더위로 아이스크림 판매량 예측하기

1951 ~ 1953년, 미국인들은 왜 아이스크림을 사먹었나?

데이터 요약

- 1951 3월 18일부터 ~ 1953년 7월 11일까지,
- 4주를 묶음하여 만든 총 30개 시점의 데이터

데이터명	설명
Cons	아이스크림 소비량,(인당 파인트)
Average Income	가구당 평균 소득(달러)
Price	아이스크림의 가격(파인트당 달러)
Temp	4주간 평균 기온(화씨)

ARIMAX 모델 적합 프로시져

- 사전 백색화(Pre-whitening)

- 출력 계열을 안정된 ARIMA 모형으로 적합
- 입력 계열을 위에서 적합된 구조와 계수를 이용하여 적합
- 두 계열의 잔차를 여과

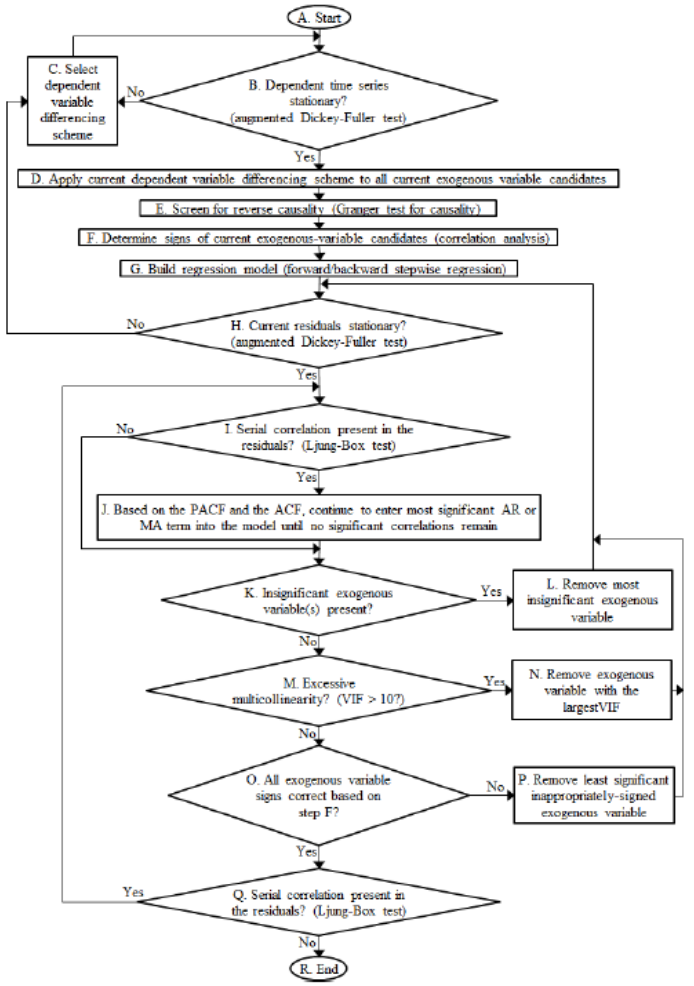
- 그레인저 인과성 검정
- 반드시 입력계열만이 출력계열에 영향을 미쳐야하며, 그 반대는 인과성 위반

- 양 쪽의 잔차들을 이용하여 CCF 구조 파악
- 최초 언제 상관관계가 발생하는가?
- CCF는 줄어드는가 늘어나는가?

- CCF에서 추정한 모수를 이용해 충격반응가중치 추정, 모델 가적합

- 가적합 모형의 잔차 검정
- -> 시계열성이 존재하는 경우, 잔차로 ARMA 모형 추가 적합
- -> 시계열성이 존재하지 않는 경우, 적합 완료

Figure 8. ARIMAX model-building algorithm.



데이터 구조

- 30개 시점의 아이스크림 판매량과 온도, 소득, 가격의 데이터
- 25개 시점은 적합용으로 쓰고, 5개는 사후 검정용으로 활용

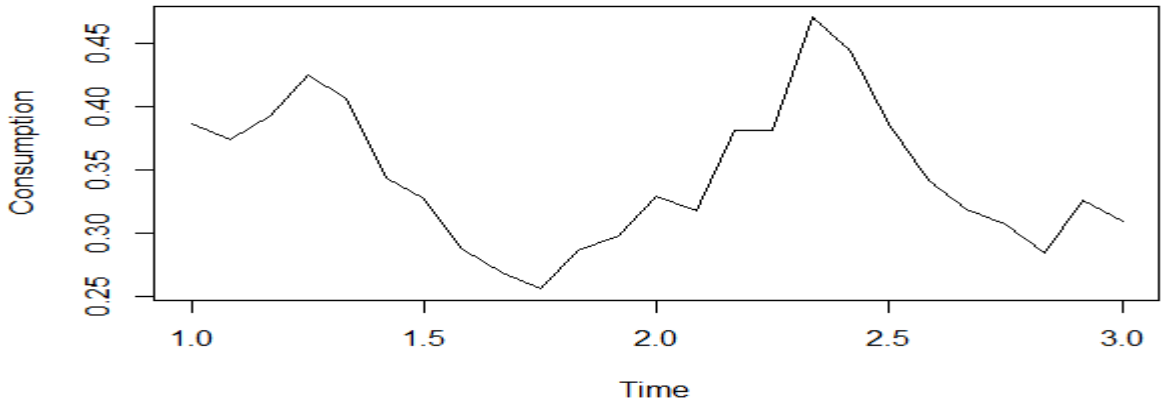
	X	cons	income	price	temp
1	1	0.386	78	0.270	41
2	2	0.374	79	0.282	56
3	3	0.393	81	0.277	63
4	4	0.425	80	0.280	68
5	5	0.406	76	0.272	69
6	6	0.344	78	0.262	65
7	7	0.327	82	0.275	61
8	8	0.288	79	0.267	47
9	9	0.269	76	0.265	32
10	10	0.256	79	0.277	24
11	11	0.286	82	0.282	28
12	12	0.298	85	0.270	26
13	13	0.329	86	0.272	32
14	14	0.318	83	0.287	40
15	15	0.381	84	0.277	55
16	16	0.381	82	0.287	63
17	17	0.470	80	0.280	72

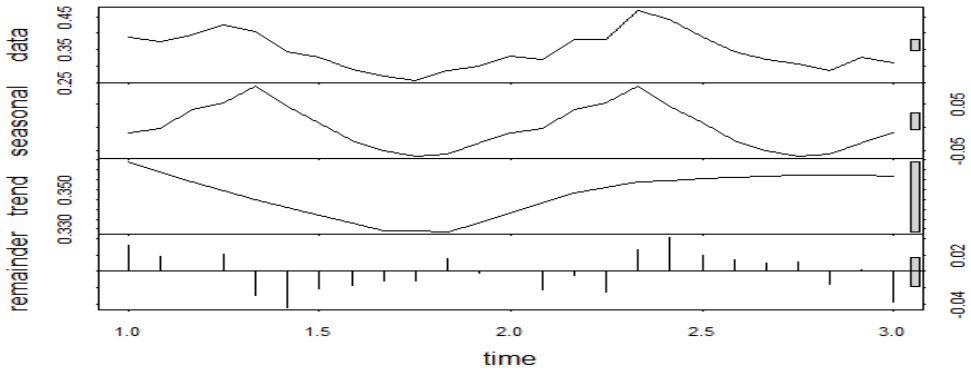
컬럼명	설명
Cons	아이스크림 소비량,(인당 파인트)
Income	가구당 평균 소득 (달러)
Price	아이스크림의 가격 (파인트당 달러)
Temp	4주간 평균 기온 (화씨)

사전 백색화

- ARIMA 모형 적합

- 소비량(cons)



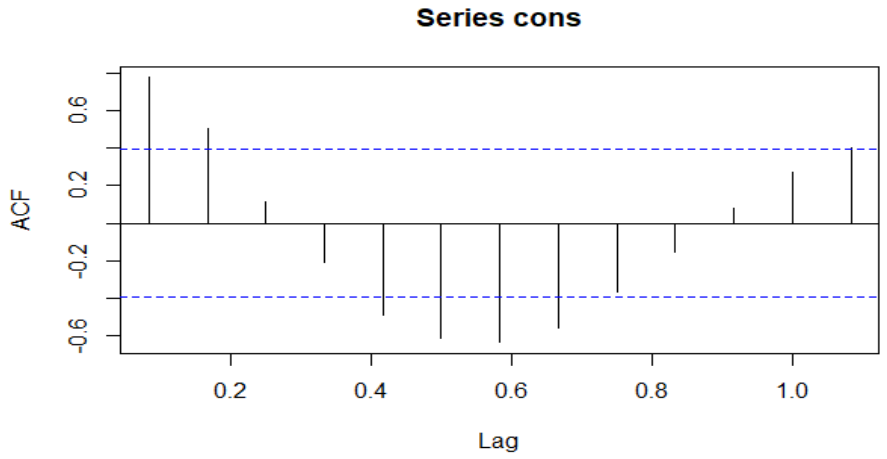


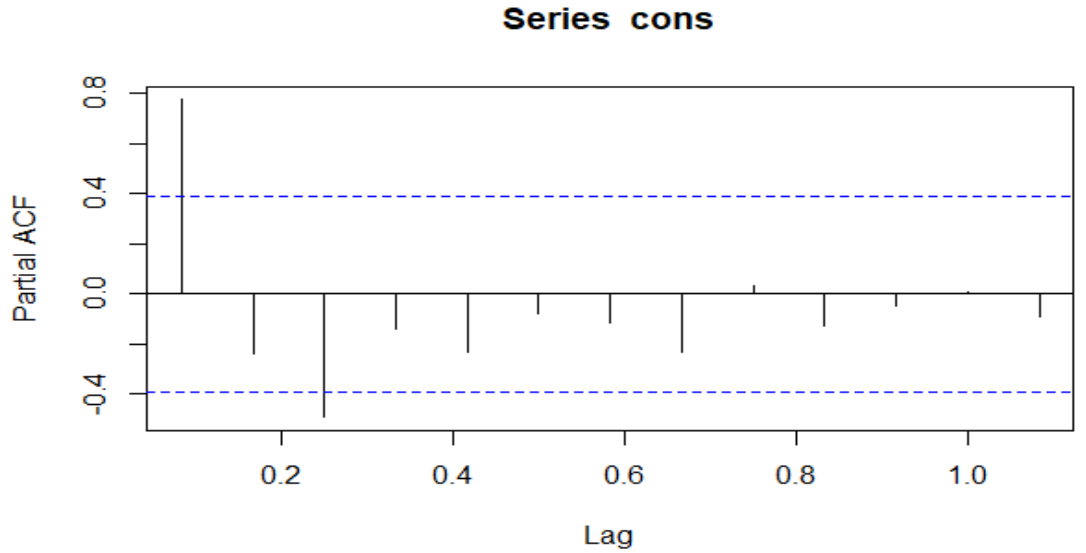
- 추세는 존재하지 않는 것으로 보인다
- 하지만, 계절성은 분명히 존재한다.

사전 백색화

- ARIMA 모형 적합

- 소비량(cons)





- ACF 그래프가 지수함수적으로 줄어들지 않는다.
- 단위근이 존재하는 것으로 의심할 수 있다.

사전 백색화

- ARIMA 모형 적합
  - 소비량(cons)

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag    ADF p.value
[1,]   0 -0.674   0.428
[2,]   1 -0.581   0.461
[3,]   2 -0.662   0.432
Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -1.61   0.470
[2,]   1 -1.89   0.371
[3,]   2 -2.91   0.062
Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -1.59   0.720
[2,]   1 -1.87   0.609
[3,]   2 -2.82   0.254
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
> |
```

(1) BOX-test

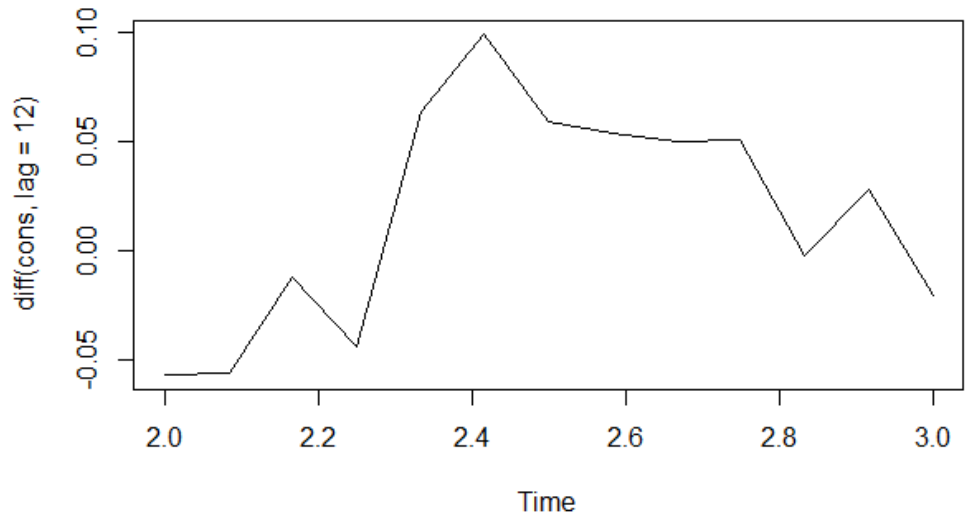
1) 테스트 결과, 상수항은 있고 추세는 없는 모형의 경우 p.value = 0.470

2) 이는 단위근이 존재한다는 귀무가설을 기각하기에는 매우 높은 확률

사전 백색화

- ARIMA 모형 적합

- 소비량(cons)



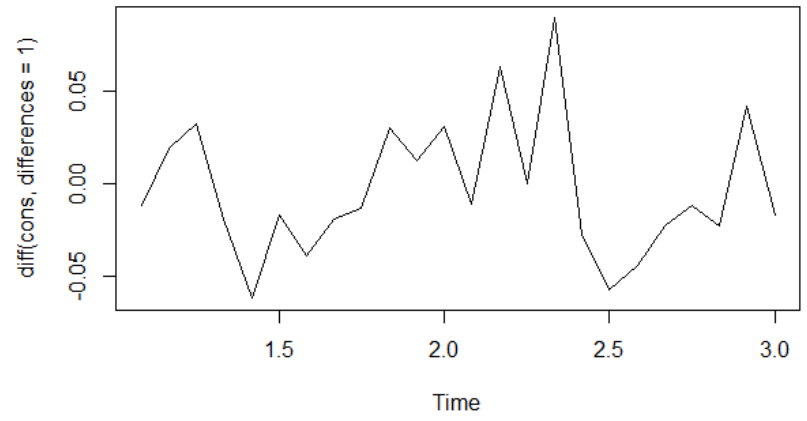

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -1.73  0.0819
[2,]   1 -1.46  0.1486
[3,]   2 -1.19  0.2458
Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -1.95   0.350
[2,]   1 -1.93   0.357
[3,]   2 -1.59   0.477
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -1.321   0.828
[2,]   1 -1.043   0.915
[3,]   2 -0.736   0.956
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
>
```

- 계절 차분후 다시 ADF 테스트를 해본 결과,
- 상수항이 없고 추세가 없는 모형에서 p-value는 0.08로, 충분히 감당 가능한 p-value가 도출되었다.

사전 백색화

- ARIMA 모형 적합

- 소비량(cons)




```
Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -4.04  0.0100
[2,]   1 -2.34  0.0218
[3,]   2 -2.71  0.0100
Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -3.97  0.0100
[2,]   1 -2.31  0.2187
[3,]   2 -2.72  0.0891
Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -3.87  0.0307
[2,]   1 -2.24  0.4619
[3,]   2 -2.65  0.3146
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
> |
```

- 한편, 일반 1계차분의 경우 더 파워풀한 결과가 나왔다.
- 이후 모형 적합은 계절차분이 아닌 1계차분을 기준으로 진행

사전 백색화

- ARIMA 모형 적합

- 소비량(cons)

```
cons_diff = diff(cons,differences=1)
income_diff = diff(income,differences = 1)
price_diff = diff(price,differences = 1)
temp_diff = diff(temp,differences = 1)
```

```
a <- arima(cons_diff, order=c(1,0,0))
```

```
b = arima(income_diff,order=c(1,0,0),fixed=c(a$coef))
c = arima(price_diff,order=c(1,0,0),fixed=c(a$coef))
d = arima(temp_diff, order=c(1,0,0),fixed=c(a$coef))
```

1) ARIMA(1,0,0),(0,1,0)[12]로 모형을 적합하고,

2) 그 구조를 투입계열 변수들에도 똑같이 적용한다.

그래인저 인과성 검정

- 개요

- 자기 자신의 AR모형과, 투입 변수를 포함하는 다변량 AR 모형의 유의함 여부를 검정

- F통계량을 활용하여 검정함

그래인저 인과성 검정

- 검정 결과

- Cons ~ Income(계절차분)

```
Model 1: cons_diff ~ Lags(cons_diff, 1:3) + Lags(income_diff, 1:3)
Model 2: cons_diff ~ Lags(cons_diff, 1:3)
  Res.Df Df      F Pr(>F)
1     14
2     17 -3 0.5345 0.6661
```

```
Model 1: income_diff ~ Lags(income_diff, 1:3) + Lags(cons_diff, 1:3)
Model 2: income_diff ~ Lags(income_diff, 1:3)
  Res.Df Df      F Pr(>F)
1     14
2     17 -3 0.7978 0.5154
> |
```

양 방향 모두에 그래인저 인과성이 존재하지 않는다.

- Cons ~ Price (계절 차분)

```
Model 1: cons_diff ~ Lags(cons_diff, 1:3) + Lags(price_diff, 1:3)
Model 2: cons_diff ~ Lags(cons_diff, 1:3)
  Res.Df Df      F Pr(>F)
1     14
2     17 -3 1.2481 0.3298
```

```
Model 1: price_diff ~ Lags(price_diff, 1:3) + Lags(cons_diff, 1:3)
Model 2: price_diff ~ Lags(price_diff, 1:3)
  Res.Df Df      F Pr(>F)
1     14
2     17 -3 1.2693  0.323
> |
```

양 방향 모두에 그래인저 인과성이 존재하지 않는다.

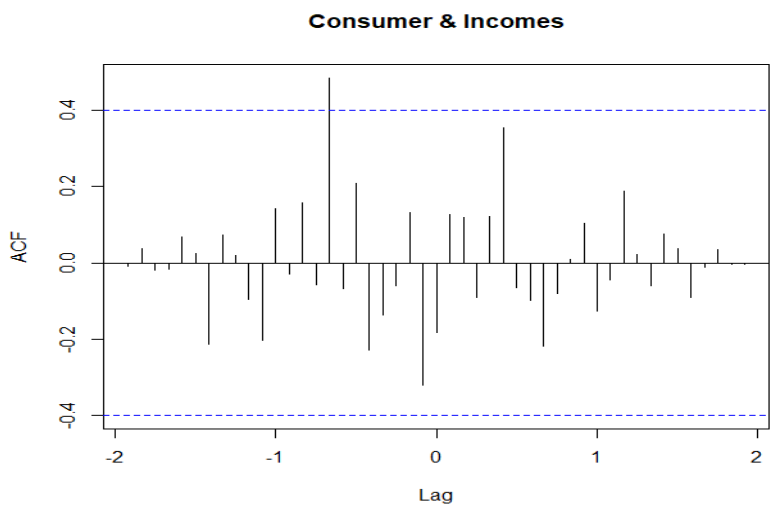
- Cons ~ Temp(계절차분)

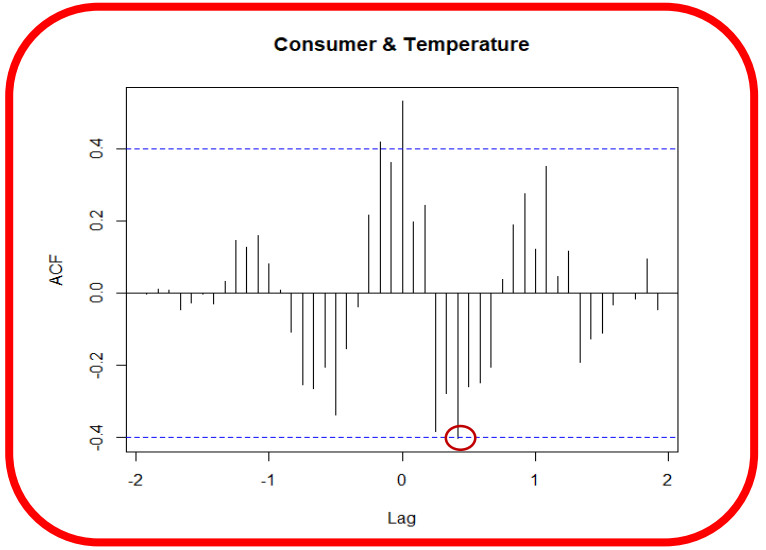
```
Model 1: cons_diff ~ Lags(cons_diff, 1:3) + Lags(temp_diff, 1:3)
Model 2: cons_diff ~ Lags(cons_diff, 1:3)
  Res.Df Df      F  Pr(>F)
1     14
2     17 -3 3.8173 0.03442 *
```

```
Model 1: temp_diff ~ Lags(temp_diff, 1:3) + Lags(cons_diff, 1:3)
Model 2: temp_diff ~ Lags(temp_diff, 1:3)
  Res.Df Df      F Pr(>F)
1     14
2     17 -3 1.1083 0.3787
```

Temp -> cons 방향으로 그래인저 인과성이 존재한다.

CCF 구조 확인





- 유일한 상관관계는 판매량 – 온도에서 나타났으며,
- 지연모수는 0이고 투입계열 모수 0인 충격반응가중치를 상정할 수 있다.

$$Y_t = \frac{w_0 B_0}{1 - \delta_5 B_5} X_t + \varphi$$

모형 적합

- 가적합

```
Call:
arimax(x = su1[, 1], order = c(1, 0, 0), fixed = c(NA, NA, 0, 0, 0, 0, NA, NA),
    xtransf = su1[, 2], transfer = list(c(5, 0)))

Coefficients:
          ar1  intercept  T1-AR1  T1-AR2  T1-AR3  T1-AR4   T1-AR5  T1-MA0
      -0.1594    -0.0020       0       0       0       0  -0.1501  0.0023
s.e.   0.3141     0.0054       0       0       0       0   0.8228  0.0008

sigma^2 estimated as 0.0008078:  log likelihood = 51.39,  aic = -94.77
>
```

- 가적합 잔차를 ARMA로 적합하는 것을 Y와 잔차의 ARIMA로 보는 외국 논문에 맞게 적합을 시도

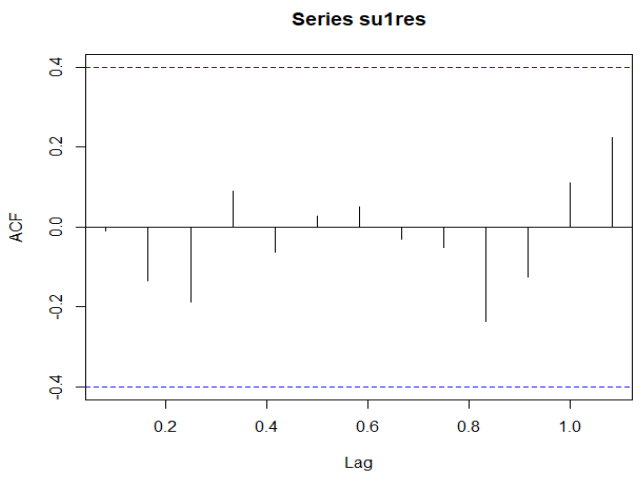
- 잔차 검정 결과 가적합 잔차 e는 ARMA(1,0)을 의심할 수 있으므로, 이를 order에 반영한다.

모형 적합

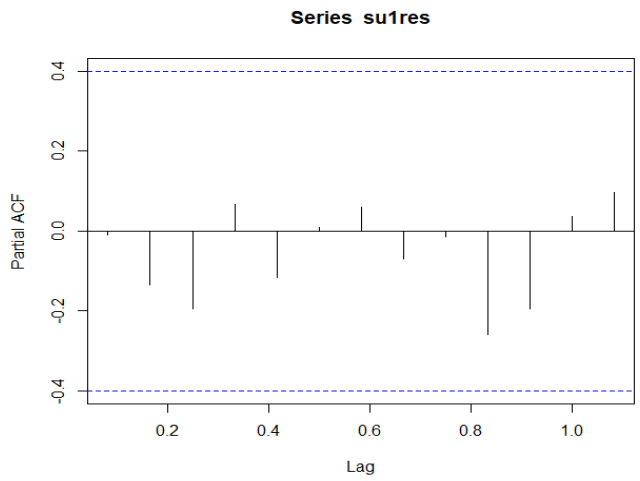
- Box – test 및 잔차 검정

```
        Box-Ljung test

data:  su1res
X-squared = 0.0025168, df = 1, p-value = 0.96
```

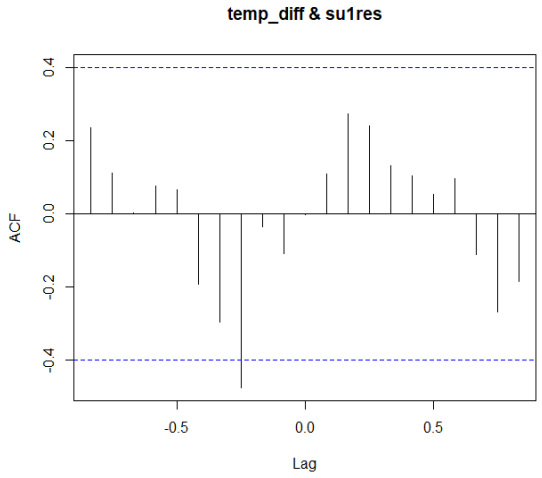
<잔차의 acf>



<잔차의 PACF>



<투입계열과 잔차의 CCF>



BOX-test 결과 잔차의 시계열성은 제거되었고, 완전한 백색잡음이 되었다.

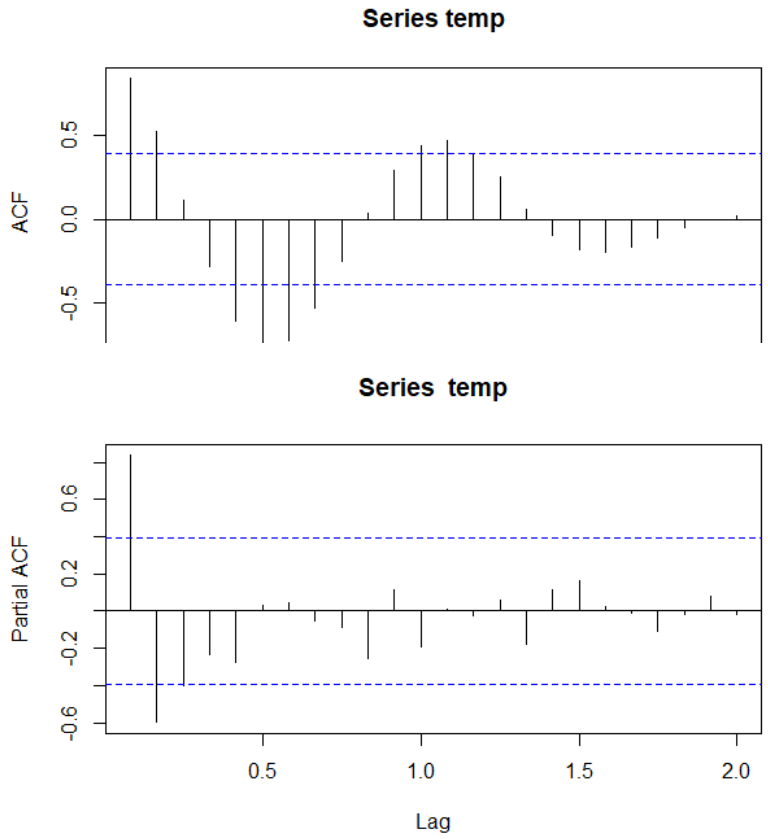
모형 테스트

- 예측값 생성

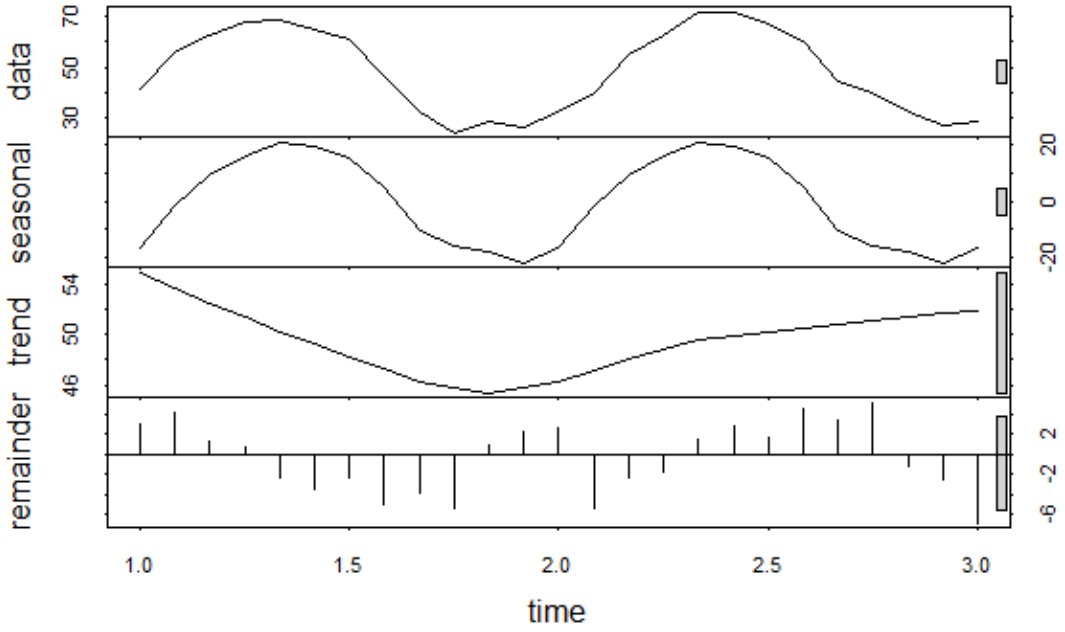
- 예측을 위해 투입계열인 temp에 대한 5차시 예측값을 생성한다.
- 이를 미리 빼둔 26 ~ 30차시의 원래 데이터와 비교하여 오차율을 본다.

모형 테스트

- ARIMA 모형 적합
  - 온도(temp)



산출계열인 cons와 마찬가지로
온도 역시 강력한 계절성을 의심해볼 수 있다.



모형 테스트

- ARIMA 모형 적합

- 온도(temp)

<계절 차분>

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -1.30  0.2074
[2,]   1 -1.83  0.0673
[3,]   2 -2.16  0.0332
Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -1.33   0.569
[2,]   1 -2.16   0.272
[3,]   2 -2.34   0.208
Type 3: with drift and trend
     lag    ADF p.value
[1,]   0 -0.408   0.979
[2,]   1  0.312   0.990
[3,]   2 -0.161   0.990
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

<일반 1계차분>

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -2.46  0.0174
[2,]   1 -1.97  0.0487
[3,]   2 -2.52  0.0151
Type 2: with drift no trend
     lag   ADF p.value
[1,]   0 -2.46   0.163
[2,]   1 -1.98   0.338
[3,]   2 -2.48   0.159
Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 -2.21   0.474
[2,]   1 -1.82   0.628
[3,]   2 -2.36   0.422
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
> |
```

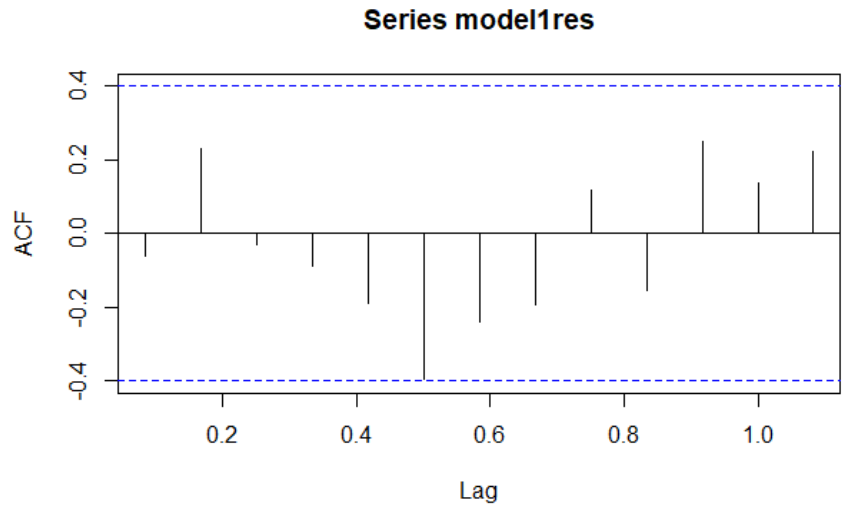
가설과는 달리, 계절차분은 단위근을 제거하지 못했다. 1계 차분을 이용하자

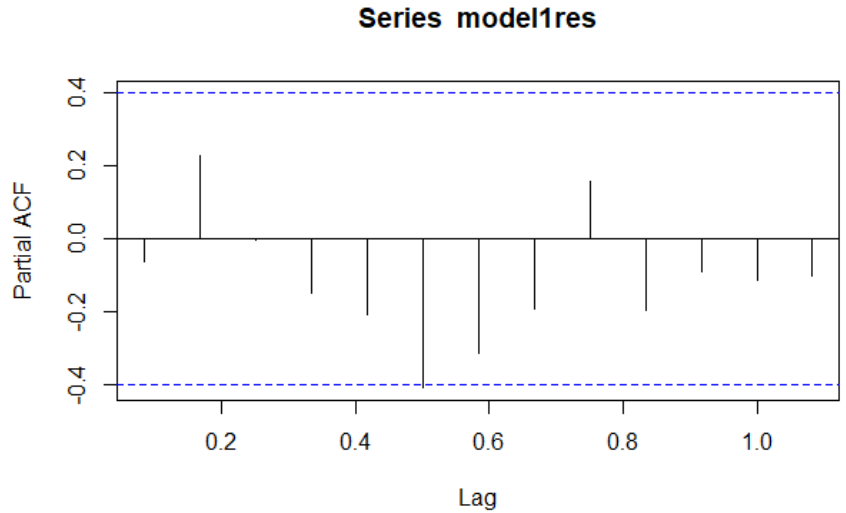
모형 테스트

- ARIMA 모형 적합

- 온도(temp)

```
        Box-Ljung test

data:  model1res
X-squared = 0.10284, df = 1, p-value = 0.7484

> |
```





적합된 모형의 잔차에서 시계열성은 완전히 제거되었고, 잘 적합되었다.

모형 테스트

- Summary
- - 오차율

	26차시	27차시	28차시	29차시	30차시
실제값	0.359	0.376	0.416	0.437	0.548
예측	0.308	0.357	0.373	0.422	0.433
오차율	14.04%	4.91%	10%	3.2%	20%

MAPE = 10.6763

Chart 24. ARIMAX four-quarter rolling MAPEs and MADs (Q1 1988–Q4 2012).

Estimate Type	Goodness-of-Fit Measures	n	Mean	Std. Dev.	Min.	Max.
Fit	MAPE	100	3.20	1.62	0.79	9.14
	MAD	100	0.10	0.06	0.02	0.32
Forecast	MAPE	100	4.06	2.24	0.62	12.07
	MAD	100	0.12	0.07	0.02	0.37

참조 – 참고논문의 MAPE

시사점

- 변수들의 Log변환 필요

- 1차 가적합으로 여과된 잔차에 추가 투입변수를 적용하여 예측력이 개선된 2차 이상의 다변량 모형 적합 도전 필요

감사합니다.