개별 연구

선행연구 조사

학　　과	컴퓨터공학과
학　　번	2017112292
이　　름	김준하

1. 시계열 데이터의 특징

- 시계열 데이터는 시간별로 구성된 값의 집합이다.
- 시계열은 반드시 고정된 시간 구간의 관측치이어야 한다. 불규칙한 시간 구간이면 안된다.
- 시계열 데이터는 크게 두 가지로 나눌 수 있다.
  - → 정상 시계열: 평균과 표준편차가 일정하다는 조건이 선행되어야 분석이 가능하다. 대표적인 예시로는 ARIMA모델 이 있다.
  - → 비정상 시계열: 차분이나 log함수를 씌워 정상시계열로 변환 후 분석을 해야 한다.
- 시계열 구간을 작은 범위에서 큰 구간으로는 변환할 수 있지만 반대로는 불가능하다.
  - → 'Monthly' -> 'Quarterly -> 'Yearly' (변환 가능)
  - → 'Yearly' -> 'Quarterly' -> 'Monthly' (변환 불가능)
- 이벤트가 발생하고 처리를 위해 도착하는 순서대로 구성된다. (임시 순서 지정)
- 일반적으로 시간순으로 도착하며, 데이터 저장소에 삽입되고 업데이트되는 경우는 거의 없다.
- 시계열 데이터에는 타임스탬프가 있으며, 시간은 데이터를 보거나 분석하기 위한 의미 있는 축이다. 시계열 데이터는 분산형 또는 꺾은선형 차트를 사용하여 가장 잘 시각화된다.

2. 시계열 변동 모형

- 패턴에 따른 시계열 변동 모형

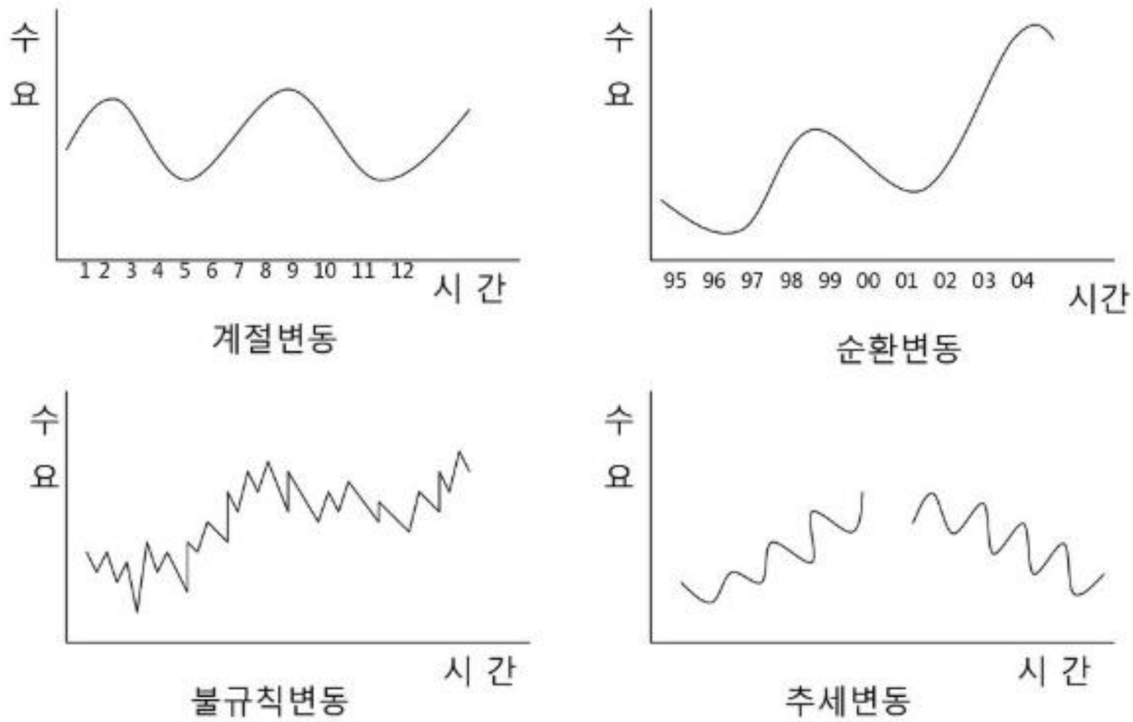
→ 추세(Trend) 변동: 상승과 하락이 있는 변동

→ 계절(Season) 변동: 1년안에 월, 분기로 반복되는 패턴

→ 순환(Circulation) 변동: 경기변동이라고도 하며, 5년, 10년처럼 장기간 동안 간격을 두고 상승, 하락이 주기적으로 반복되는 패턴을 말한다. 이 때는 데이터가 크며 추세변동과 결합해 주로 분석을 진행한다.

→ 불규칙(irregular) 변동: 위의 세 가지 변동으로는 설명할 수 없는 패턴. 모형은 두 가지로 나누어진다.

1. 승법(곱셈) 모형: 추세*계절*순환*불규칙 변동
2. 가법(덧셈) 모형: 추세+계절+순환+불규칙 변동



3. 시계열 데이터의 예측 방법

- 양적 예측 방법: 과거에 대한 정보(양적 자료)를 이용해 예측에 필요한 경험적 법칙을 추정해 예측하는 방법이다. 과거의 패턴이 미래에도 지속될

것으로 예측하는 것이다.

→ 평활법과 분해법 : 주어진 데이터를 잘 설명하는 것에 초점을 맞춘 방법이다.

→ 확률적 시계열 분석 : ARIMA(시간영역), Fourier 분석(주파수 영역)의 2 가지가 존재한다.

- 질적 예측 방법 : 미래 예측을 위해 전문가들의 주관적 견해를 사용한다. 또는 과거의 정보가 없거나 불충분한 경우에 사용을 한다. 대표적인 방법으로는 델파이 기법과, 시나리오 기법이 있다.

4. 시계열 데이터의 예측평가와 기준

1. 예측한 값이 적절한지를 판단, 평가하는 방법
    1. 사전평가와 모형추정을 같이 진행
    2. 모형추정을 먼저 하고 사전평가를 시행
2. 예측평가의 기준 – 기본적으로 error(= 실제값 – 예측값)을 기준
    1. 평균 제곱 오차(MSE)
    2. 평균 제곱근 오차(RMSE)
    3. 평균 절대 오차(MAE)
    4. 평균 절대 백분비 오차(MAPE)
    5. 타일의 불일치계수

5. 시계열 모델 종류

1. AR/MA (Autoregressive Moving Average) 조합
    - Autoregressive Model
    - Moving Average Model
    - ARAM & ARIMA
    - SARIMA
2. Exponential Smoothing (ES)
    - Simple Exponential Smoothing (SES)
    - Holt-Winters' Model
3. Vector AR/MA 조합
    - VAR, VMA, VARMA
    - VARMAX
4. 이외
    - XGBoost
    - Prophet
    - CNN, RNN, LSTM, GRU (딥러닝) + RNN 기반 Autoencoder Multistep
    - DeepAR
    - N-BEATS
    - Temporal Fusion Transformer

위 모델들 중에서 대표적인 ARIMA 모델에 대해 알아보았습니다.

ARIMA (Autoregressive Integrated Moving Average)

- 차분을 적용한 데이터에 AR 모형과 MA 모형을 합친 모형이다.

  - AR (Autoregression) 모형

    자기회귀모형. 자기상관성을 포함하고, 예측하려는 특정 변수의 과거 관측값의 선형 결합을 통해 해당 변수의 미래값을 예측한다. 이전 관측값이 이후 관측값에 영향을 준다는 개념이 담겨 있다.

    AR(p) 모형의 식은 다음과 같다.

    $$y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \varepsilon_t$$

    (y_t는 t 시점의 관측값, c는 상수, Φ는 가중치, ε_t는 오차항을 의미한다.)

  - MA 모형

    t시점의 오차와 과거의 예측 오차들을 이용하여 미래값을 예측하는 모형이다.

    MA(q) 모형의 식은 다음과 같다.

    $$y_t = c + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q} + \varepsilon_t$$

  - 차분 (differencing)

    현 시점의 데이터에서 d시점 이전의 데이터를 뺀 것.

    연이은 관측값들의 차이를 계산하는 것으로, 시계열의 수준에서 나타나는 변화를 제거하여 시계열의 평균 변화를 일정하게 만들어준다. non stationary 인 데이터들을 stationary 데이터들로 바꾸어 주는 방법이다.

    1차 차분: $Y_t = X_t - X_{t-1}$

    n차 차분은 차분을 n번 거듭한 것.

  - ARIMA 모형

    ARIMA(p,d,q) 모형은 d차 차분한 데이터에 위의 두 모형을 합친 모형으로, 식은 다음과 같다.

    $$y'_t = c + \Phi_1 y'_{t-1} + \Phi_2 y'_{t-2} + \ldots + \Phi_p y'_{t-p} + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \ldots + \theta_q\varepsilon_{t-q} + \varepsilon_t$$

    y': d차 차분을 구한 시계열

p: 자기회귀 부분의 차수, AR 모델의 parameter 개수
d: 차분의 차수
q: 이동평균 부분의 차수, MA 모델의 parameter 개수

- 정상성(stationarity)을 가정하는 모델로, 정상성 확인이 필요하다.
  - ACF(Autocorrelation Function) plot를 사용한다. ACF plot은 x축이 lag, y축이 ACF 값으로 이루어진다.
  - Lag: 현재의 데이터와 lag값 만큼 미룬(이전의) 데이터
    ex) ACF plot에서 Lag 1 이라는 것은 현재 데이터와 한 시점 미룬 (바로 이전의) 데이터와의 correlation 값을 의미한다.
  - ACF plot이 특정 패턴 없이 random하게 나타나거나 빠르게 감소하면 stationary하다고 할 수 있다. non stationary 한 데이터를 사용하여 ACF plot을 그려 보면 전반적으로 서서히 감소한다.
  - 원래 데이터가 stationary: 차분 필요 없음.
    지속적으로 증가 또는 감소: 1차 차분으로 충분
    더 복잡한 trend: 2차 차분 이상 (대부분 2차 차분이면 충분, 3차 차분 이상으로 가야 하면 ARIMA 모델에 적합하지 않다고 볼 수 있음.)
- ARIMA 모델링 과정
  1. 데이터 전처리
     i. stationarity 확인
     ii. 데이터가 stationary하지 않다면 전처리(transformation, differencing) 과정을 통해 stationary하게 바꾸어준다.
  2. 시범적으로 시행해 볼 만한 모델 찾기
     여러 방법 중 한 가지로 'Graphical method'가 있는데, 방법은 다음과 같다.
     i. 데이터를 사용하여 ACF, PACF plot을 생성하고 그 패턴

으로부터 어떠한 모델을 사용할 지 선택.

ii. ACF, PACF plot이 다음과 같은 형태일 때, 각각 MA, AR, ARMA 모델이 적합하다고 알려져 있음.
패턴을 파악하는 과정이 주관적일 수 있음.

모델	ACF	Partial ACF
MA(q)	q시차 이후 0으로 급감	지수적으로 감소, 소멸하는 sine함수 형태
AR(p)	지수적으로 감소, 소멸하는 sine함수 형태	p시차 이후 0으로 급감
ARMA(p,q)	시차 (q-p)이후 급감	시차 (q-p)이후 급감

(시차는 lag를 의미)

위와 같이 graphical 한 방법을 사용할 수도 있지만 이는 ACF, PACF plot을 보고 주관적으로 판단하는 것이기 때문에 어떤 데이터들은 명확히 모델을 선정하기 어려울 수 있다. 그래서 보통은 graphical 방법보다도 다음과 같은 방법을 사용한다.

A. ARIMA(p,d,q) 에서 d는 거의 3이상 넘어가지 않으므로 1 또는 2로 설정한다.
B. p와 q의 범위를 설정해서 범위 내의 모든 경우의 조합을 통해 여러 개의 ARIMA(p,d,q)를 만든다.
C. 각 모델들에 대해 AIC값 또는 testing data의 예측 정확도를 통해 가장 좋은 모델을 선정한다.

3. parameter 추정
4. 모델이 괜찮은 지 확인, 적합하지 않으면 2번부터 반복

i. 모델을 사용해서 이미 알고 있는 데이터를 예측해 보고, residual을 구한 후 residual에 대한 ACF plot을 생성한다.
residual이란 모델을 사용하여 예측한 값(y-hat)과 실제 값의 차를 의미한다.

ii. 대부분의 residual 값이 bound 안에 들어와 있고, 40개 중 두세개의 데이터만 bound를 벗어나면 괜찮은 모델로 평가한다. Bound는 residual의 분산에 3을 곱하여 더하고 뺀 값이다. (Three-sigma limits)

5. 예측 모델로 사용

6. ARIMA 모델의 통계적 접근

(수식의 유도 과정은 생략하였습니다.)

기본적인 통계 지식

- $X_1, X_2, \ldots, X_t$: a sequence of a random variable
- X: random variable, 확률 변수
- $F_X(x) = P(X \le x)$: cdf(cumulative distribution Function), 누적 분포 함수
- $E(X) = \mu_X$: 기댓값
- $V(X) = E[(X - \mu_X)^2] = \sigma_X^2$: 분산
- $Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)] = \sigma_{X1X2}$: covariance, 공분산
  - $Cov(X1, X1) = V(X1) = \sigma_{X1}^2$
- $Corr(X_1, X_2)$: correlation, 상관 계수

$$\begin{aligned}
\rho_{X,Y} &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} && (1)\\
&= \frac{\mathbf{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} && (2)\\
&= \frac{\mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)}{\sigma_X \sigma_Y} && (3)
\end{aligned}$$

Function 정의

Definition 1.8 *The* **autocovariance function of a stationary time series** *will be written as*

$$\gamma(h) = \mathrm{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu)(x_t - \mu)]. \tag{1.22}$$

Definition 1.9 *The* **autocorrelation function (ACF) of a stationary time series** *will be written using (1.14) as*

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}. \tag{1.23}$$

- Autocovariance function: 특정 시점 이후의 자기 자신과의 covariance
  - 다음과 같은 특징을 가짐.

$$Cov(X_t, X_{t+h}) = \gamma_X(h)$$

$$① \ \gamma_X(0) = Cov(X_t, X_t) = V(X_t) = \sigma_{X_t}^2$$

$$\begin{aligned} ② \ \gamma_X(-h) &= Cov(X_t, X_{t-h}) \\ &= Cov(X_{t-h}, X_t) \\ &= Cov(X_{t-h}, X_{(t-h)+h}) \\ &= \gamma_X(h) \end{aligned}$$

$\Rightarrow \gamma_X(h) = \gamma_X(-h)$ for all h. (대칭성)

- Autocorrelation function:
  - 다음과 같은 특징을 가짐.

$$\rho_X(h) = \frac{Cov(X_t, X_{t+h})}{\sqrt{V(X_t)\cdot V(X_{t+h})}} = \frac{\gamma_X(h)}{\sqrt{\gamma_X(0)\cdot\gamma_X(0)}} = \frac{\gamma_X(h)}{\gamma_X(0)}$$

$$① \ \rho_X(0) = \frac{\gamma_X(0)}{\gamma_X(0)} = 1 \rightarrow Corr(X_t, X_t)$$

$② \ \rho_X(-h) = \rho_X(h)$ for all h

$$③ \ -1 \le \rho_X(h) \le 1$$

- Backward Shift Operator (function)
  - B라고 표기

$$B \cdot X_t = X_{t-1}$$
$$B^2 \cdot X_t = X_{t-2}$$
$$\vdots$$
$$B^m X_t = X_{t-m}$$

$$AR(1) \rightarrow X_t = \frac{a_t}{1-\phi B} = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \phi^3 a_{t-3} + \cdots$$

$$AR(2) \rightarrow X_t = \frac{1}{1-\phi_1 B - \phi_2 B^2} \cdot a_t = \frac{1}{(1-\alpha_1 B)(1-\alpha_2 B)} \cdot a_t$$

$$= a_t + (\alpha_1+\alpha_2)a_{t-1} + (\alpha_1^2+\alpha_2^2+\alpha_1\alpha_2)a_{t-2} + \cdots$$

White Noise

- White Noise (백색잡음, 백색노이즈)

  a_t WN (a_t)

  ① $E(a_t) = 0, \ \forall t$

  ② $V(a_t) = \sigma_a^2, \ \forall t$

  → 시점에 관계없이 일정

  ③ $Corr(a_t, a_s) = 0, \ t \neq s$

  ④ $\gamma_a(h) = Cov(a_t, a_{t+h}) = \begin{cases} \sigma_a^2, & h=0 \\ 0, & h \neq 0 \end{cases}$

  ⑤ $\rho_a(h) = \begin{cases} 1, & h=0 \\ 0, & h \neq 0 \end{cases}$

ARIMA 모델을 위한 stationary의 정의

- ARIMA

  $X_1, X_2, \ldots, X_t$

  $E(X_t) = \mu$, $V(X_t) = \sigma_x^2$, for $t = 1, 2, \cdots$ → 시점에 관계없이 일정

  ⇒ constant probability distribution over time

  ⇒ stationary time series (stationary process)

  ↳ 시점에 관계없이 평균과 분산이 일정하면 stationary.

MA, AR, ARMA 각 모델의 개념

- 특정 파라미터에 대한 모델

  ➢ MA(1) 모델

  Moving Average (1), MA(1) → White Noise로 모델링

  $X_t = a_t - \theta a_{t-1}$, $a_t \sim N(0, \sigma_a^2)$

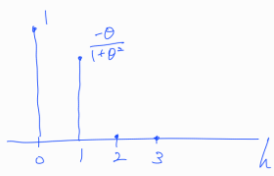
  $E(X_t) = 0$

  $$\gamma_x(h) \Rightarrow \begin{cases} h=0: \gamma_x(0) = (1+\theta^2)\sigma_a^2 = V(X_t). \\ h=1: \gamma_x(1) = -\theta\sigma_a^2. \\ h \geq 2: \gamma_x(h) = 0. \end{cases}$$

  → MA(1) 모델의 Auto covariance function

  $$\rho_x(h) \Rightarrow \begin{cases} h=0: \rho_x(0) = 1 \\ h=1: \rho_x(1) = \dfrac{-\theta}{1+\theta^2} \\ h \geq 2: \rho_x(h) = 0. \end{cases}$$

  → MA(1) 모델의 Autocorrelation function

  이론적 MA(1) 모델의 ACF

  ➢ MA(2) 모델

MA(2)

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

$$E(X_t) = 0$$

$$\gamma_X(h) : \; = Cov(X_t, X_{t-h}) \begin{cases} h=0 : \gamma_X(0) = (1+\theta_1^2+\theta_2^2)\sigma_a^2 \\ h=1 : \gamma_X(1) = (-\theta_1+\theta_1\theta_2)\sigma_a^2 \\ h=2 : \gamma_X(2) = -\theta_2\sigma_a^2 \\ h \geq 3 : \gamma_X(h) = 0. \end{cases}$$

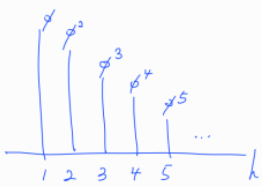
$$\rho_X(h) : \begin{cases} h=0 : 1 \\ h=1 : \dfrac{\gamma_X(1)}{\gamma_X(0)} = \dfrac{-\theta_1+\theta_1\theta_2}{1+\theta_1^2+\theta_2^2} \\ h=2 : \dfrac{\gamma_X(2)}{\gamma_X(0)} = \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2} \end{cases}$$

$$h \geq 3 : 0$$

- AR(1) 모델

AR(1). → 자기자신의 과거 값을 사용하여 모델링

- $X_t = \phi X_{t-1} + a_t$
- $E(X_t) = 0$
- $V(X_t) = \sigma_X^2 = \dfrac{\sigma_a^2}{1-\phi^2} = \gamma_X(0)$
- $\gamma_X(h) = \phi^h \cdot \gamma_X(0)$, $(h \geq 1)$
- $\rho_X(h) = \dfrac{\gamma_X(h)}{\gamma_X(0)} = \phi^h$, $(h \geq 1)$ ⟶



- AR(2) 모델

AR(2)

- $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t$,
- $E(X_t) = 0$.
- $Cov(X_t, X_{t-h}) = E(X_t \cdot X_{t-h})$
- $\gamma_X(h) = \phi_1 \gamma_X(h-1) + \phi_2 \gamma_X(h-2)$, $(h \geq 1)$
- $\rho_X(h) = \dfrac{\gamma_X(h)}{\gamma_X(0)} = \phi_1 \rho_X(h-1) + \phi_2 \rho_X(h-2)$

↳ Yule - Walker equation

- ARMA(1,1) 모델

ARMA(1,1)

$\cdot\ X_t = \phi X_{t-1} + a_t + \theta a_{t-1}$

$$\cdot\ \gamma_X(h) : \begin{cases} \gamma_X(0) = \dfrac{1+2\theta\phi+\theta^2}{1-\phi^2}\cdot\sigma_a^2 \\ \gamma_X(1) = \dfrac{(1+\theta\phi)(\phi+\theta)}{1-\phi^2}\cdot\sigma_a^2 \\ h\geq 2 :\ \gamma_X(h) = \phi^{h-1}\cdot\dfrac{(1+\theta\phi)(\phi+\theta)}{1-\phi^2}\cdot\sigma_a^2 \end{cases}$$

↳ Auto covariance function of ARMA(1,1)

- AR, MA, ARMA process

  - AR process, AR(p)

AR process, AR(p)

$\cdot\ X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + a_t$

$\quad = \phi^{-1}(B)\cdot a_t \quad , \ (\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)$

↳ White Noise a_t를 함수 $\phi^{-1}(B)$에 통과시켜 나온 결과가 AR process 이다.

  - MA process, MA(q)

MA(q)

$\cdot\ X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$

$\quad = \theta(B)\cdot a_t \quad , \ (\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)$

↳ White Noise a_t를 $\theta(B)$ 함수에 통과시키면 MA 모델이 됨.

  - ARMA process, ARMA(p,q)

ARMA(p,q)

$$X_t = \underbrace{\phi_1 X_{t-1} + \cdots + \phi_p X_{t-p}}_{AR} + \underbrace{a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}}_{MA}$$

$$\phi(B) X_t = \theta(B) \cdot a_t$$

$$X_t = \phi^{-1}(B) \cdot \theta(B) \cdot a_t$$

White Noise a_t가 이 함수를 통과하면
ARMA(p,q) process로 바뀜

미래값의 예측, Prediction (forecasting)

- Prediction의 개념

Prediction (forecasting)

- $X_1, X_2, \ldots, X_t$ 가 주어졌을 때, X_{t+1}을 예측하는 방법
- error : $X_{t+1} - \hat{X}_{t+1}$

  실제값 예측값 → 이것을 잘 예측하는가 핵심!
- $E[(X_{t+1} - \hat{X}_{t+1})^2]$이 최소가 되도록 하는 $\hat{X}_{t+1}$을 구한다.

  ↳ $\boxed{\hat{X}_{t+1} = E[X_{t+1} \mid X_1, X_2, \cdots, X_t]}$ → conditional expectation

- 각 모델의 예측

  - AR(1)

※ AR(1) : $X_t = \phi X_{t-1} + a_t + \mu$

↳ 일반적인 모델의 상수항

$$X_{t+1} = \phi X_t + a_{t+1} + \mu$$

$$\Rightarrow \hat{X}_{t+1} = E[X_{t+1} \mid X_1, X_2, \cdots, X_t]$$

$$= \phi X_t + \mu$$

$$\therefore \boxed{\hat{X}_{t+1} = \phi X_t + \mu}$$

- Prediction Interval

$X_{t+1} \rightarrow \hat{X}_{t+1}$

X_{t+1}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\Rightarrow \hat{X}_{t+1} \pm \underbrace{C.V}_{\text{constant value}}(\alpha)\sqrt{Var(\hat{X}_{t+1})}$

$a_t \sim N(0, \sigma_a^2)$을 가정한다면,

$\Rightarrow \boxed{\hat{X}_{t+1} \pm Z(1-\frac{\alpha}{2})\cdot \sigma_a}$

- $\hat{X}_{t+2} = \phi \hat{X}_{t+1} + \mu$ → 점 예측값

- $a_t \sim N(0, \sigma_a^2)$을 가정할 때, X_{t+2}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\boxed{\hat{X}_{t+2} \pm Z(1-\frac{\alpha}{2})\cdot \sqrt{(1+\phi^2)\sigma_a^2}}$

➢ AR(2)

* AR(2)

- $X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + a_t$

- $\boxed{\hat{X}_{t+1} = \mu + \phi_1 X_t + \phi_2 X_{t-1}}$ → 점 예측값

- $a_t \sim N(0, \sigma_a^2)$을 가정할 때, X_{t+1}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\boxed{\hat{X}_{t+1} \pm Z(1-\frac{\alpha}{2})\cdot \sigma_a}$

- $X_{t+2}, X_{t+3}, \cdots$에 대해서도 다음과 같이 구할 수 있다.

$\hat{X}_{t+h} = E[X_{t+h} \mid X_1, X_2, \cdots, X_t]$

X_{t+h}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\hat{X}_{t+h} \pm Z(1-\frac{\alpha}{2})\cdot \sqrt{\underbrace{V(\hat{X}_{t+h})}_{\text{분산}}}$

➢ ARMA(1,1)

ARMA (1,1)

- $X_t = \phi X_{t-1} + a_t + \theta a_{t-1} + \mu$
- $\hat{X}_{t+1} = \phi X_t + \theta a_t + \mu$ → 점 예측값

추정해야 함.

$a_t = X_t - \phi X_{t-1} - \theta a_{t-1} - \mu$

$\tilde{a}_t = X_t - \phi X_{t-1} - \theta \tilde{a}_{t-1} - \mu$

a_t의 tilde

$\Rightarrow \boxed{\hat{X}_{t+1} = \mu + \psi_1 a_t + \psi_2 a_{t-1} + \psi_3 a_{t-2} + \cdots}$, $\psi_1 = \phi + \theta$, $\psi_2 = \phi^2 + \phi\theta$, …

- X_{t+1}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\boxed{\hat{X}_{t+1} \pm Z(1-\frac{\alpha}{2})\cdot \sigma_a}$

X_{t+2}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\boxed{\hat{X}_{t+2} \pm Z(1-\frac{\alpha}{2})\cdot \sqrt{(1+\psi_1^2)\cdot \sigma_a^2}}$

→ $\psi_1 = \phi + \theta$. $\psi_2 = \phi^2 + \phi\cdot\theta$, …

X_{t+3}에 대한 $(1-\alpha)\cdot 100\%$ prediction interval

$\boxed{\hat{X}_{t+3} \pm Z(1-\frac{\alpha}{2})\cdot \sqrt{(1+\psi_1^2+\psi_2^2)\cdot \sigma_a^2}}$

⋮

+ SARIMA 모델의 개념

Seasonal ARIMA 모델 (SARIMA): 기존 ARIMA 모델에 계절 변동을 반영

- 각 계절에 따른 독립적인 ARIMA 모델이 합쳐져 있는 모형이다.
- 기존 ARIMA(p,d,q) 모형에 계절성 주기를 나타내는 차수 s가 추가적으로 필요하기 때문에 ARIMA(p,d,q)(P,D,Q)s 로 표기한다.
- s의 값은 월별 계절성을 나타낼 때는 s=12 가 되고, 분기별 계절성을 나타낼 때는 s=4가 된다.

$$\text{ARIMA }(p,d,q)(P,D,Q)_s$$

$$\underbrace{\phi_p(B)}_{\text{비계절 AR(p)}}\underbrace{\Phi_p(B^s)}_{\text{계절 AR(p)}}\underbrace{(1-B)^d}_{\text{비계절 차분 d}}\underbrace{(1-B^s)^D}_{\text{계절차분 D}}y_t = \underbrace{\theta_q(B)}_{\text{비계절 MA(q)}}\underbrace{\Theta_Q(B^s)}_{\text{계절 MA(q)}}a_t$$

$$\mathrm{ARIMA}\ (1,1,1)(1,1,1)_4$$

$$\underbrace{(1-\emptyset_1 B)}_{\text{비계절 AR(1)}}\underbrace{(1-\Phi_1 B^4)}_{\text{계절 AR(1)}}\underbrace{(1-B)^1}_{\text{비계절 차분 1}}\underbrace{(1-B^4)^1}_{\text{계절차분 4}}y_t=\underbrace{(1+\theta_1 B)}_{\text{비계절 MA(1)}}\underbrace{(1+\Theta B^4)}_{\text{계절 MA(1)}}a_t$$

추가. 시계열 모델 제작 과정

1. Data ETL (Extract, Transform, Load)

→ 데이터 시간 index 확인 및 format 설정

✓ 데이터의 index가 시간으로 설정되어 있는지 확인

✓ python 각 패키지마다 요구하는 index의 형태 확인

→ Data Quality 확인 (Missing Values, Duplicate Values, …)

→ 도표를 통해 기초 통계 정보 분석(Trend, Seasonality, …)

✓ Trend

I. 데이터에서 제외해야 한다.

II. 시간이 지남에 따라 뚜렷한 상하 추세가 있는지 확인해야 한다. random한 움직임 안에도 상승 또는 하강의 추세가 있을 수 있다. Trend의 유무에 따라 사용할 수 있는 모델이 달라진다.

III. 대부분의 모델은 De-trending이나 differencing을 통해 trend를 제외하고 모델링을 시작한다.

✓ Causality

I. Multivariate(다변량) 시계열 모델을 활용할 때에만 적용된다.

II. 한 변수의 시계열 데이터를 활용해서 다른 변수의 시계열 데이터를 예측하고자 할 때에는 그 두 변수 간의 상관관계가 존재하는지 확인하는 것이 중요하다.

III. 여러 개의 독립변수가 존재하는 경우 독립변수 간의 상관관계 역시 중요하다.

2. 시계열 모델 제작(Fit Model)

→ 모델 가정 확인

✓ Stationarity Test (Augmented Dickey-Fuller test, ACF/PACF 표 확인)

I. Stationarity
어느 데이터의 mean, variance, autocorrelation이 시간에 걸쳐 일정하다는 가정이다. 현실적으로는 지키기 힘들므로, 이를 극복하기 위해 실제로 stationary하지 못한 데이터가 가정을 충족시킬 수 있도록 변형하는 작업이 필요하다. 이러한 작업에는 de-trending, differencing, transforming이 있다.

II. de-trending
시계열 데이터에서 trend 부분만을 파악해서 제외하는 방법이다.

III. Differencing
시간 t의 값과 시간 t-1 값의 차이를 이용해 모델을 만드는 방법이다. 데이터를 $Y_t – Y_{t-1}$ 형식으로 재정의해 trend를 제외한다.

IV. Transforming
예측하고자 하는 변수를 log 등의 형태로 변형하는 방법이다.

✓ Seasonality 확인

I. Seasonality
계절 변화에 따른 추세의 존재 여부이다. 여름에 아이스크림의 수요가 늘어나는 것 또는 연말정산 시기에 특정 수요가 늘어나는 등이 이러한 예가 될 수 있다.
이를 확인하는 가장 쉬운 방법은 데이터를 선 그래프화 해서 시각적으로 확인하는 것이지만, 계절 효과와 기타 trend가 겹쳐

분명히 구분하기 어려울 때가 있다. 이러한 경우 trend를 제거한 데이터를 이용한 periodogram plot을 사용해야 한다.

→ 적합한 모델 종류 선택

- ✓ 계절의 영향을 심하게 받는 데이터: SARIMA, Exponential Smoothing
- ✓ multivariate(다변량) 데이터: Vector ARIMA Model
- ✓ 여러 모델들을 사용해 보고 가장 좋은 것을 선택하는 것이 가장 좋은 방법이다.

→ 여러 parameter 비교 및 선택 (AIC, BIC)

3. 수요 예측(Forecast Demand)