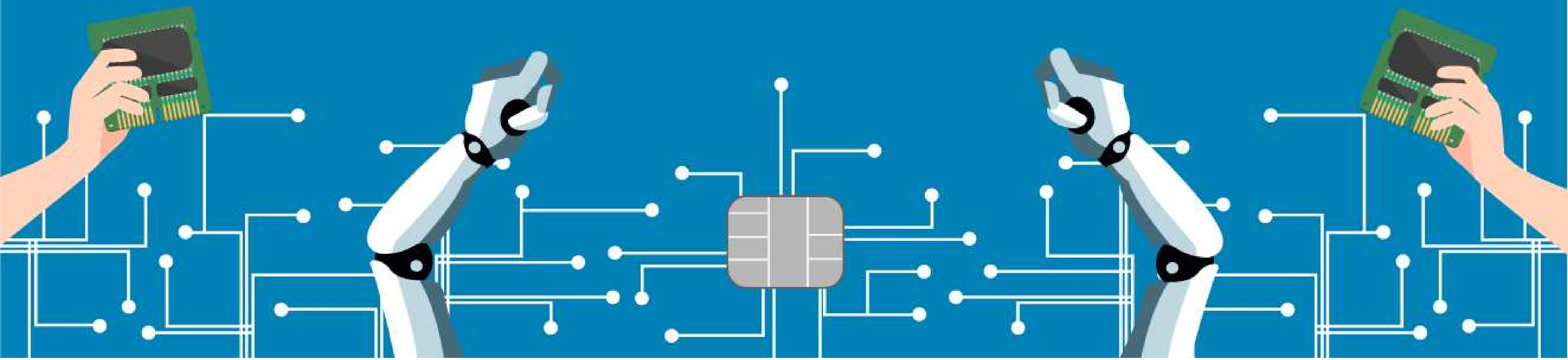
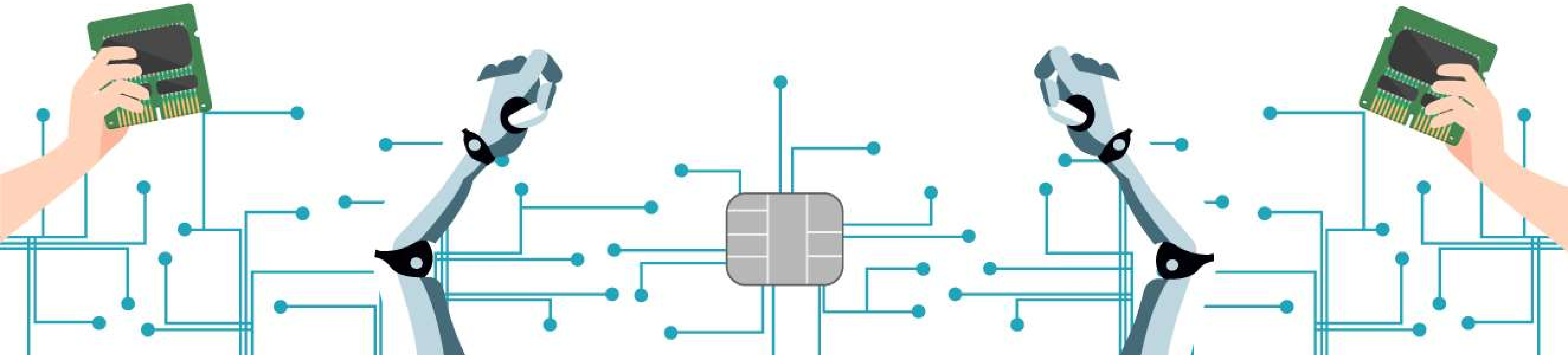
Kuggle

2020_02 Kuggle 정규세션_W7

2020.11.03



CONTENTS



CONTENTS

01.시계열 자료

- 시계열 자료
- 시계열 자료의 형태

01 시계열 자료

1. 시계열 자료

시계열 자료 : 시간의 흐름에 따라 관측된 자료

Ex) 국민 총생산, 물가지수, 강우량 등

*시계열은 분석할 자료들이 시차에 따라 서로 독립이 아닌 경우가 많으므로, 기존의 자료 분석 방법으로는 분석하기 곤란한 경우가 대부분이다. 따라서 이에 맞는 적절한 분석법이 필요하다.

01 시계열 자료

2.시계열 자료의 형태

시계열에서 나타나는 성분

불규칙 성분 : 시간에 따른 규칙적인 움직임과 무관하게 랜덤한 원인에 의해 나타나는 변동 성분

체계적 성분 (추세성분, 계절성분, 순환성분)

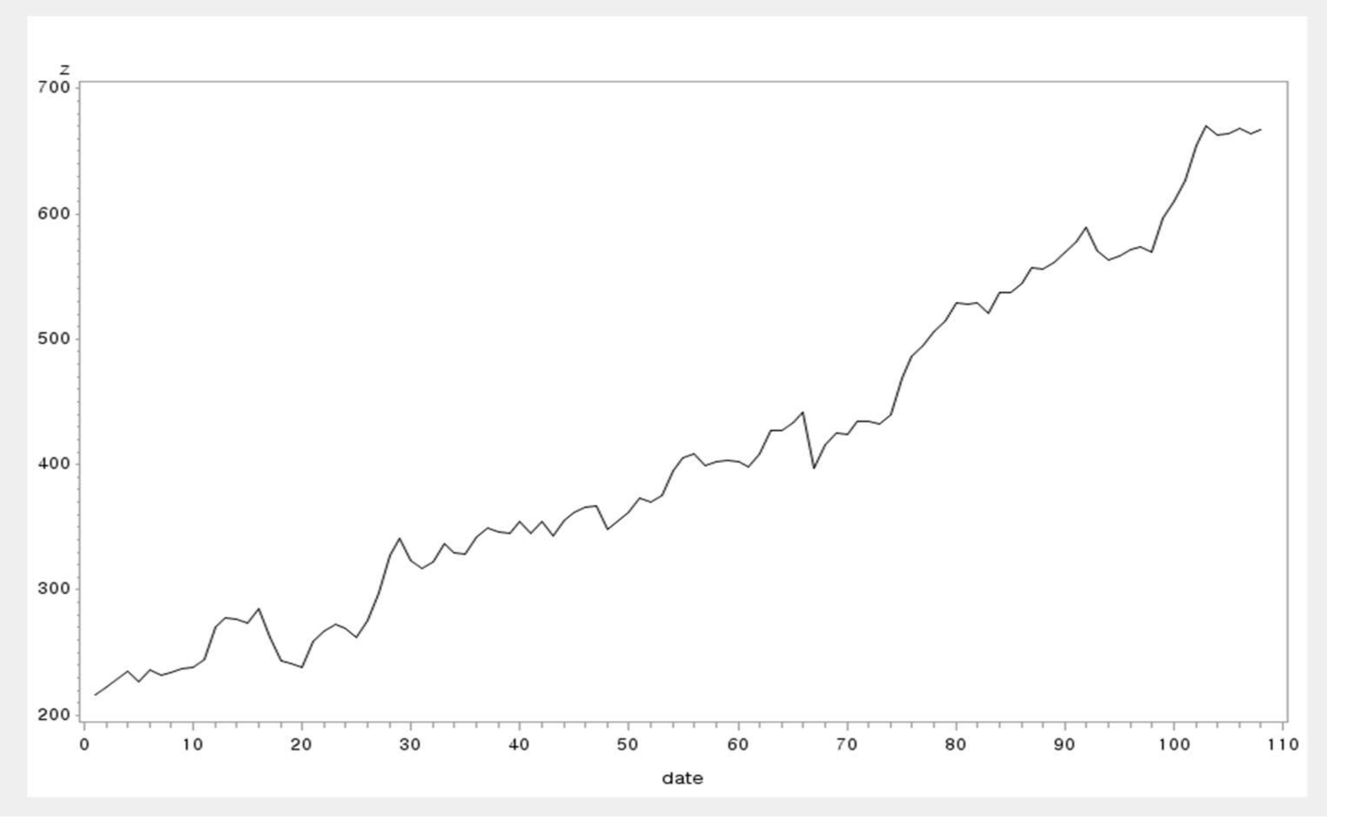
01 시계열 자료

2.시계열 자료의 형태

시계열 자료 분석 시
가장 먼저 시계열도를
그려봐야한다.
(산점도와 비슷)

추세 성분

* 시간이 경과함에 따라 관측값이 지속적으로 증가하거나 감소하는 추세를 갖는 경우의 변동

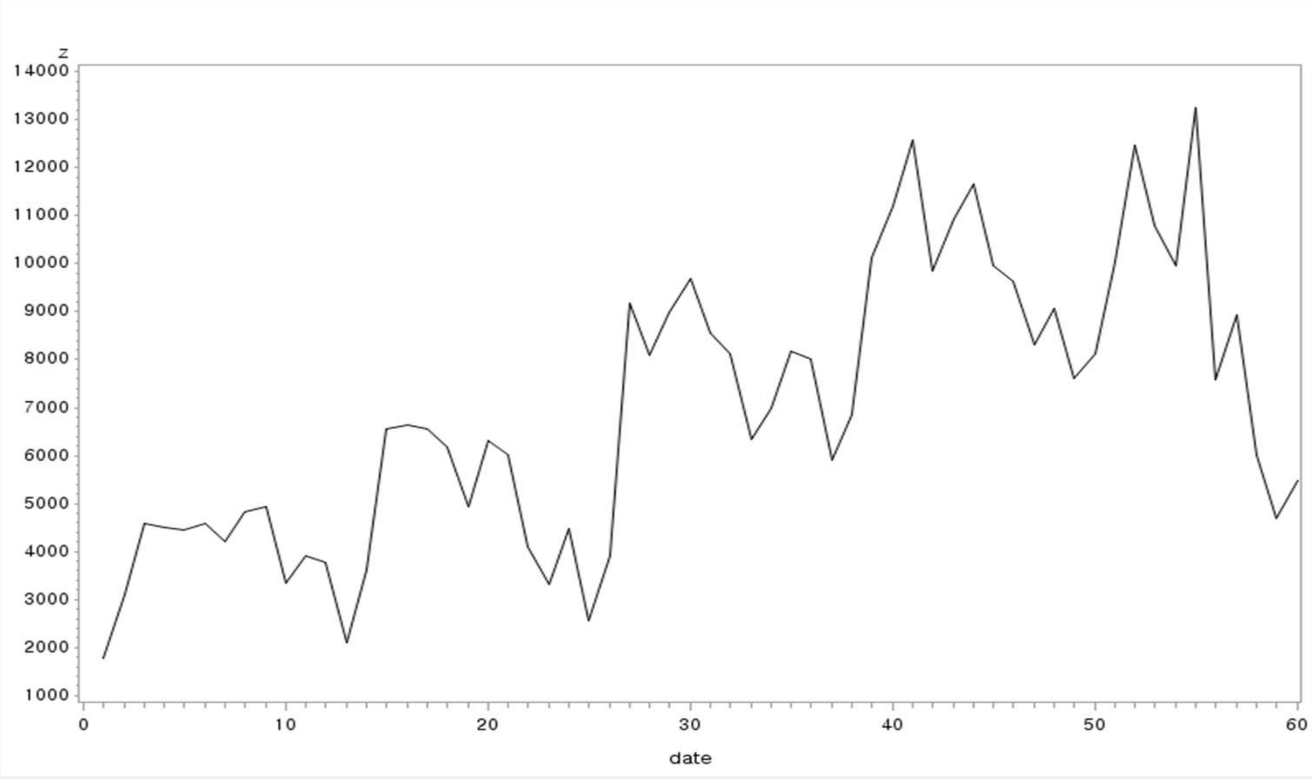


01 시계열 자료

2.시계열 자료의 형태

계절 성분

주별, 월별, 계절별과 같은
주기적인 성분에 의한 변동

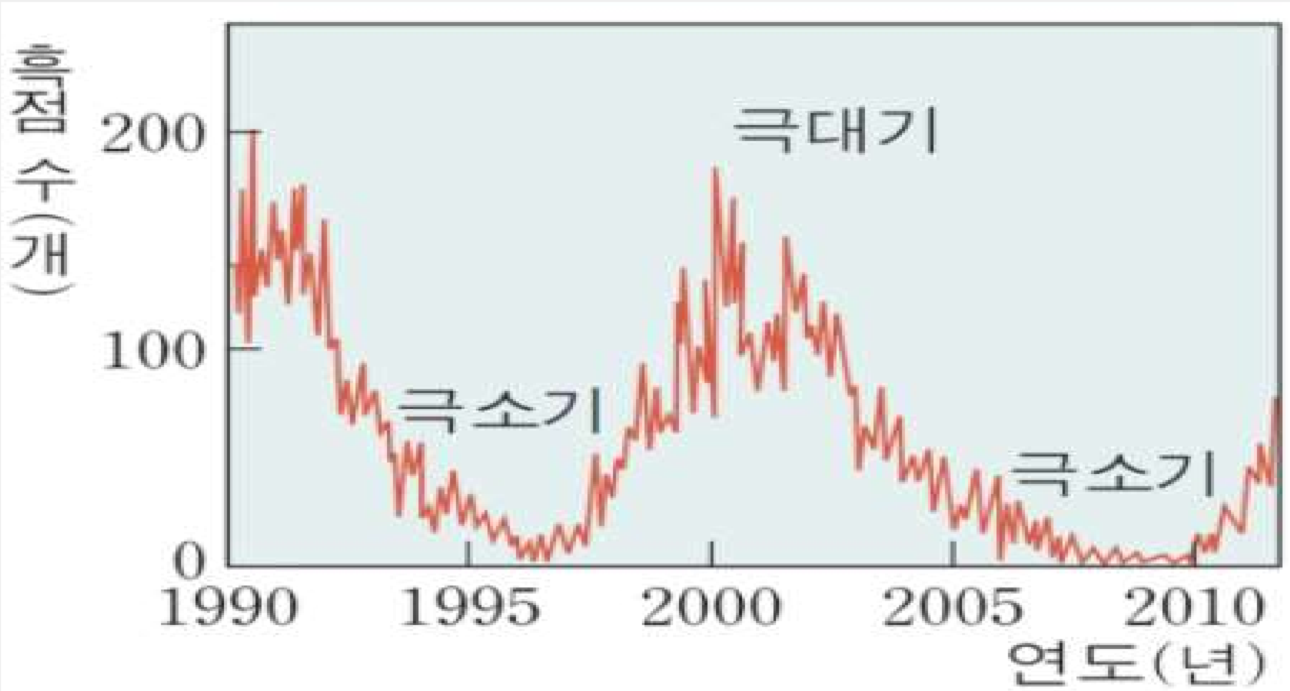


01 시계열 자료

2.시계열 자료의 형태

순환 성분

* 계절 성분과 같이 주기적인 변화를 가지지만, 그 변화가 계절에 의한 것이 아니고 주기가 긴 경우의 변동



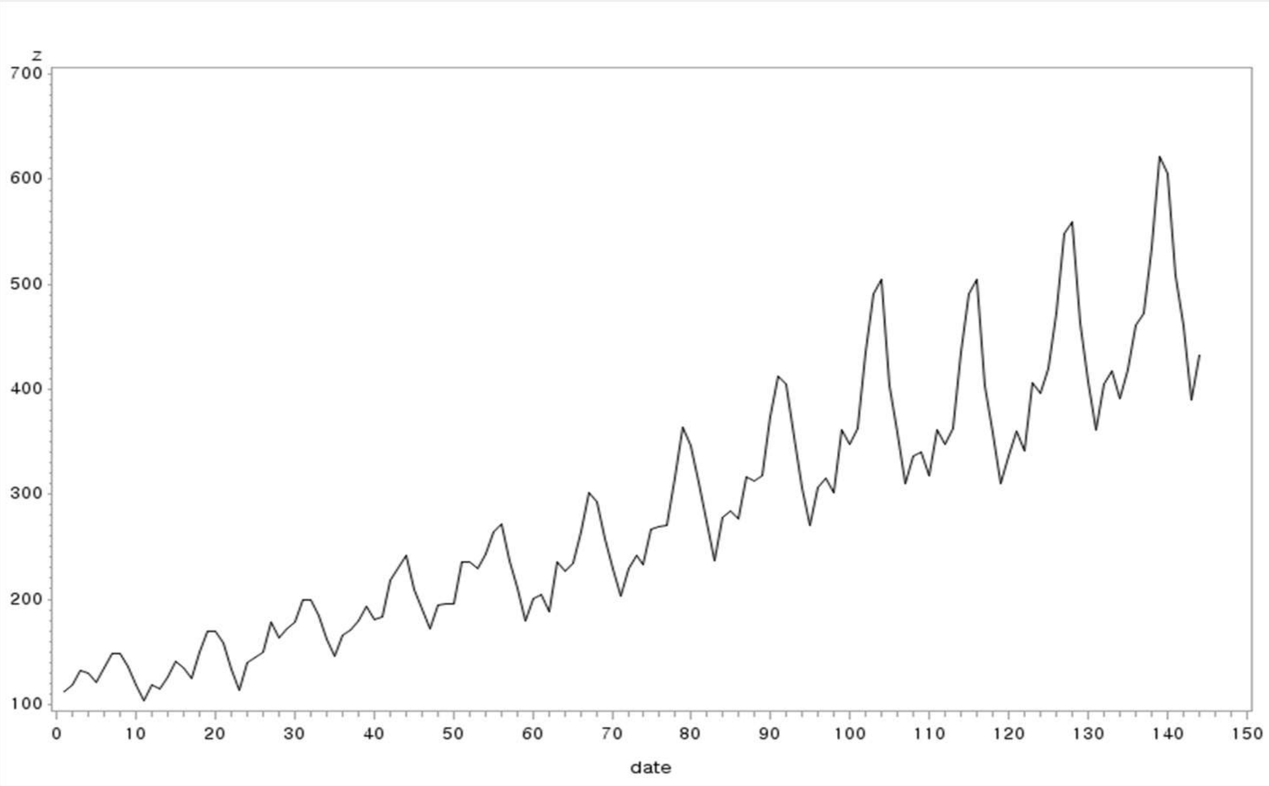
출처 : http://study.zum.com/book/14576

01 시계열 자료

2.시계열 자료의 형태

시간의 흐름에 따라 변동폭이 커지는(이분산성) 시계열의 예

로그 변환을 해주자



CONTENTS

02.추세분석

- 다항 추세모형
- 자기회귀오차모형

02 추세분석

1. 다항 추세 모형

추세모형

관측값을 시간의 함수로서 표현하는 방법. $Z_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \epsilon_t$

추세모형을 이용한 예측법은 다항회귀모형과 유사한 모형을 가정하고 모수의 추정을 통해 예측값을 구함.

회귀모형과의 차이점은 설명변수로 시간의 함수(t)를 사용한다는 점이다.

회귀분석과 마찬가지로 등분산 가정을 만족시키지 못한 경우에는 위에서 언급한 로그변환 등과 같은 분산안정화 변환 필요.

또한, 자료들이 시차에 따라 서로 독립이 아닌 경우가 많으므로, 잔차분석을 통해 추정된 모형이 주어진 자료에 적합한지를 진단해야한다.

02 추세분석

1. 다항 추세 모형

불규칙 성분만을 갖는 경우 : 상수평균모형

$Z_t = \beta_0 + \epsilon_t$, ϵ_t은 서로 독립이고 평균 0, 분산 σ_ϵ^2을 갖는 오차항, 정규분포 가정

$$\hat{\beta_0} = \frac{1}{n}\sum_{t=1}^{n} Z_t = \bar{Z}$$

추세성분만을 갖는 경우 : 선형추세 또는 2차 추세모형

선형추세모형(linear trend model) : $Z_t = \beta_0 + \beta_1 t + \epsilon_t$

$\beta_0 + \beta_1 t$: 선형추세요인

계절성분만을 갖는 경우 : 계절추세모형(삼각함수 or 지시함수 사용)

지시함수를 사용하는 경우

$$Z_t = \beta_0 + \sum_{i=1}^{s} \beta_i \mathrm{IND}_{ti} + \epsilon_t, \quad \text{단}, \ \mathrm{IND}_{ti} = \begin{cases} 1, & t = i \ (\mathrm{mod}\ s) \\ 0, & \text{그 외의 경우} \end{cases}$$

β_0항이 모형에 포함되어 있으면 다음 중 하나의 제약을 준다.

다중공선성을 피하기 위해서

(a) $\beta_0 = 0$: β_i는 i번째 계절의 평균

(b) $\sum_{i=1}^{s} \beta_i = 0$: β_i는 전체평균과 i번째 계절의 평균과의 차이

(c) $\beta_s = 0$: $\beta_i (i \neq s)$는 s번째 계절의 평균과 i번째 계절의 평균과의 차이

주로 a의 제약을 많이 씀.

02 추세분석

1. 다항 추세 모형

추세 및 계절 성분을 동시에 갖는 경우 : 선형계절추세모형

선형추세요인과 계절요인을 동시에 내포하고 있는 경우

$$Z_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \epsilon_t$$

또는

$$Z_t = \beta_0 + \beta_1 t + \sum_{i=1}^{8} \beta_{si} \mathrm{IND}_{ti} + \epsilon_t$$

※시간의 흐름에 따라 그 변화의 폭이 커지는 특징이 있는 경우
먼저 로그변환을 한 후 모형을 적합

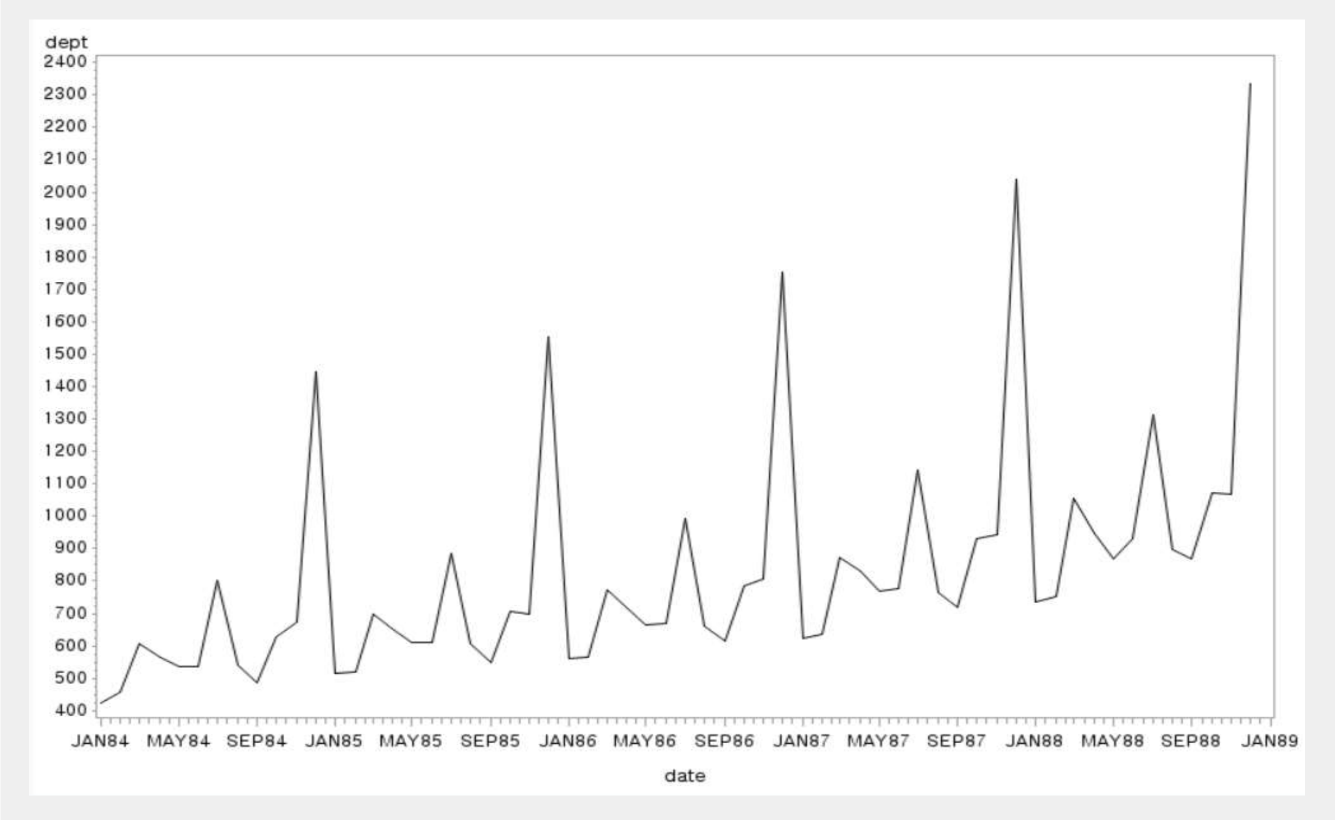
$$\ln Z_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \epsilon_t$$

02 추세분석

1. 다항 추세 모형

선형계절추세모형 예제

시계열도를 그려보면 추세 성분과 계절 성분을 갖는 것을 확인할 수 있으며, 시간이 지날 수록 변동폭이 커지는 이분산성 또한 확인할 수 있다. > 로그변환

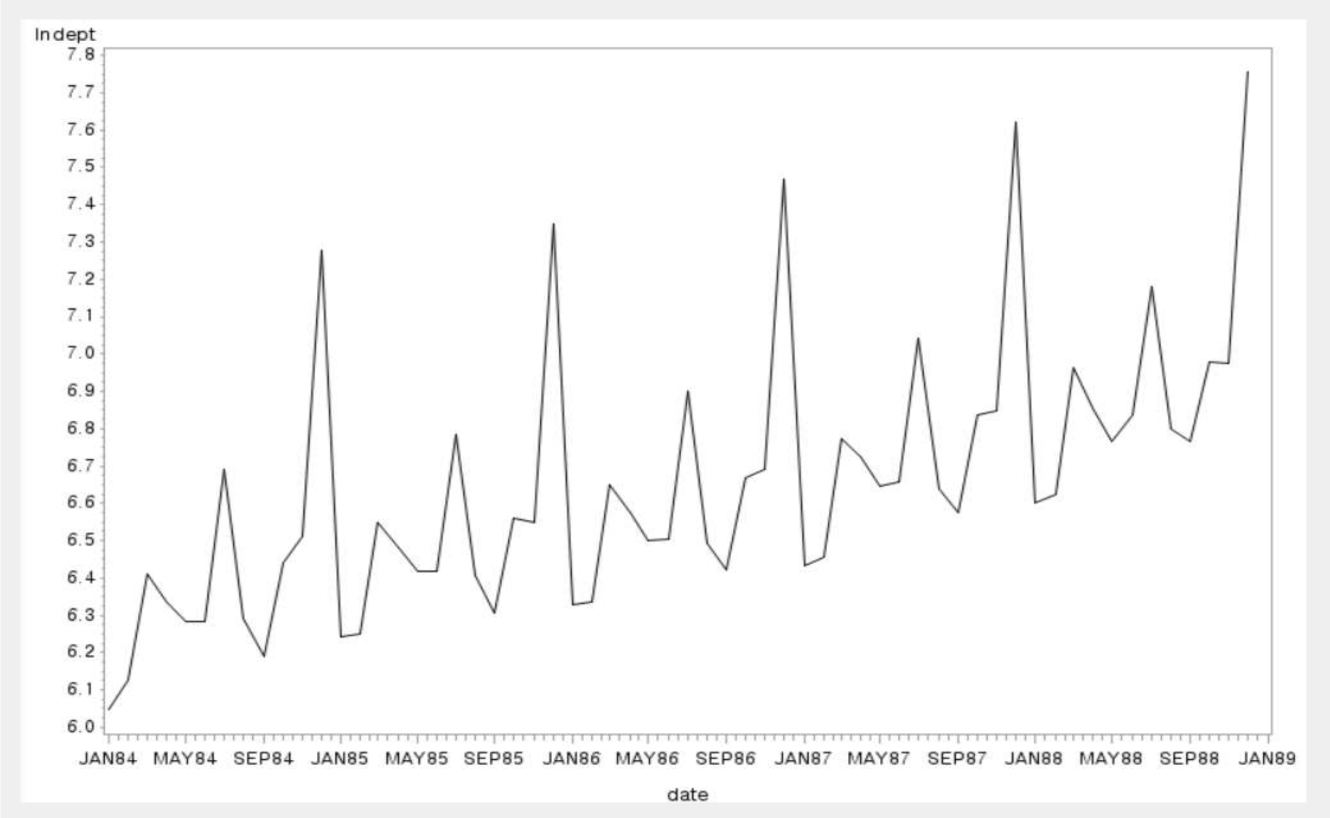


02 추세분석

1. 다항 추세 모형

선형계절추세모형 예제

로그 변환 결과 체계적 성분만 남고, 이분산성이 사라진 것을 확인할 수 있다. 따라서 로그변환한 자료를 선형계절추세모형에 적합



선형계절추세모형 예제

모형을 적합시킨결과, 모두
유의하게 나타남.
따라서 모형식은
ln( ) = 0.01066t+...+7.11048i_t,12

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
t	1	0.01066	0.00019247	55.39	<.0001
i1	1	6.06419	0.01230	493.22	<.0001
i2	1	6.08080	0.01237	491.50	<.0001
i3	1	6.38112	0.01245	512.50	<.0001
i4	1	6.29535	0.01253	502.32	<.0001
i5	1	6.21324	0.01262	492.47	<.0001
i6	1	6.21978	0.01270	489.64	<.0001
i7	1	6.58851	0.01279	515.07	<.0001
i8	1	6.18428	0.01288	480.06	<.0001
i9	1	6.10011	0.01298	470.13	<.0001
i10	1	6.33345	0.01307	484.55	<.0001
i11	1	6.34171	0.01317	481.60	<.0001
i12	1	7.11048	0.01327	535.93	<.0001

02 추세분석

1. 다항 추세 모형

선형계절추세모형 예제

더빈왁슨 통계량/ 잔차 시계열도 분석

더빈 왁슨 통계량이 0.826으로 나옴

변수가 13개(t1개+지시변수12개) 이고 자료가 60개이다.

0.965~1.964 사이에 D 값이 존재하면
Ho : 오차들이 서로 독립이다. 를 기각할 수 없다. 0.965보다 작거나 1.964보다 커야 기각 가능. 0.826은 0.965보다 작으므로 **오차들이 서로 독립이 아니라는 것을 알 수 있다.**

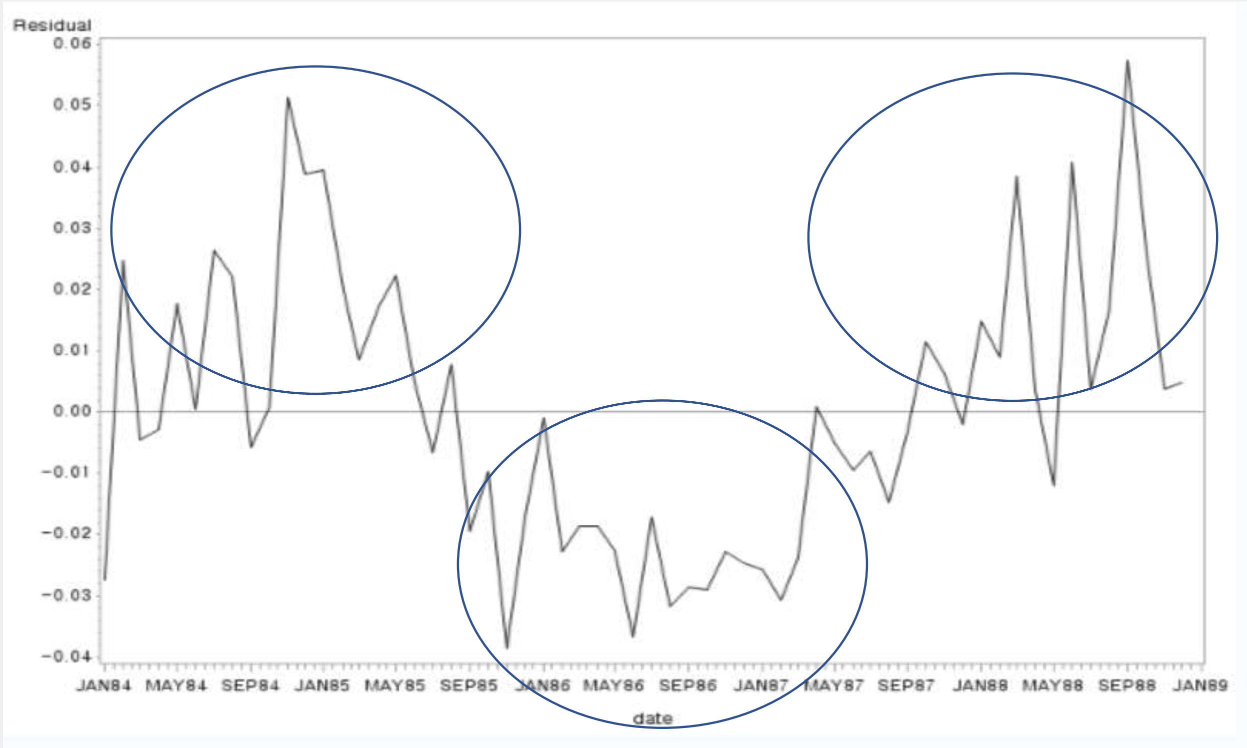
Durbin-Watson D	0.826
Number of Observations	60
1st Order Autocorrelation	0.574

n\k	11		12		13	
31	0.531	2.248	0.475	2.367	0.422	2.487
32	0.558	2.216	0.503	2.330	0.450	2.446
33	0.585	2.187	0.530	2.296	0.477	2.408
34	0.610	2.160	0.556	2.266	0.503	2.373
35	0.634	2.136	0.581	2.237	0.529	2.340
36	0.658	2.113	0.605	2.210	0.554	2.310
37	0.680	2.092	0.628	2.186	0.578	2.282
38	0.702	2.073	0.651	2.164	0.601	2.256
39	0.723	2.055	0.673	2.143	0.623	2.232
40	0.744	2.039	0.694	2.123	0.645	2.210
45	0.835	1.972	0.790	2.044	0.744	2.118
50	0.913	1.925	0.871	1.987	0.829	2.051
55	0.979	1.891	0.940	1.945	0.902	2.002
60	1.037	1.865	1.001	1.914	$\boxed{0.965}$	$\boxed{1.964}$

* 더빈 왁슨(Durbin Watson) 검정
오차항이 독립성을 만족하는지를 검정하기 위해 사용된다. 왜냐하면 오차항의 독립성은 회귀분석의 중요한 가정이기 때문이다. 0 ~ 4 사이의 값을 가지며 판단방법은 다음과 같다.
값이 0에 가까울수록 → 양의 상관관계
값이 4에 가까울수록 → 음의 상관관계
값이 2에 가까울수록 → 오차항의 자기상관이 없음

선형계절추세모형 예제

잔차 시계열도를 그려보아도 한 곳에 오래 머무르는 경향이 있어 양의 상관관계가 있음을 알 수 있다.
>> 자기회귀오차모형



02 추세분석

2. 자기회귀오차모형

자기회귀오차모형

오차의 독립성 가정이 지켜지지 않고 자기상관관계가 존재할 경우 최소제곱법을 이용한 회귀분석은 모수추정의 효율이 떨어지고 bias를 갖게된다.

자기회귀오차모형은 오차들이 서로 독립이 아니고 자기상관관계를 가지며 오차들끼리 자기회귀과정(AR)을 따르는 모형

02 추세분석

2. 자기회귀오차모형

선형계절추세모형 예제

위의 예제를 통해 오차들이 서로 독립이 아니고, 잔차들 간의 양의 상관관계가 존재하므로 자기회귀오차모형을 적합하면,

$$\ln(\widehat{dept_t}) = 0.0108t + 6.0617i1 + \cdots + 7.1020i12 + \epsilon_t$$

$$\epsilon_t = 0.379627\epsilon_{t-1} + 0.389151\epsilon_{t-3}$$

이러한 모형식을 얻을 수 있다.

Estimates of Autoregressive Parameters

Lag	Coefficient	Standard Error	t Value
1	0.379627	0.128388	-2.96
3	0.389151	0.128388	-3.03

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
t	1	0.0108	0.000458	23.67	<.0001
i1	1	6.0617	0.0180	337.38	<.0001
i2	1	6.0766	0.0181	335.90	<.0001
i3	1	6.3764	0.0183	348.86	<.0001
i4	1	6.2910	0.0184	342.52	<.0001
i5	1	6.2084	0.0185	335.95	<.0001
i6	1	6.2141	0.0186	334.08	<.0001
i7	1	6.5829	0.0187	352.42	<.0001
i8	1	6.1790	0.0188	328.60	<.0001
i9	1	6.0931	0.0189	323.18	<.0001
i10	1	6.3261	0.0188	336.45	<.0001
i11	1	6.3370	0.0190	333.96	<.0001
i12	1	7.1020	0.0190	374.30	<.0001

CONTENTS

03.확률과정

- 정상확률과정
- 확률과정의 예
- 자기상관함수(ACF)
- 부분자기상관함수(PACF)

03 확률과정

1. 정상확률과정

정상성

시계열모형이 무수히 많아 이들 중에서 어떤 특정한 성질을 가진 일부만을 고려하자는 취지에서 나온 개념

시계열의 정상성이란 시계열의 확률적인 성질들이 시간의 흐름에 따라 불변이라는 것을 의미.

대부분의 시계열 이론들이 정상성을 가정하고 전개되기 때문에 정상적이 아닌 시계열은 정상 시계열로 변환해 주어야 함

예를 들어 추세변동이나 계절변동 등과 같은 비정상적인 요소들을 갖고 있는 시계열의 경우 이러한 요소들을 먼저 제거하면 정상적이라고 생각 할 수 있다.

엄격한 의미의 정상성 : $f(Z_{t1},Z_{t2},\dots,Z_{tn}) = f(Z_{t1+k},\dots Z_{tn+k})$: 시간 축을 k만큼 이동하여도 pdf가 동일하다는 것을 의미 (거의 만족x)

약한 의미의 정상성 : 평균과 분산이 각각 상수로서 시간 t에 관계없이 동일하고, 자기공분산은 시차 k에만 의존하고 t와는 무관

03 확률과정

2. 확률과정의 예

백색잡음과정

정상적인 확률과정 중에서 가장 중요한 과정이 서로 독립이고 동일한 분포를 따르는 확률변수들의 계열로 구성된 확률과정으로서 이를 백색잡음과정이라고 부른다.

$$E(Z_t) = E(\epsilon_t) = 0$$

$$\gamma_0 = Var(Z_t) = Var(\epsilon_t) = E(\epsilon_t^2) = \sigma_\epsilon^2,$$

$$\gamma_k = Cov(Z_t, Z_{t+k}) = Cov(\epsilon_t, \epsilon_{t+k}) = 0. \qquad K \neq 0$$

※ 기대값과 분산이 시간 t 에 무관 => 정상확률과정

03 확률과정

2. 확률과정의 예

확률보행과정

- 절편(drift)이 없는 **확률보행과정** 또는 **임의보행과정**

$$Z_t = Z_{t-1} + \epsilon_t, \quad Z_0 = 0, \quad t = 1, 2, \dots$$

원점($Z_0 = 0$)에서 출발, $Z_1 = \epsilon_1$이고 $Z_t = \sum_{i=1}^{t} \epsilon_i$는 t 시간 후의 위치

$$E(Z_t) = 0,$$

$$Var(Z_t) = Var\left(\sum_{i=1}^{t} \epsilon_i\right) = t\sigma_\epsilon^2$$

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E\left\{\left(\sum_{i=1}^{t} \epsilon_i\right) \cdot \left(\sum_{i=1}^{t+k} \epsilon_i\right)\right\} = E\left(\sum_{i=1}^{t} \epsilon_i^2\right) = t\sigma_\epsilon^2$$

분산과 자기공분산이 시간의 함수 => **비정상(nonstationary) 확률과정**

- **절편이 있는 확률보행과정**($Z_0 = 0$)

$$Z_t = \delta + Z_{t-1} + \epsilon_t, \qquad t = 1, 2, \dots$$

$$Z_1 = \delta + \epsilon_1, \ \cdots, \ Z_t = t\delta + \sum_{i=1}^{t} \epsilon_i$$

$E(Z_t) = t\delta$: 시간 t의 함수 : 시간의 흐름에 따라 수준이 점차로 증가

$Var(Z_t) = t\sigma_\epsilon^2$: 시간 t의 함수 : 시간의 흐름에 따라 분산이 점차로 증가

03 확률과정

2. 확률과정의 예

이동평균과정

1차 이동평균과정 $Z_t = \mu + \epsilon_t - \theta\epsilon_{t-1}$, $t = 1, 2, \ldots$

$$E(Z_t) = \mu$$

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E[(\epsilon_t - \theta\epsilon_{t-1})(\epsilon_{t+k} - \theta\epsilon_{t+k-1})]$$

$$= \begin{cases} (1+\theta^2)\sigma_\epsilon^2, & k=0 \\ -\theta\sigma_\epsilon^2, & k=1 \\ 0, & \text{다른경우} \end{cases}$$

: 평균과 분산 및 자기공분산은 시간 t와는 무관, 시차 k만의 함수

=> 정상확률과정. 1차-**이동평균과정**

03 확률과정

2. 확률과정의 예

선형과정

$$Z_t = \mu + \epsilon_t + \psi_1 \epsilon_{t-1} + \psi_2 \epsilon_{t-2} + \cdots = \mu + \sum_{j=0} \psi_j \epsilon_{t-j}, \quad \psi_0 = 1$$

Z_t : 서로 독립인 확률변수들의 선형결합

$E(Z_t) = \mu$,

$$Var(Z_t) = \sigma_\epsilon^2 \sum_{j=1}^{\infty} \psi_j^2$$

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E\left(\sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \psi_i \psi_j \epsilon_{t-i} \epsilon_{t+k-j}\right) = \sigma_\epsilon^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$$

선형과정이 정상적이기 위한 조건(stationarity condition)

시점 t 와는 무관, 시차 k 만의 함수이기 위한 조건 : $\sum_{j=0}^{\infty} \psi_j^2 < \infty$

$$\because \ |\gamma_k| \le \sqrt{Var(Z_t) \cdot Var(Z_{t+k})} = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2$$

무한합이 유한합임을 보여야 하므로 위와 같은 조건

03 확률과정

2. 확률과정의 예

자기회귀과정

1차 AR과정 $Z_t - \mu = \phi(Z_{t-1} - \mu) + \epsilon_t$, $E(Z_t) = \mu < \infty$

※ 1차 AR과정은 마코프(Markov)과정이라고도 함

Let $Z_t = \delta + \phi Z_{t-1} + \epsilon_t$ 또는 $\dot{Z}_t = \phi \dot{Z}_{t-1} + \epsilon_t$, 단, $\mu = \delta/(1-\phi)$이고 $\dot{Z}_t = Z_t - \mu$

$$\dot{Z}_t = \phi(\phi \dot{Z}_{t-2} + \epsilon_{t-1}) + \epsilon_t = \phi^2(\phi \dot{Z}_{t-3} + \epsilon_{t-2}) + \epsilon_t + \phi\epsilon_{t-1}$$

$$\vdots$$

$$= \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}.$$

정상성의 조건

$$\gamma_0 = Var(Z_t) = E(Z_t - \mu)^2 = \sum_{j=0}^{\infty} \phi^{2j} E(\epsilon_{t-j}^2) = \sigma_\epsilon^2 \sum_{j=0}^{\infty} \phi^{2j}$$

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E(Z_t - \mu)(Z_{t+k} - \mu) = \mathrm{E}(\dot{Z}_t \dot{Z}_{t+k})$$

$$= E[(\sum_{j=0}^{\infty} \phi^j \epsilon_{t-j})(\sum_{i=0}^{\infty} \phi^i \epsilon_{t+k-i})]$$

$$= \sigma_\epsilon^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+k} = \phi^k \sigma_\epsilon^2 \sum_{j=0}^{\infty} \phi^{2j} \ (= \phi^k \gamma_0)$$

$$|\phi| < 1 \ \Longrightarrow \ \sum_{j=0}^{\infty} \phi^{2j} < \infty$$

03 확률과정

3. 자기상관함수

자기상관함수 – 시간에 따른 상관 정도를 나타내기 위해.
이 함수의 그래프를 통해 뒤에 나오는 AR,MA,ARMA 모형의 차수를 정함

정의 자기공분산 함수(autocovariance function)

$$\gamma_k = Cov(Z_t, Z_{t+k}) = E[(Z_t - \mu)(Z_{t+k} - \mu)]$$

: 시간에 따른 상관정도의 척도

정의 자기상관 함수(autocorrelation function)

$$\rho_k = Corr(Z_t, Z_{t+k}) = \frac{Cov(Z_t, Z_{t+k})}{\sqrt{Var(Z_t) \cdot Var(Z_{t+k})}} = \frac{\gamma_k}{\gamma_0}$$

정의 표본상관도표(sample correlogram)

- $SACF$ $\hat{\rho}_k$, $k = 0, 1, 2, \ldots$ 와 표본부분자기상관계수($SPACF$)의 그림
- 주어진 시계열자료가 어느 확률과정의 모형으로부터 생성된 것인지를 판단하는데 이용되며 $ARMA(p,q)$ 모형의 적합시 모형의 차수 p 와 q 의 값을 결정하는 데 주로 이용됨

03 확률과정

4. 부분자기상관함수

부분자기상관함수

정의 부분자기상관 함수(autocorrelation function, PACF)

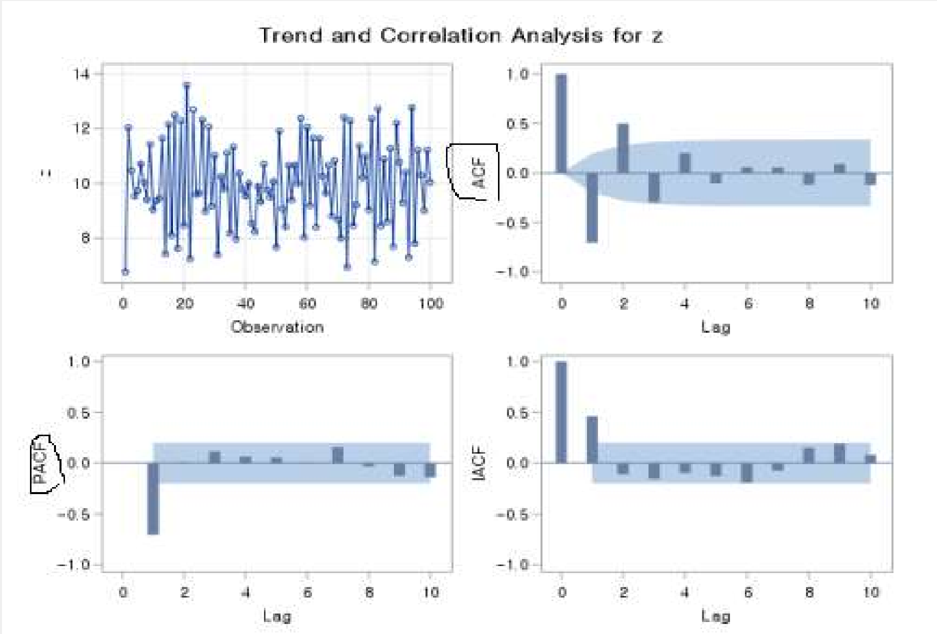
- 표기: ϕ_{kk}
- Z_t 와 Z_{t+k} 에서 $Z_{t+1}, Z_{t+2}, \ldots, Z_{t+k-1}$ 의 효과를 제거한 후의 상관계수
- $PACF \quad \phi_{kk} = Corr\{ Z_t^*, Z_{t+k}^* \}$

03 확률과정

3. 자기상관함수

자기상관계수, 부분자기상관계수의 그래프

> ACF, PACF 의 절단점, 형태 등을 통해
모형의 차수를 정할 수 있다.(뒤에서 자세히)



CONTENTS

04.AR,MA 모형

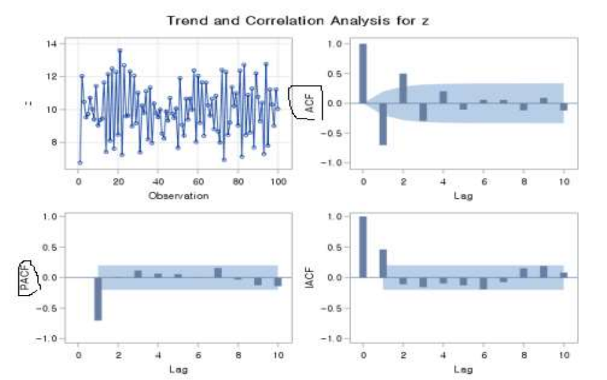
- AR 모형
- MA 모형
- ARMA 모형

04 AR, MA 모형

1. AR 모형

자기회귀과정

시계열과정에서 현재의 상태가 과거의 상태에 의존한다면 현재 관측값들의 함수 형태로 나타낼 수 있다.



Zt = f(Zt-1,Zt-2,...) + et 의 관계를 만족할 때 자기회귀과정이라고 부른다.

가장 간단한 형태는 바로 직전의 값 1개가 다음 값에 영향을 미치는 AR(1) 모형

X(t) = {a*X_t-1+c}+u*Et 형태. a : 자기 상관계수

임의의 이전 시점 p부터 모델에 넣으면 AR(p) 모형이 된다.

AR(p) 과정이 정상이기 위한 조건은 특성함수의 근들이 단위원 밖에 있어야 함

AR과정의 ACF는 지수적, 싸인함수 형태로 감소하고, PACF는 차수(p)에 해당하는 차수 이후에는 0이 되는 성질을 갖고 있다.

04 AR, MA 모형

2. MA 모형

이동평균과정

위의 선형과정에서 유한개의 ψ_j 만이 0이 아닌 확률과정이 이동평균과정

가장 간단한 형태는 직전의 값 1개에서 발생한 오차가 다음 데이터에 영향을 준다고 가정한 MA(1) 모형

X(t) = {a*E_(t-1)+c} + u*E_t : MA(1) 모형

유한개의 항을 갖는 MA과정은 항상 정상과정. MA과정은 가역성 조건을 만족해야함

MA(q) 과정이 가역성을 만족하려면 특성함수의 근의 절대값이 1보다 커야함

가역성 조건을 부과하는 이유는 1. 하나의 ACF에 하나의 모형이 대응되도록 하기위해, 관측 불가능한 확률오차를 관측값들을 이용하여 표현하기 위해

MA과정의 PACF는 지수적, 싸인함수 형태로 감소하고, ACF는 차수(p)에 해당하는 차수 이후에는 0이 되는 성질을 갖고 있다.

04 AR, MA 모형

3. ARMA 모형

자기회귀이동평균과정

시계열 자료를 AR이나 MA 만으로 설명하려면 p,q의 차수가 너무 커질 수 있고, 추정해야할 모수가 많아 효율성이 떨어지고 해석도 쉽지 않다.

AR과 MA 부분을 동시에 포함하는 확률과정을 ARMA 과정이라고 한다.

ARMA(p,q) 의 이론적인 ACF는 시차 (q-p) 이후에는 지수적으로 감소하거나 소멸하는 싸인함수 형태이고,
PACF는 시차 (p-q) 이후에는 지수적으로 감소하거나 소멸하는 싸인함수 형태이다.

CONTENTS

05.ARIMA 모형

- 비정상시계열의 정상화
- ARIMA 과정

05 ARIMA 모형

1. 비정상시계열의 정상화

비정상시계열

시간이 흐름에 따라 시계열의 관측값이 추세를 갖고 증가하거나, 산포의 정도가 변하는 경우 혹은, 추세를 갖고 증가하지 않더라도 국지적으로는 그 수준이 같거나 산포의 정도가 동일하나 혹은 계절요인에 의해 시계열이 주기적으로 반복되는 경우 비정상적이라고 한다.

대표적인 특성 : 시계열의 수준이 시간대에 따라 다르다, 시계열이 추세를 갖는다, 시계열이 계절성을 보인다, 시계열의 분산이 시간대에 따라 변한다.

비정상성의 중요한 요인인 추세에는 결정적 추세와 확률적 추세가 있다. 추세가 결정적이고 동시에 영원히 지속된다면 결정적 추세라 부르고, 인접 자료들 간에 강한 양의 상관관계 때문에 어떤 추세가 있는 것처럼 보이는 것을 확률적 추세라고 한다.

결정적 추세는 시계열도를 그려보면 파악할 수 있지만, 확률적 추세는 시계열도 뿐만 아니라 SACF를 함께 그려보아야 한다.

시계열자료의 분석과 관련된 대부분의 이론들이 정상성을 가정하고 있으므로 시계열이 비정상을 보이는 경우에는 로그변환이나 차분을 통해 시계열이 정상성을 만족하게 해주어야 한다.

05 ARIMA 모형

1. 비정상시계열의 정상화

분산이 일정하지 않은 경우의 정상화

비정상 시계열의 한 가지 예로 시간의 흐름에 따라 변동폭이 커지는 경우를 들 수 있는데, 이처럼 시계열의 분산이 일정하지 않고 시간대에 따라 다르다면 분산 안정화 변환(로그변환, 제곱근변환)을 통해 시간의 흐름에 따라 변하지 않고 일정하도록 해주어야한다.

05 ARIMA 모형

1. 비정상시계열의 정상화

수준이 일정하지 않은 경우의 정상화

결정적 추세와는 달리 추세가 확률적인 경우에는 추세모형으로 해결할 수 없다. 이러한 경우 비정상시계열을 차분하여 기대값 등을 t에 의존하지 않게하여 정상시계열 형태로 바꾸어 주어야한다.

차분 : $\nabla Y_t = (1-B) = Y_t - Y_{t-1}$ 을 하여

일반적으로 차수가 d인 다항추세를 갖는 비정상시계열로부터 정상시계열을 얻기 위해서는 d차 차분을 해야한다.

계절차분 : $\nabla_s = (1-B^s) = I_t - I_{t-s}$

시계열도 확인 결과 추세를 갖고 있을때, 혹은 천천히 감소하는 SACF를 갖고 있을때 차분 사용

05 ARIMA 모형

2. ARIMA 과정

ARIMA

d차 차분된 $W_t = (1 - B)^d Z_t$가 수준이 μ인 정상 ARMA(p,q)과정을 따를 때 Z_t는 자기회귀누적이동평균과정
즉, ARIMA(p,d,q)과정을 따른다고 한다.

차분의 차수 d는 주로 시계열 그림과 SACF를 참고하여 결정하며, 시계열이 추세를 갖고 있는 경우 혹은 SACF가 서서히 감소하고 있는 경우에는 차분이 필요하다.

일반적으로 d=1 또는 d=2가 널리 쓰인다.

지나친 차분은 ACF를 복잡하게만들고 분산을 크게 만들 수 있으므로 적절하게 차분을 해야한다.

감사합니다

Kuggle