예측방법

목차

I. 개요

1. History

17 세기에 태양의 흑점 자료나 밀 가격 지수 변동을 나타내는 함수로 Sine, Cosine 곡선을 이용하였다. Yule(1926)은 ARMA 에 대한 개념을 제시하였고 Walker(1937)는 ARMA 모형을 제안하였다. ARMA 모형에 대한 추정은 Durbin(1960), 그리고 Box & Jenkins(1970)에 의해 이루어졌다. Holt(1957)는 지수 평활법(exponential smoothing)을, Winter(1960)는 계절성(seasonal) 지수 평활법을 제안하였다. 미국 Bureau of the Census 는 경기지수에 대한 계절 변동으로 1967 년 X-11 을 제안하였다. X-11 은 이동 평균 개념을 사용하므로 초기 관측치와 마지막 관측치를 사용할 수 없는 문제점을 안고 있어, 이에 대한 해결책으로 1975 년 캐나다는 X11-ARMA 방법을 제안하였다.

2. 시계열 데이터

시계열(time series) 데이터는 관측치가 시간적 순서를 가지게 된다. 일정 시점에 조사된 데이터는 횡단(cross-sectional) 자료라 한다. ○○전자 주가, △△기업 월별 매출액, 소매물가지수, 실업률, 환율 등이 시계열 자료이다. $\{Y_t; t = 1,2,...,T\}$

(분석목적)

가장 중요한 목적은 미래 값을 **예측** : trend analysis, smoothing, decomposition, ARMA model

시스템 시계열 데이터 이해와 특성 파악 : spectrum analysis, intervention analysis, transfer function analysis

(방법)

frequency domain : Fourier 분석에 기초, spectrum density function

time domain : 자기상관함수 이용, 관측값들의 시간적 변화 탐색

3. 시계열 데이터 4 가지 component $\{Y_t; t = 1,2,..,T\}$

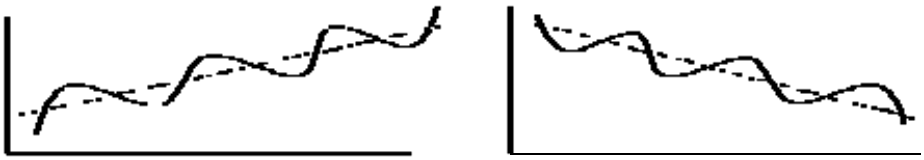
경향(Trend): 데이터가 증가(감소)하는 경향이 있는지 혹은 안정적인지 알 수 있다. 직선의 기울기가 있는가?

주기(cycle): 일정한 주기(진폭)마다 유사한 변동이 반복된다. (sine, cosine 곡선)

계절성(seasonality): 주별, 월별, 분기별, 년별 유사 패턴이 반복된다.

불규칙성(irregular): 일정한 패턴을 따르지 않는다.

$Y_t = Trend + Cycle + Seasonality + Irregular$



시계열 형태

①white noise process : 평균이 0 이고 분산이 σ^2 인 동일분포로부터 독립적으로(iid) 얻어진 시계열 데이터 $\{Y_t\}$을 백색 잡음(white noise) process 라 한다. 백색 잡음 데이터의 평균 수준을 μ 라 하면 이 시계열 데이터의 모형은 $Y_t = \mu + e_t$ 라 쓸 수 있다.

만약 $Y_0 = \mu$ 라 하면 $Y_t = Y_0 + e_1 + e_2 + ... + e_t$ 가 되며 $\{Y_t\}$ 을 random walk process 라 한다. $\{Y_t\}$ 는 동일한 분포를 가지며 서로 독립이라는 가정이다.

②stationary process : $F(y_{t_1}, y_{t_2}, ..., y_{t_n}) = F(y_{t_1+k}, y_{t_2+k}, ..., y_{t_n+k})$ 이면 시계열 데이터 $\{Y_t\}$를 strongly stationary process(강한 정상성)이라 한다. 일정한 기간의 종속변수 결합밀도함수는 동일한 분포를 가진다는 것을 의미한다.

다음 조건을 만족하는 시계열 데이터 $\{Y_t\}$는 weakly stationary process(약한 정상성)라 정의한다.

(1)평균이 일정하다. $E(Y_t) = \mu$

(2)분산이 존재하며 일정하다. $V(Y_t) = \gamma(0) < \infty$

(3)두 시점 사이의 자기 공분산(auto-correlation)은 시간의 차이에 의존한다.

$COV(Y_t, Y_{t-j}) = COV(Y_s, Y_{s-j}) = \gamma(j), for j \neq s$

Ⅱ. Time Plot

시계열 자료 $\{Y_t; t = 1,2,..,T\}$ 는 자료가 시간적 순서를 가지므로 Y 축은 $\{Y_t\}$ 값, X 축을 시간이므로 하여 산점도를 그릴 수 있다. 이를 시간도표(time plot)이라 한다.

- 시계열 자료의 구조를 파악하는데 도움이 되며 **시계열 분석의 시작**이다.

- 시계열 데이터 4 가지 성분 진단 가능 : 시각적 도움

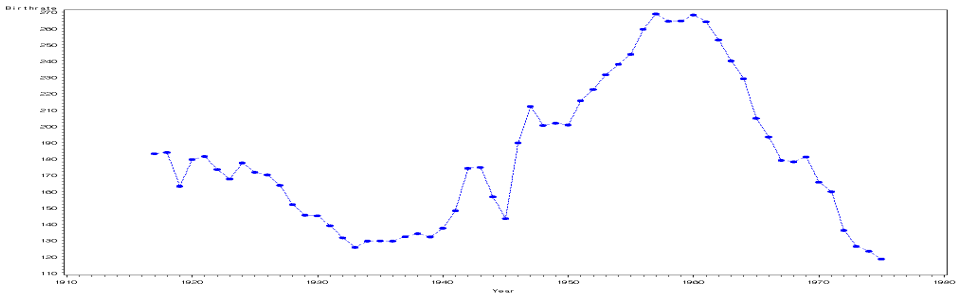
백색잡음 : $Y_t = e_t \sim N(0, \sigma^2)$

정상성 시계열 : $F(y_{t_1}, y_{t_2}, \ldots, y_{t_n}) = F(y_{t_1+k}, y_{t_2+k}, \ldots, y_{t_n+k})$

$E(Y_t) = \mu$, $V(Y_t) = \gamma(0) < \infty$ $COV(Y_t, Y_{t-j}) = COV(Y_s, Y_{s-j}) = \gamma(j), for j \neq s$

- econometrics (계량 경제) : 종속변수의 등분산 가정 체크

Example data http://lib.stat.cmu.edu/DASL/Datafiles/Birthrates.html



LOAD_DATA

2013 년 10 월 1 일부터 3 개월간 전력소비량(일별) 측정한 것이다. 최대기온, 바람세기, 일조량, 휴일여부도 조사하였다.

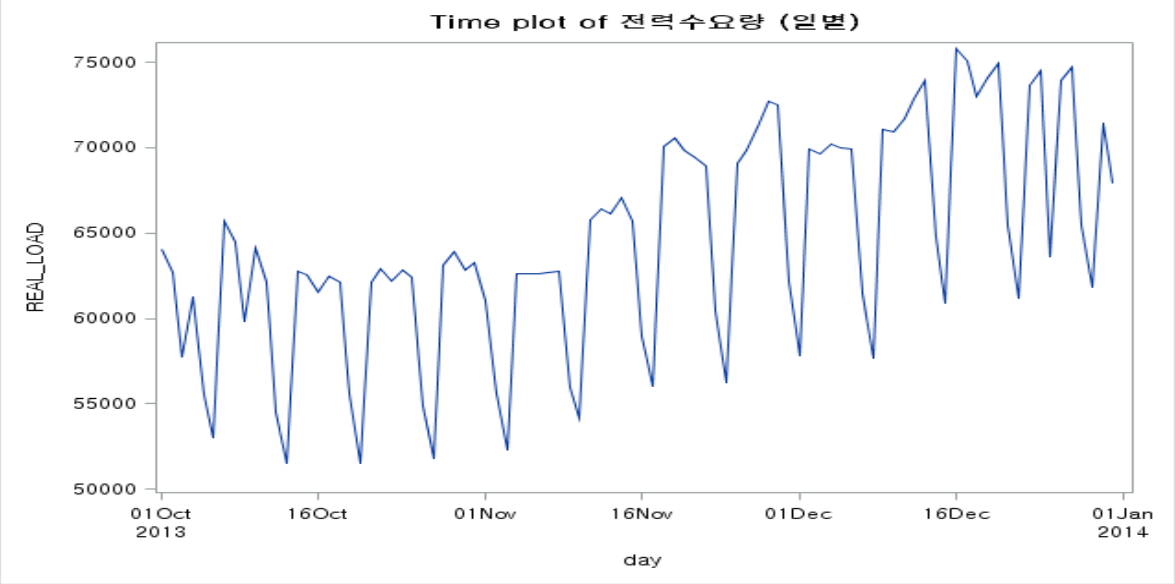
```
data load0;
   set load;
   format yr 4.;format mm 2.;format dd 2.;
   format day date9.;
   yr=substr(compress(load_date),1,4);
   mm=substr(compress(load_date),5,2);
   dd=substr(compress(load_date),7,2);
   day=mdy(mm,dd,yr);
   week=weekday(day);
   QTR=QTR(day);
   month=month(day);
   log_y=log(real_load);
   sqrt_y=sqrt(real_load);
   drop mm dd;
run;
```

compress() – 문자 데이터 공백 없애기

substr(문자변수,s,l) – 문자를 s 부터 시작 l 길이만큼 가져오기

mdy() - 월 2 자리, 날짜 2 자리, 연도 2 자리

weekday() – 요일, 1=일요일, 2=월요일, …



직선으로 증가하는 경향 (trend) / 1 주 주기 계절성 (seasonality)

In R – ICECREAM.csv – 아이스크림 판매량, 가격, 소득, 온도 (주별 데이터)

제목 없음 - R 편집기

```
ds.br=read.csv("data name", header=T)
attch(ds.vr)
plot.ts(time, y)
```

판매량	가격	주별소득	평균온도
0.386	0.27	78	41
0.374	0.282	79	56
0.393	0.277	81	63
0.425	0.28	80	68
0.406	0.272	76	69
0.344	0.262	78	65
0.327	0.275	82	61

Ⅲ. Moving average 이동평균법

자신의 m 개 관측치 평균으로 시계열 자료 $\{Y_t\}$의 패턴 인식

가중치는 1/m 으로 동일하다.

이를 이용하여 미래 값 $\{Y_{t+1}\}$ 예측한다.

$$MA = \frac{\sum \text{최근 } m\text{개자료}}{m}$$

$$\hat{Y}_{t+1} = MA_{t,m} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-m+1}}{m}$$

(다음 1 기만 예측 가능)

M 의 결정

일반적으로 주기를 m 으로 놓는다.

주가의 경우 5 일 (단기), 20 일, 60 일(중기), 120 일(장기), … 이동평균을 주로 사용한다

이동평균법 특징

m 이 클수록 주기의 영향은 없어지고 직선에 가까워짐, Trend(경향)을 보는데 활용

작은 m 은 단기 예측, 큰 주기 m 은 장기 예측에 사용

주가 예측에 가장 많이 이용, 그러나 예측보다는 (실제 예측 가능은 다음 1 기) 추세분석에 가까움

```
ts.hj=ts(ds.olympic[2:2], start=c(2001,3), frequency=4)
plot(ts.hj,type="o",col="blue",lty="dashed")
```

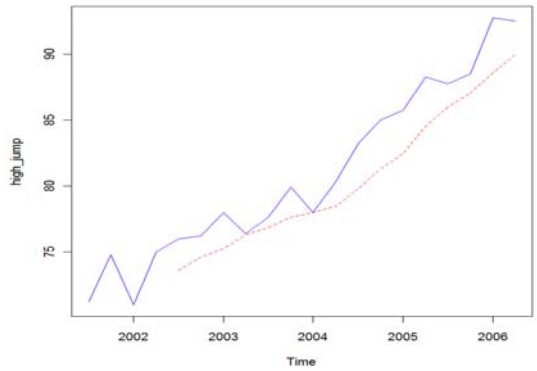
(시계열 데이터 변형)

```
fit.hj=filter(ts.hj,sides=1,rep(0.2,5))
plot(ts.hj,col="blue")
lines(fit.hj,col="red",lty="dashed")
```

주기 5 의 이동평균법, sides=1 은 이동평균법, sides=2 는 lag 0 을 중심으로 한 양측 이동평균법을 의미한다.

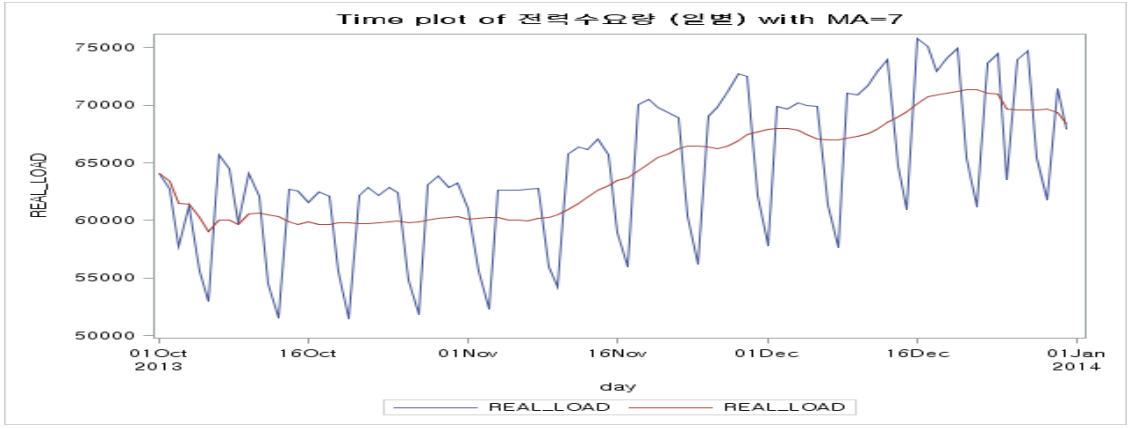
$$fjj(t) = \tfrac{1}{8}\,jj(t-2) + \tfrac{1}{4}\,jj(t-1) + \tfrac{1}{4}\,jj(t) + \tfrac{1}{4}\,jj(t+1) + \tfrac{1}{8}\,jj(t+2)$$

```
library(TTR)
SMA(ts.hj,n=5)
```




```
PROC EXPAND data=load0 OUT=MA;
    CONVERT real_load = MA7 / TRANSFORM=( MOVAVE 7 );
RUN;

title "Time plot of 전력수요량 (일별) with MA=7";
proc sgplot data=MA;
   series x=day y=real_load;
   series x=day y=ma7;
run;
```

주기 m=7



™ 우리나라 2005 년 1 월부터 월 별 실업률 1 년치를 찾아 3, 6, 12, 이동평균법 추정치를 하나의 그래프에 그리시오.

Ⅳ. 지수평활법 개요

o 모든 관측치에 동일한 가중치를 부여하는 이동평균법과는 달리 최근 관측치에 높은 가중치, 멀어질수록 지수적으로 가중치 값 감소

o 이동평균법(동일한 가중치로 평활하여 계절성분, 불규칙성 제거하여 추세 명확하게)은 분해법에서 계절조정을 하는데 주로 사용하며, 지수평활법은 예측에 사용

1. Simple Exponentially Smoothing 단순지수 평활법

o 단순지수평활법은 경향이나 계절성이 없을 때 사용한다. (가중평균 weighted mean)

o 평활 가중치 값의 설정이 다소 주관적이나, 계산이 간편함

상수모형

$Y_t = \beta_0 + \varepsilon_t$: 사인 곡선, 시간 추세 없음

시간 변동 모형

$Y_t = \beta_{0,t} + \varepsilon_t$

Locally 동일한 평균을 가지나 globally 평균 차이 보임

예측치(추정치) 및 추정 오차

다음 차기 Y_{t+1}의 예측치는 S_t에 의해 추정된다.

$\hat{Y}_{t+1} = S_t = wY_t + (1-w)S_{t-1}$ ---①

(a) S_t는 t 시점에서 평활 된 값이고 가중치 $0 < w < 1$

(b) 이전 값이 계속 필요한 것이 아니라 최근 값과 평활 값 만으로 예측이 갱신

(c) 추정오차 : $e_t = Y_t - \hat{Y}_t$

(d) 다음 차기 하나만 예측 가능

식 ①을 시계열 데이터 $\{Y_t\}$ 표시하면 다음과 같으므로 가중치가 지수적으로 감소하여 이를 지수평활법이라 한다.

$\hat{Y}_{t+1} = S_t = wY_t + w(1-w)Y_{t-1} + w(1-w)^2 Y_{t-2} + \ldots$ --- 가중치의 합은 $\sum w_i = 1$이다.

단순지수 평활 통계량 S_t 활용

1) Y_{t+1} 예측치

2) $\beta_{0,t}$ 의 추정치

3) Y_{t+1} 예측치인 S_t 의 신뢰구간은 가중최소제곱법의 특수한 경우가 지수평활법 예측이므로

초기치 평활값 선택

초기평활 값 $S_0 = \frac{\sum_{i=1}^{T} y_i}{T}$ 이고 일반적으로 $T = 6$ 혹은 $T = n/2$ 을 사용한다.

가중치 결정

이제 가중치를 결정하는 문제를 생각해 보자. 일반적으로 지수평활법은 현재에 가까운 관측치에 높은 가중치를 주기 위하여 0.05 에서 0.3 사이의 값을 준다. (다른 측면에서 보면 μ 가 시간에 따른 변화가 느리기 때문이다) 그럼 어떤 값이 가장 적절할까?

(가중치 범위)

클수록 최근 관측치 영향이 크다.

일반적으로 0.05 와 0.3 사이의 값

SAS 의 default 값: 1-0.8^(1/trend) Montgomery and Johnson (1976)

가중치 선택 : 모형 적합 정도를 나타내는 통계량을 이용하여 trial and error 방법으로 어떤 가중치가 좋은가를 판단하는 기준은 많으나 가장 많이 사용되는 것은 다음과 같다.

최적 ARIMA : 데이터에 ARIMA(0, 1, 1) 모형을 적합시켜 계산하는 기본 가중치를 사용

주관적 : 가중치 값이 크면 최근 관측값 반영이 크므로 예측 변동이 심하며, 가중치 값이 작으면 예측 변동이 완만하다. 이동평균법 $M = \frac{2-w}{w}$ (m=7 인 경우 0.25 가 적절)

시계열 모형 적합도

관측치 Y_t 와 예측치 $\hat{Y}_t$ 차이로 측정, 작을수록 적합 정도 높음

-MAPE (Mean Absolute Percent Error): 평균 절대 퍼센트 오차

-MAD (Mean Absolute Deviation): 평균 절대편차

-MSD (Mean Squared Deviation): 평균 제곱오차

-SSE(Sum of Squared Error): 오차 제곱 합

-MSE(Mean Square prediction Error): 평균 오차 자승

$$\text{SSE} = \sum (Y_t - \hat{Y}_t)^2$$

$$\text{MSE} = \frac{\sum (Y_t - \hat{Y}_t)^2}{T}$$

$$\text{MAPE} = \frac{\sum_{t=1}^{T} |(y_t - \hat{y}_t)/y_t|}{T} \times 100$$

$$\text{MAD} = \frac{\sum_{t=1}^{T} |y_t - \hat{y}_t|}{T}$$

$$\text{MSD} = \frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{T}$$

🖫LOAD_DATA

2013 년 10 월 1 일부터 3 개월간 전력소비량(일별) 측정한 것이다. 이를 이용하여 향후 1 주일 전력소비량을 예측하시오.

```
title '단순지수평활법';
proc forecast data=load0 lead=7 method=expo trend=1
out=pred outest=est outfull;
var real_load;
id day;
run;
proc print data=est;run;
proc print data=pred;run;
```

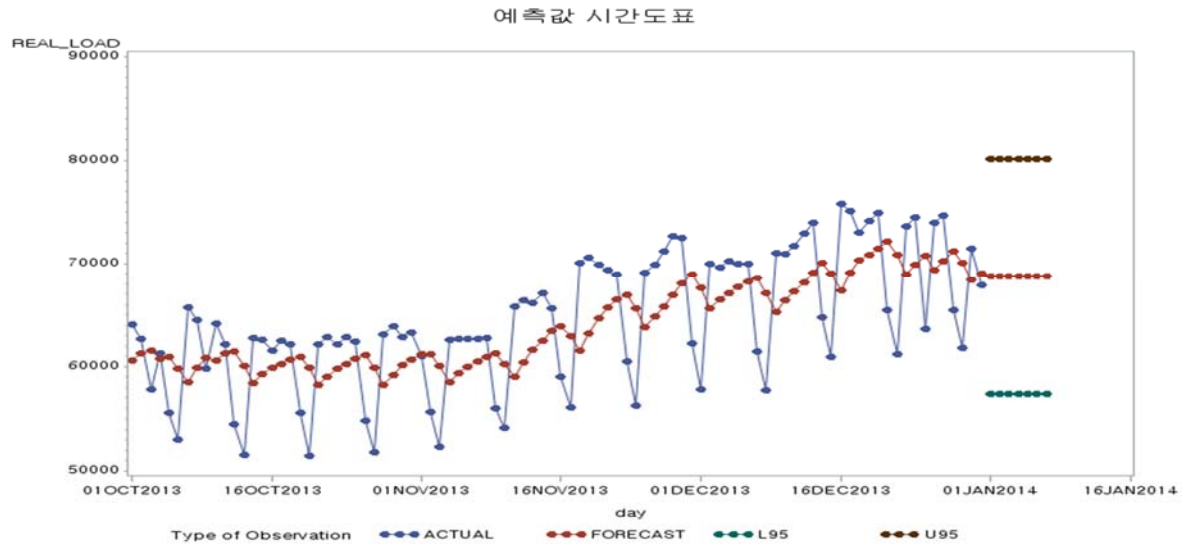
- lead=7 - 미래 7 개 관측값을 예측 / trend=1 - 단순 지수평활법
- out=pred - 예측결과를 저장하는 SAS 이름 - outfull 의 의미는 현재 주기 관측값도 예측하라는 의미 / outset=est - 추정 내용, 모형 적합도 관련 내용 저장
- weight 옵션을 사용하지 않는 경우 $1 - 0.8^{1/trend}$ 사용
- 단순지수 평활법은 lead=1 까지만 예측 가능, 그 이후에는 lead=1 의 관측치가 없음

```
title '예측값 시간도표';
proc gplot data=pred;
   plot real_load * day = _type_ ;
   symbol1 i=join v=dot; /* for _type_=ACTUAL */
   symbol2 i=join v=circle; /* for _type_=FORECAST */
run;
```

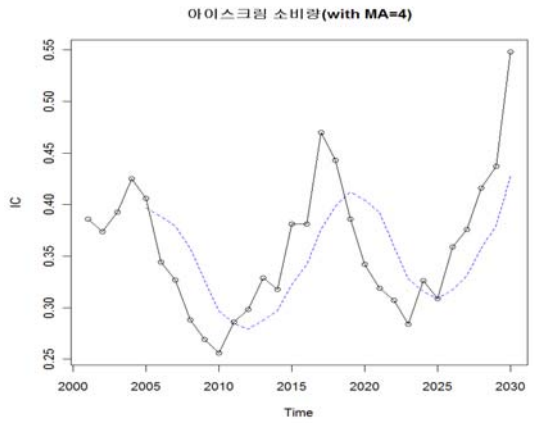
- symbol 문장은 시간 도표의 점들의 속성 지정, i=interpolate (보간법) 옵션으로 join 은 직선 연결이고 곡선 연결은 spline, 연결을 원치 않으면 none
- v=value 점에 대한 것으로 circle 은 동그라미, dot 는 점

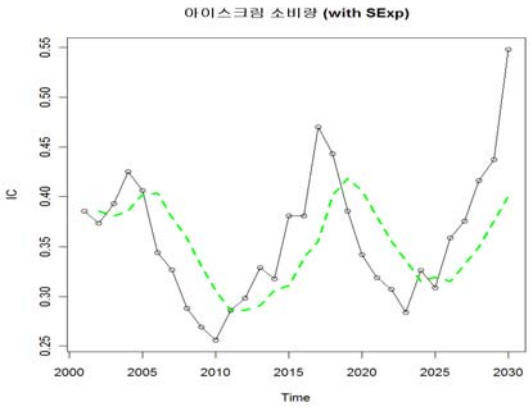
TYPE	day	REAL_LOAD
N	31DEC2013	92
NRESID	31DEC2013	92
DF	31DEC2013	91
WEIGHT	31DEC2013	0.2
S1	31DEC2013	68768.316
SIGMA	31DEC2013	5303.6848
CONSTANT	31DEC2013	68768.316
SST	31DEC2013	3.6617E9
SSE	31DEC2013	2.55975E9
MSE	31DEC2013	28129072
RMSE	31DEC2013	5303.6848
MAPE	31DEC2013	7.5562138
MPE	31DEC2013	0.0571277
MAE	31DEC2013	4752.8674
ME	31DEC2013	445.73726
RSQUARE	31DEC2013	0.3009407

day	_TYPE_	_LEAD_	REAL_LOAD
01JAN2014	FORECAST	1	68768.32
02JAN2014	FORECAST	2	68768.32
03JAN2014	FORECAST	3	68768.32
04JAN2014	FORECAST	4	68768.32
05JAN2014	FORECAST	5	68768.32
06JAN2014	FORECAST	6	68768.32
07JAN2014	FORECAST	7	68768.32




```
#시간도표 그리기
ds=read.csv("ic.csv",header=T)

ts=ts(ds[2:2], start=c(2001,1), frequency=1)

#이동평균법
fit=filter(ts,sides=1,rep(0.2,5))
plot(ts, type="o",main="아이스크림 소비량(with MA=4)")
lines(fit,col="blue",lty="dashed")

library(TTR)
fit2=SMA(ts,n=5)
plot(ts, type="o",main="아이스크림 소비량")
lines(fit2,col="red",lty="dashed")
```





Holt Winters Exponential Smoothing

Simple Exponential : $S_t = \alpha Y_t + (1-\alpha)S_{t-1}$

Double Exponential : trend only $G_t = \beta(S_t - S_{t-1}) + (1-\beta)G_{t-1}$

Triple Exponential : trend and seasonality $R_t = \gamma(Y_t - R_t) + (1-\gamma)R_{t-L}$

- **승법모형** multiplicative 사용 : 계절성이 추세와 함께 변동이 있을 때
- **가법모형** additive 사용 : 계절성이 추세에 관계없이 일정할 때

```
#단순지수평활법
fit3=HoltWinters(ts, gamma=FALSE, beta=FALSE, alpha=0.4)
plot(ts, type="o",main="아이스크림 소비량 (with SExp)")
lines(fitted(fit3)[,1], col="green", lty="dashed", lwd=3)
```

우리나라 가전제품 (anything) 월별 판매량을 데이터(2005 년 1 월~) 찾아 이동평균법, 단순지수평활법으로 다음 월의 판매량을 예측하시오. 그리고 향후 판매량의 추이를 해석하시오.

4. Double Exponentially Smoothing

시간 추세 모형 $Y_t = \beta_{0,t} + \beta_{1,t}t + \varepsilon_t$ ---(a)

Locally 동일한 평균을 가지나 globally 평균 차이 보임

β_1은 추세 기울기

단순지수 평활법 적용

$$E(\hat{Y}_{t+1}) = E(S_t) = w\sum_{j=0}(1-w)^j Y_{t-j} = w\sum_{j=0}(1-w)^j(\beta_{0,t} + \beta_{1,t}(t-j))$$

$$= \beta_{0,t} + \beta_{1,t}(t+1) - \frac{1}{w}\beta_{1,t}$$

불편 추정량

이중지수 평활법

단순지수평활값을 평활하여 얻음

$$S_t^{[1]} = wY_t + (1-w)S_{t-1}^{[1]}$$

$$S_t[2] = wS_t^{[1]} + (1-w)S_{t-1}^{[2]} = w\sum_{0}^{t-1}(1-w)^j S_{t-j}^{(1)} + (1-w)^n S_0^{(2)}$$ (일모수 이중지수 평활법)

Holt 제안

$$S_t^{[1]} = w_1 Y_t + (1-w_1)S_{t-1}^{[1]}$$

$$S_t[2] = w_2 S_t^{(1)} + (1-w_2)S_{t-1}^{[2]}$$

추정치

$$\hat{Y}_t = 2S_t^{(1)} - S_t^{(2)}$$

L 기 이후 예측치

$$\hat{Y}_{t+L} = (2 + \frac{w}{1-w}L)S_t^{(1)} - (1 + \frac{w}{1-w}L)S_t^{(2)}$$

초기치 $S_0^{(1)}, S_0^{(2)}$ 선택

수식 (a) OLS 추정치 => $\hat{\beta}_{0,0}, \hat{\beta}_{1,0}$ 을 이용하여

$$S_0^{(1)} = \hat{\beta}_{0,0} - \frac{1-w}{w}\hat{\beta}_{1,0}$$
$$S_0^{(2)} = \hat{\beta}_{0,0} - 2\frac{1-w}{w}\hat{\beta}_{1,0}$$

가중치 w 선택

Brown (1962) 0.03~0.16 권장- 일반적으로 이 범위를 벗어나는 값들이 선택된다.

- 이중 지수 평활은 각 기간에서의 수준 성분과 추세 성분을 사용합니다. 또한 두 개의 가중치 또는 평활화 모수를 사용하여 각 기간의 성분을 업데이트합니다. 이중 지수 평활 방정식은 다음과 같습니다.

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1})$$
$$T_t = \gamma[L_t - L_{t-1}] + (1-\gamma)T_{t-1}$$
$$\hat{Y}_t = L_{t-1} + T_{t-1}$$

- 여기서 L_t는 시간 t 에서의 수준 성분이고 α는 수준 성분에 대한 가중치입니다. T_t는 시간 t 에서의 추세 성분이고 γ는 추세 성분에 대한 가중치입니다. Y_t는 시간 t 에서의 데이터 값이고 $\hat{Y}_t$는 시간 t 에서의 적합치 또는 한 단계 전 예측값입니다.

최적 ARIMA 가중치

Minitab 에서는 오차 제곱의 합을 최소화하기 위해 데이터에 ARIMA(0,2,2) 모형을 적합시킵니다. 추세 성분과 수준 성분이 후방 예측을 통해 초기화됩니다.

```
title '이중지수평활법';
proc forecast data=load0 lead=7 method=expo trend=2
out=esm_pred2 outest=est2 outfull;
var real_load;
id day;
run;
proc print data=est2;run;

title '예측값 시간도표';
proc gplot data=esm_pred2;
   plot real_load * day = _type_ ;
   symbol1 i=join v=dot; /* for _type_=ACTUAL */
   symbol2 i=join v=circle; /* for _type_=FORECAST */
run;
```

Obs	_TYPE_	day	REAL_LOAD
1	N	31DEC2013	92
2	NRESID	31DEC2013	92
3	DF	31DEC2013	90
4	WEIGHT	31DEC2013	0.1055728
5	S1	31DEC2013	69102.732
6	S2	31DEC2013	69039.063
7	SIGMA	31DEC2013	5376.2885
8	CONSTANT	31DEC2013	69166.4
9	LINEAR	31DEC2013	7.5150681
10	SST	31DEC2013	3.6617E9
11	SSE	31DEC2013	2.6014E9
12	MSE	31DEC2013	28904478
13	RMSE	31DEC2013	5376.2885
14	MAPE	31DEC2013	7.4616206
15	MPE	31DEC2013	-0.546167
16	MAE	31DEC2013	4653.2129
17	ME	31DEC2013	33.567325
18	RSQUARE	31DEC2013	0.2895641

```
#단순/이중 지수평활법
ds=read.csv("ic.csv",header=T)
ts=ts(ds[2:2], start=c(2001,1), frequency=1)
fit.s=HoltWinters(ts, gamma=FALSE, beta=FALSE)
fit.d=HoltWinters(ts, gamma=FALSE)
plot(ts, type="o",main="아이스크림 소비량(단순/이중)",
     xlim=c(2001,2035),ylim=c(0,1))
lines(fitted(fit.s)[,1], col="blue", lty="dashed")
lines(fitted(fit.d)[,1], col="red", lty="dashed")
lines(predict(fit.d,n.ahead=4),col="green")
fit.s;fit.d
fit.s$SSE;fit.d$SSE
predict(fit.d,n.ahead=4)
```

Holt Winters Exponential Smoothing

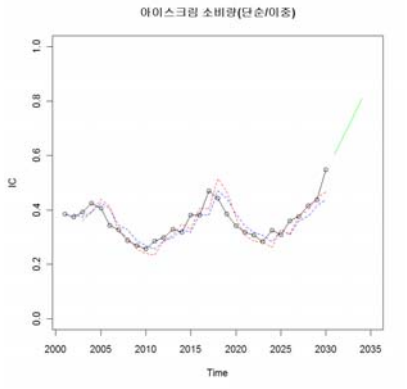
Simple Exponential : $\hat{Y}_{t+1} = S_t = \alpha Y_t + (1-\alpha)S_{t-1}$

Double Exponential : trend only $\hat{Y}_{t+1} = G_t = \beta(S_t - S_{t-1}) + (1-\beta)G_{t-1}$

Triple Exponential : trend and seasonality $\hat{Y}_{t+1} = R_t = \gamma(Y_t - R_t) + (1-\gamma)R_{t-L}$

```
Call:
 HoltWinters(x = ts, gamma = FALSE)

Smoothing parameters:
 alpha:  0.8876731
 beta :  0.5495471
 gamma:  FALSE

Coefficients:
        [,1]
a 0.53868503
b 0.06759742
```

```
> fit.s$SSE;fit.d$SSE
[1] 0.0484111
[1] 0.04762695
```




```
> predict(fit.d,n.ahead=4)
Time Series:
Start = 2031
End = 2034
Frequency = 1
          fit
[1,] 0.6062824
[2,] 0.6738799
[3,] 0.7414773
[4,] 0.8090747
```

™ 여러분 우리나라 가전제품 (anything) 월별 판매량을 데이터(2005 년 1 월~)를 이중지수평활법으로 다음 6 개월 판매량을 예측하고 해석하시오.

5. Triple Exponentially Smoothing

시간 추세 모형

$$Y_t = \beta_{0,t} + \beta_{1,t}t + \beta_{2,t}t^2/2 + \varepsilon_t \quad \text{--- (b)}$$

β_1, β_2 은 추세 기울기

$\beta_0, \beta_1, \beta_2$ 의 추정치는 식 (b)의 OLS 추정치

삼중지수 평활법

단순지수평활값을 평활하여 얻음

$$S_t^{[1]} = wY_t + (1-w)S_{t-1}^{[1]}$$

- $S_t^{[2]} = wS_t^{[1]} + (1-w)S_{t-1}^{[2]}$
- $S_t^{[3]} = wS_t^{[2]} + (1-w)S_{t-1}^{[3]}$
-
- 초기치 $S_0^{(1)}, S_0^{(2)}, S_0^{(3)}$ 선택 – 수식 (b) OLS 추정치 $\hat{\beta}_{0,0}, \hat{\beta}_{1,0}, \hat{\beta}_{2,0}$을 이용하여 얻음
- 추정치와 L 기 이후 예측치 – 이전과 동일한 절차
- 가중치 w 선택 : Brown (1962) 0.02~0.11 권장

```
title '이중지수평활법';
proc forecast data=load0 lead=7 method=expo trend=2
out=esm_pred2 outest=est2 outfull;
var real_load;
id day;
run;
proc print data=est2;run;

title '예측값 시간도표';
proc gplot data=esm_pred2;
   plot real_load * day = _type_ ;
   symbol1 i=join v=dot; /* for _type_=ACTUAL */
   symbol2 i=join v=circle; /* for _type_=FORECAST */
run;
```

Obs	_TYPE_	day	REAL_LOAD
1	N	31DEC2013	92
2	NRESID	31DEC2013	92
3	DF	31DEC2013	89
4	WEIGHT	31DEC2013	0.0716822
5	S1	31DEC2013	69210.069
6	S2	31DEC2013	70933.469
7	S3	31DEC2013	81053.927
8	SIGMA	31DEC2013	45479.812
9	CONSTANT	31DEC2013	75883.727
10	LINEAR	31DEC2013	1188.7545
11	QUAD	31DEC2013	50.067612
12	SST	31DEC2013	3.6617E9
13	SSE	31DEC2013	1.8409E11
14	MSE	31DEC2013	2.06841E9
15	RMSE	31DEC2013	45479.812
16	MAPE	31DEC2013	61.211316
17	MPE	31DEC2013	-60.8268
18	MAE	31DEC2013	37989.888
19	ME	31DEC2013	-37745.54
20	RSQUARE	31DEC2013	-49.27413

trend 가 직선의 경향이 있으므로 이중 지수 평활법의 모형 적합도 (MAPE, MSE)가 가장 높음

삼중지수 평활법 함수는 없음. 그러므로 차분한 데이터에 이중지수 평활법을 적용하면 된다.

```
fit.t=HoltWinters(diff(ts), gamma=FALSE)
fit.d$SSE;fit.t$SSE
predict(fit.t,n.ahead=4)
```

```
> fit.d$SSE;fit.t$SSE
[1] 0.04762695
[1] 0.06415211
> predict(fit.t,n.ahead=4)
           Qtr1       Qtr2       Qtr3       Qtr4
2008                         0.09344845 0.10630071
2009 0.11915297 0.13200523
```

```
> ts
      Qtr1  Qtr2  Qtr3  Qtr4
2001 0.386 0.374 0.393 0.425
2002 0.406 0.344 0.327 0.288
2003 0.269 0.256 0.286 0.298
2004 0.329 0.318 0.381 0.381
2005 0.470 0.443 0.386 0.342
2006 0.319 0.307 0.284 0.326
2007 0.309 0.359 0.376 0.416
2008 0.437 0.548
```

6. Winters 계절 지수 모형

o 추세와 계절성이 있는 경우 활용

o 분산이 시간의 흐름에 따라 일정하면 => 가법계절모형 additive seasonal model

o 분산이 시간의 흐름에 따라 변동하면 => 승법계절모형 multiplicative seasonal model

가법계절모형 – addwinters 모형

$Y_t = \beta_0 + \beta_1 t + S_t + \varepsilon_t$

o β_0 : 고정 성분

o β_1 : 추세 선형 기울기

o S_t : 가법 추세 성분

(전체 평활) $\bar{R}_t = \alpha(y_t - \bar{S}_{t-L}) + (1-\alpha) * (\bar{R}_{t-1} + \bar{G}_{t-1})$, *) L 은 계절 주기임

(추세 요인 평활) $\bar{G}_t = \beta * (\bar{S}_t - \bar{S}_{t-1}) + (1-\beta) * \bar{G}_{t-1}$

(계절성분 평활) $\bar{S}_t = \gamma * (y_t - \bar{S}_t) + (1-\gamma) * \bar{S}_{t-L}$

승법계절모형 – winters 방법

$Y_t = (\beta_0 + \beta_1 t) S_t + \varepsilon_t$

(전체 평활) $\bar{R}_t = \alpha \frac{y_t}{\bar{S}_{t-L}} + (1-\alpha) * (\bar{R}_{t-1} + \bar{G}_{t-1})$

(추세 요인 평활) $\bar{G}_t = \beta * (\bar{S}_t - \bar{S}_{t-1}) + (1-\beta) * \bar{G}_{t-1}$

(계절성분 평활) $\bar{S}_t = \gamma * (y_t / \bar{S}_t) + (1-\gamma) * \bar{S}_{t-L}$

```
title 'ADD-winters 지수평활법';
proc forecast data=load0 lead=7 method=addwinters trend=2 seasons=day
out=add_pred outest=add_est outfull;
var real_load;
id day;
run;
proc print data=add_est run;
```

- seansons 옵션 - 계절성에 대한 지정, DAY-주 주기, Month-년 주기, HOUR-일 주기, QTR - 분기 주기
- method 옵션 -가법모형은 "ADDWINTERS", 승법모형은 "WINTERS" 지정함

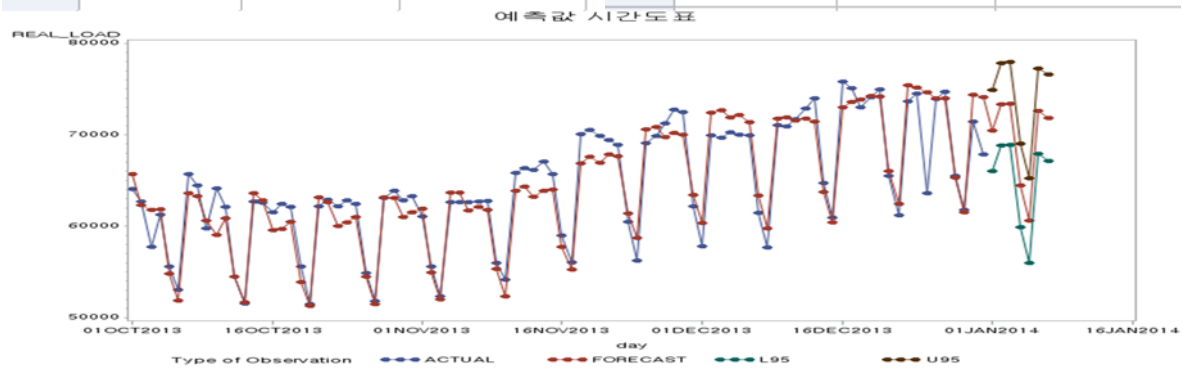
- seasons 옵션 설정하지 않으면 method=addwinters (혹은 winters) 설정하여도 일반 지수평활법 추정함.
- 승법모형과 가법모형 지정은 시간도표를 보고 관측값의 분산이 커지는 (⇔ 주기의 폭이 넓어지거나 좁아지면 분산이 일정하지 않음) 경향을 보이면 승법모형을 적용함.
- 실제 추정에서는 두 방법 모두 적용해 보고 모형 적합도가 높은 방법 적용

(**addwinters**)

19	MSE	31DEC2013	5186420.2
20	RMSE	31DEC2013	2277.3713
21	MAPE	31DEC2013	2.4321116
22	MPE	31DEC2013	0.2751887
23	MAE	31DEC2013	1579.1081

(winters)

19	MSE	31DEC2013	5201019.5
20	RMSE	31DEC2013	2280.5744
21	MAPE	31DEC2013	2.4018322
22	MPE	31DEC2013	0.2922486
23	MAE	31DEC2013	1564.1756




```
fit.wa=HoltWinters(ts, seasonal = c("additive"))
fit.wa$SSE
plot(fit.wa)
fit.wm=HoltWinters(ts, seasonal = c("multiplicative"))
fit.wm$SSE
plot(fit.wm)
```

™ 여러분 우리나라 가전제품 (anything) 월별 판매량을 데이터(2005 년 1 월~)를 Winters 평활법으로 향후 6 개월 판매량을 예측하고 해석하시오.

V. ARMA

1. 개요

o George Box, Gwilym Jenkins 제안한 시계열 모형

o 시계열 데이터는 (**Trend + Cycle + Seasonality + Irregular**) 성분이 있에 (1)설명변수 설정이 용이하지 못하거나 (2) $\{Y_t\}$에 대한 예측을 위하여(시계열 데이터 분석의 주요 목적) 설명변수에 대한 예측치(X_t)가 있어야 하는 문제가 있고 (3)독립성 가정을 만족하지 못해 이 문제를 해결하는 어려움이 있어 회귀모형에 의한 분석보다는 관측치의 이전 관측치를 활용하는 방법이 제안

o ARIMA(Auto-Regressive Integrated Moving-Average) 모형은 시계열 데이터 $\{Y_t\}$의 과거치(previous observation) $\{Y_{t-1}, Y_{t-2}, \ldots\}$가 설명변수인 AR과 과거 관측치가 설명하지 못하는 부분에 해당되는 오차항($e_{t-1}, e_{t-2}, \ldots$)들이 설명변수인 MA, 차분을 나타내는 integrate의 합성어이다.

AR 모형은 아래 가설에 의해 제안되었다.

- ○과거의 패턴이 지속된다면 시계열 데이터 관측치 Y_t는 과거 관측치 $Y_{t-1}, Y_{t-2}, Y_{t-p}, \ldots$에 의해 예측할 수 있을 것이다.
- ○어느 정도의 멀리 있는 과거 관측치까지 이용할 것인가? 그리고 멀어질수록 영향력을 줄어들 것이다. 이런 상황을 고려할 수 있는 가중치를 사용해야 하지 않을까?

Backshift Notation $B(Y_t) = Y_{t-1}$, $B^2(Y_t) = Y_{t-2}$, …, $B^p(Y_t) = Y_{t-p}$

2. ARMA 모형 적합절차

시계열 데이터 수집

모형 식별 identification

- 데이터 안정성 진단
- 상관함수 활용, p, q, d 결정

모형 추정 estimation : 계수 추정

모형 진단 diagnosis : 계수의 유의성 및 잔차의 백색잡음

예측모형 활용

3. Process

white noise process

평균이 0 이고 분산이 σ^2 인 동일분포로부터 독립적으로(iid) 얻어진 시계열 데이터 $\{Y_t\}$을 백색 잡음(white noise) process 라 한다. 백색 잡음 데이터의 평균 수준을 μ 라 하면 이 시계열 데이터의 모형은 $Y_t = \mu + e_t$ 라 쓸 수 있다.

만약 $Y_0 = \mu$ 라 하면 $Y_t = Y_0 + e_1 + e_2 + ... + e_t$ 가 되며 $\{Y_t\}$을 random walk process 라 한다. $\{Y_t\}$는 동일한 분포를 가지며 서로 독립이라는 가정이다.

whitenoise.test {normwhn.test} => whitenoise.test(x)

stationary process (정상성)

$F(y_{t_1}, y_{t_2}, ..., y_{t_n}) = F(y_{t_1+k}, y_{t_2+k}, ..., y_{t_n+k})$ 이면 시계열 데이터 $\{Y_t\}$를 strongly stationary process(강한 정상성)이라 한다. 일정한 기간의 종속변수 결합밀도함수는 동일한 분포

시계열 데이터 $\{Y_t\}$의 weakly stationary process(약한 정상성)라 정의한다.

- (1)평균이 일정하다. $E(Y_t) = \mu$
- (2)분산이 존재하며 일정하다. $V(Y_t) = \gamma(0) < \infty$
- (3)두 시점 사이의 자기 공분산(auto-correlation)은 시간의 차이에 의존한다.
- $COV(Y_t, Y_{t-j}) = COV(Y_s, Y_{s-j}) = \gamma(j), for j \neq s$

정상적 stationary 확률 모형(시계열 데이터 $\{Y_t\}$는 확률 변수)의 대표적인 것이 AR, MA, ARMA 모형이다.

4. 상관함수 Correlation Function

자기상관함수 Auto Correlation Function (ACF)

자기상관함수(ACF)는 다음과 같이 정의한다.

- $\rho(j) = \frac{\gamma(j)}{\gamma(0)} = \frac{Cov(Y_t, Y_{t-j})}{VAR(Y_t)}$ 그러므로 $\rho(0) = 1$, $\rho(j) = \rho(-j)$

부분자기상관함수 Partial Auto Correlation Function (PACF)

o 두 변수 (X, Y)의 상관관계를 시간의 효과를 제거한 후 구한 순수 상관관계

$$\rho_{XY.Z} = \frac{E(X - E(X \mid Z))E(Y - E(Y \mid Z))}{\sqrt{E(X - E(X \mid Z))^2 E(Y - E(Y \mid Z))^2}}$$

- ⇔ Z->X 장차와 Z->Y 잔차의 상관계수

o 시계열 분석 : (Y_{t-1}, Y_{t-k+1})의 효과 제외한 (Y_t, Y_{t-k})의 순수 상관계수 ϕ_k을

부분자기상관계수, 즉 $\phi_k = Corr(Y_t^z, Y_{t-k}^z)$

$$\phi_1 = \rho(1), \phi_2 = \frac{\rho(1) - \rho(1)^2}{1 - \rho(1)^2}, \ldots, \quad \phi_k = \frac{\rho(k+1) - \sum \phi_{k,j}\rho(k+1-j)}{1 - \sum \phi_{k,j}\rho(j)}$$

역자기상관함수 Inverse Auto Correlation Function (IACF)

ARMA(p, q) 모형의 IACF 는 ARMA(q, p)의 ACF 이다.

그러므로 AR(p)의 IACF 는 MA(p)의 ACF 와 같고 MA(q)의 IACF 는 AR(q)의 ACF 와 같다.

5. AR(p) 모형

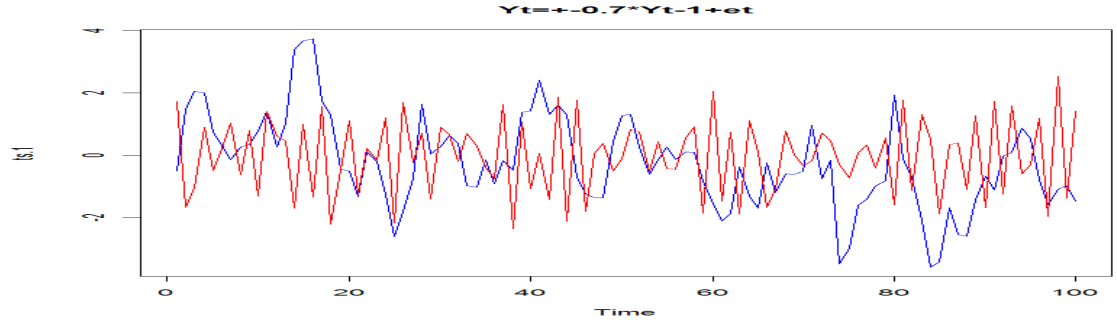
AR(1) 모형: $Y_t = a + \rho Y_{t-1} + e_t, \quad e_t \sim iid\ N(0, \sigma^2)$

o Markov process : $|\rho| < 1$ ⇔ stationary 프로세스

o 만약 시계열 데이터가 서로 독립이고 유한인 평균과 분산을 갖는 동일 분포를 따르면(iid) 이 데이는 white noise(백색 잡음)이라 한다. 만약 평균이 0, 분산이 σ^2인 정규분포를 따른다면 이를 Guassian white noise 라 한다. $\{Y_t\}$ 대신 $\{Z_t\} = \{Y_t - \mu\}$를 사용하기도 하는데 이는 평균을 0 으로 하기 위함이다.

o μ는 시계열 데이터의 총 평균(grand mean)에 해당된다.

```
#AR(1)
y0.1=c(0);y0.2=c(0);y1=c(0);y2=c(0) #초기 관측치

for (i in 1:100) {
y1[i]=0.7*y0.1+rnorm(1,0,1)
y2[i]=-0.7*y0.2+rnorm(1,0,1)
y0.1=y1[i];y0.2=y2[i]}   #관측치 100개 생성

ts.1=ts(y1, start=c(1,1), frequency=1)
ts.2=ts(y2, start=c(1,1), frequency=1)
plot(ts.1, col="blue", type="l",main="Yt=+-0.7*Yt-1+et")
lines(ts.2, col="red") #시간도표 그리기
```



평균 $E(Y_t) = a + \rho E(Y_{t-1})$ => $\mu = \frac{a}{1-\rho}$

분산 $V(Y_t) = \rho^2 \gamma(0) + \sigma^2$ => $V(Y_t) = \gamma(0) = \frac{\sigma^2}{1-\rho^2}$

자기상관함수

AR(1) 모형을 이를 다시 쓰면 다음과 같다. 즉 AR(1) 모형이더라도 과거의 흔적을 모두 모함하고 있다.

$$Y_t = \mu + e_t + \rho e_{t-1} + \rho^2 e_{t-2} + \rho^3 e_{t-3} + \ldots + \rho^{t-1} e_1 + \rho^t (Y_o - \mu)$$

그리고 $|\rho| < 1$ (stationary)이면, 자수적 감소 (

MA(∞) 모형: $$Y_t = \mu + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \beta_3 e_{t-3} + \ldots = \mu + \sum_{j=0}^{\infty} \beta_j e_{t-j}$$

$\gamma(j) = COV(Y_t, Y_{t-j}) = \rho^j \sigma^2 / (1-\rho^2)$ => (ACF) $\rho(k) = \rho^k$ 지수적으로 감소

부분자기상관함수

o 1 차 이후 회귀계수가 0 이므로 1 차 PACF 는 $\phi_1 = \rho$ 이고, 2 차부터 이후는 0 이다.

```
library(Hmisc)
rcorr(as.matrix(cbind(ts.1,lag(ts.1))))
rcorr(as.matrix(cbind(ts.2,lag(ts.2))))
```

```
> rcorr(as.matrix(cbind(ts.1,lag(ts.1))))
          ts.1 lag(ts.1)
ts.1      1.00      0.73
lag(ts.1) 0.73      1.00

> rcorr(as.matrix(cbind(ts.2,lag(ts.2))))
           ts.2 lag(ts.2)
ts.2       1.00     -0.68
lag(ts.2) -0.68      1.00
```

(추정) AR(1)

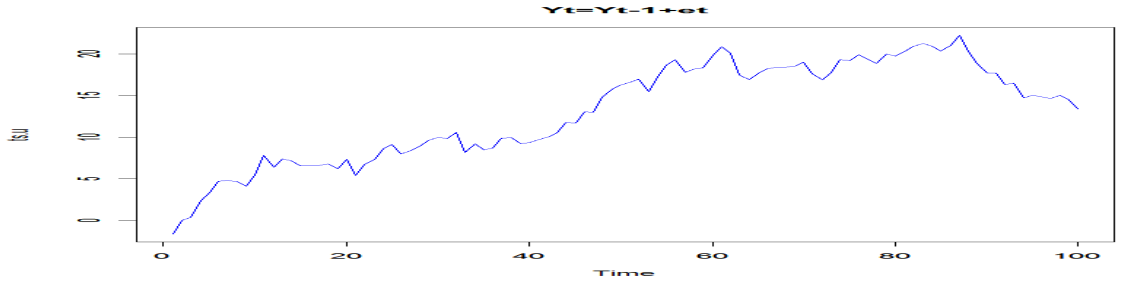
```
> arima(ts.1,order=c(1,0,0))

Call:
arima(x = ts.1, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.7231    -0.2805
s.e.  0.0678     0.3479
```

Unit-Root 검정

AR(1) 모형을 갖는 시계열 데이터의 경우 UNIT root 문제는 ($Y_t = \mu + \alpha Y_{t-1}, \alpha = 1$)임을 의미한다. Unit-root 갖는 데이터는 안정적이지 못하므로 모형 설정의 의미가 없다.

test 방법 : augmented Dickey-Fuller 검정 방법, Phillips-Perron 검정 방법 등이 있음

```
#AR(1) Unit root
u0=c(0);u=c(0) #초기 관측치

for (i in 1:100) {
u[i]=u0+rnorm(1,0,1)
u0=u[i]}    #관측치 100개 생성

ts.u=ts(u, start=c(1,1), frequency=1)
plot(ts.u, col="blue", type="l",main="Yt=Yt-1+et")
```




```
> arima(u,order=c(1,0,0))

Call:
arima(x = u, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.9918     7.9988
s.e.  0.0089     6.7675
```

```
library(fUnitRoots)
urppTest(y1)
urppTest(y2)

acf(y1)$acf
pacf(y1)$acf
```

(단일근 검정)

```
library(tseries)
pp.test(ts.u,alternative=c("explosive"))
```

```
        Phillips-Perron Unit Root Test

data:  ts.u
Dickey-Fuller Z(alpha) = -4.4565, Truncation lag parameter = 3,
p-value = 0.1417
alternative hypothesis: explosive
```

Stationarity (정상성)

AR 모형 $Y_t = u + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + e_t$ 은 $1 - \alpha_1 M - \alpha_2 M^2 - \ldots - \alpha_p M^p = 0$ 의 방정식을 만족하는 근들의 절대값이 모두 1 보다 클 경우 stationary 하다.

정상적인 AR(p) 모형은 MA(∞) 모형으로 변환할 수 있음을 의미

⇔ 정상적인 process 인 경우

- $\{Y_t\}$ 는 $e_t, e_{t-1}, e_{t-2}, \ldots$ 으로 표현할 수 있으며,
- $\{Y_t\}$ 에 대한 $e_t, e_{t-1}, e_{t-2}, \ldots$ 들의 영향은 시점이 멀어질수록 줄어든다.
- 그러므로 Y_{t+1} 에 대한 예측치를 구할 경우 $e_0 = 0$ 으로 사용해도 무방하다.

AR(p) 모형: $Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + e_t$, $e_t \sim iid\ N(0, \sigma^2)$

o 설명변수의 개수 p 개

o AR(p)도 MA(∞) 모형으로 쓸 수 있으므로 정상적인 AR(p)의 자기상관함수는 지수적으로 감소하며, 부분자기상관함수는 p 차 이후부터 0 이다.

자기상관함수

Stationary 시계열 데이터의 AR(p)의 ACF 는 AR(1)과 동일하게 지수적으로 감소한다.
자기상관함수 $\rho(k)$ 는 Yule-Walker 방정식에 의해 구한다. (complicated)

부분자기상관함수

o $\phi_{k=}\alpha_k$ for $k \le p$

o p 차부터 이후는 0 이다.

 AR(2) 모형 대하여,

stationary 진단, white noise 진단 whitenoise.test(x)

관측치 100 개 생성하고, 자기상관함수 3 차까지 추정하시오. 그리고 stationary 검정하시오.

모형	ACF (이론)	PACF (이론)		
		1 차	2 차	3 차
$Y_t = 0.7Y_{t-1} + e_t$	$\rho(j) = 0.7^j$	0.7	0	0
$Y_t = -0.7Y_{t-1} + e_t$	$\rho(j) = (-0.7)^{-j}$	-0.7	0	0
$Y_t = 0.3Y_{t-1} + 0.4Y_{t-2} + e_t$	$\rho(1) = 0.5$ $\rho(j) = 0.3\rho(j-1) + 0.4\rho(j-2)$	0.3	0.4	0
$Y_t = 0.7Y_{t-1} - 0.49Y_{t-2} + e_t$	$\rho(1) = 0.4698$ $\rho(j) = 0.7\rho(j-1) - 0.49\rho(j-2)$	0.7	-0.49	0

6. MA(q) 모형

MA(1) 모형: $Y_t = e_t - \beta_1 e_{t-1}$, $e_t \sim iid\ N(0,\sigma^2)$

o 평균은 0 이다.

o $\gamma(0) = V(Y_t) = (1+\beta_1^2)\sigma^2$, $\gamma(1) = COV(Y_t, Y_{t-1}) = -\beta_1\sigma^2$,

o $\gamma(2) = \gamma(3) = \gamma(4) = \ldots = 0$

자기상관함수

$\gamma(0) = \sigma^2(1+\beta_1^2)$, $\gamma(1) = -\frac{\beta_1}{1+\beta_1^2}$: 1 차 이후 0 이다.

부분자기상관함수

o invertibility 에 의해 AR(∞)로 변환가능하다.

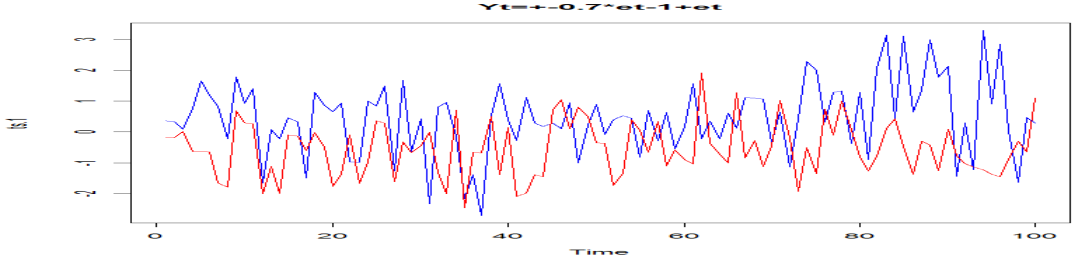
o $\phi_k = \frac{-\beta_1^k(1-\beta_1^2)}{1-\beta_1^{2(k+1)}}$

Invertibility

$Y_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} + \ldots - \beta_q e_{t-q}$ MA(q) 모형에서 $1 - \beta_1 M - \beta_2 M^2 - \ldots - \beta_q M^q = 0$ 의 방정식을 만족하는 근들의 절대값이 모두 1 보다 클 경우 MA 모형은 Invertibility 하다. 이 말은 AR(∞)모형으로 변환할 수 있다는 것이다.

- $\{Y_t\}$를 AR(∞)로 표현할 수 있으며, 즉 $Y_{t-1}, Y_{t-2}, \ldots$들로 표현되며
- $\{Y_t\}$에 대한 $Y_{t-1}, Y_{t-2}, \ldots$들의 영향은 시점이 멀어질수록 줄어든다.

```
#MA(1)
y0.1=c(0);y0.2=c(0);y1=c(0);y2=c(0) #초기 관측치

for (i in 1:100) {
e1=rnorm(1,0,1);e2=rnorm(1,0,1)
y1[i]=0.7*y0+e1
y2[i]=-0.7*y0+e2
y0.1=e1;y0.2=e2}

ts.1=ts(y1, start=c(1,1), frequency=1)
ts.2=ts(y2, start=c(1,1), frequency=1)
plot(ts.1, col="blue", type="l",main="Yt=+-0.7*et-1+et")
lines(ts.2, col="red")
```



자기상관함수

MA(1) with $\beta_1 = 0.7$: $V(Y_t) = (1+\beta_1^2)\sigma^2$, $\gamma(1) = \frac{Cov(Y_t, Y_{t-1})}{V(Y_t)} = \frac{0.7}{1+0.49} = 0.47$

2 차부터 0 이다. $\gamma(2) = \gamma(3) = \ldots = 0$

부분자기상관함수

MA(1)가 invertibility 하면, $\phi_1 = 0.47$

MA(q) 모형: $Y_t = e_t - \beta_1 e_{t-1} - \beta_2 e_{t-2} - \ldots - \beta_q e_{t-q}$, $e_t \sim iid\ N(0, \sigma^2)$

o 과거 오차항 $e_{t-1}, e_{t-2}, \ldots$ 의미 : 이전 관측치 $Y_{t-1}, Y_{t-2}, \ldots$ 에 포함되어 있지 않은 정보

o 시계열 데이터 $\{Y_t\}$ 에서 시점 t 의 관측치 Y_t 가 과거 오차 $e_{t-1}, e_{t-2}, \ldots, e_{t-q}$ 들에 의해 설명될 때 MA(q) (차수가 q 인 Moving-Average 이동평균) 모형을 따른다고 한다.

o MA(∞) 모형은 언제나 정상적(stationary)이다.

자기상관함수

o $\rho(k) = \dfrac{-\beta_k + \beta_1\beta_{k+1} + \ldots + \beta_{q-k}\beta_q}{1 + \beta_1^2 + \beta_2^2 + \ldots + \beta_k^2}$, $k \le q$

o $\gamma(q+1) = \gamma(q+2) = \ldots = 0$, q 차 이후 0 이다.

부분자기상관함수

o invertibility 에 의해 AR(∞)로 변환가능하다. 그러므로 MA(q) 모형의 PACF 는 Invertibility 조건 하에서 지수적으로 감소한다.

 시뮬레이션

모형	ACF (이론)	PACF (이론)		
		1 차	2 차	3 차
$Y_t = 0.8e_{t-1} + e_t$	$\rho(1) = 0.4878$, $\rho(j) = 0, j \ge 2$	0.49	-0.31	0.22
$Y_t = e_t$ (white noise)	$\rho(0) = 1$, $\rho(j) = 0, j \ge 1$	0	0	0
$Y_t = -0.3e_{t-1} - 0.4e_{t-2} + e_t$	$\rho(1) = -0.144$, $\rho(2) = -0.32$	-0.14	-0.35	-0.13

7. ARMA(p, q) 모형

ARMA(p,q) 모형: $Y_t = e_t - \beta_1 e_{t-1} - \ldots - \beta_q e_{t-q} + \alpha_1 Y_{t-1} + \ldots + + \alpha_p Y_{t-p}$, $e_t \sim iid\ N(0, \sigma^2)$

o AR 모형과 MA 모형의 결합이다. 그러므로 AR(∞), MA(∞)로 표현될 수 있음.

o 일반적으로 (2, 2)가 최대

자기상관함수(acf) 부분자기상관함수(pacf)

지수적으로 감소

 시뮬레이션

모형	ACF (이론)	PACF (이론)

		1 차	2 차	3 차
$Y_t = 0.6Y_{t-1} + 0.4e_{t-1} + e_t$	$\rho(1) = 0.7561$ $\rho(j) = 0.6\rho(j-1)$	0.7561	-0,276	0.109

VI. 차분 Difference (계절성 및 추세 성분)

o ARMA 모형은 시계열 데이터 중 사이클 (cycle) 성분에 대한 패턴을 표현하게 된다.

o 물론 불규칙 irregularity 성분은 오차항으로 커버한다.

o 그럼 추세 trend, 계절성 seasonality 성분은 어떻게 하지? 차분이 답이다.

o 차분은 추세나 계절성 성분을 제외시키는 효과가 있다.

o 차분에 의해 추세나 계절성 성분을 제외하면 주기와 불규칙 성분만 남아 수평 상태로 사이클만 존재하게 된다.

정의

1 차 차분 : $Y_t^* = \nabla Y_t = Y_t - Y_{t-1}$ => 직선 추세성분 해결

2 차 차분 : $\nabla Y_t = Y_t^* - Y_{t-1}^*$ => 이차형식 추세성분 해결

d 차 차분 : $(Y_t - Y_{t-d})$ => 주기 d 계절성 성분

차분 필요성 진단

o PACF 에서 차분이 필요한 주기에서 Peak 가 발생하며, ACF 는 지수적으로 감소

ARMA 모형 진단 표

	AR(p)	MA(q)	ARMA(p, q)
ACF	T	D(q)	T
PACF	D(p)	T	T
IACF	D(p)	T	T

*) T: Tail off exponentially 지수적으로 감소

*) D(p): Drop off after p 차수 p 이후 0 의 값

VII. ARMA 모형 적합 절차 ()

1) 시간도표

(1) 주기, 계절성 확인 (실제 진단은 상관함수 이용) => plot() 함수

(2) 안정성 stationary process

(a) 평균의 이동 => 평균이 이동하는 경우에는 시계열 데이터 분리하여 모형 적합

(b) 분산의 크기 변동 ⇔ 주기의 폭이 변함 => 분산 안정화, LN 혹은 제곱근(SQRT) 변환

2) 모형 적합 가능성 진단

(1) white noise 데이터는 모형 적합 불가 => whitenoise.test 함수

MN 검정통계량

test value 유의확률

(2) 또 다른 백색 잡음 검정 수정 Ljung Box-Pierce Q 통계량 $n(n+2)\sum_{j=1}^{k}\frac{\gamma(j)}{(n-j)} \sim \chi^2(k)$. Q-

=> Box.test(type="Ljung-Box") 검정

(3) unit root (단일근) 검정 => pp.test() 함수 ⇔ 시계열 데이터의 안정성 stationary

```
/*비계절성 모형 진단 */
proc arima data=load0;
   identify var=real_load nlag=28;
run;
```

3) 모형 진단

ACF, PACF 활용하여 (p, q, d) 결정

4) 모형추정

회귀계수 추정 => arima() 함수

method = c("CSS-ML", "ML", "CSS")

maximum likelihood / minimize conditional sum-of-squares.

5) 모형 적합성

(1) 회귀계수의 유의성 검정

(2) 잔차의 백색 잡음 Ljung Box-Pierce Q 통계량 : 오차의 분산 추정량인 잔차 $r_t = Y_t - \hat{Y}_t$

차분이 필요한 경우 잔차가 백색잡음 형태가 아니라 AR 시리즈 데이터 PACF 가짐

6) 예측모형 활용

(1) 여러 모형 중 가장 적합한 모형 : AIC, SBC 작은 값의 모형이 더 적합

o AIC (Akaike Information Criterion) $AIC = -\log \hat{\sigma}_e^2 + 2(p+q)$

o SBC (Schwartz Bayesian Criterion) $SBC = n \log \hat{\sigma}_e^2 + (p+q)\log(n)$

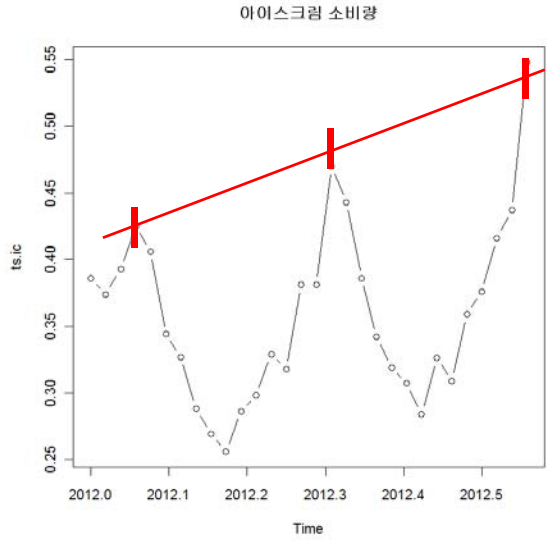
$\hat{\sigma}_e^2$ 은 오차의 분산 σ^2 의 추정치로 MSE 이다.

(2) 향후 필요한 주기까지 최종 모형을 활용하여 관심 변수 예측값 추정

ICECREAM.csv

1) 시간도표

```
ds=read.csv("icecream.csv")
ts.ic=ts(ds$IC, start=c(2012,1), frequency=52)
plot(ts.ic, type="b", main="아이스크림 소비량")
```



(1) 추세는 선형을 보인다. / 주기 13 의 계절성을 보인다.

(2) 평균 이동이나 분산 변동은 보이지 않음 => 데이터 분리하여 분석하거나 분산 안정화 변환 필요 없음

2) 모형 적합 가능성 진단

(1) white noise 검정 (1)

귀무가설 : 시계열 자료는 백색잡음이다.

```
library(normwhn.test)
whitenoise.test(ts.ic)
```

```
[1] "tMN"
[1] -7.736887
[1] "test value"
[1] 0.0001095323
```

유의확률이 0.01%이므로 귀무가설이 기각되어 시계열 데이터는 백색잡음 (no pattern) 아니므로 ARMA 모형 적용이 가능하다.

(2) white noise 검정 검정 (2)

```
Box.test(ts.ic, type=c("Ljung-Box"))
```

```
> Box.test(ts.ic, type=c("Ljung-Box"))

        Box-Ljung test

data:  ts.ic
X-squared = 14.5389, df = 1, p-value = 0.0001373
```

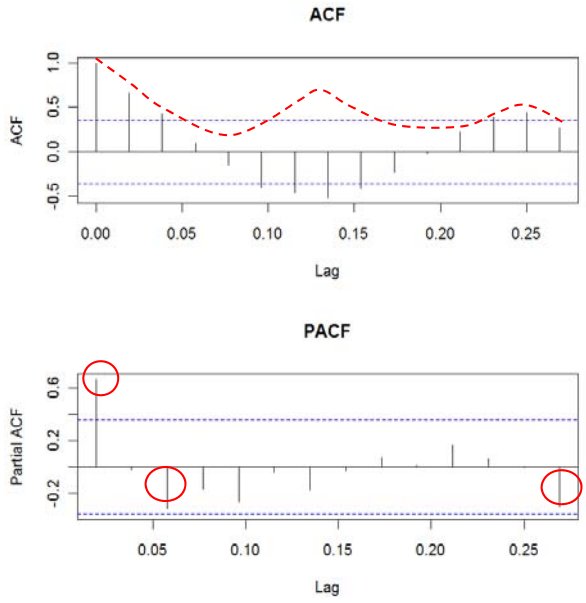
=> 동일한 결론

(3) unit root (단일근) 검정 => pp.test() 함수 ⇔ 시계열 데이터의 안정성 stationary

```
> library(tseries)
> pp.test(ts.ic, alternative=c("explosive"))

        Phillips-Perron Unit Root Test

data:  ts.ic
Dickey-Fuller Z(alpha) = -5.0509, Truncation lag parameter = 2,
p-value = 0.1912
alternative hypothesis: explosive
```

3) 모형 진단

```
par(mfrow=c(2,1))
acf(ts.ic, main="ACF")
pacf(ts.ic, main="PACF")
```



o AR(1) 이 적절해 보인다. 문제는 ACF 주기 5~8 에서 유의한 모습을 보인다.

o 만약 AR(1) 모형이 적합하다면 실제로는 지수적으로 감소하여 유의한 선 아래로 떨어짐

o PACF 주기 (3, 14)에 peak 가 발생하는 것으로 보임, 차분이 필요?

4) 모형추정

회귀계수 추정 => arima() 함수

- method = c("CSS-ML", "ML", "CSS")
- maximum likelihood / minimize conditional sum-of-squares.

```
fit.ic=arima(ts.ic, order=c(1,0,0))
fit.ic
```

```
Call:
arima(x = ts.ic, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.8679     0.3922
s.e.  0.1034     0.0503

sigma^2 estimated as 0.001585:  log likelihood = 53.44,  aic = -100.88
```

- o AR(1) 추정 결과 : $IC_t = 0.392 + 0.8679 IC_{t-1}$

5) 모형 적합성

(1) 회귀계수의 유의성 검정

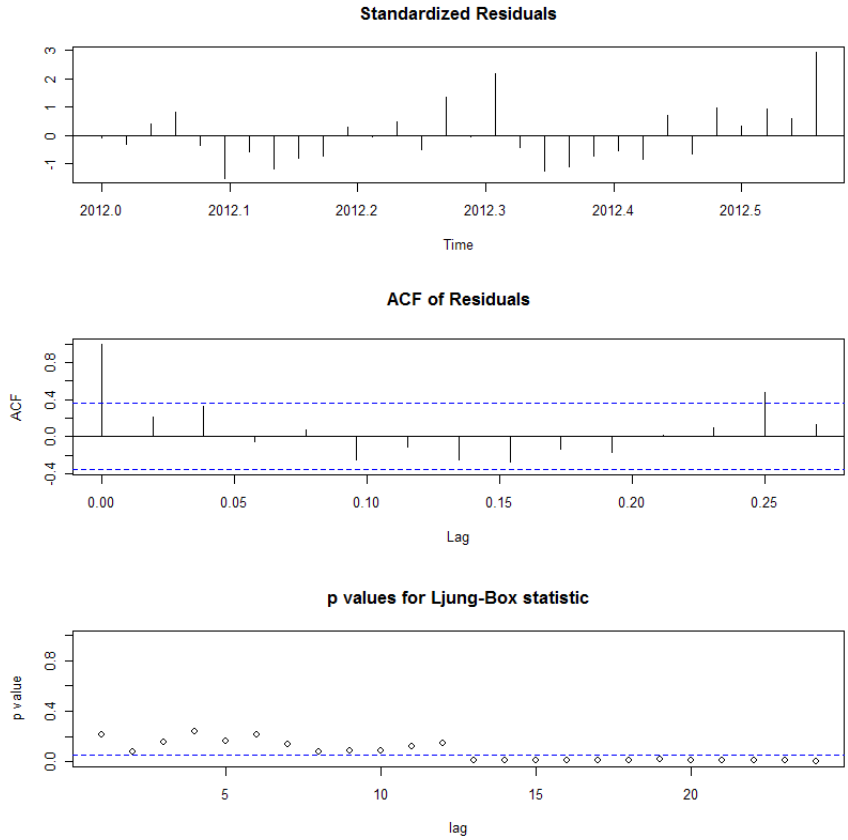
```
t=fit.ic$coef[1:1]/sqrt(fit.ic$var.coef)[1:1]
p_value=1-pt(t, length(ts.ic)-2)
cat("t-통계량",t, "유의확률",p value)
```

```
> cat("t-통계량",t, "유의확률",p_value)
t-통계량 8.394274 유의확률 1.971794e-09>
```

- o IC_{t-1}의 회귀계수는 유의하므로 AR(1) 모형은 적절하다.

(2) 잔차의 백색 잡음 Ljung Box-Pierce Q 통계량 : 오차의 분산 추정량인 잔차 $r_t = Y_t - \hat{Y}_t$

```
> tsdiag(fit.ic,gof.lag=24)
```



o 잔차는 유의확률 기준 선을 벗어나므로 백색 잡음이 아님 => 모형 설정 잘못

o 잔차의 백색잡음 검정 (Ljung-Box 방법 사용) 결과 주기가 13 이후 white noise 경향에서 벗어나고 있음 => AR(1) 모형 설정에 문제가 있음, 차분이 답이다.

(이런 계절성이 문제)

```
fit.ic2=arima(ts.ic, order=c(1,13,0), method=c("CSS")
fit.ic2
tsdiag(fit.ic2)
```

```
arima(x = ts.ic, order = c(1, 13, 0), method = c("CSS"))

Coefficients:
          ar1
      -0.9874
s.e.   0.0733

sigma^2 estimated as 538.1:  log likelihood = -77.57,  aic = NA
```

o 이런 이제는 단일근 문제가 발생한다.

o 그러므로 1 차 차분이 필요하다. => 그런 후 MA(1) 모형 적용

```
fit.ic2=arima(diff(ts.ic), order=c(0,13,1), method = c("CSS"))
fit.ic2
tsdiag(fit.ic2)
```

```
arima(x = diff(ts.ic), order = c(0, 13, 1), method = c("CSS"))

Coefficients:
          ma1
      -0.9081
s.e.   0.0888
```

$$\nabla Y_t = (Y_t - Y_t - 1)$$
$$(\nabla Y_t - \nabla Y_{t-13}) = 0.9081 e_{t-1}$$

6) 예측모형 활용

(1) 여러 모형 중 가장 적합한 모형 : AIC, SBC 작은 값의 모형이 더 적합

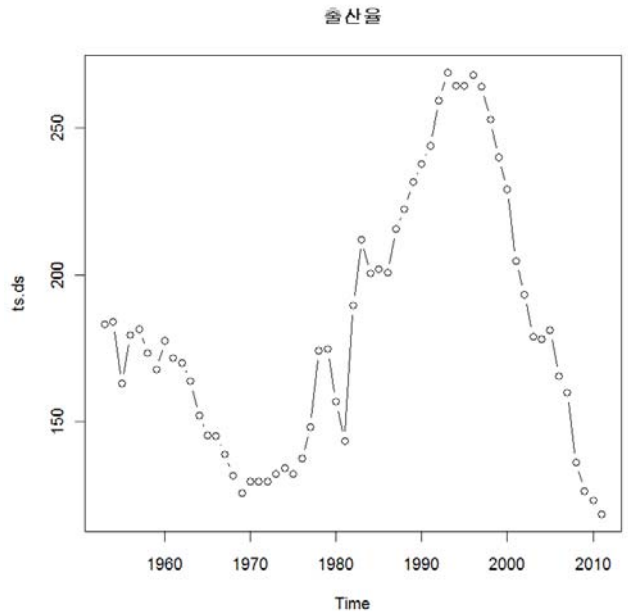
- o AIC (Akaike Information Criterion) $AIC = -\log \hat{\sigma}_e^2 + 2(p+q)$
- o SBC (Schwartz Bayesian Criterion) $SBC = n \log \hat{\sigma}_e^2 + (p+q)\log(n)$

$\hat{\sigma}_e^2$ 은 오차의 분산 σ^2 의 추정치로 MSE 이다.

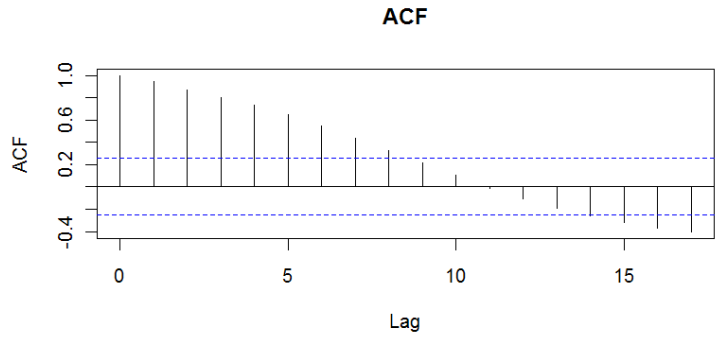
(2) 향후 필요한 주기까지 최종 모형을 활용하여 관심 변수 예측값 추정

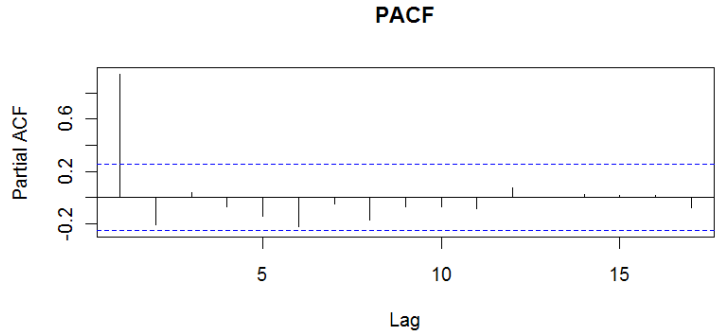
```
> predict(fit.ic2, n.ahead=8)$pred
```

```
ds=read.csv("BR.csv")
ts.ds=ts(br$Birth_rate, start=c(1953,1), frequency=1)
plot(ts.ds,type="b", main="출산율")

library(normwhn.test)
whitenoise.test(ts.ds)

Box.test(ts.ds, type=c("Ljung-Box"))

par(mfrow=c(2,1))
acf(ts.ds, main="ACF")
pacf(ts.ds, main="PACF")

fit.ds=arima(ts.ds, order=c(1,0,0))
fit.ds
tsdiag(fit.ds)
```




```
[1] "tMN"
[1] 28.3132
[1] "test value"
[1] 1.70022e-45
>
> Box.test(ts.ds, type=c("Ljung-Box"))

        Box-Ljung test

data:  ts.ds
X-squared = 55.4049, df = 1, p-value = 9.814e-14
```






```
arima(x = ts.ds, order = c(1, 0, 0))

Coefficients:
         ar1  intercept
      0.9611   167.4573
s.e.  0.0297    30.2607

sigma^2 estimated as 143.9:  log likelihood = -231.6,  aic = 469.2
```

(단일근 검정)

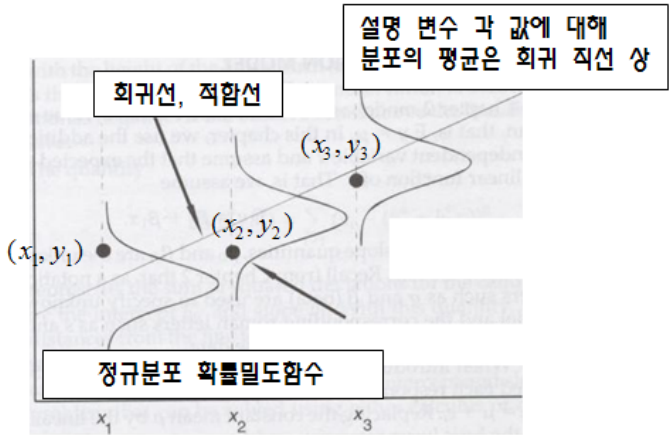
```
> pp.test(ts.ds)

        Phillips-Perron Unit Root Test

data:  ts.ds
Dickey-Fuller Z(alpha) = -1.2263, Truncation lag parameter = 3,
p-value = 0.9816
alternative hypothesis: stationary
```

IX. 계량경제 회귀모형

1. 개념

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i, \quad i = 1.2\ldots., n$$

(가정) $e_i \sim iidN(0, \sigma^2)$ (독립성 : 시계열 데이터) (정규성) (등분산성)



행렬

$$\underline{y} = X\underline{\beta} + \underline{e}, \quad \underline{e} \sim MN(\underline{0}, \sigma^2 I)$$

그러므로 $E(\underline{y}) = X\underline{\beta}, \quad V(\underline{y}) = \sigma^2 I$

추정 : OLS 구하기

$$\min_{\underline{\beta}} e_i^2 = \min_{\underline{\beta}} \underline{e}'\underline{e} = \min_{\underline{\beta}} (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})$$

$$Q = (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta}) = \underline{y}'\underline{y} - \underline{y}'X\underline{\beta} - (X\underline{\beta})'y + (X\underline{\beta})'X\underline{\beta}$$

$$\frac{\partial Q}{\partial \underline{\beta}} = -X'\underline{y} - X'y + \frac{(X'X)}{2}\underline{\beta} = 0 \quad \Leftrightarrow \quad \underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y} \text{ when } (X'X)^{-1} \text{ exist}$$

$(X'X)^{-1}$가 존재한다는 의미 ⇔ $X'X$ 가 full rank ⇔ 설명변수들의 상관계수가 1 인 경우는 없거나 다른 설명변수의 선형 결합으로 임의의 설명변수가 표현될 수 없음

분산 σ^2 추정량

$\sigma^2 = E(\underline{y} - X\underline{\beta})^2$ => $SSE = (\underline{y} - X\hat{\underline{\beta}})'(\underline{y} - X\hat{\underline{\beta}}) = \underline{y}'[I - X(X'X)^{-1}X']\underline{y}$

$SSE = \underline{y}'\underline{y} - \hat{\underline{\beta}}'X\underline{y}$ (easy form)

$MSE = \frac{SSE}{(n-k-1)} = \hat{\sigma}^2$ (Mean Square of Errors : 평균자승합)

성질 : $E(SSE) = \sigma^2(n-k-1)$ => 그러므로 $E(\frac{SSE}{(n-k-1)}) = \sigma^2$

분산분석적 접근

$SST = \sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 = SSE + SSR$

$SSR = \hat{\underline{\beta}}X'\underline{y} - n\bar{y}^2$

결정계수 Multiple Determinant Coefficients : $R^2 = \frac{SSR}{SST}$

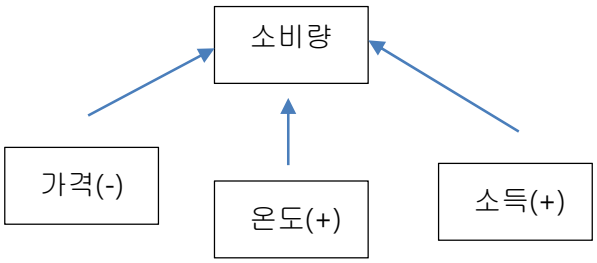
- 설정된 설명변수가 종속변수의 총변동을 (SST) 설명하는 정도로 모형의 설명력
- $0 \le R^2 \le 1$, 일반적으로 70% 이상이면 적절한 설명변수 선택하였음

2. 예제 데이터

Icecream Data

종속변수 Y : 아이스크림 소비량

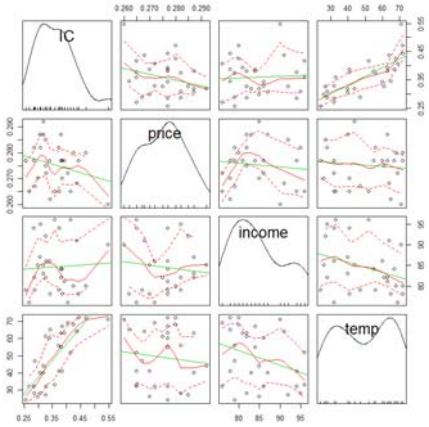
설명변수 X : (가격, 소득, 온도)



3. 회귀분석 순서

순서 1: 산점도 행렬 그리기

```
ds.ic=read.csv("ic.csv")
library(car)
scatterplot.matrix(~IC+price+income+temp,ds.ic)
```



(이상치, 영향치 존재) (유의한 변수 사전 진단)

순서 2 : Model 추정

```
fit.ic=lm(IC~price+income+temp,data=ds.ic)
summary(fit.ic)
```

```
lm(formula = IC ~ price + income + temp, data = ds.ic)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065302 -0.011873  0.002737  0.015953  0.078986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1973151  0.2702162   0.730  0.47179
price       -1.0444140  0.8343573  -1.252  0.22180
income       0.0033078  0.0011714   2.824  0.00899 **
temp         0.0034584  0.0004455   7.762  3.1e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03683 on 26 degrees of freedom
Multiple R-squared: 0.719,      Adjusted R-squared: 0.6866
F-statistic: 22.17 on 3 and 26 DF,  p-value: 2.451e-07
```

o 가격 변수 유의하지 않음

순서 3 : 변수선택

이유

적은 정보로 동일한 수준의 정보를 얻음

종속변수의 변동을 설명하는 정도가 낮은 설명변수는 삭제

가정 적절한 방법

수작업 backward => 유의확률이 가장 높은 (설명력이 가장 낮은) 설명변수 순으로 하나씩 제거하면서 일정 수준의 유의확률 (10%) 이하인 설명변수만 남을 때까지

(후진제거)

모든 설명변수를 고려한 모형에서 유의하지 않은 설명변수를 하나씩 제거하는 방법이다.

(전진삽입)

고려된 설명변수 중 설명력 (종속변수와 상관관계 가장 높음)이 가장 높고 설명력이 유의하면 변수를 선택한다.

(단계삽입 stepwise)

Forward 방법과 유사하지만 한 번 선택된 설명 변수에 대해서는 유의성 검정을 다시 실시한다는 점이 다르다.

결정 계수($R_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST}$)

R_p^2는 설명 변수들의 설명력의 정도를 나타내는 수치이므로 변수 선택의 지표가 된다. 설명변수의 수가 같은 경우 어떤 변수 그룹이 설명력이 높은가를 쉽게 알아보는 사용할 수 있다. 또한 R_p^2는 설명변수의 수(p)가 증가할 때 마다 항상 증가

수정 결정계수 이용 $R_{adj}^2 = 1 - \frac{SSE_p/(n-p-1)}{SST/(n-1)}$

수정(adjusted) 결정계수 R_{adj}^2는 R_p^2의 문제점(유의하지 않은 설명변수가 삽입되어도 항상 증가)을 해결하였으므로 R_{adj}^2 값이 가장 큰 설명 변수 그룹을 선택하면 된다.

```
fit2.ic=lm(IC~income+temp,data=ds.ic)
summary(fit2.ic)
```

```
> summary(fit2.ic)

Call:
lm(formula = IC ~ income + temp, data = ds.ic)

Residuals:
      Min        1Q    Median        3Q       Max
-0.065420 -0.022458  0.004026  0.015987  0.091905

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.113195   0.108280  -1.045  0.30511
income       0.003530   0.001170   3.017  0.00551 **
temp         0.003543   0.000445   7.963 1.47e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03722 on 27 degrees of freedom
Multiple R-squared: 0.7021,     Adjusted R-squared:  0.68
F-statistic: 31.81 on 2 and 27 DF,  p-value: 7.957e-08
```

순서 4 : 다중공선성

설명변수들간 높은 상관관계로 인하여 $|X'X| \approx 0$이 되고 $(X'X)^{-1}$의 값이 불안정

추정회귀계수 $(X'X)^{-1}X'\underline{y}$와 그의 분산 $MSE(X'X)^{-1}$이 불안정해져 추정 회귀계수의 부호까지 바뀌는 문제 발생한다.

진단방법

상관계수 이용

o 상관계수의 부호와 회귀계수의 부호가 다른 경우 다중공선성 문제 발생

o 산점도 행렬의 기울기 부호와 추정 회귀식의 부호가 일치하므로 문제 없음

VIF 분산팽창지수 Variance Inflation Index

o $VIF_k = \frac{1}{1-R_k^2}$ R_k^2는 설명변수 X_k를 종속변수로 하고 나머지 다른 변수들을 설명변수로 하여 계산된 결정계수

o 일반적으로 3 이상(어떤 이는 10 이상)이면 문제

o 두 변수간 (pairwise) 문제를 발견하지 못하는 문제가 있다. 여러 설명변수가 동시에 고려되므로... 이에 대한 보완으로 상태지수가 있음.

```
library(car)
vif(fit2.ic)>3
```

```
> vif(fit2.ic)>3
income   temp
 FALSE  FALSE
```

해결방법

o 문제가 되는 설명변수 중 하나 삭제

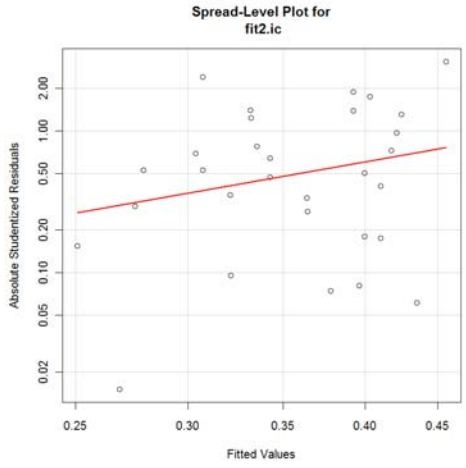
o 주성분변수 활용

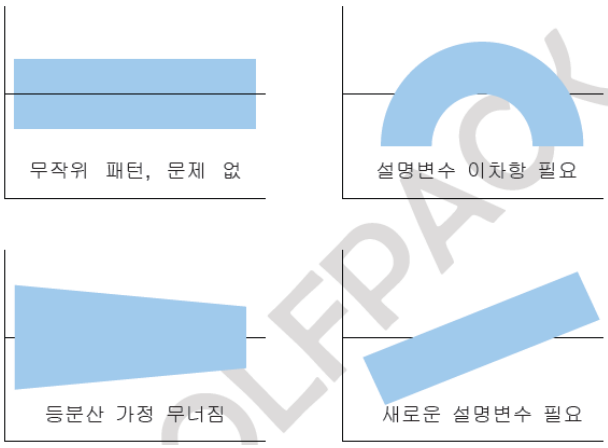
순서 5 : 잔차 진단

오차 가정 진단

(1) 정규성, 등분산성, 선형성

```
spreadLevelPlot(fit2.ic)
```





o 등분산 가정 무너짐 : 종속변수 로그 변환

o 설명변수 이차항 ? 산점도행렬에서 사전 진단 가능

(2) 독립성 (다음 시간)

오차의 독립성은 Durbin and Watson(1951) 통계량의 의해 검정한다. DW 통계량은 오차 자기상관 존재여부를 판단한다. $e_t = \rho e_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iidN(0, \sigma_*^2)$ 자기상관이 존재한다는 것은 회귀계수 ρ 가 0 이 아니라는 것이다. 다음은 자기상관을 검정하는 DW 검정통계량이다.

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

만약 자기상관이 존재하지 않으면 DW 는 2 에 근사한다.(why? DW 검정통계량에 $e_t = \rho e_{t-1} + \varepsilon_t$ 을 넣고 $\rho = 0$ 으로 해 보자) 임계치 D_L 과 D_U 는 자료의 개수와 설명변수의 개수 p 에 의존하며 표가 따로 주어진다. 만약 $D_L \le DW \le D_U$ 이면 귀무가설 채택한다. 그렇지 않으면 귀무가설 기각한다.

DW 검정통계량에 대한 유의확률이 주어지지 않으므로 표를 찾아야 하는 번거로움이 있다. 오차의 자기상관계수($Corr(e_t, e_{t-1})$) r 과 $DW \approx 2(1-r)$ 의 관계가 있으므로 오차(잔차, 오차의 추정치)의 자기상관계수를 이용하여 독립성을 검정할 수 있다. DW 통계량 표는 강의노트에서 다운받기 바란다. 다음은 그 일부분이다. p 는 설명변수의 개수이다.

$(0, D_L)$	(D_L, D_U)	$(D_U, 4-D_U)$	$(4-D_U, 4-D_L)$	$(4-D_L, D_L)$
귀무가설 기각 양의 자기상관	미결정 H_0 기각도 채택도 하지 않음	귀무가설 채택 자기상관 없음	미결정 H_0 기각도 채택도 하지 않음	귀무가설 기각 음의 자기상관

데이터($n=30, p=2$) DW 통계량은 1.003 이었는데 DW-통계표에서 $(D_L = 1.13, D_U = 1.26)$ 이므로 오차의 자기상관이 존재한다. $(0, D_L)$ 사이에 있으므로 양의 자기상관이 존재한다

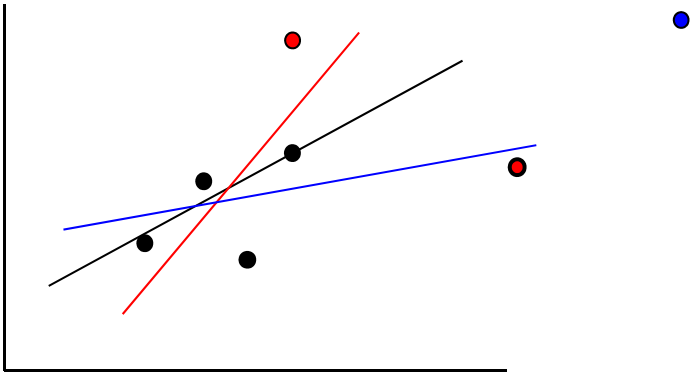
n	$p=1$		$p=2$		$p=3$		$p=4$		$p=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60

순서 6 : 영향치/이상치 진단

(이상치)는 종속변수와 설명변수의 선형 관계식에서 멀리 떨어진 관측치 (Y-축 기준), 삭제함을 원칙으로 함

(**영향치**)는 선형모형에 영향을 주는 관측치, 다른 관측치와 설명변수 범위 면에서 (X-축 기준) 떨어진 관측치 : 결정계수를 높이는 역할, 실제 사이 설명변수 구간의 관측치를 더 수집한 후 결과를 냄

종속변수 Y



설명변수 X

잔차 residual $r_i = y_i - \hat{y}_i$

o 관측치와 추정치의 차이 : 오차의 추정치 $\underline{r} = \underline{\hat{e}} = \underline{y} - \underline{\hat{y}} = (I - H)\underline{y}$

o $\sigma^2(\underline{z}) = \sigma^2(I - H)$ => 추정치 $\hat{\sigma}^2(\underline{z}) = s^2(\underline{z}) = MSE(I - H)$

o hat 행렬 : $H = X'(XX)^{-1}X$ => 대각원소 h_{ii}

스튜던트 잔차 (Studentized Residual) : 이상치 진단

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{MSE/1 - h_{ii}}}, \quad h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i$$

t-분포를 따르는 통계량으로 만든 것으로 ± 2 이면 이상치(혹은 영향치)로 판단하게 된다.

영향치 진단 통계량 : Leverage $h_{ii} \Leftarrow H = X(X'X)^{-1}X$ 의 대각 원소

H 행렬 대각 원소 h_{ii} 는 i 번째 관측치가 설명 변수들의 중심점으로부터 얼마나 떨어져 있는가를 나타낸다.

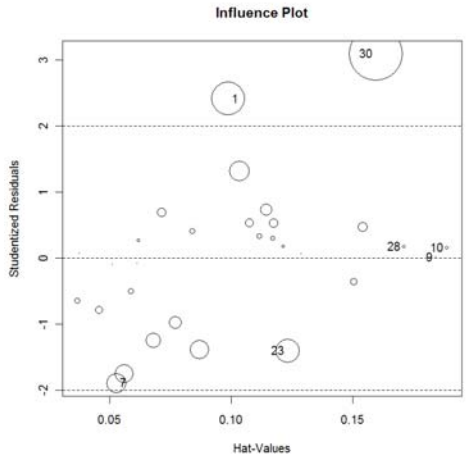
COV Ratio $CovRatio = \frac{MSE_{(i)} |(X'_{(i)}X_{(i)})^{-1}|}{MSE\,|(X'X)^{-1}|}$

i-번째 관측치를 제외했을 때 추정치의 분산이 커진다면 이 관측치는 회귀 선상에 있고 다른 관측치와 떨어져 있을 것이다. 기준 값은 $|CovRatio-1| \geq 3(p+1)/n$ 이며 이 값이 크다면 일반적으로 Leverage 값도 크다.

Cook's Distance $C_i = \frac{\sum_{j=1}^{n}(\hat{Y}_{j.f} - \hat{Y}_{j(i)})^2}{(p+1)MSE}$

Leverage 통계량은 설명 변수들간의 관계만으로 영향치를 판단하지만 Cook's 거리 통계량은 추정 회귀 모형에서 판단된다.

```
influencePlot(fit2.ic, id.method="identity", main="Influence Plot", id.n=3)
```



1 번째, 30 번째 관측치는 이상치, 영향치 없음.

```
ds.ic2=rbind(ds.ic[2:27,],ds.ic[28:29,])
fit3.ic=lm(IC~income+temp,data=ds.ic2)
summary(fit3.ic)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0886859  0.0926372  -0.957  0.34756
income       0.0033043  0.0009860   3.351  0.00256 **
temp         0.0033071  0.0003763   8.789 4.05e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 

Residual standard error: 0.02926 on 25 degrees of freedom
Multiple R-squared: 0.7557,     Adjusted R-squared: 0.7361
F-statistic: 38.67 on 2 and 25 DF,  p-value: 2.235e-08
```

I. 일별 Daily 예측모형 Time Plot (Daily)

(시각적 판단)

1) Trend

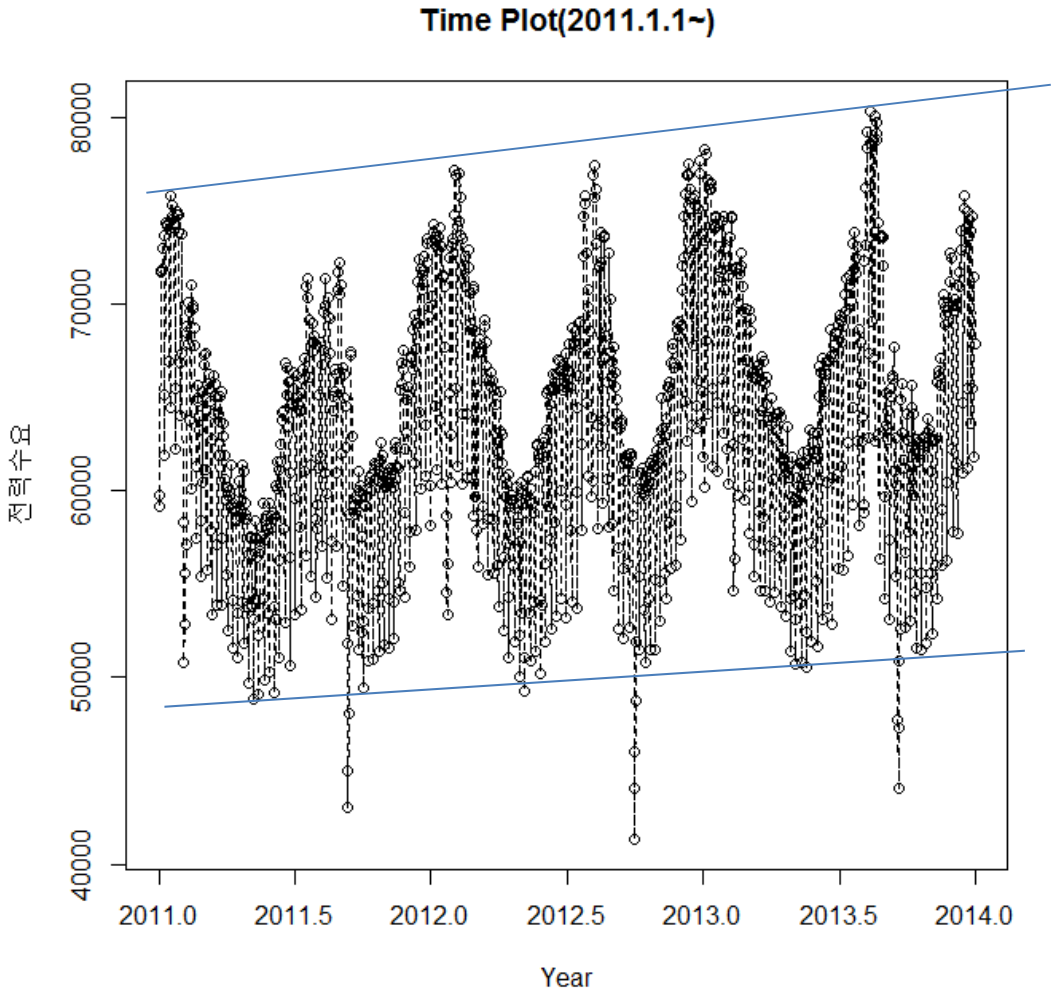
- 2009~, 2010~ time plot 에는 직선 증가 경향(linear trend) 여전히 존재하는 것으로 보임. (시각적 판단)
- 2011~ time plot 역시 직선 증가 경향이 보임.

2) 이분산 문제

- time plot 을 보면 문제가 없음.

3) stationary 검정

- Augmented Dickey-Fuller 검정 => 3 시계열 모두 “Stationary”



D-1. 비계절형 ARIMA model (Daily) $ARMA(7,1,0)(0,1,0)_7$

0. 시계열 데이터

1) 분석 활용 데이터 : 2011.01.01~2013.12.31 (3 년 일별 전력수요량)

2) stationary 정상성

- 정상적 시계열 데이터만 ARIMA 적용 가능함.
- 유의확률=0.99 로 2011.01.01~ 시계열 데이터는 stationary 함.

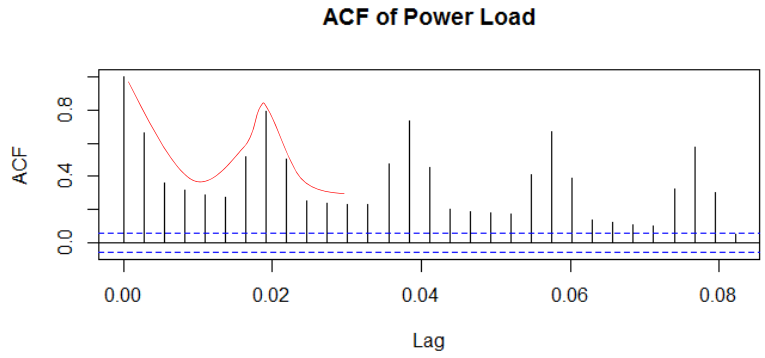
```
        Augmented Dickey-Fuller Test

data:  ts.ds11
Dickey-Fuller = -4.0423, Lag order = 10, p-value = 0.99
alternative hypothesis: explosive
```

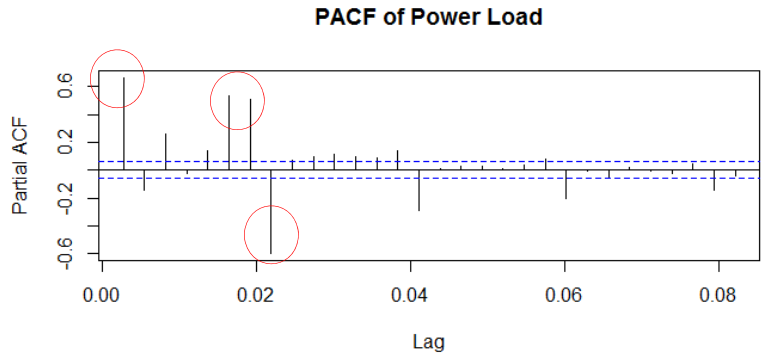
;

1. 모형진단

1) 원 시계열 Y ACF, PACF 활용

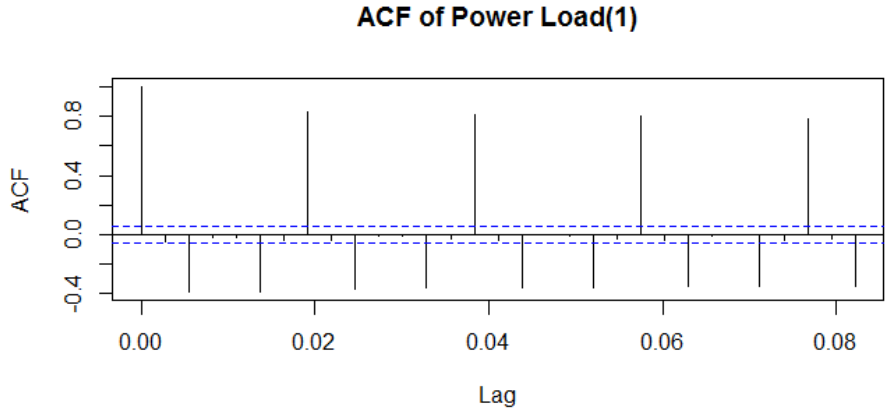
- ACF: 선형적으로 감소하다가 증가, 주기 7 에 의한 peak 발생
- PACF: 주기 7 까지 증가하다가 peak 발생
- 결론적으로 주기 7 에 의한 차분 후 ARMA(p,q) 모형이 적절해 보임

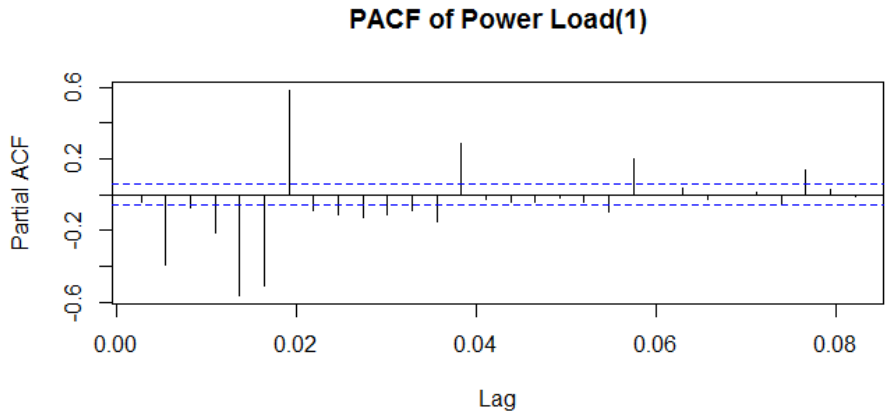




2) 1 차 차분 Y(1) ACF, PACF 함수

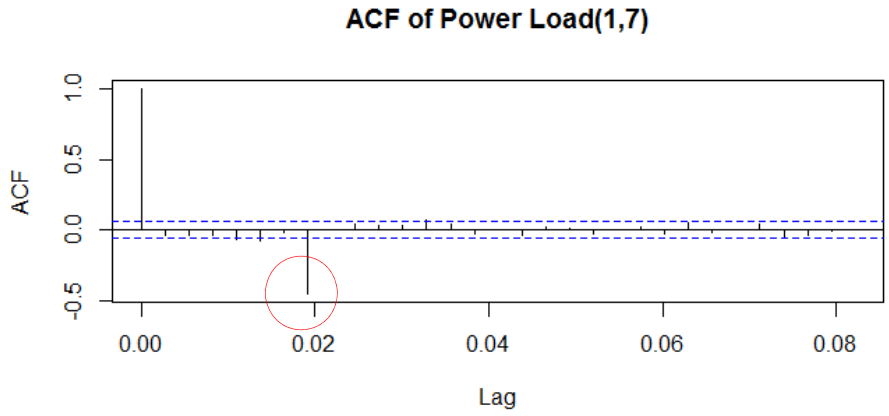
- ACF: 주기 7 에 의한 peak 발생
- PACF: 주기 7 까지 증가하다가 peak 발생
- 결론적으로 주기 7 에 의한 차분이 한 번 더 필요함.

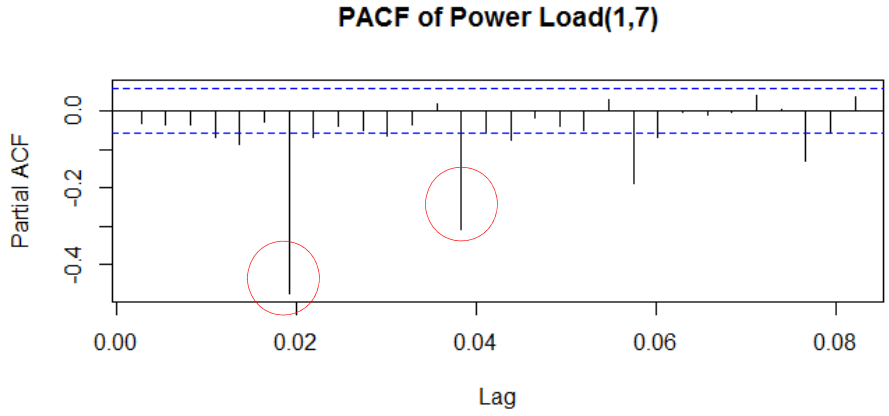




2. 전력수요(1,7) 데이터 ARMA 모형 추정

- 1 차 차분 후, 주 s=7 차분 : $\nabla_{(1,7)}Y_t = (Y_t - Y_{t-1}) - (Y_{t-7} - Y_{t-8})$
- ACF s=7, PACF s=7, 14, 21 에 peak 가 존재함
- 전력수요(1,7) 시계열 데이터 acf(주기=7 에서 peak), pacf(주기=7, 14, .. peak) 함수 아래와 같으므로 비계절형 ARMA 모형은 계절형 ARMA 모형 적합이 적절함.





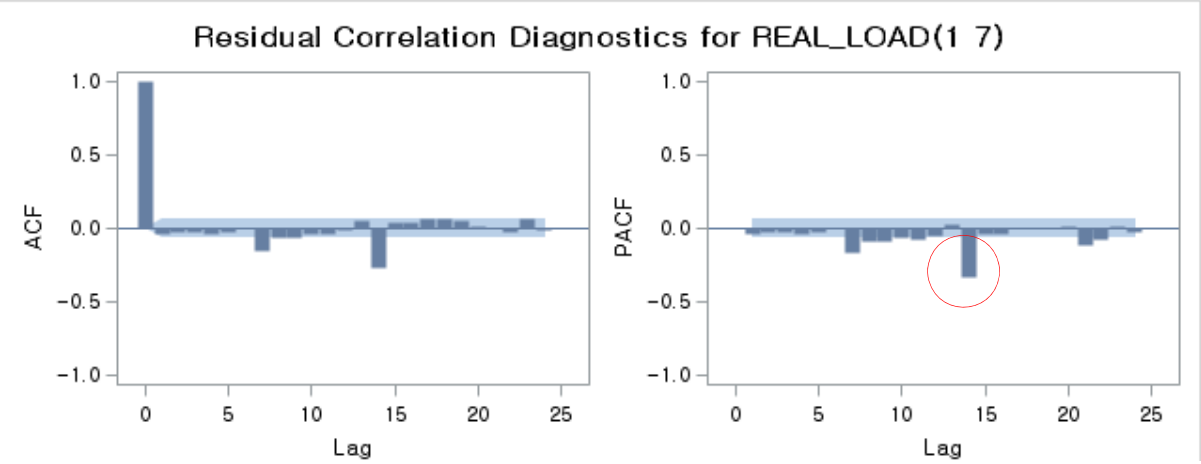
*비계절형 ARMA(p=7, q=0) , q>=1 인 경우 추정이 converge 되지 않음.

ARIMA (p=7, q=0) 1차 차분_7차 차분 데이터

The ARIMA Procedure

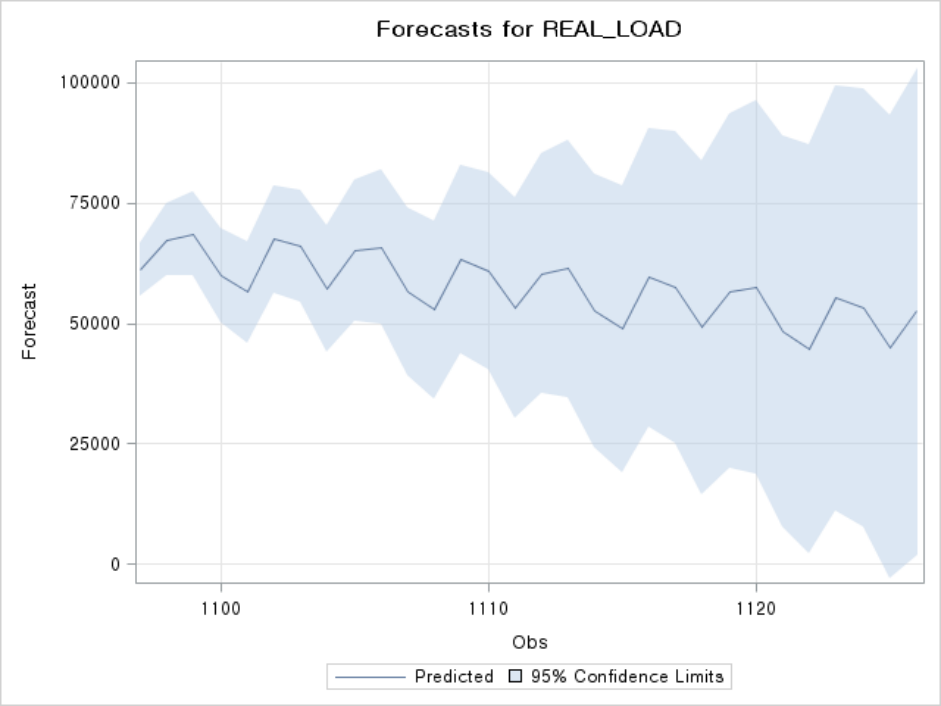
Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	-6.91436	43.08527	-0.16	0.8725	0
AR1,1	-0.05901	0.02673	-2.21	0.0273	1
AR1,2	-0.08715	0.02676	-3.26	0.0011	2
AR1,3	-0.07737	0.02668	-2.90	0.0037	3
AR1,4	-0.09350	0.02664	-3.51	0.0004	4
AR1,5	-0.10914	0.02669	-4.09	<.0001	5
AR1,6	-0.04609	0.02685	-1.72	0.0861	6
AR1,7	-0.47869	0.02693	-17.77	<.0001	7



Analysis Variable : mape
Mean
0.0293215

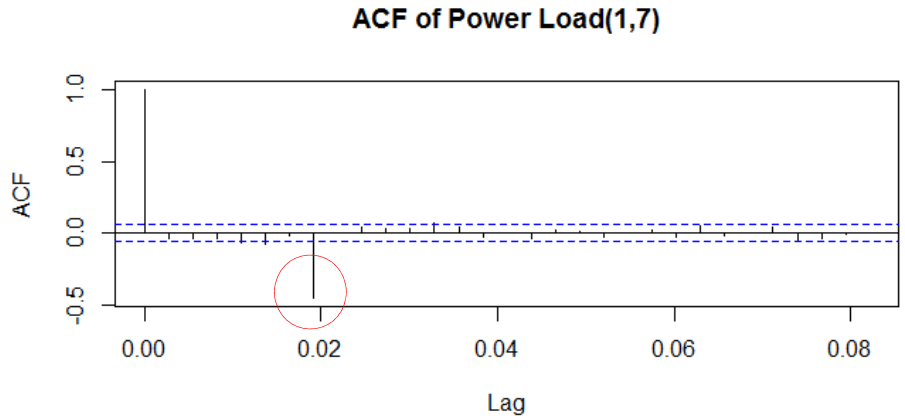
MAPE=2.93%

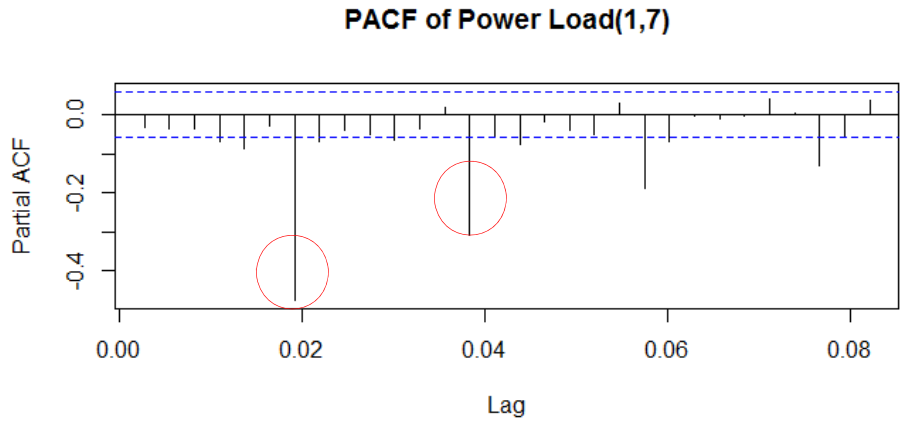


D-2. 계절형 ARIMA model (Daily) **ARMA(0,1,0)(2,1,1)$_7$**

1. 전력수요(1,7) 시계열 모형진단

   - 전력수요(1,7) 시계열 데이터 acf(주기=7 에서 peak), pacf(주기=7, 14, .. peak) 함수 아래와 같음.
   - p=(7, 14), q=(7) 모형을 적합하는 것이 적절해 보임





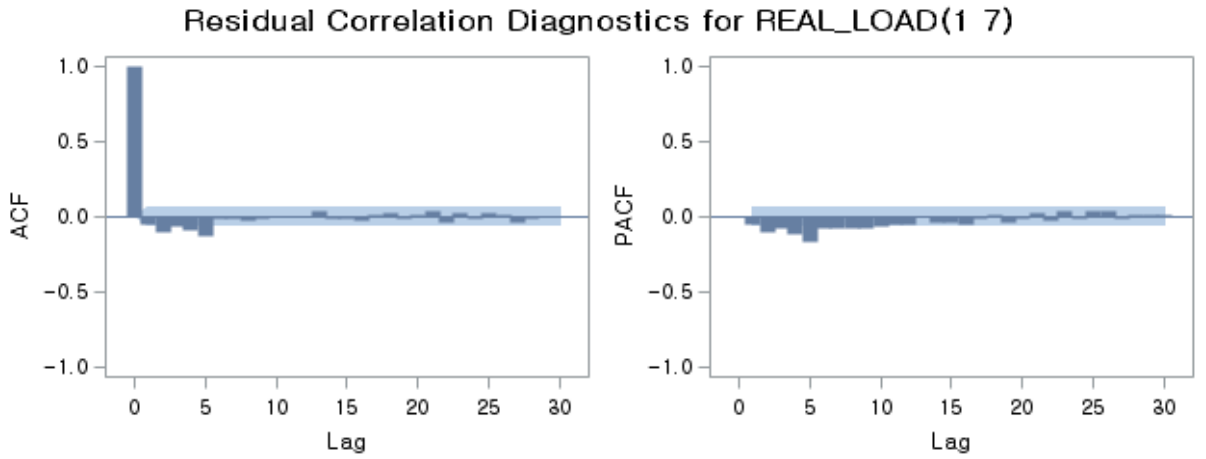
2. 전력수요(1,7) 시계열 p=(7, 14), q=(7) 계절형 모형 추정

   (1) 회귀계수 유의성 검정 => pass

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.97385	0.01093	89.13	<.0001	7
AR1,1	0.14441	0.03201	4.51	<.0001	7
AR1,2	0.06483	0.03191	2.03	0.0422	14

(2) 잔차

- acf, pacf 함수에는 특별한 패턴이 없음 => pass
- 잔차 white noise 검정 => fail, 아직도 잔차에는 추정되어야 하는 pattern 이 남아 있음.



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	45.68	3	<.0001	-0.055	-0.098	-0.069	-0.094	-0.124	-0.013
12	46.34	9	<.0001	-0.005	-0.018	-0.013	0.007	0.004	0.003
18	50.45	15	<.0001	0.042	-0.008	-0.016	-0.027	0.017	0.025
24	55.18	21	<.0001	-0.017	0.008	0.042	-0.034	0.031	-0.009

3. 예측 모형 적합도

1) 추정계수

Model for variable REAL_LOAD	
Period(s) of Differencing	1,7

No mean term in this model.

Autoregressive Factors	
Factor 1:	1 - 0.14441 B**(7) - 0.06483 B**(14)

Moving Average Factors	
Factor 1:	1 - 0.97385 B**(7)

2) MAPE

Analysis Variable : mape
Mean
0.0252248

MAPE=2.52%

3) 이상치

a. additive outlier : 한 시점에서 유의한 크기의 일정 값이 더해진 경우

$$Y_t = \begin{cases} Y_t, t \neq T \\ Y_t + a, t = T \end{cases}$$

b shift outlier : 한 시점 이후 관측치에 일정한 기간 동안 영향을 미침.

Outlier Details				
Obs	Type	Estimate	Chi-Square	Approx Prob>ChiSq
958	Additive	-10343.3	106.04	<.0001
725	Additive	-10257.9	105.70	<.0001
1090	Additive	-10061.6	97.49	<.0001
733	Shift	13726.0	95.23	<.0001
593	Additive	-9340.8	88.29	<.0001

2013/08/15

2012/12/25

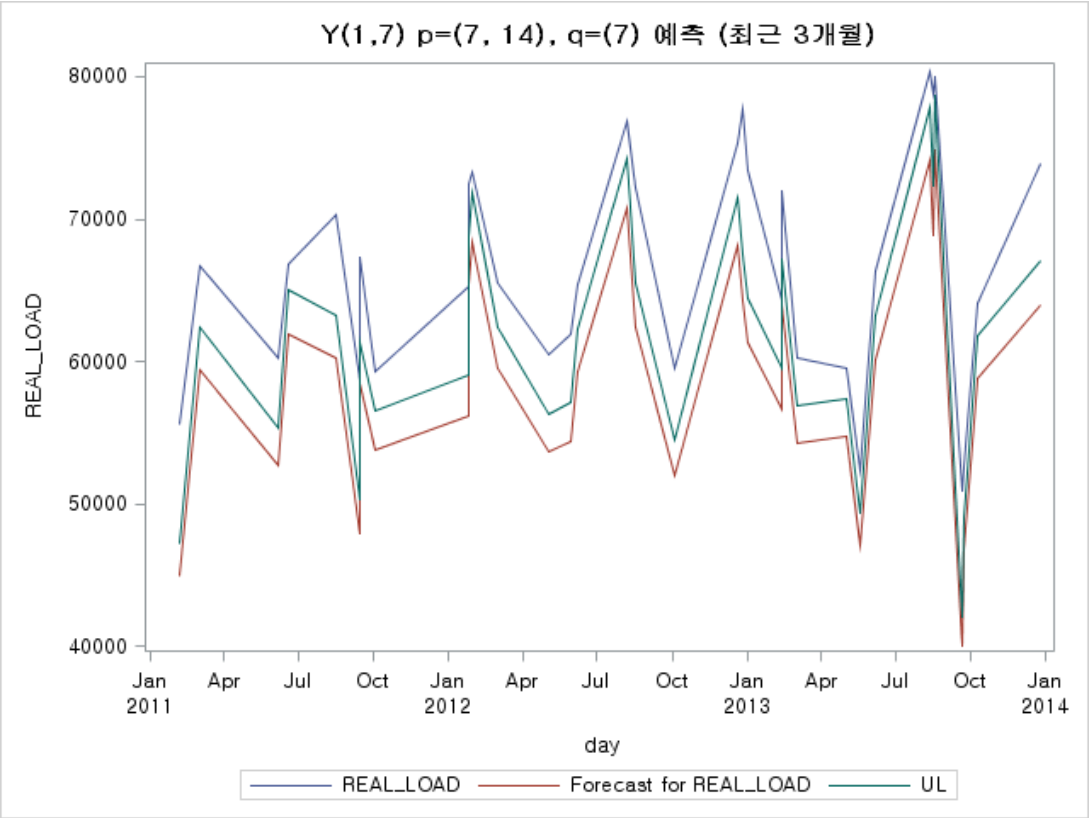
2013/12/25

2013/01/02

4) 예측값 5% 상한, 95%신뢰구간을 벗어난 관측치

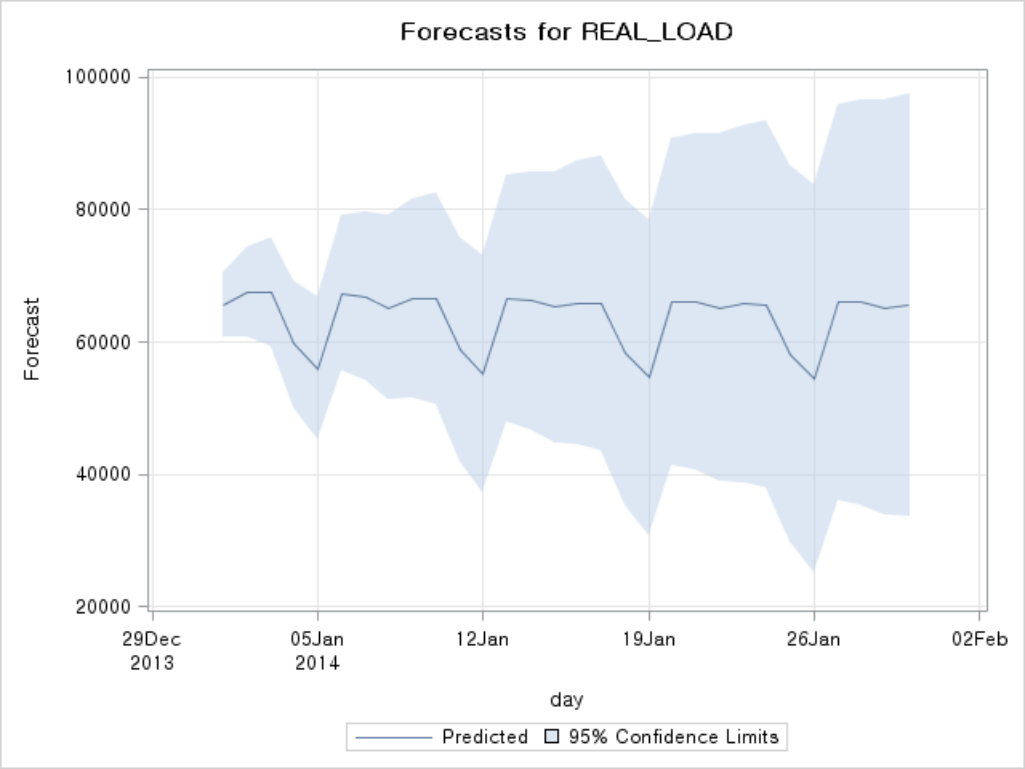
flag05	Frequency	Percent
NO	69	6.17
OK	1049	93.83

flag95	Frequency	Percent
NO	34	3.04
OK	1084	96.96



	요일	전력수요	예측치	5%*상한	95%상한	95%하한
21	02JAN2013	73387.00	61336.13	64402.94	66134.12	56538.15
22	12FEB2013	64266.00	56688.67	59523.11	61486.52	51890.82
23	13FEB2013	71969.00	63952.16	67149.77	68750.01	59154.31
24	02MAR2013	60310.00	54242.18	56954.29	59040.00	49444.36
25	02MAY2013	59540.00	54699.01	57433.96	59496.70	49901.31
26	18MAY2013	52409.00	47049.41	49401.88	51847.09	42251.74
27	07JUN2013	66363.00	60180.24	63189.25	64977.88	55382.59
28	12AUG2013	80366.00	74135.68	77842.47	78933.27	69338.10
29	16AUG2013	78753.00	68864.31	72307.52	73661.89	64066.72
30	19AUG2013	80048.00	74899.48	78644.46	79697.06	70101.90
31	21SEP2013	50921.00	40002.27	42002.38	44799.83	35204.70
32	22SEP2013	52615.00	46673.47	49007.14	51471.03	41875.90
33	10OCT2013	64113.00	58843.98	61786.18	63641.54	54046.42
34	26DEC2013	73938.00	63935.00	67131.75	68732.53	59137.46

5) 향후 30 일간 예측값 및 95% 신뢰구간



D-3. 개입모형 (Daily) ARMA(0,1,0)(2,1,1)$_7$ w/shift 3 개, 휴일, 계절

ARMA 모형은 일정 장기간 기간(본 연구에서는 3 년 일별 데이터)의 시계열 데이터의 패턴을 이용하여 미래 값을 예측하므로 그 기간 중 여러 요인으로 인하여 패턴의 변화가 생길 수 있음.

시계열 패턴에 영향을 줄 수 있는 요인들이 발생하는 시점을 아는 경우 이를 고려한 모형분석을 개입분석이라 하고, 시점을 알지 못하는 경우는 모형 추정 후 사후적으로 알게 되는 "이상점"이다.

개입변수의 형태

a. 지시함수 indicator : 발생 시점에만 영향 ⇔ additive
b. 계단함수 step 발생 시점 이후 일정기간 ⇔ shift

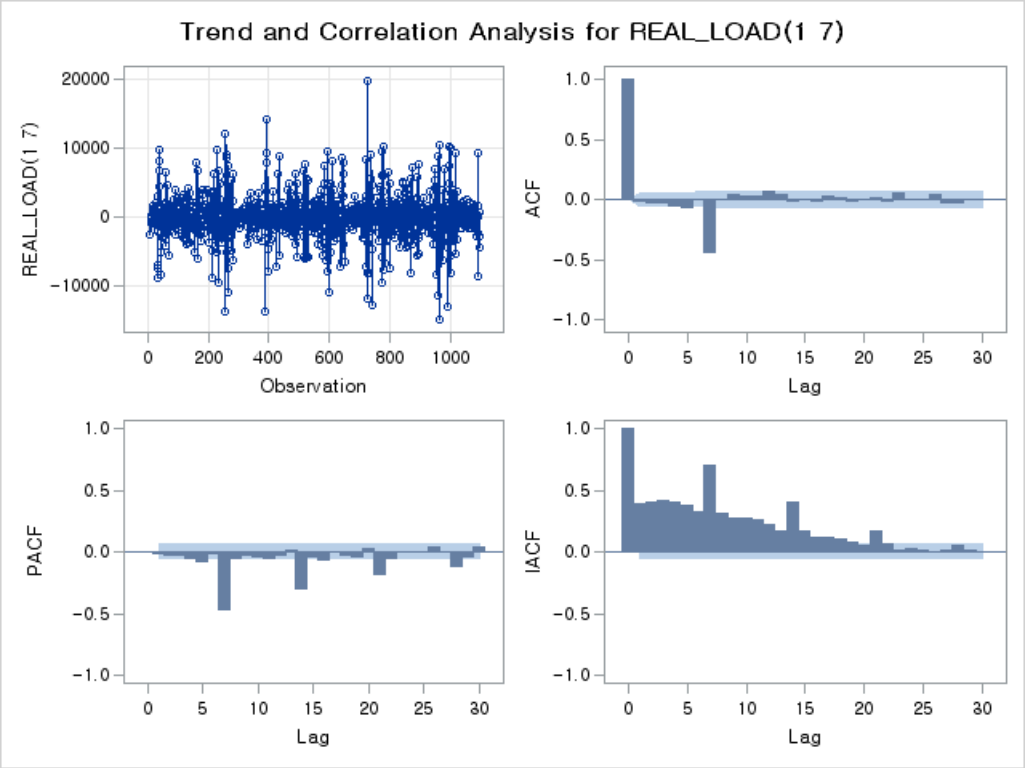
0. 개입 발생 시점 진단

- ARMA 예측모형에서 진단된 shift 이상점 : 2013 년 1 월 2 일
- 휴일 유무
- 전력수요 시계열 패턴이 변화하리라 예상되는 기간

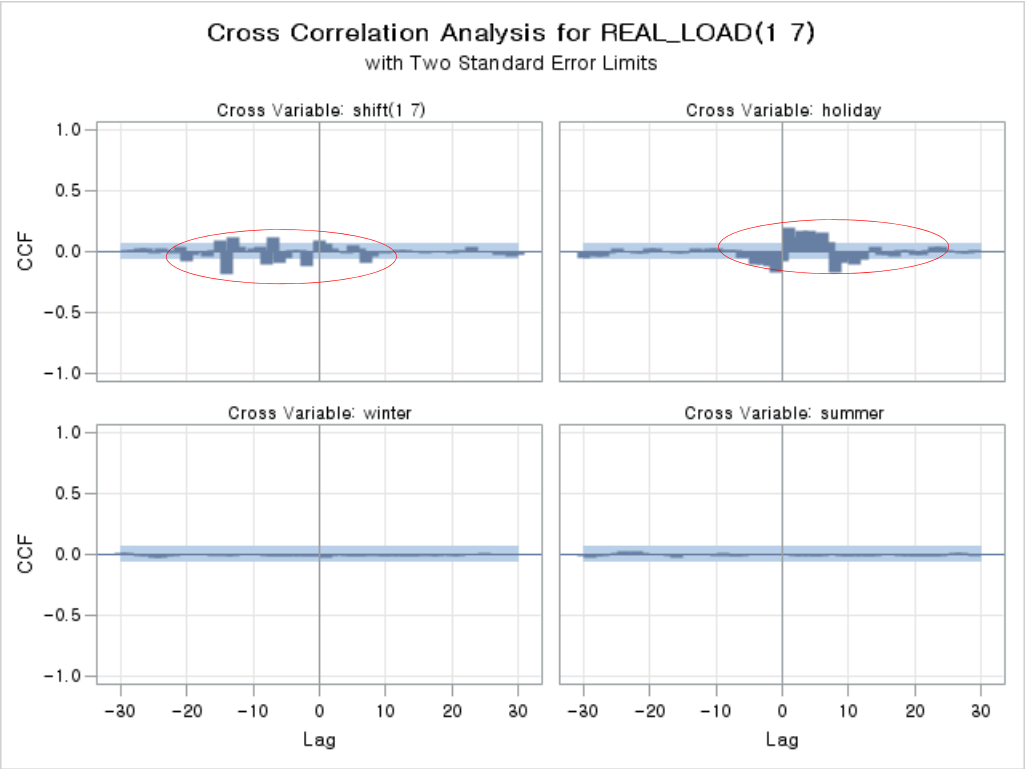
  - 여름(7, 8 월), 겨울(12 월~1 월) : 가장 유의성 높음.
  - 분기 (사분기)

1. 모형진단

- D-2 절의 계절형 ARMA 모형 Y(1,7) (1 차 차분 후 s=7 차분 데이터) => P=(7,14), Q=(7) 예측모형에 개입모형 적용
- 그러므로 D-2 의 Y(1,7)의 acf 와 pacf 함수 형태는 동일함.
- 하여 ARMA 추정 모형은 ARMA(0,1,0)(2,1,1)$_{m=7}$ 동일함.



- 시계열과 개입변수의 상관관계가 유의 ⇔ 빈도가 커 보임.
- Shift 변인과 휴일 변인만 개입변수로 유의함.



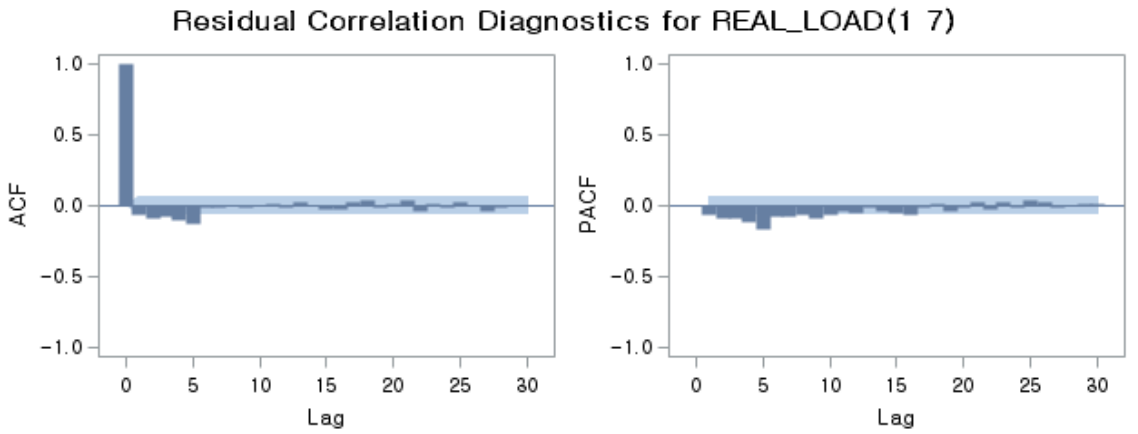
2. 모형 추정

- 분기 효과는 유의하지 않았음.
- Shift, 휴무 변인은 유의수준 5%에서 유의하고 여름 (유의수준=35%), 겨울 (유의수준=15%) 개입은 유의하지 않으나 추정의 정확도를 높이기 위하여 예측모형에 삽입하였음.

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.97533	0.01116	87.40	<.0001	7	REAL_LOAD	0
AR1,1	0.11498	0.03248	3.54	0.0004	7	REAL_LOAD	0
AR1,2	0.08325	0.03223	2.58	0.0098	14	REAL_LOAD	0
NUM1	12719.7	2404.5	5.29	<.0001	0	shift	0
NUM2	173.07347	88.29698	1.96	0.0500	0	holiday	0
NUM3	-49.56661	34.27572	-1.45	0.1481	0	winter	0
NUM4	-29.42968	29.79519	-0.99	0.3233	0	summer	0

- 모형 추정 후 잔차 acf, pacf 함수에는 특이 패턴이 보이지 않으나 잔차의 백색잡음 검정에서는 fail 하여 모형 추정이 완전하게 되지는 않음.



Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	47.40	3	<.0001	-0.059	-0.088	-0.076	-0.097	-0.129	-0.015
12	47.97	9	<.0001	-0.006	0.003	-0.016	0.007	0.010	-0.007
18	52.32	15	<.0001	0.025	-0.005	-0.024	-0.030	0.026	0.034
24	56.35	21	<.0001	-0.015	0.013	0.038	-0.035	0.020	-0.011

3. 모형 적합성

Variance Estimate	5837622
Std Error Estimate	2416.117
AIC	20063.83
SBC	20098.77
Number of Residuals	1088

Analysis Variable : mape
Mean
0.0249225

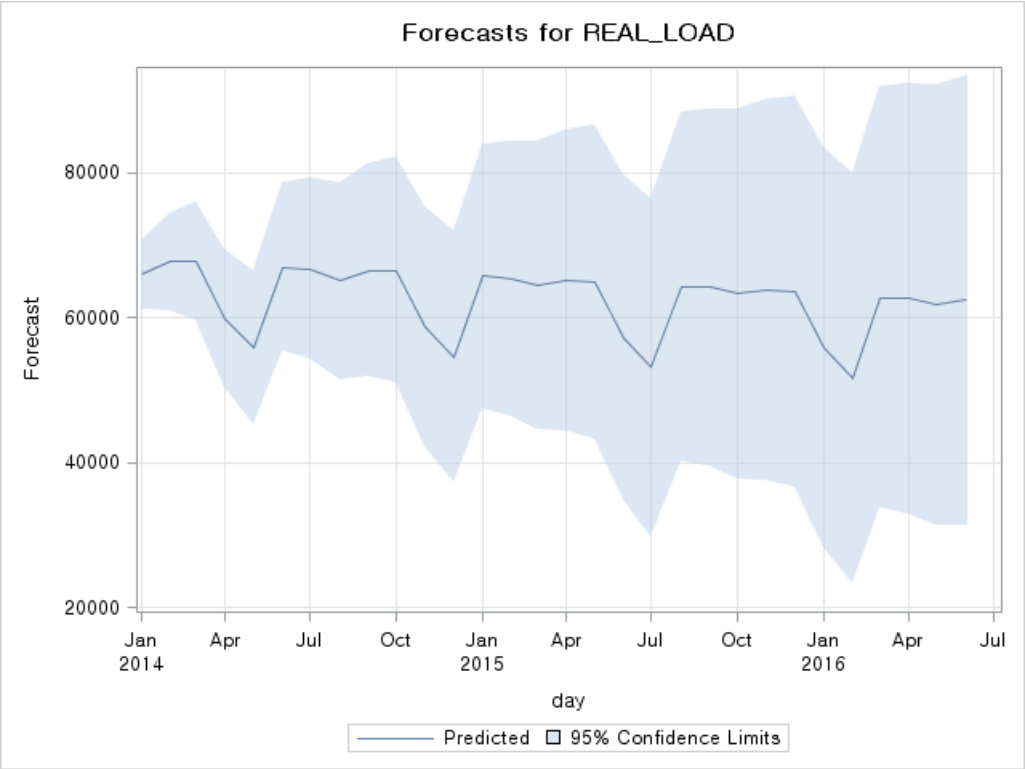
=> MAPE=2.49%

Outlier Details				
Obs	Type	Estimate	Chi-Square	Approx Prob>ChiSq
725	Additive	-11085.7	128.78	<.0001
958	Additive	-10225.3	107.07	<.0001
1090	Additive	-10022.6	101.74	<.0001
593	Additive	-9506.8	95.66	<.0001
388	Shift	-12605.4	85.42	<.0001

2012/12/25
2013/08/15
2013/12/25
2012/08/15
2012/01/23

Shift 이상점 388 ⇔ 2012 년 1 월 23 일 (*)

(향후 30 일간 예측치)



4) 모형 추정 후 shift2 변인 활용 모형 추정 및 적합성

- 2012 년 1 월 23 일=Shift2 변인을 활용하여 개입모형을 추정한 결과 2011 년 9 월 12 일에 다시 shift 가 발생하였음.
- 개입변수 = shift 3 개와 휴일유무, 여름, 겨울을 활용

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.97896	0.01144	85.61	<.0001	7	REAL_LOAD	0
AR1,1	0.12712	0.03247	3.91	<.0001	7	REAL_LOAD	0
AR1,2	0.10211	0.03224	3.17	0.0015	14	REAL_LOAD	0
NUM1	12743.9	2341.8	5.44	<.0001	0	shift	0
NUM2	-12798.3	2291.1	-5.59	<.0001	0	shift2	0
NUM3	-12234.6	2294.5	-5.33	<.0001	0	shift3	0
NUM4	183.39742	86.14705	2.13	0.0333	0	holiday	0
NUM5	-54.44884	33.87314	-1.61	0.1080	0	winter	0
NUM6	-29.57904	29.60429	-1.00	0.3177	0	summer	0

Model for variable REAL_LOAD	
Period(s) of Differencing	1,7

No mean term in this model.

Autoregressive Factors	
Factor 1:	1 - 0.12712 B**(7) - 0.10211 B**(14)

Moving Average Factors	
Factor 1:	1 - 0.97896 B**(7)

Variance Estimate	5547182
Std Error Estimate	2355.246
AIC	20010.92
SBC	20055.85
Number of Residuals	1088

Analysis Variable : mape	
Mean	
0.0244216	

MAPE=2.44%

Outlier Details				
Obs	Type	Estimate	Chi-Square	Approx Prob>ChiSq
725	Additive	-11085.6	130.15	<.0001
958	Additive	-10243.7	110.15	<.0001
1090	Additive	-10038.1	102.47	<.0001
593	Additive	-9547.4	97.58	<.0001
992	Shift	-12191.4	79.45	<.0001

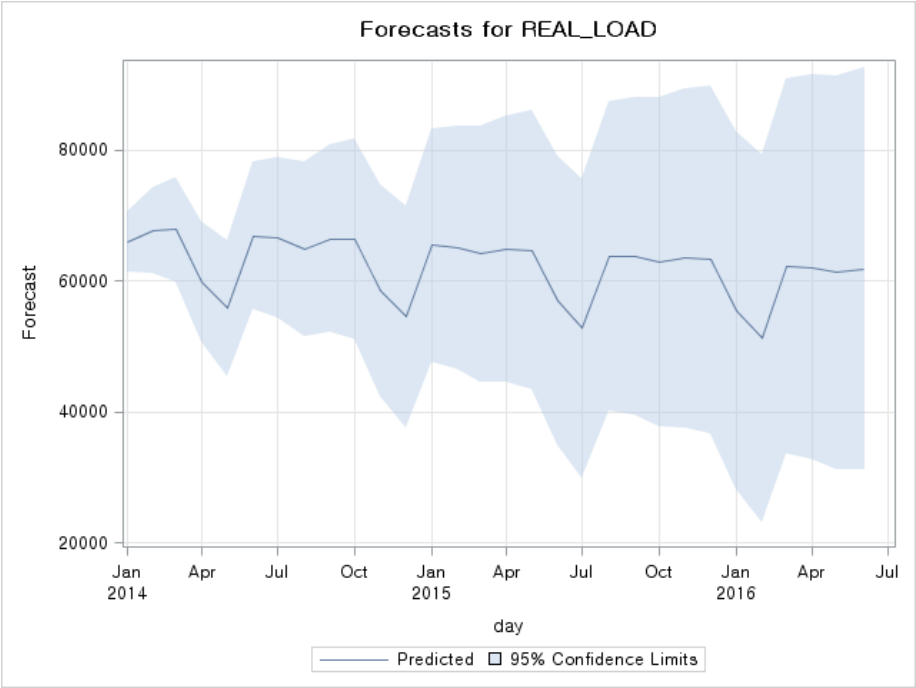
2012/12/25

2013/08/15

2013/12/25

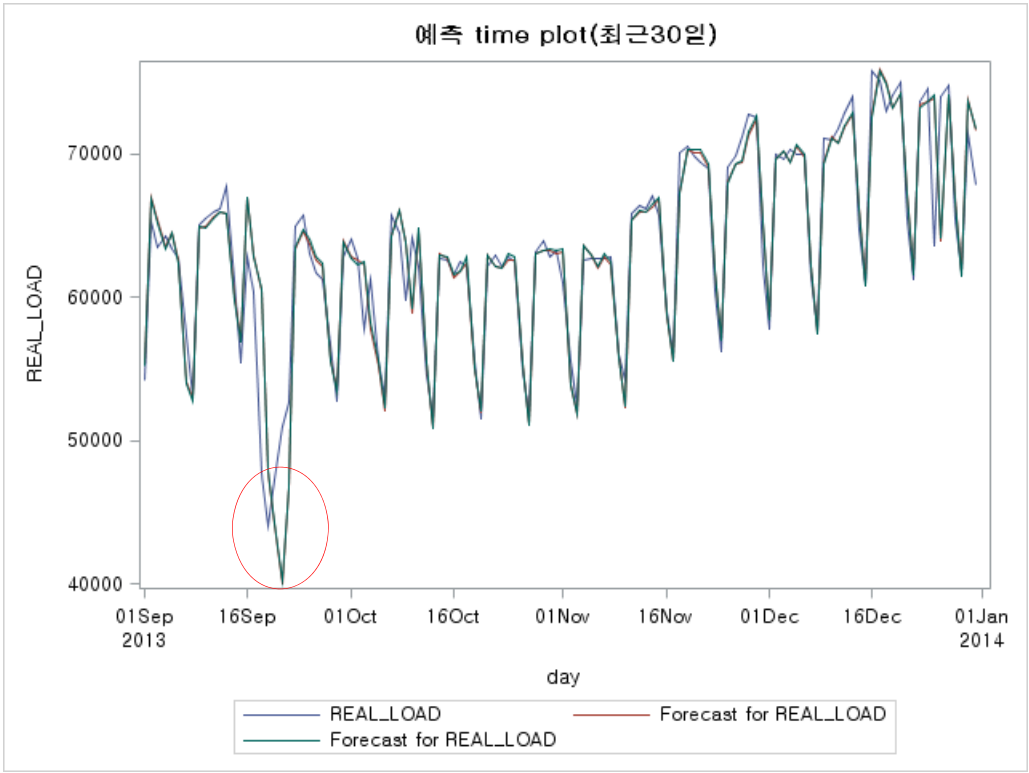
2012/08/15

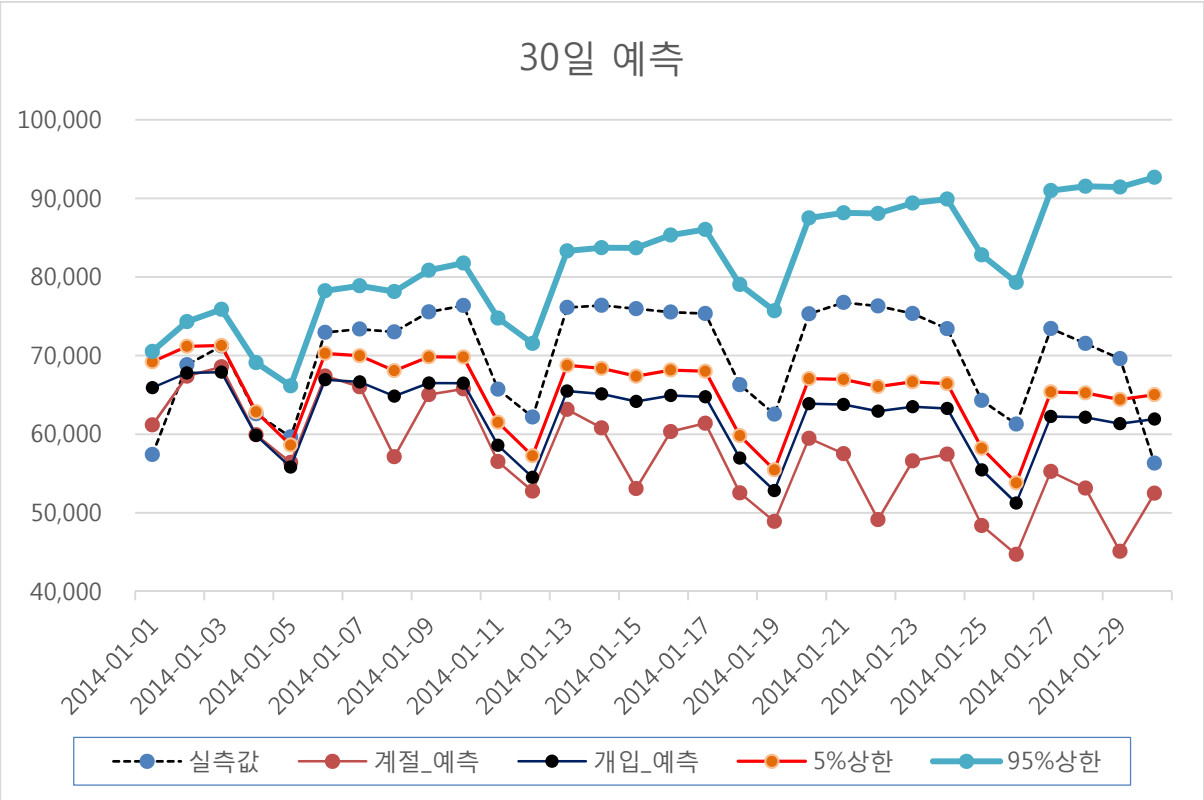
❍ 관측값 992 = 2013 년 9 월 18 일, shift 개입변인으로 개입모형을 추정한 결과 유의하지 않았음.



5. 향후 예측값

(계절 ARMA _ 개입 모형 비교) 최근 30 일





D-4. 동적 회귀모형 (Daily)

0. 개요

- ARMA 모형은 전력수요량의 과거 관측값(AR)과 과거 관측값으로 설명되지 않는 항(MA)에 의해 미래 값을 예측함.
- 동적모형은 설명변수(온도, 습도) 등을 고려함.

1. 모형 설정

목표변수 : 전력수요량

설명변수

- 전력수요량 시차, lag=1, 7
- 피크 시각
- 온도 : 최대값, 겨울에는 최저 온도 사용.
- 습도 : 최대값
- 바람 : 최대값
- 휴일여부
- 분기, 계절(겨울, 여름) 등

2. 모형 추정

❍ 유의수준 10%에서 제거된 변수

Summary of Backward Elimination							
Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
WIND_SPEED	WIND_SPEED	14	0.0000	0.7696	14.0002	0.00	0.9902
q2		13	0.0000	0.7696	12.0012	0.00	0.9744
lag_h		12	0.0001	0.7695	10.3978	0.40	0.5286
winter		11	0.0005	0.7691	10.5849	2.19	0.1390

❍ 전력수요량 높이는 변인 : 전날 전력수요, 일주일 전 전력수요, 어제 온도, 2013 년 1 월 이후, 여름, 1 분기, 3 분기

❍ 전력수요량 낮추는 변인 : 온도 (겨울의 영향), 습도, 휴일, **피크시각 (엑셀 데이터)**

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	11475	1835.12265	6.25	<.0001	0
lag1	1	0.23548	0.02012	11.70	<.0001	0.23545
lag7	1	0.63798	0.01904	33.50	<.0001	0.63917
TEMPERATURE	1	-137.10945	38.52442	-3.56	0.0004	-0.24253
lag_t	1	89.17416	38.76211	2.30	0.0216	0.15793
peak_hr	1	-53.22973	24.66892	-2.16	0.0312	-0.03810
HUMIDITY	1	-33.12828	9.79743	-3.38	0.0007	-0.05840
holiday	1	-4918.79934	397.98458	-12.36	<.0001	-0.19152
shift	1	391.60809	219.70143	1.78	0.0750	0.02695
summer	1	2375.90818	414.83563	5.73	<.0001	0.13115
q1	1	831.46759	311.20038	2.67	0.0077	0.05270
q3	1	992.89067	277.89627	3.57	0.0004	0.06332

3. 모형적합성

Root MSE	3304.41111	R-Square	0.7691

Analysis Variable : mape
Mean
0.0380527

=> MAPE=3.81%