시계열자료분석팀

5팀 김태훈 이소율 강희균 정희주 마채영

INDEX

1

TIME SERIES

시계열이란?

확률변수 $X_1, X_2, \cdots, X_t$ 의 집합 $\{x_t, t \in T_0\}$

TIME SERIES

시계열이란?

확률과정

시계열

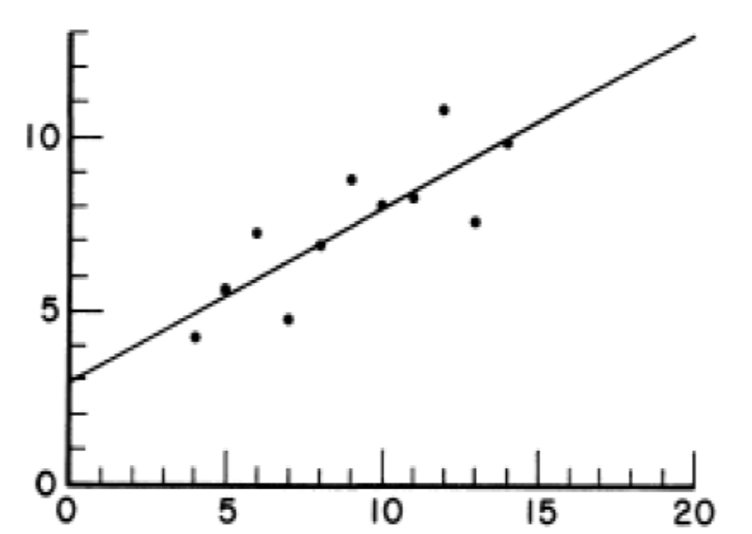
t = time

확률변수 $X_1, X_2, \cdots, X_t$ 의 집합 $\{x_t, t \in T_0\}$

TIME SERIES

시계열이란?

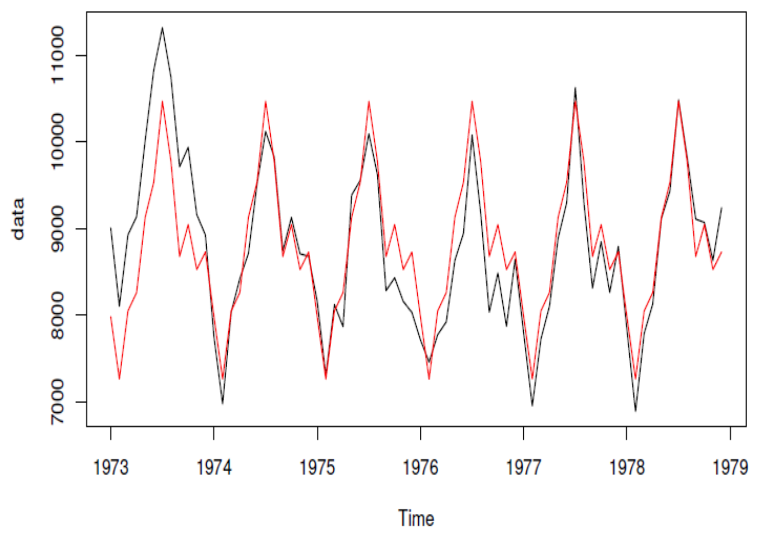
회귀



$$cov(\varepsilon_i, \varepsilon_j) = 0$$

X_i and X_j *are* *independent*

시계열



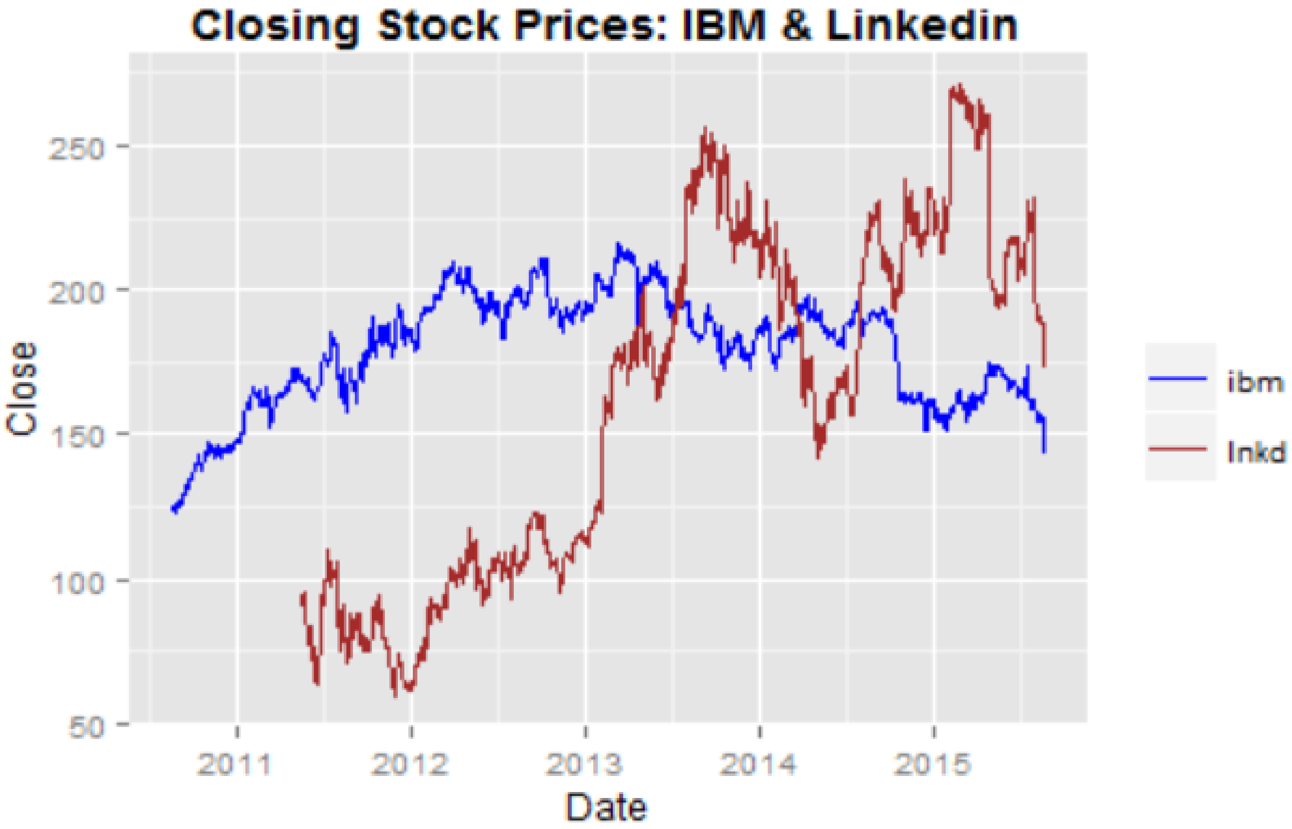
X_i and X_j *are* *dependent*

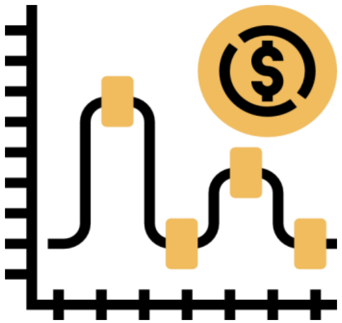
1 TIME SERIES

시계열 자료에는 어떤 것이 있을까요?

주식

이율



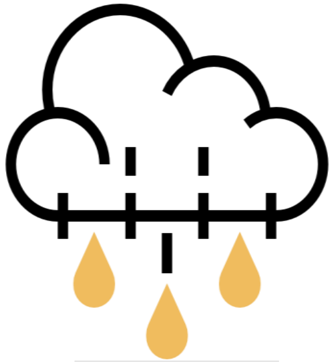


주가변동성

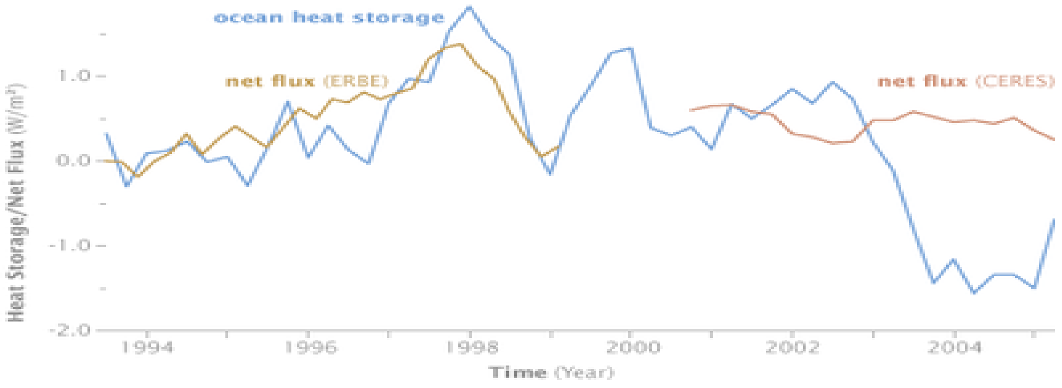
환율

TIME SERIES

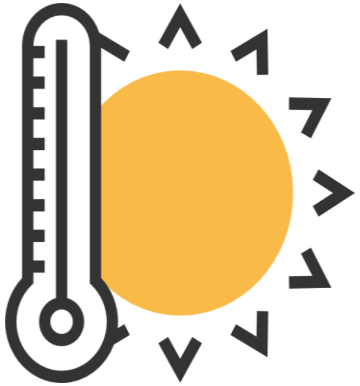
시계열 자료에는 어떤 것이 있을까요?



강우

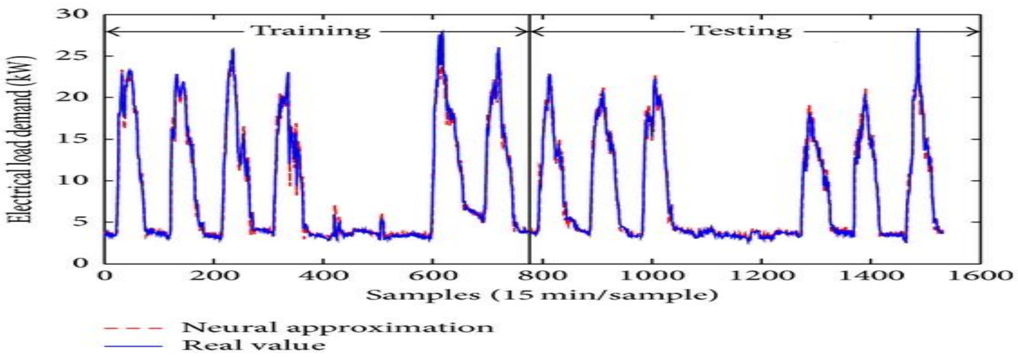


OCEAN HEAT FLUX

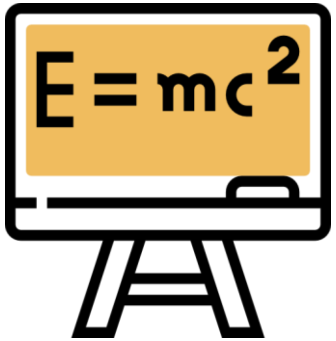


지구온난화

전자기장



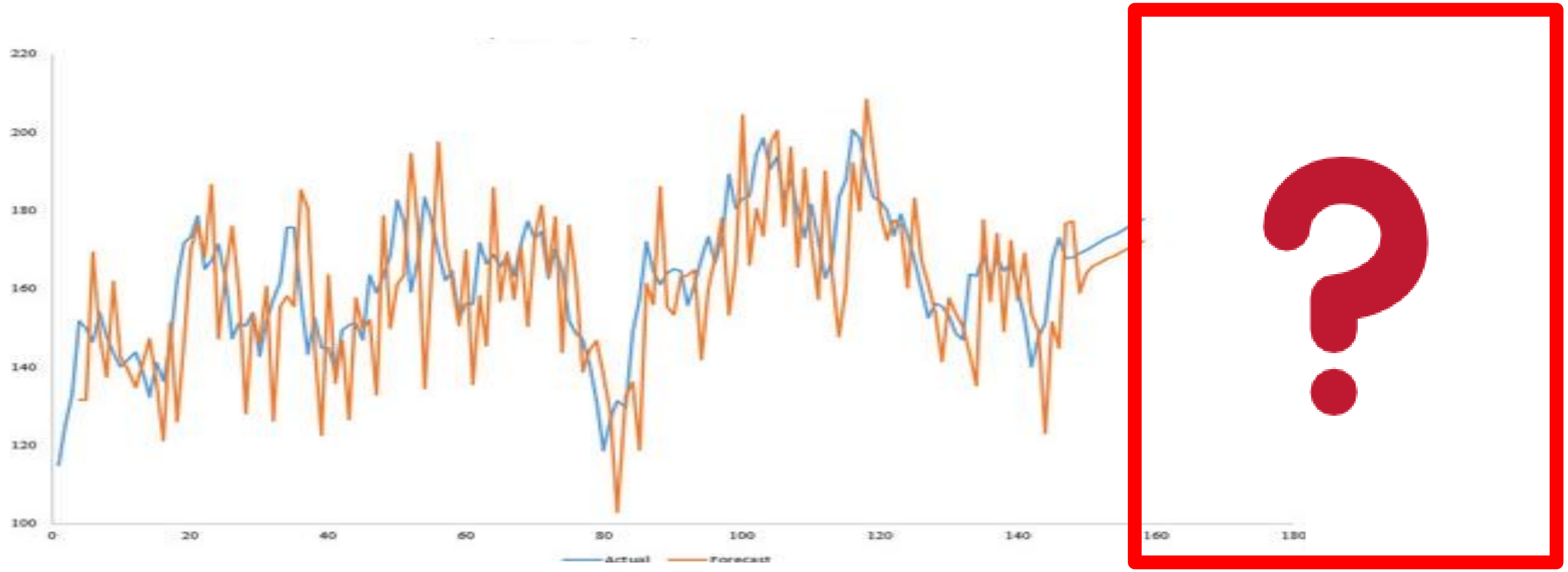
ELECTRICAL LOAD



노이즈

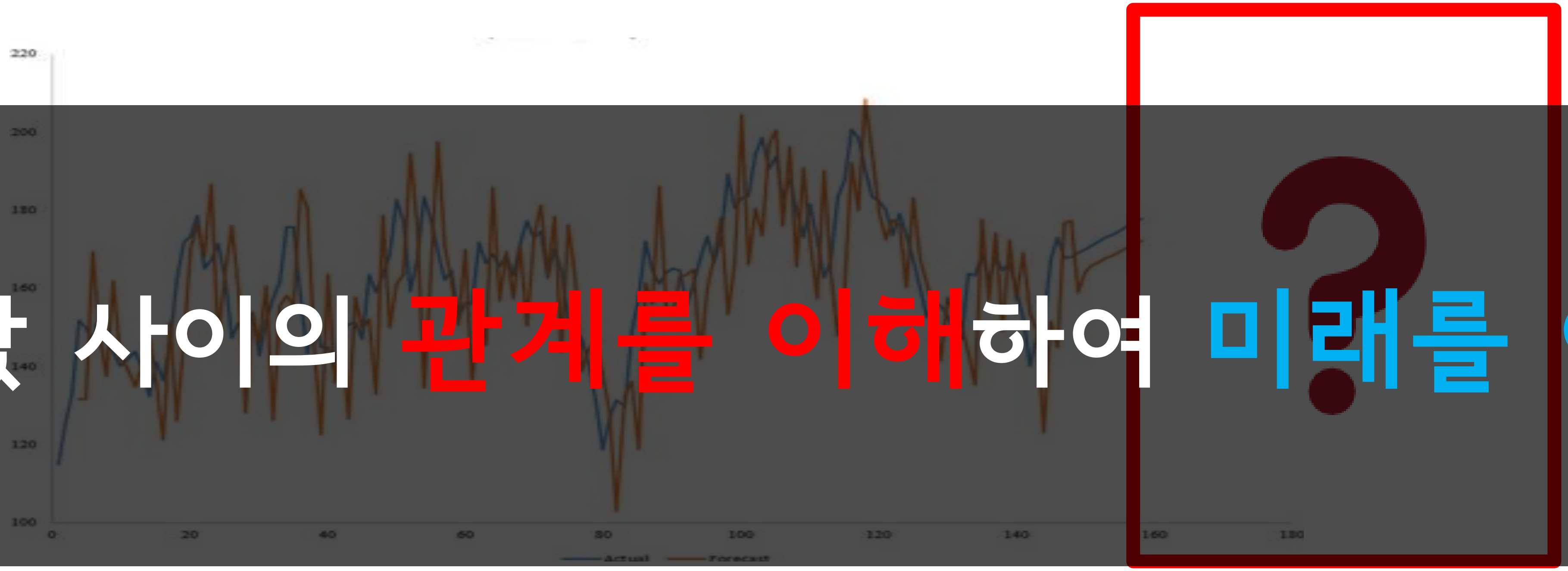
TIME SERIES

시계열 자료 분석



자연적인 역학관계를 이해하고 미래를 예측하기 위해 시계열 자료를 분석

시계열 자료 분석



관측값 사이의 관계를 이해하여 미래를 예측

자연적인 역학관계를 이해하고 미래를 예측하기 위해 시계열 자료를 분석

TIME SERIES

시계열 자료 분석

X의 집합
전체에 대한 이해

미래의 값이 포함된
모든 X에 대한
결합분포함수

무한한 차원

시계열 자료 분석

X의 집합
전체에 대한 이해

미래의 값이 포함된
모든 X에 대한
결합분포함수

무한한 차원

다른가정이 필요!

STATIONARITY

2

STATIONARITY

STATIONARITY

STRICT STATIONARITY(강정상성)

$\{X_t, t \in \mathbb{Z}\}$ *is strictly stationary if for all* n *and* h,

$$(X_{t_1}, \dots, X_{t_n}) \stackrel{d}{=} (X_{t_1+h}, \dots, X_{t_n+h})$$

시차 h

- If $n = 1$, it means that $X_1 \stackrel{d}{=} X_2 \stackrel{d}{=} X_3 \dots$.
- If $n = 2$, then

$$(X_1, X_2) \stackrel{d}{=} (X_2, X_3) \stackrel{d}{=} (X_5, X_6) \stackrel{d}{=} \dots$$

$$(X_1, X_3) \stackrel{d}{=} (X_2, X_4) \stackrel{d}{=} (X_3, X_5) \stackrel{d}{=} \dots$$

분포의 특징이 Lag(시차); h에 의존

STATIONARITY

STRICT STATIONARITY(강정상성)

$\{X_t, t \in \mathbb{Z}\}$ *is strictly stationary if for all* n *and* h,

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h})$$

시차 h

현실적으로 모든 X에 대한 결합분포함수를 구하는 게 불가능!

- If $n = 1$, it means that $X_1 \stackrel{d}{=} X_2 \stackrel{d}{=} X_3 \ldots$
- If $n = 2$, then

조건을 완화한 것이 WEAKLY STATIONARITY

$$(X_1, X_2) \stackrel{d}{=} (X_2, X_3) \stackrel{d}{=} (X_5, X_6) \stackrel{d}{=} \ldots$$

$$(X_1, X_3) \stackrel{d}{=} (X_2, X_4) \stackrel{d}{=} (X_3, X_5) \stackrel{d}{=} \ldots$$

분포의 특징이 Lag(시차); h에 의존

STATIONARITY

WEAKLY STATIONARITY(약정상성)

$$i)\ E[\,|X_t|^2\,] < \infty$$

$$\text{ii)}\ E[X_t] \text{ is constant}$$

$$\text{iii)}\ \gamma_x(r,s) = \gamma_x(r+h, s+h)$$

평균과 **공분산**만 알면 됨!

앞으로 언급할 정상성은 약정상성을 의미

STATIONARITY

개념 : ACVF / ACF / PACF

ACVF (Autocovariance Function)	시차 h에서 $\{X_t\}$의 자기공분산함수	$\gamma_x(\mathrm{h}) = \mathrm{Cov}(X_t, X_{t+h})$
ACF (Autocorrelation Function)	시차 h에서 $\{X_t\}$의 자기상관함수	$\rho_x(h) = \frac{\gamma_x(\mathrm{h})}{\gamma_x(0)} = \mathrm{Corr}(X_t, X_{t+h})$
PACF (Partial Correlation Function)	Z의 영향력을 제외한 X와 Y의 부분상관	$\rho_{x,y.z} = \mathrm{Corr}(\mathrm{X}, \mathrm{Y} \mid \mathrm{Z})$

STATIONARITY

ACVF / ACF의 특징

1) $\rho_x(0) = \frac{\gamma_x(0)}{\gamma_x(0)} = 1$

2) $\gamma_x(0) = \mathrm{Cov}(x_t, x_t) = \mathrm{Var}(x_t) \geq 0$

3) $|\rho_x(\mathrm{h})| \leq \rho_x(0)$ <=> $-1 \leq \rho_x(h) \leq 1$

4) **우함수(Even function) :** $\gamma(h) = \gamma(-h)$

STATIONARITY

PACF

X와 Y의 관계를 볼 때 Z의 영향력을 배제시킨다!

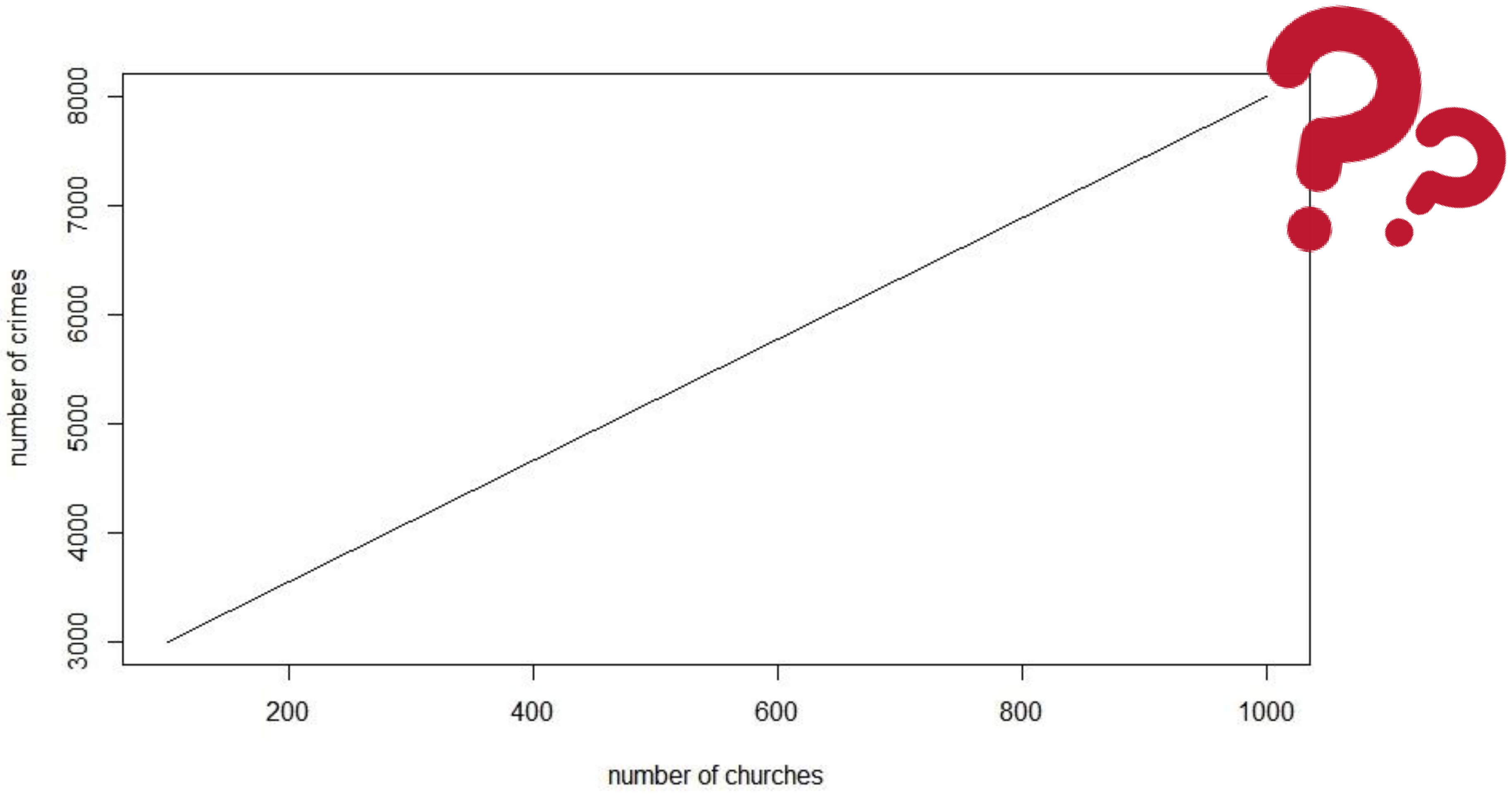
EX

교회의 수가 증가하면
발생하는 범죄 수도 증가할까?

STATIONARITY

PACF

EX

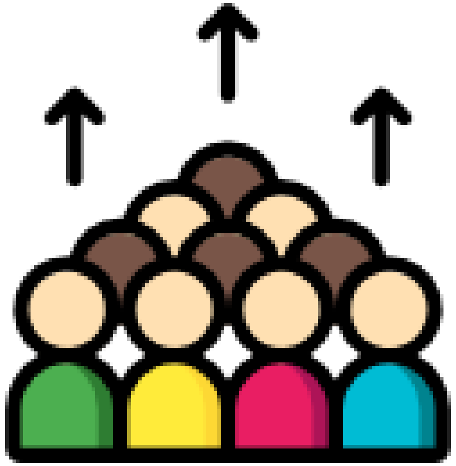


STATIONARITY

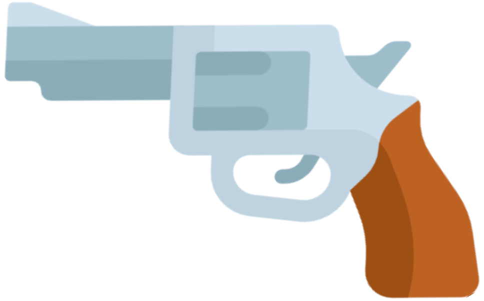
PACF

EX

교회가 많다



인구가 많다



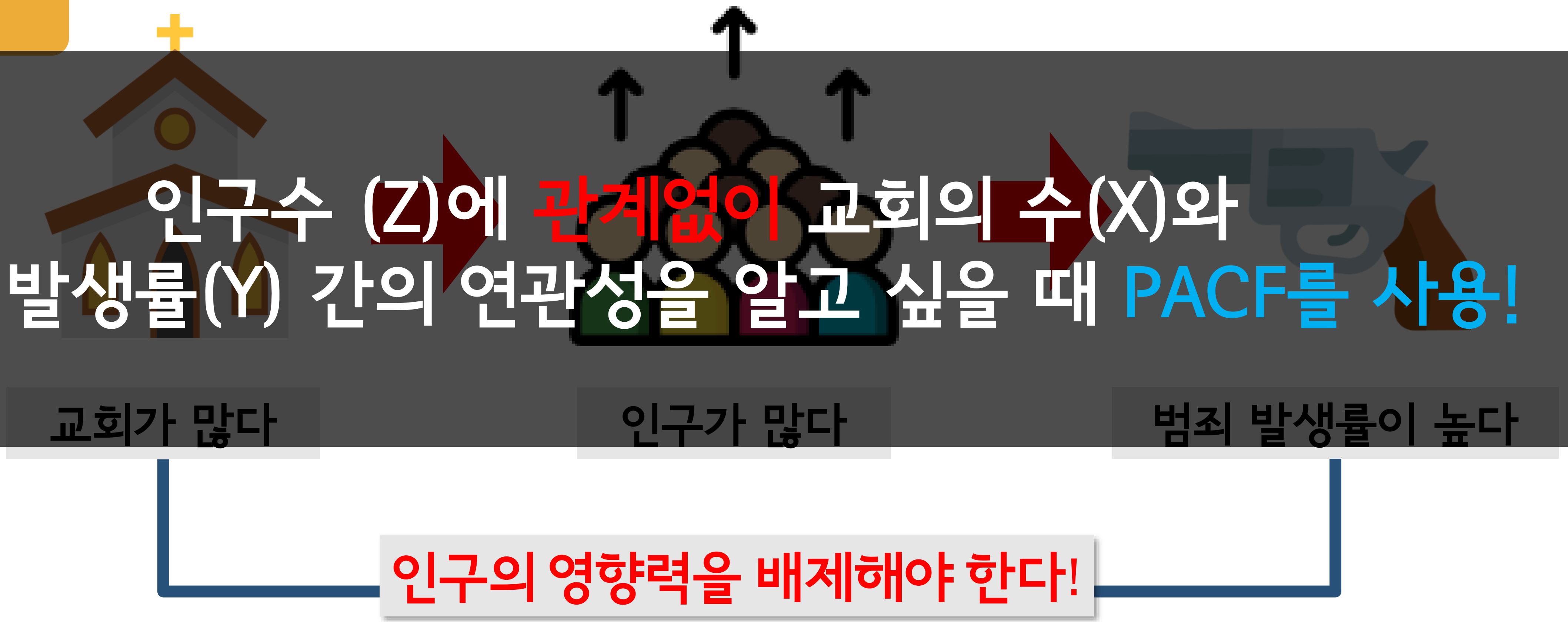
범죄 발생률이 높다

인구의 영향력을 배제해야 한다!

STATIONARITY

PACF

EX



인구수 (Z)에 관계없이 교회의 수(X)와
범죄 발생률(Y) 간의 연관성을 알고 싶을 때 PACF를 사용!

STATIONARITY

PACF

$$\rho_{x,y.z} = \mathrm{Corr(X,Y|Z)}$$

$$X = \alpha \cdot Z + \ error_X\ , Y = \ \beta \cdot Z + \ error_Y$$

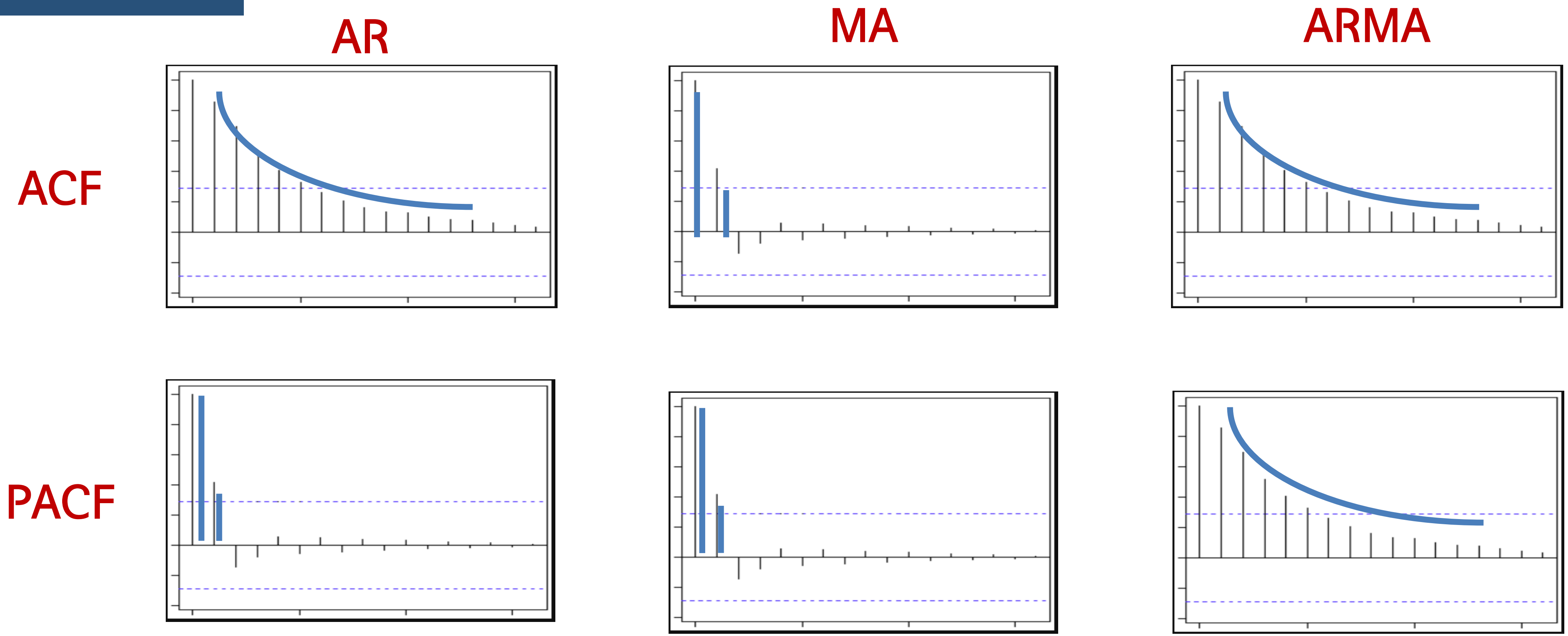
$$error_X = X \ - \ \alpha \cdot Z\ , error_Y = Y \ - \beta \cdot Z$$

$$\mathrm{Corr(X,Y|Z)} = \mathrm{Corr}(X \ - \ \alpha \cdot Z\ , Y \ - \beta \cdot Z)$$

$$\rho_{X,Y.Z} = \frac{\rho_{XY} - \rho_{XZ} \cdot \rho_{YZ}}{\sqrt{1 - \rho_{XZ}{}^2}\sqrt{1 - \rho_{YZ}{}^2}}$$

STATIONARITY

PACF



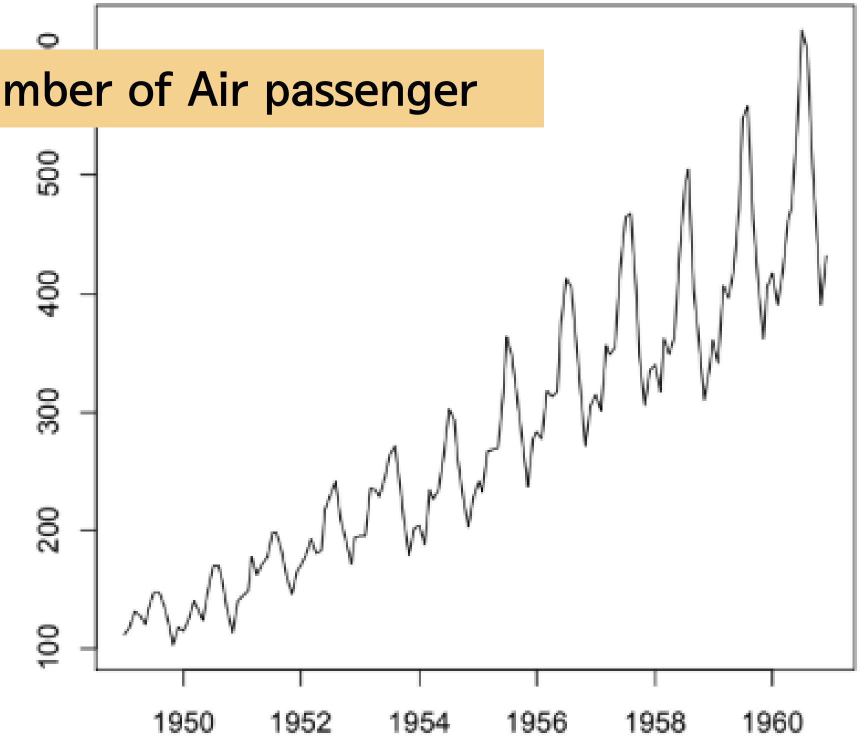
ACF와 PACF의 그림을 보고 어떤 모형을 적용할지 결정 하는데 사용될 것

STATIONARITY

비정상성(Non-stationarity)

But! 우리 주변의 대부분의 데이터들은 정상성을 따르지 않는다.

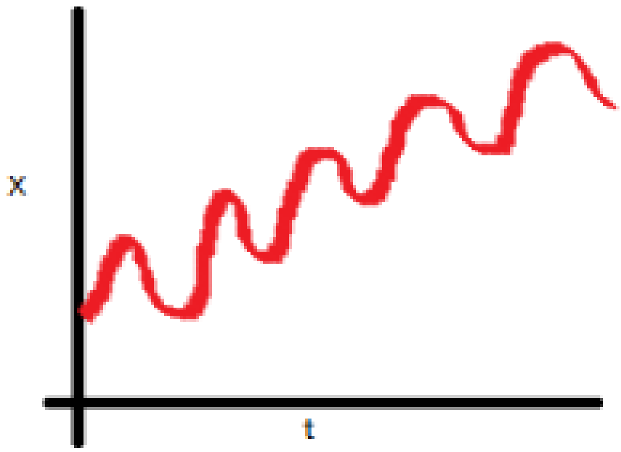
The number of Air passenger



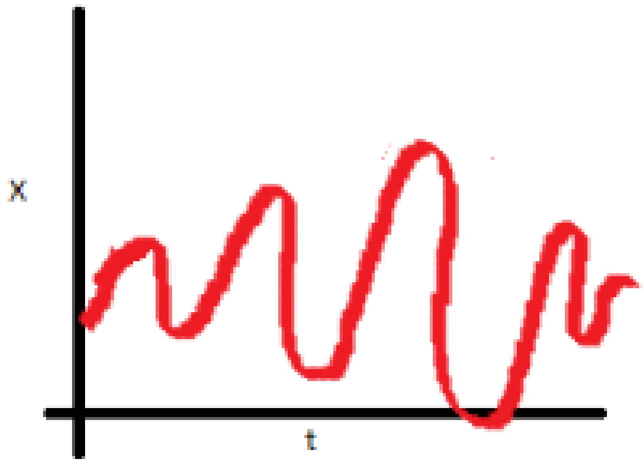
STATIONARITY

비정상성(Nonstationarity)

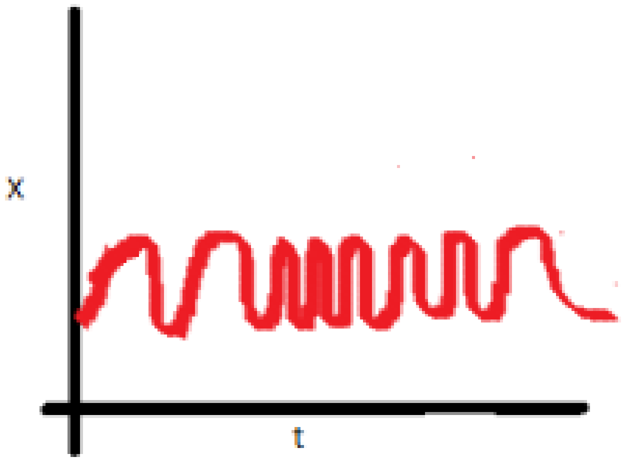
Non-constant mean



Non-constant variance



Time dependent covariance



3

Trend Estimation

Trend Estimation

분해(Decomposition)

$$X_t = \overbrace{m_t + s_t}^{Trend} + Y_t$$

Non Stationary part

Stationary Residuals

Trend Estimation

추세(Trend)

우리가 가진 데이터

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

추세의 다항식

$$m_t = c_0 + c_1 t + \ldots + c_p t^p$$

$$(\hat{c}_0, \ldots, \hat{c}_p) = \underset{c}{\operatorname{argmin}} \sum_{t=1}^{n} (X_t - m_t)^2$$

Trend Estimation

OLS

Moving Average Filter

Exponential Smoothing

Smoothing Splines

Kernel Smoothing

추세(Trend)

우리가 가진 데이터 $X_t = m_t + Y_t, \quad E(Y_t) = 0$

추세의 다항식 $m_t = c_0 + c_1 t + \ldots + c_p t^p$

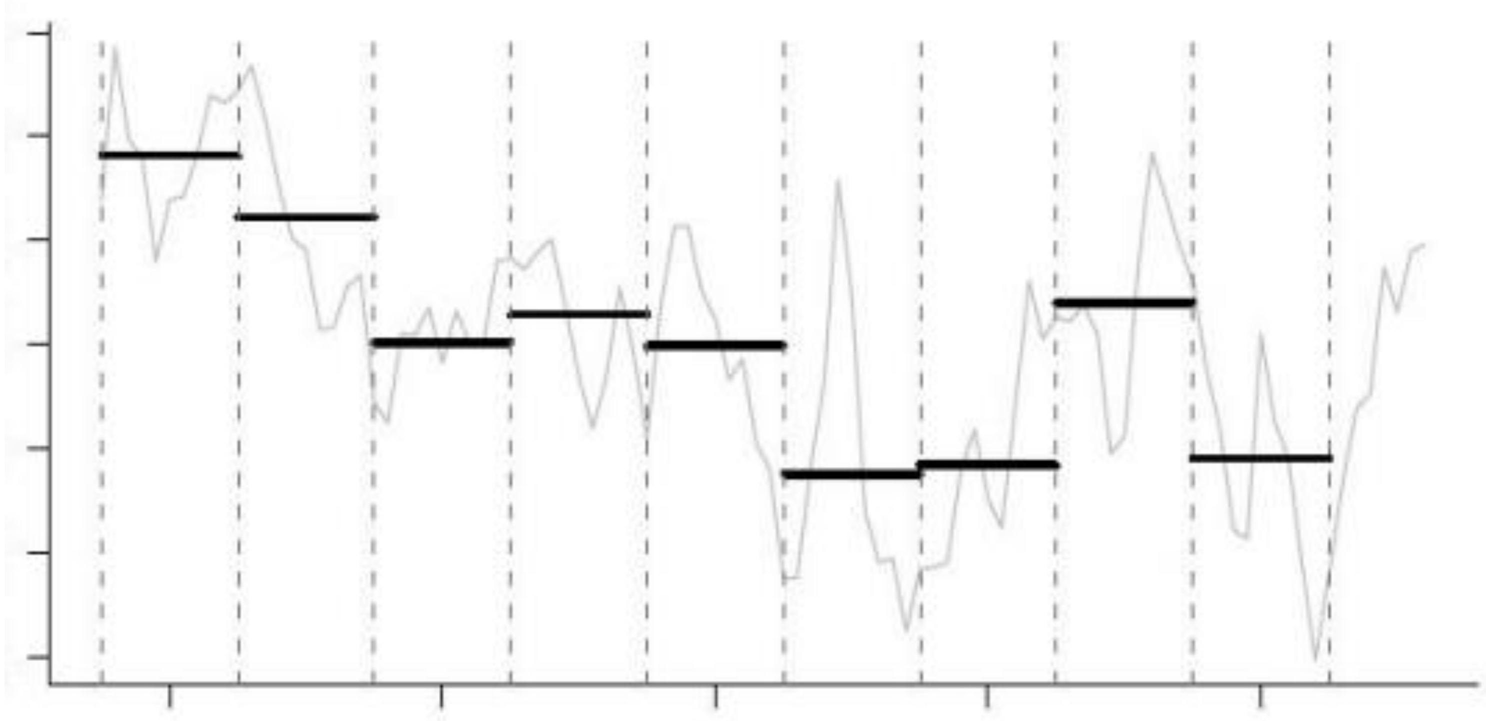
X_t 사이에 자기 상관성이 있기 때문에

"오차항의 공분산은 0"이라는 OLS의 기본 가정을 위배

$$(\hat{c}_0, \ldots, \hat{c}_p) = \underset{c}{\operatorname{argmin}} \sum_{t=1}^{n} (X_t - m_t)^2$$

Trend Estimation

Smoothing



자료를 일정 기간을 나누어 평균을 사용하여 매 측정 순간마다 값에 영향을 미치는 TREND를 보정

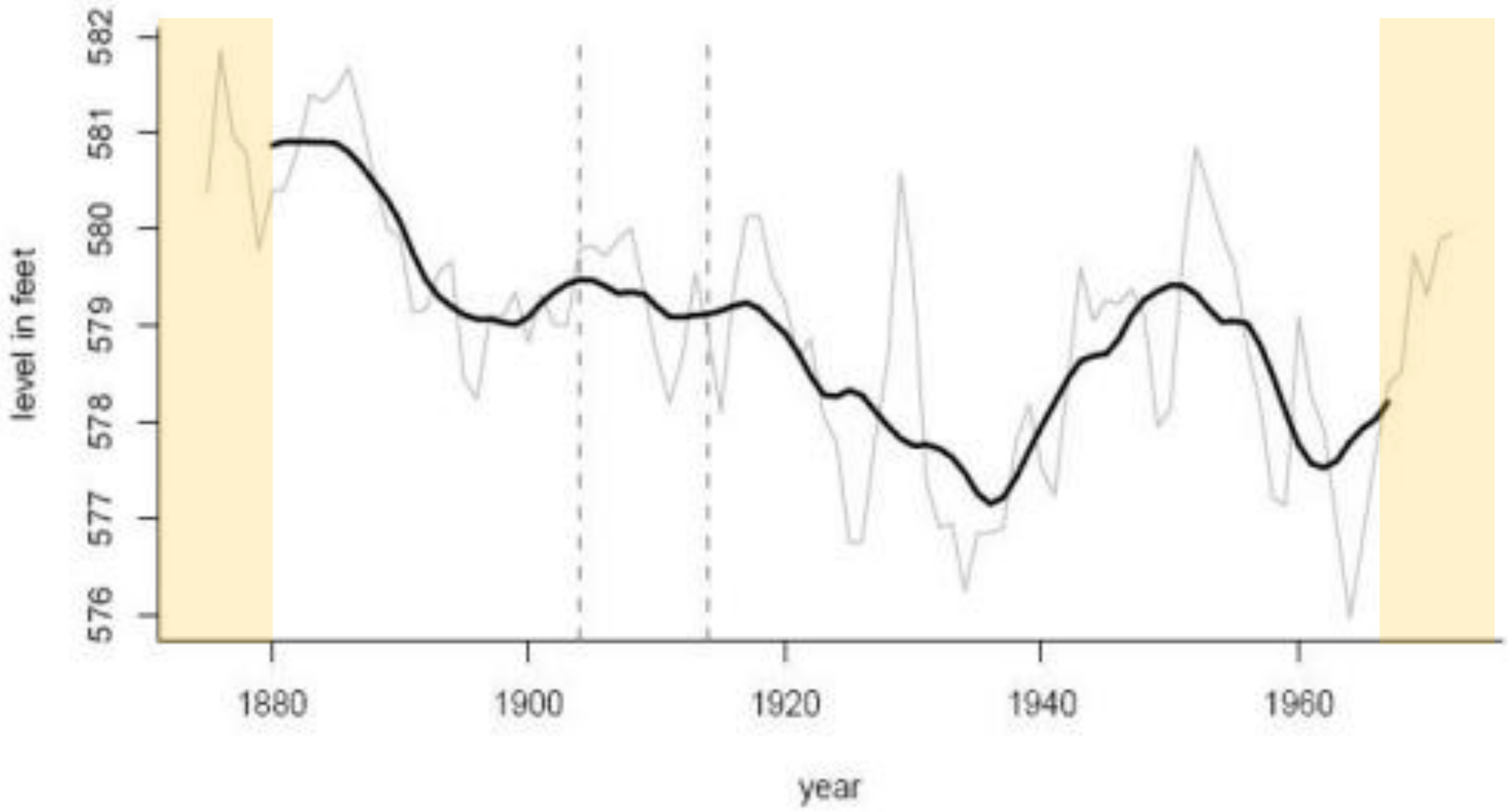
Trend Estimation

$$W_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}$$

주변 과거(-q)와 미래의 값(+q)으로 평활화

Trend Estimation

- OLS
- **Moving Average Filter**
- Exponential Smoothing
- Smoothing Splines
- Kernel Smoothing



자료의 시작 지점과 끝 지점에서 **추세를 추출할 수 없다**

Trend Estimation

우리가 가진 데이터

추세(Trend)

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

Moving Average Filter

$$w_t = \frac{1}{2q+1} \sum_{j=-q}^{q} X_{t-j}$$

$$w_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t-j} + \boxed{\frac{1}{2q+1} \sum_{j=-q}^{q} y_{t-j}} \quad E(Y_t) = 0$$

Ex) $m_t = c_0 + c_1 t$ 라고 할 때, $\frac{1}{2q+1} \sum_{j=-q}^{q} m_{t-j} = c_0 + c_1 t = m_t$

Trend Estimation

추세(Trend)

우리가 가진 데이터

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

Moving Average Filter

$$w_t = \frac{1}{2q+1}\sum_{j=-q}^{q} X_{t-j}$$

그렇다면 q의 크기는 어떻게 정해야 할까?

$$w_t = \frac{1}{2q+1}\sum_{j=-q}^{q} m_{t-j} + \boxed{\frac{1}{2q+1}\sum_{j=-q}^{q} y_{t-j}} \quad E(Y_t) = 0$$

Ex) $m_t = c_0 + c_1 t$ 라고 할 때, $\frac{1}{2q+1}\sum_{j=-q}^{q} m_{t-j} = c_0 + c_1 t = m_t$

Trend Estimation

OLS
Moving Average Filter
Exponential Smoothing
Smoothing Splines
Kernel Smoothing

q ↑

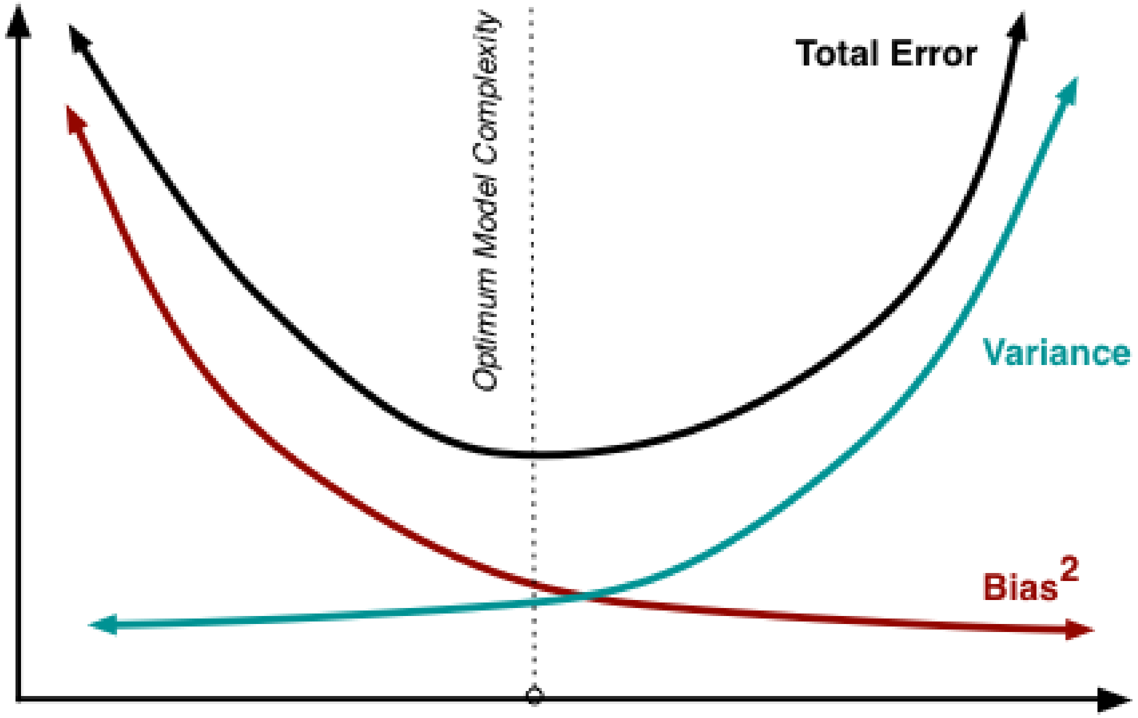
1. 큰 범위를 보기 때문에 추세를 놓칠 수 있음. (Bias 증가)
2. 안정적인 추세선을 찾을 수 있음. (Variance 감소)

q ↓

1. 작은 범위를 보기 때문에 작은 추세까지지도 찾음. (Bias 감소)
2. 변동적인 추세선을 가짐. (Variance 증가)

Bias와 Variance를 모두 고려한 q값!

$$MSE(\hat{\theta}) = V_{ar}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2$$



Trend Estimation

과거의 데이터만 가지고 데이터 예측

$$\begin{cases} \hat{m}_t = \alpha X_t + (1-\alpha)\hat{m}_{t-1} \\ \hat{m}_1 = X_1 \end{cases}$$

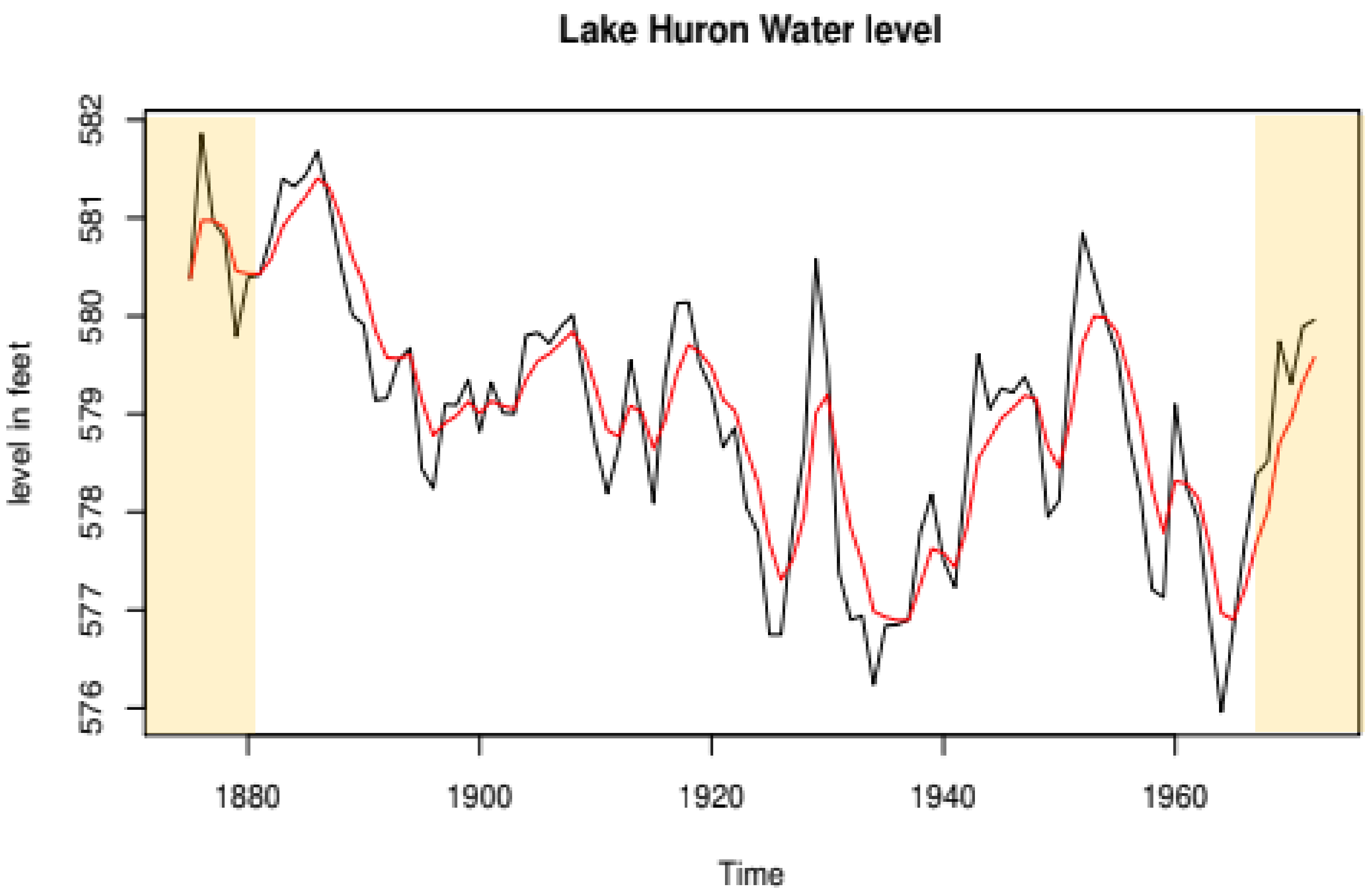
α 값이 최근 값의 비중을 결정

$$\hat{m}_t = aX_t + (1-a)\hat{m}_{t-1}$$

$$= \sum_{j=0}^{t-2} a(1-a)^j X_{t-j} + (1-a)^{t-1} X_1$$

Trend Estimation

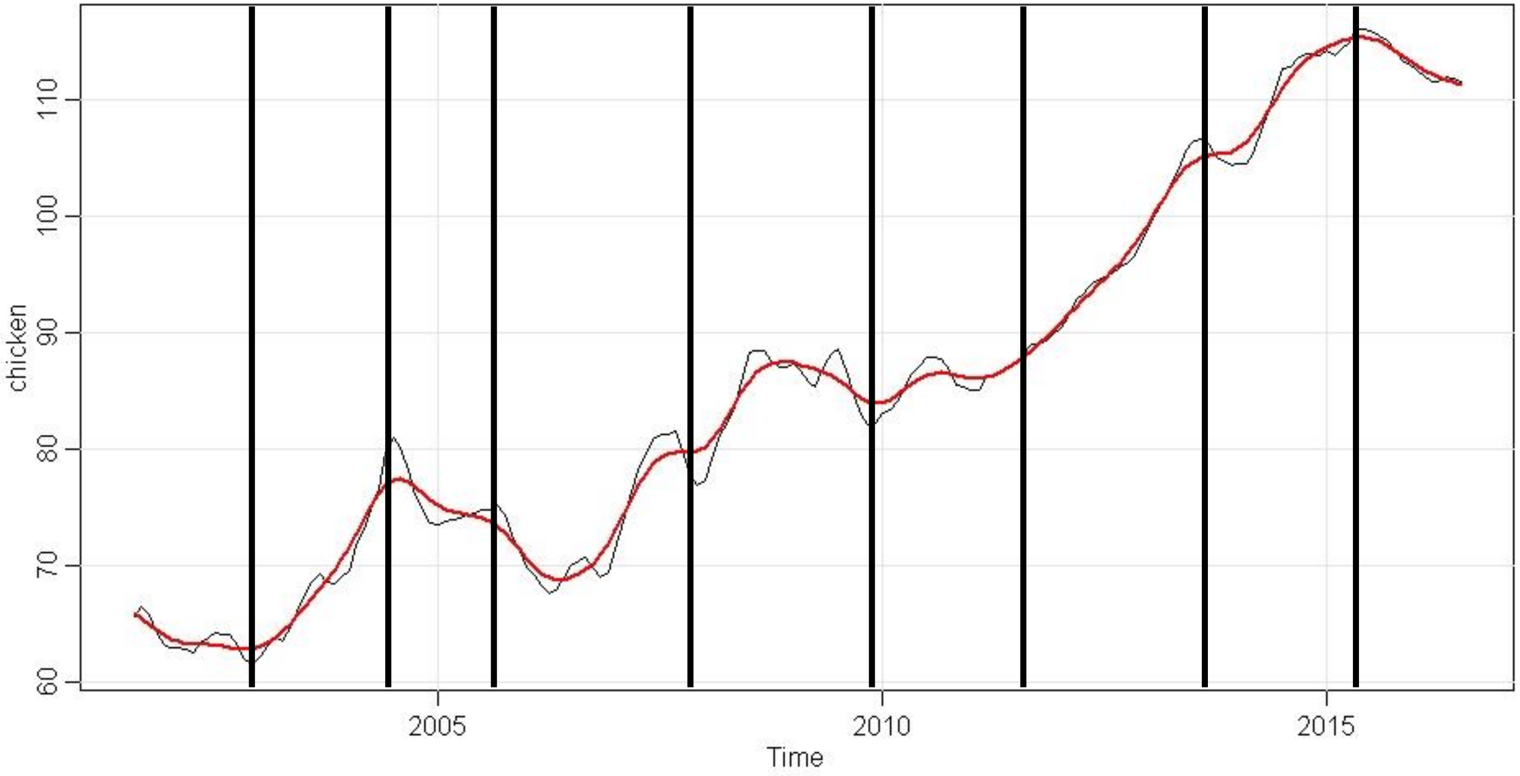
- OLS
- Moving Average Filter
- **Exponential Smoothing**
- Smoothing Splines
- Kernel Smoothing



과거의 자료를 가지고 분석하기 때문에 처음과 끝 모두 추세 분석 가능

Trend Estimation

두 기간을 삼차식으로 연결한 추세 Estimation!



$$\sum_{t=1}^{n}[X_t - f_t]^2 + \alpha \int (f_t{}'')^2 dt$$

모형의 복잡한 정도에 대한 Penalty α

OLS
Moving Average Filter
Exponential Smoothing
Smoothing Splines
Kernel Smoothing

MA와 유사하지만 데이터의 근접성을 고려한 가중치 모델

$$\hat{m}_t = \sum_{i=1}^{n} w_i(t)\, x_i$$

추세는 X_t의 가중 평균

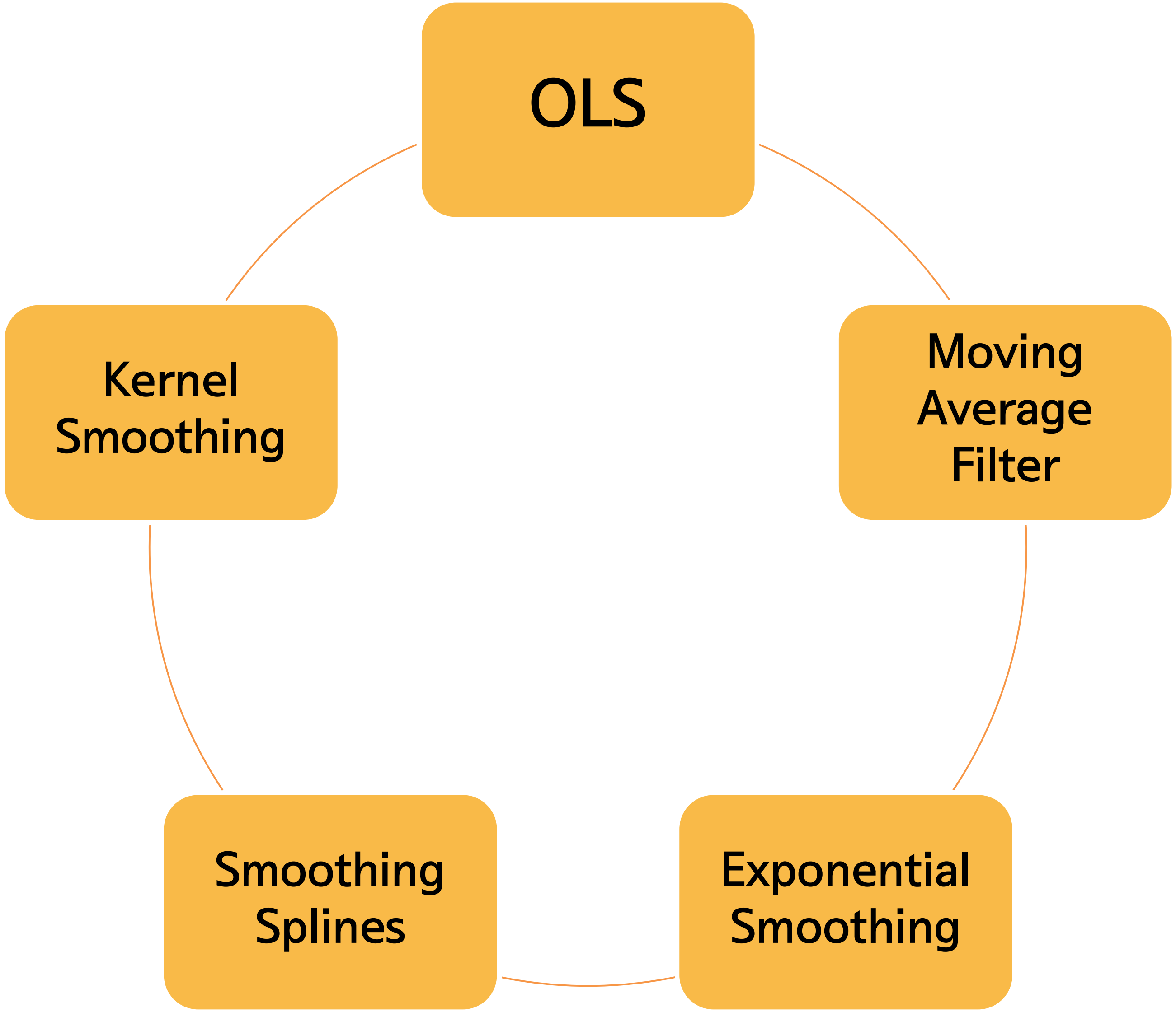
$$w_i(\mathrm{t}) = \frac{K(\frac{t-i}{b})}{\sum_j K(\frac{t-j}{b})}$$

i로부터 멀어질수록 가중치가 줄어든다!

b 값이 최근 값에 가중치를 얼마나 둘 것인지를 결정!

* Kernel function $K(z) = \frac{1}{\sqrt{2\pi}} exp^{-\frac{z^2}{2}}$

Trend Estimation



Trend Estimation

OLS

Kernel Smoothing

Moving Average

Smoothing Splines

Exponential Smoothing

한번에 추세를 제거하는 방법은 없을까?

Trend Estimation

차분(Differencing)

Backshift Operator "B"

$$BX_t = X_{t-1}$$

1차 차분 (Lag-1 Differencing)

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t$$

Trend Estimation

차분(Differencing)

추세 (Trend)

EX

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

if $m_t = c_0 + c_1 t,$

$$\begin{aligned}
\nabla X_t &= X_t - X_{t-1} \\
&= (m_t + Y_t) - (m_{t-1} + Y_{t-1}) \\
&= (m_t - m_{t-1}) + (Y_t - Y_{t-1}) \\
&= (c_0 + c_1 t) - (c_0 + c_1(t-1)) + \nabla Y_t \\
&= c_1 + \nabla Y_t
\end{aligned}$$

추세가 제거됨!

Trend Estimation

차분(Differencing)

EX

추세 (Trend)

$$X_t = m_t + Y_t, \quad E(Y_t) = 0$$

if $m_t = c_0 + c_1 t,$

K번 차분 → K차승의 다항식 Trend까지 제거 가능

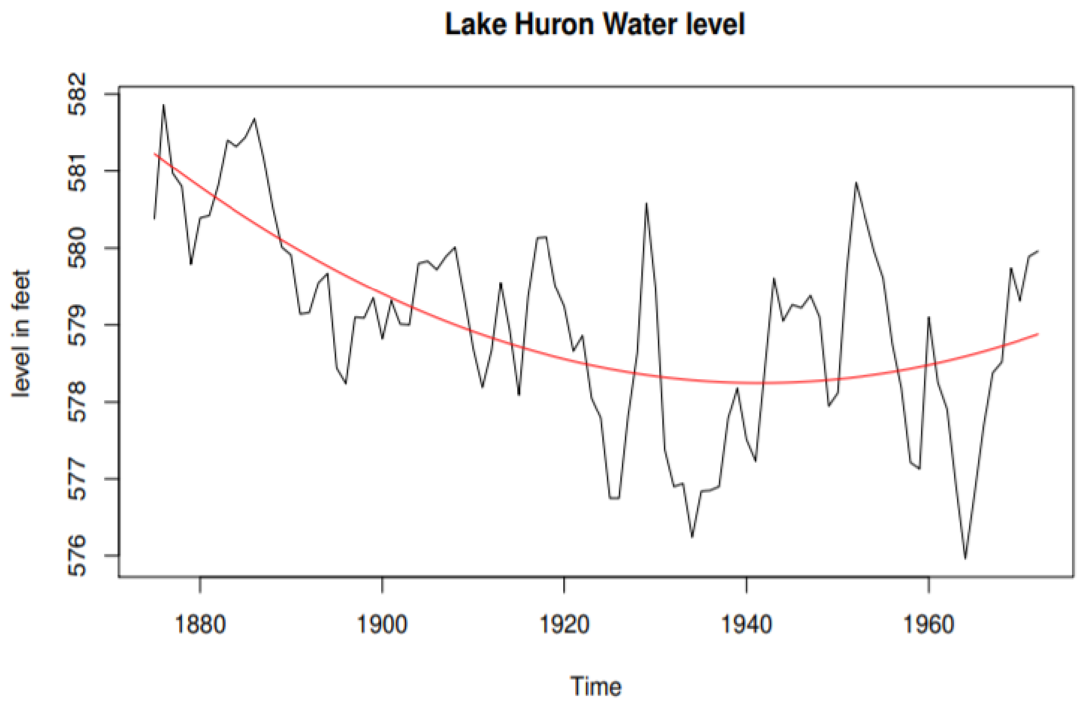
$$\begin{aligned}\nabla X_t &= X_t - X_{t-1} \\ &= (m_t + Y_t) - (m_{t-1} + Y_{t-1}) \\ &= (m_t - m_{t-1}) + (Y_t - Y_{t-1}) \\ &= (c_0 + c_1 t) - (c_0 + c_1(t-1)) + \nabla Y_t \\ &= c_1 + \nabla Y_t\end{aligned}$$

추세가 제거됨!

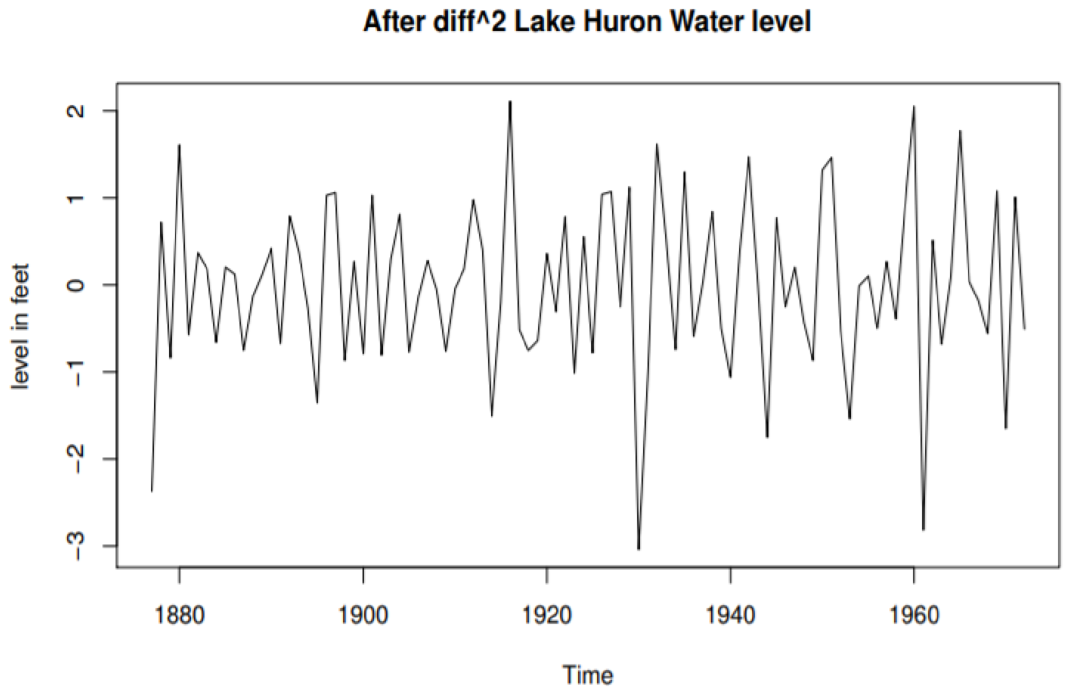
Trend Estimation

차분(Differencing)

EX



2차 차분 후



차분(Differencing)

추세와 계절성이 잘 제거되었는지 어떻게 **확인**하지?

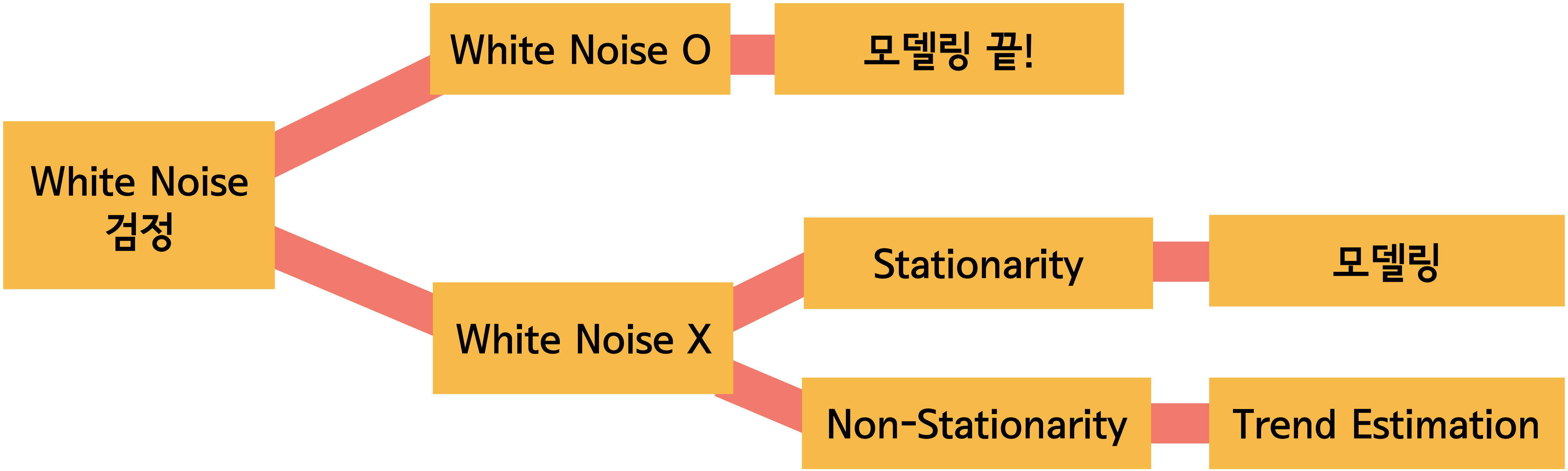
$$\hat{Y}_t = X_t - \hat{m}_t - \hat{s}_t$$

Residual $\hat{Y}_t$이 **정상성을 만족하는지 CHECK!**

4

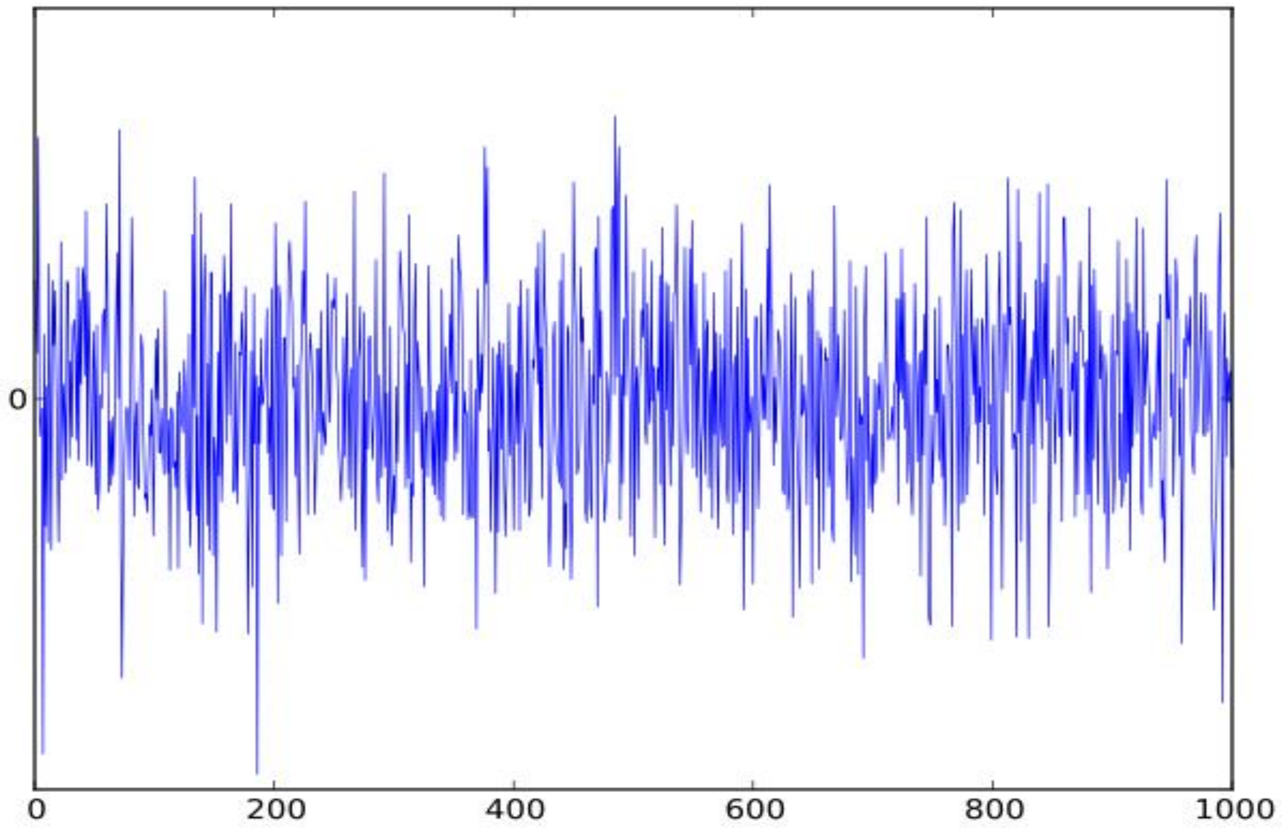
White Noise

White Noise


White Noise 검정
White Noise O
모델링 끝!
White Noise X
Stationarity
모델링
Non-Stationarity
Trend Estimation

White Noise

White Noise란?



When $\boldsymbol{X_t \sim WN\left(0, \sigma_x{}^2\right)}$,

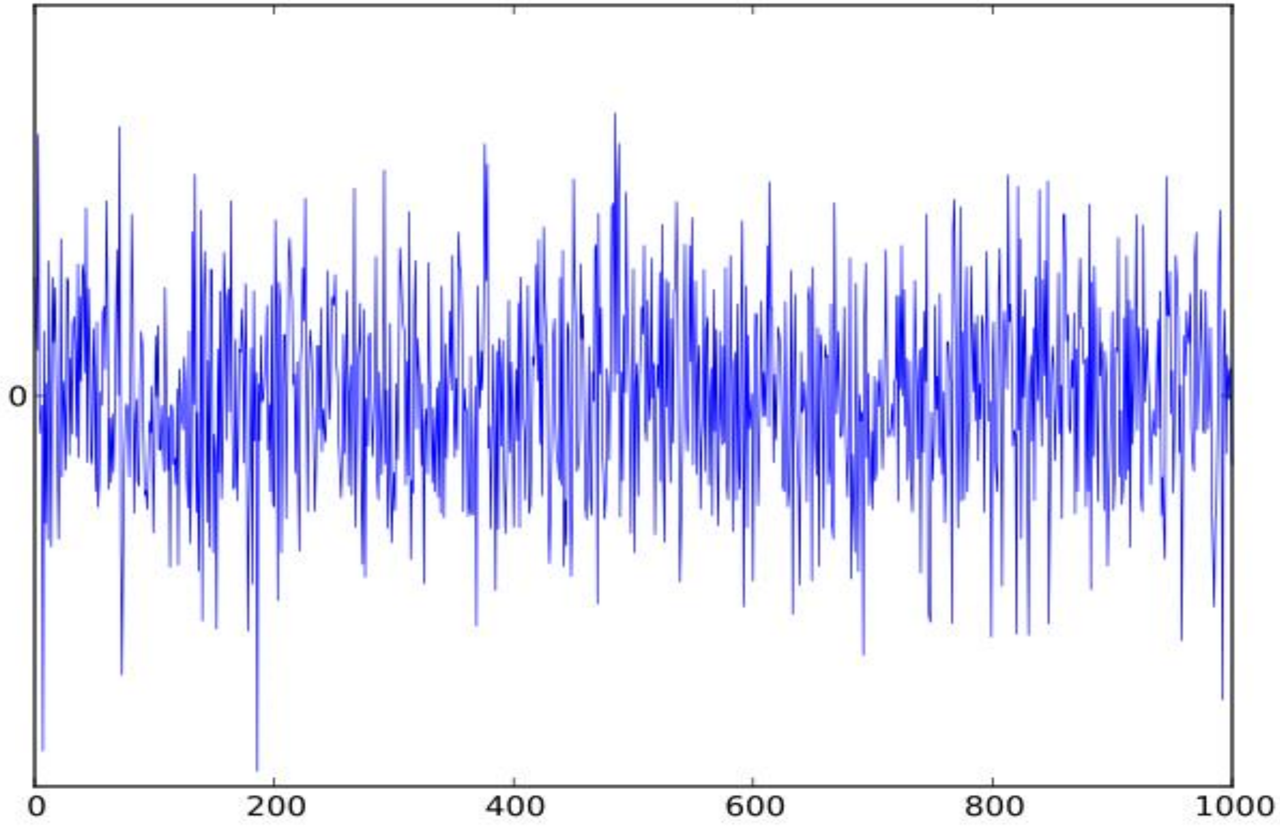
$$i)\ \ E[X_t] = 0$$

$$\text{ii)}\ \ \mathrm{Var}[X_t] = \sigma_x{}^2$$

$$iii)\ \ \gamma_x(r,s) = 0$$

White Noise

White Noise란?



When $\boldsymbol{X_t \sim WN(0, \sigma_x^2)}$,

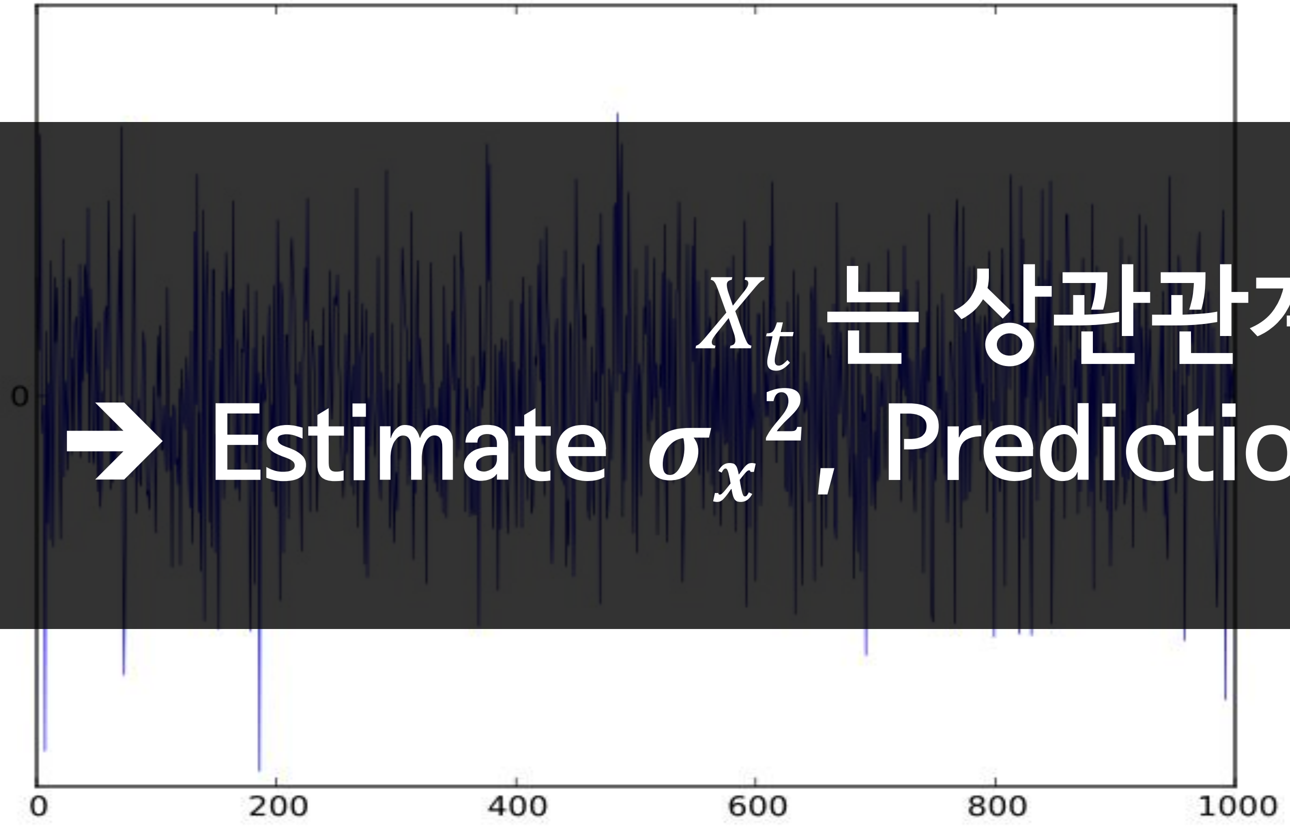
i) $E[X_t] = 0$

ii) $\mathrm{Var}[X_t] = \sigma_x^2$

iii) $\gamma_x(r, s) = 0$

White Noise

White Noise란?



When $X_t \sim WN(0, \sigma_x^2)$,

i) $E[X_t] = 0$

ii) $\mathrm{Var}[X_t] = \sigma_x^2$

iii) $\gamma_x(r, s) = 0$

X_t 는 상관관계가 없다!

→ Estimate σ_x^2, Prediction 불가, Just Random!

White Noise

Test for White Noise

$$H_0 : \rho(h) = 0 \quad \text{vs} \quad H_1 : \rho(h) \neq 0$$

그렇다면 $\rho(h)$의 분포를 알아야겠네?

Test for White Noise

By Central Limit Theorem,

$$\hat{\rho}(h) = \frac{1}{n-h}\sum_{t=h+1}^{n} a_t a_{t-h} \ , \ a_t \sim wn(0,1)$$

$$E[\hat{\rho}(h)] = \frac{1}{n-h} E\left[\sum_{t=h+1}^{n} a_t a_{t-h}\right] = 0$$

$$Var[\hat{\rho}(h)] = \left(\frac{1}{n-h}\right)^2 Var\left[\sum_{t=h+1}^{n} a_t a_{t-h}\right] = \left(\frac{1}{n-h}\right)^2 \times (n-h) = \frac{1}{n-h} \approx \frac{1}{n}$$

Test for White Noise

If errors are WN, then

$$\hat{\rho}(h) \approx \mathcal{N}\left(0, \frac{1}{n}\right).$$

$\hat{\rho}(h)$가 $\frac{1.96}{\sqrt{n}}$ **내**에 있으면 귀무가설(H_0)을 **기각하지 못함**

Y_t는 **Uncorrelated!**

White Noise

Test for White Noise

$$\hat{\rho}(j) \approx N\left(0, \frac{1}{n}\right) \Rightarrow \sqrt{n}\hat{\rho}(j) \sim N(0,1)$$

$$Q = n \sum_{j=1}^{H} \hat{\rho}_{(j)}^2 \approx x_H^2$$

Thus, we reject

H_0: errors are iid vs H_1: not H_0

If $Q > x_H^2(1-\alpha)$

Trend Estimation

Test for White Noise

Portmanteau test

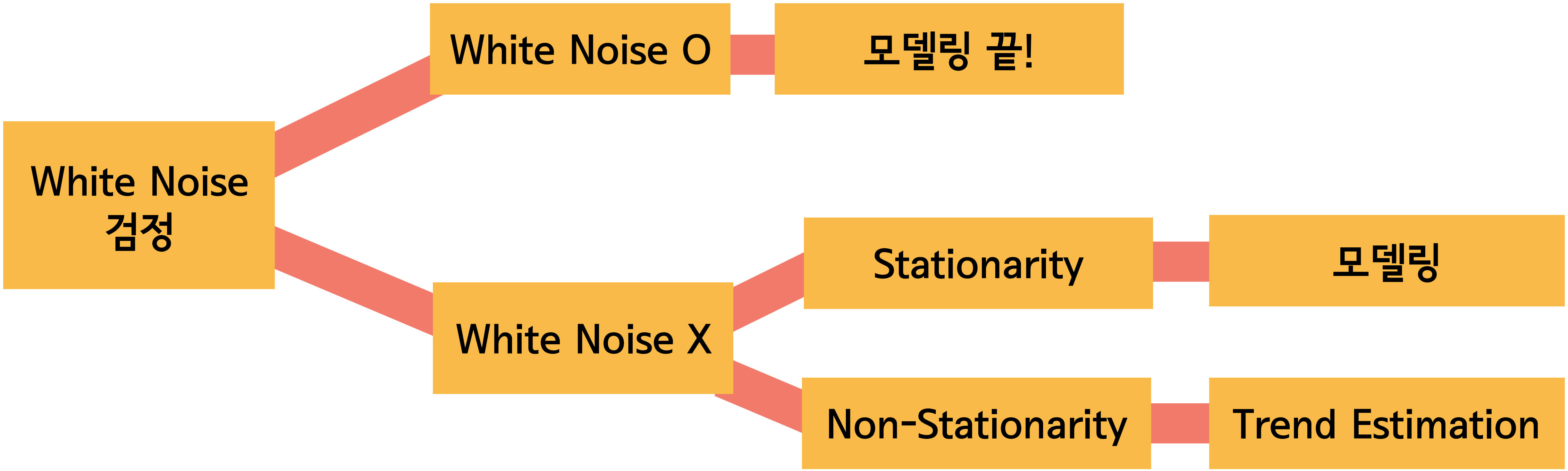
Ljung-Box test

McLeod and Li test

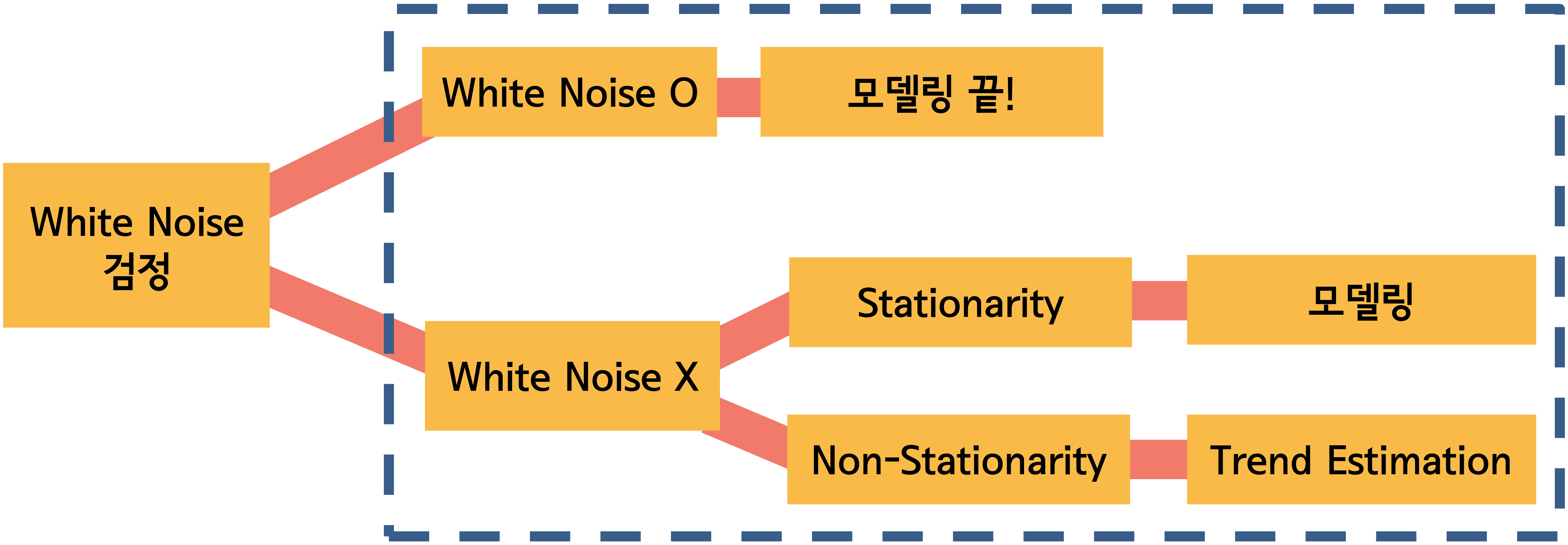
5

Preview

Preview


White Noise 검정
White Noise O
모델링 끝!
White Noise X
Stationarity
모델링
Non-Stationarity
Trend Estimation

Preview


White Noise 검정
White Noise O
모델링 끝!
White Noise X
Stationarity
모델링
Non-Stationarity
Trend Estimation

THANK YOU

부록1: Trend Estimation Code

```
### Kernel Smoothin
ksmooth <- ksmooth( time(chicken), chicken, kernel = 'normal', bandwidth = 0.5 )
ksmooth$y
# larger the bandwidth, trend estimation will be more rough

### smoothing spline
spline <- smooth.spline( chicken, spar = .5)
spline$y
#spar is smoothing parameter lambda

###difference
diff( chicken )

###exponenital smoothing
exp <- smooth.exp( chicken, a = .4 )
#choose a by CV

###moving average
ma <- smooth.ma( chicken , q = 10 )
#choose q by CV
```

부록2: Estimate Seasonality

Seasonal Smoothing

$$\widehat{S_k} = \frac{1}{m} = \left(x_k + x_{k+d} + x_{k+(m-1)d}\right)$$

$$= \frac{1}{m}\sum_{j=0}^{m-1} x_{k+jd}$$

d = one cycle term, m = #of obs,

cycle의 평균이 sesonality의 값일 것이다!

```
season.avg <- season( USA, d = 12 )
#d is cycle
```

부록2: Estimate Seasonality

Seasonal Differencing

$$\nabla_d X_t = (1 - B^d) X_t, t = 1, \ldots, n$$

d = one cycle term, $\nabla_d X_t = s_t - s_{t-d} + Y_t - Y_{t-d} = 0 + error$
Cycle의 주기만큼의 앞 시점(12가 cycle이라면 12시점 전의 값을 빼준다) 차분을 진행

```
diff12 = diff( USA, lag = 12 )
```