요인을 말하며 보통 계절에 따라 순환하며 변동하는 특성을 지님
      * 계절변동이 순환변동과 다른 점은 순환주기가 짧다는 점임. 그러나 대부분의 경제관련 시계열들은 추세와 계절요인을 동시에 포함함. 이는 경제성장에 따라 백화점의 판매액, 해외여행자수, 청량음료, 전력소비향 등과 같이 계절상품 판매량 자료들이 시간의 변화에 따라 증가하기 때문임
      * 예, 설, 추석 등 명절요인 등

라인, 그래프, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

**1.3. 시계열 자료 종류**

* 비정상성/정상성 시계열 자료
  + 시계열 자료는 시계열 자료의 정상성 유무에 따라 비정상성/정상성 시계열 자료로 구분
  + 비정상성 시계열 자료: 정상성을 따르지 않는 시계열 자료로써 대부분의 시계열 자료가 이에 해당함; 시계열 분석을 위해서는 비정상성 시계열을 핸들링해 다루기 쉬운 정상성 시계열 자료로 변환해야 함
  + 정상성 시계열 자료: 정상성을 따르는 시계열 자료
* 연속/이산 시계열 자료
  + 시계열 자료는 생성되는 특성에 따라 연속/이산 시계열 자료로 구분
  + 연속 시계열 자료(continuous time series): 연속적으로 생성되는 시계열 자료
  + 이산 시계열 자료(discrete time series): 이산적 시점에서 생성되는 시계열 자료

**1.4. 시계열 자료 형태 (component of time series)**

* 시계열 그림은 시간의 경과에 따라 시계열 자료의 값이 변하는 것을 나타낸 그림으로 시간 t를 가로축, 시계열 관측값 Zt를 세로축에 나타냄
* 시계열 그림을 도출하는 이유는 시계열의 특징을 쉽게 파악할 수 있어 해당 자료에 적합한 분석방법 선택에 도움이 되기 때문임
* 추세와 계절변동 요인을 갖는 시계열

라인, 그래프, 텍스트, 폰트이(가) 표시된 사진

자동 생성된 설명

* 시간의 변화에 따라 변동폭이 커지는 시계열

텍스트, 라인, 폰트, 그래프이(가) 표시된 사진

자동 생성된 설명

* 두개의 추세선을 갖는 시계열

라인, 그래프, 도표, 경사이(가) 표시된 사진

자동 생성된 설명

**텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명**

**2. 시계열 (자료) 모형**

**2.1. 시계열 (자료) 모형 정의**

* 시계열 모형은 시간의 흐름에 따라 관측되는 데이터를 설명하고 예측하기 위해 사용되는 수학적 모델
* 시계열 모형은 주어진 데이터의 패턴, 추세, 계절성, 불규칙한 변동 등을 설명하고 예측하는데 사용됨

**2.2. (단변량) 자기회귀 모형(Autoregressive model, AR모형)**

* P 시점 전의 자료가 현재 자료에 영향을 주는 모형

* AR(1) 모형: , 직전 시점 데이터로만 분석
* AR(2) 모형: , 연속된 2시점 데이터로 분석
* 자기상관함수는 빠르게 감소, 부분자기함수(PACF)는 어느 시점에서 절단점을 가진다 (?)

**2.3. 이동평균 모형(Moving average model, MA모형)**

* 유한한 개수의 백색잡음의 결합이므로 언제나 정상성을 만족

* MA(1) 모형: , 시계열이 같은 시점의 백색잡음과 바로 전 시점의 백색잡음의 결합으로 이루어진 모형
* MA(2) 모형: *,* 바로 전 시점의 백색잡음과 시차가 2인 백색잡음의 결합으로 이루어진 모형

**2.4. 자기회귀이동평균 모형(Autoregressive moving average, ARMA(p,q) 모형)**

* 자기회귀와 이동평균 다항식으로 약한 정상성을 가진 확률적 시계열을 표현하는데 사용
* p는 AR 모형, q는 MA 모형과 관련이 있는 차수

**2.5. 자기회귀누적이동평균 모형(Autoregressive integrated moving average model, ARIMA(p,d,q) 모형)**

* ARIMA 모형은 비정상시계열 모형
* ARIMA 모형을 차분이나 변환을 통해 AR모형이나 MA 모형, 이 둘을 합친 ARMA 모형으로 정상화할 수 있음
  + 시계열 데이터에 대해 d차 차분한 데이터로 AR(p)+MA(q) 모형을 선형 합한 모형
  + p는 AR 모형, q는 MA 모형과 관련이 있는 차수
  + 단변량 예측을 위해서는 적합한 step size(p & q)를 설정해야 함
  + d=0 이면 ARMA(p,q) 모형이라 부르고, 이 모형은 정상성을 만족한다
  + p=0이면 IMA(d,q) 모형이라 부르고, d번 차분하면 MA(q)모형을 따른다
  + q=0이면 ARI(p,d) 모형이라 부르고, d번 차분한 시계열일 AR(p) 모형을 따른다

**2.6. 벡터자기회귀 모형(Vector Autoregressive모형)**

* 단변량 AR모형의 다변량 확장 모형으로 설명변수 간 시간에 따른 영향변화 고려
* Z1~ZN: N개의 변량조건에서 시점 t의 X1에 대해 t-1 시점부터 t-p시점까지의 X1~XN의 계수배의 벡터 선형연산

**3. 시계열 자료 분석 방법**

**3.1. 단변량 시계열 분석 방법**

단변량 시계열 분석은 시간(t)을 설명변수로 하여 회귀모형주가, 소매물가지수 등 하나의 변수에만 초점을 맞춰 분석하는 방법론

1. 시계열 시각화

데이터를 그래프나 차트로 시각화하여 시간에 따른 패턴과 트렌드를 살펴봄. 이는 데이터의 이상치나 경향을 파악하는데 도움

1. 기술통계 분석

데이터의 평균, 분산, 자기상관함수(ACF), 부분자기상관함수(PACF) 등을 계산하여 시계열 데이터의 특성을 파악

1. 이동평균법(moving average method)

* 이동평균: 어떤 것이 방향성을 가지고 움직일 때 이동하면서 구해지는 평균
* 과거로부터 현재까지의 시계열 자료를 대상으로 일정 기간별 이동평균을 계산하고 이들의 추세를 파악하여 다음 기간을 예측하는 방법
* 단순이동평균법(simple moving average, SMA)
  + 최근 몇 개 관측값의 단순평균값을 다음 기간의 예측값으로 추정하는 방법
  + 이는 시계열 자료에서 계절변동과 불규칙변동을 제거하여 추세변동과 순환변동만 가진 시계열 자료로 변환하는 방법으로 사용됨
  + 최근 n개의 관측값 Zt, Zt-1, …, Zt-n+1을 이용하여 계산한 이동평균

Mt는 t+1시점의 예측값으로 사용됨

* + n개의 시계열 데이터를 m기간으로 이동평균하면 n-m+1개의 이동평균 데이터가 생성됨
  + 이동평균법을 사용할 때 중요한 문제는 이동평균을 계산하기 위해 사용하는 과거자료의 적정개수 m을 결정하는 것임. 일반적으로 시계열 자료에 뚜렷한 추세가 나타나거나 불규칙변동이 심하지 않은 경우에는 작은 m의 개수를 사용하고 그렇지 않은 경우에는 m의 개수를 크게 함
  + 한편, 예측오차를 비교하여 m의 개수를 결정하기도 함; 예측오차를 비교하는 기준으로 평균절대편차(mean absolute deviation, MAD)와 평균제곱오차(mean squared error, MSE)가 사용됨

평균절대편차와 평균제곱오차를 비교하여 예측오차가 작은 경우의 m의 개수로 선택하는 것이 적절함

* + 이동평균법은 m개의 최근 자료로 다음 기간의 자료를 예측하는 계산방법상의 특성 때문에 처음의 일부 기간에 대한 예측값을 구할 수 없고 구간에 포함되지 않은 이전의 자료는 무시됨
  + 이동평균의 기간 m이 짝수인 경우 인접한 두 이동평균의 평균을 계산하여 이동평균을 중심화하는 중심이동평균을 구해야 함
  + 이동평균의 계산에 사용되는 관측값들에 대해 동일한 가중치를 적용하여 최근 자료와 오래된 자료의 중요성을 동일시하고 있다는 점을 고려해야 함. 즉 불규칙 변동요인이 나타나는 경우 정확하게 추출하기 어려움
* 누적이동평균법(cumulative moving average, CMA)
  + 단순이동평균법과 같이 산술평균을 구하는 것은 동일하나, 윈도우(m값)을 정해놓지 않음. 따라서 모든 데이터를 고려하고 싶을 때 사용함
  + 새로운 값이 들어올때마다 전체 평균을 새롭게 구하는 것과 동일함
* 가중이동평균법(weighted moving average, WMA)
  + 단순이동평균법과 동일하지만 관측값마다 다른 가중치를 적용한다는 차이가 있음

* 지수가중이동평균법(exponentially weighted moving average, EWMA)
  + 가중이동평균법과 동일하나 지수적으로 감소하는 가중치를 곱해준다는 의미만 다름

1. 지수평활법(exponential smoothing method)

* 과거의 모든 자료를 사용하여 평균을 구하면서 최근의 자료에 더 높은 가중치를 부여하는 방법임
* 단기간에 발생하는 파동이나 충격을 완화하는 평활법(smoothing method)

* 지수평활계수 α는 가중치 역할을 하는 것으로 불규칙변동이 큰 자료에서는 작은 값의 α를 적용하고, 불규칙변동이 작은 자료에서는 큰 값의 α를 적용함. 일반적으로 α는 0.05와 0.3 사이의 값을 사용함
* 이동평균법과 마찬가지로 예측오차를 비교하여 예측오차가 작은 α값을 선택하는 것이 바람직함
* 지수평활법에서 가중치는 과거로 갈수록 지수적으로 감소하게 됨. 그러므로 지수평활법에서는 최근 과거값에 가장 큰 가중치를 부여하므로 일종의 가중이동평균법이라 할 수 있음
* 지수평활법은 불규칙변동의 영향을 제거하는 효과가 있으며, 중기 예측 이상에 주로 사용됨. 단, 단순지수평활법의 경우, 장기추세나 계절변동이 포함된 시계열의 예측에는 적합하지 않음

1. 분해 시계열

* 시계열 데이터를 추세, 계절성, 주기, 잔차 등의 구성요소로 분해하여 각 구성요소를 분석
* 분해식의 일반적 정의

,

**3.2. 다변량 시계열 분석 방법**

다변량 시계열 분석은 둘 이상의 변수가 시간에 따라 상호작용하는 데이터를 분석하는 방법으로, 여러 변수 간의 관계를 이해하고 각 변수의 시간적 변화를 모델링하고 예측하는데 사용함

**4. 참고자료**

[1] KOCW시계열 자료, http://elearning.kocw.net/contents4/document/lec/2013/Konkuk/Leegiseong/5.pdf

[2]ADsP 데이터분석 준전문가 교재, 빅데이터교육전문기관

[3]