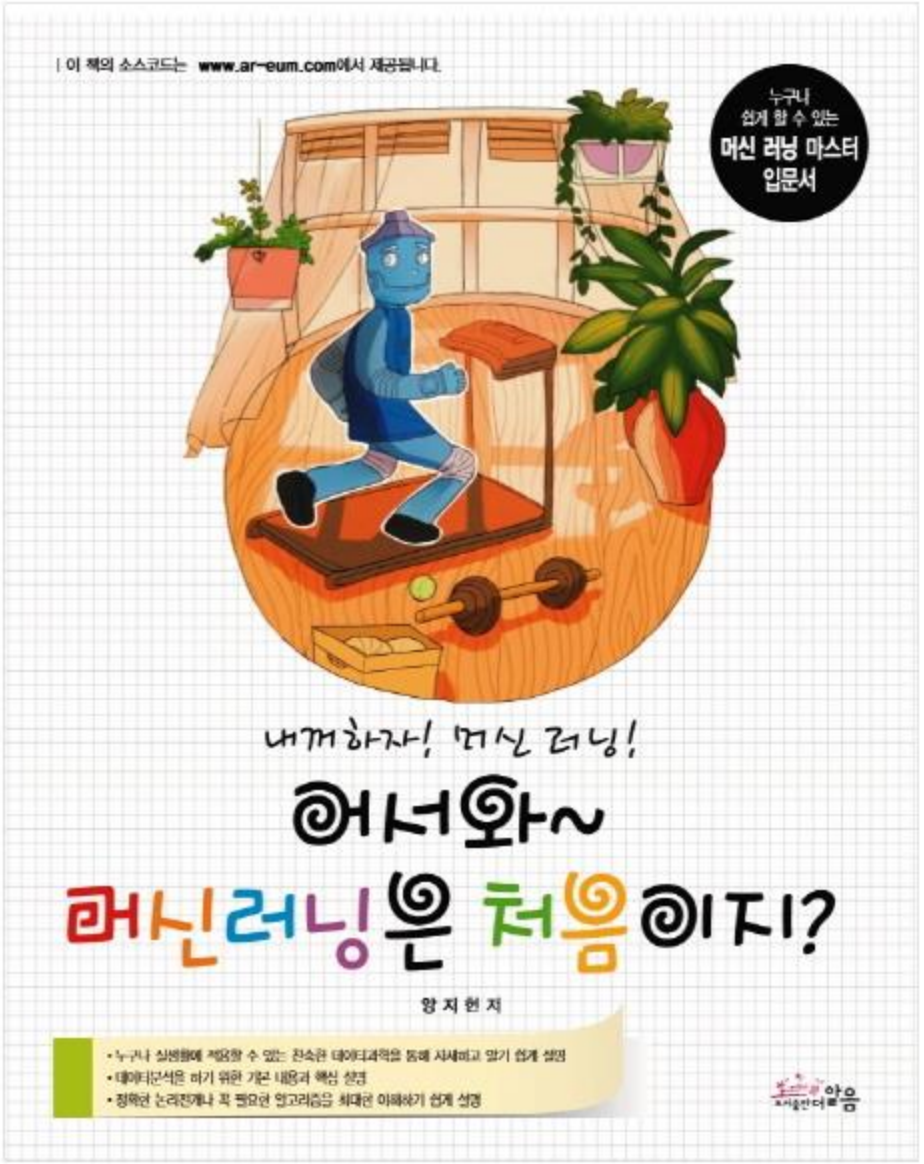
[어서와~ 머신러닝은 처음이지?]

7장. Time Series Analysis

- 장형석
- 국민대학교 빅데이터경영MBA과정 교수
- 숙명여자대학교 빅데이터센터 연구소장
- chjang1204@nate.com
- 010-3302-5543



- **저자 : 양지헌**
  - 숭실대학교 물리학과 졸업
  - 국민대학교 빅데이터경영MBA과정 석사졸업
  - 국민대학교 데이터사이언스 박사과정
  - 전) VTW 컨설턴트
  - 전) 글로벌텔레콤 IOT 분석팀장
  - 데이터 분석 전문가(ADP) 자격 보유
- **출판사 : 더알음**
- **출간일 : 2016년 12월 21일**
- **ISBN : 9791195484737**
- **http://www.ar-eum.com**

1. 생활에서 만나는 문제

1) 예측 패턴?

효성이의 어머니는 재래시장에서 반찬가게를 하고 있다. 워낙 음식솜씨가 좋으셔서 동네에 소문이 많이 났지만 특히 김치가 인기가 가장 많다. 그런데 김치는 미리 배추를 사서 담가 놓아야 하는데 어느 정도의 수요가 될지 예측이 힘들다. 너무 많이 만들어 놓고 버리거나 너무 적게 만들어서 팔고 싶어도 못 판다면 아마도 어느 쪽이나 손해가 되는 일일 것이다. 딸 효성이는 전반적인 김치판매량에 대한 예측모델을 만들어서 대략이라도 수요를 예상해보기로 하였다. 그리고 우선은 상공회의소에서 김치판매량에 대한 데이터를 무료로 구하였다. 이제는 시계열 예측모델 모델을 만들어 볼 차례이다. 시간에 따라서 김치의 판매량이 변동이 된다? 정말 그곳에 어떤 패턴이 존재하는 걸까? 그리고 그 패턴에 따라서 예측이 어느 정도나 가능한 것일까?

1. 생활에서 만나는 문제

2) 데이터셋

효성이는 다음과 같은 데이터를 만들었다.

```
> kimchi <- read.csv("kimchi.csv" , header = T)
> head(kimchi)
  YYWW 주마지막일자 대형마트수량 대형마트금액 백화점수량 백화점금액 수퍼수량 수퍼금액 편의점수량 편의점금액
1 1301     20130106        27916    233968900       11971    99796735      1795 11561690        1603     2264200
2 1302     20130113        23057    194593960       11678   103106940      1832 11493710        2149     3073450
3 1303     20130120        25153    216322950       11634   106922870      1783 11541050        2277     3368720
4 1304     20130127        24645    218053670       12102    95765955      1949 11773330        2169     3172410
5 1305     20130203        26603    236095130       11404   102729615      1988 12808900        2085     3147500
6 1306     20130210        25509    245607800       11689   117911615      1829 12558230        2185     3178160
```

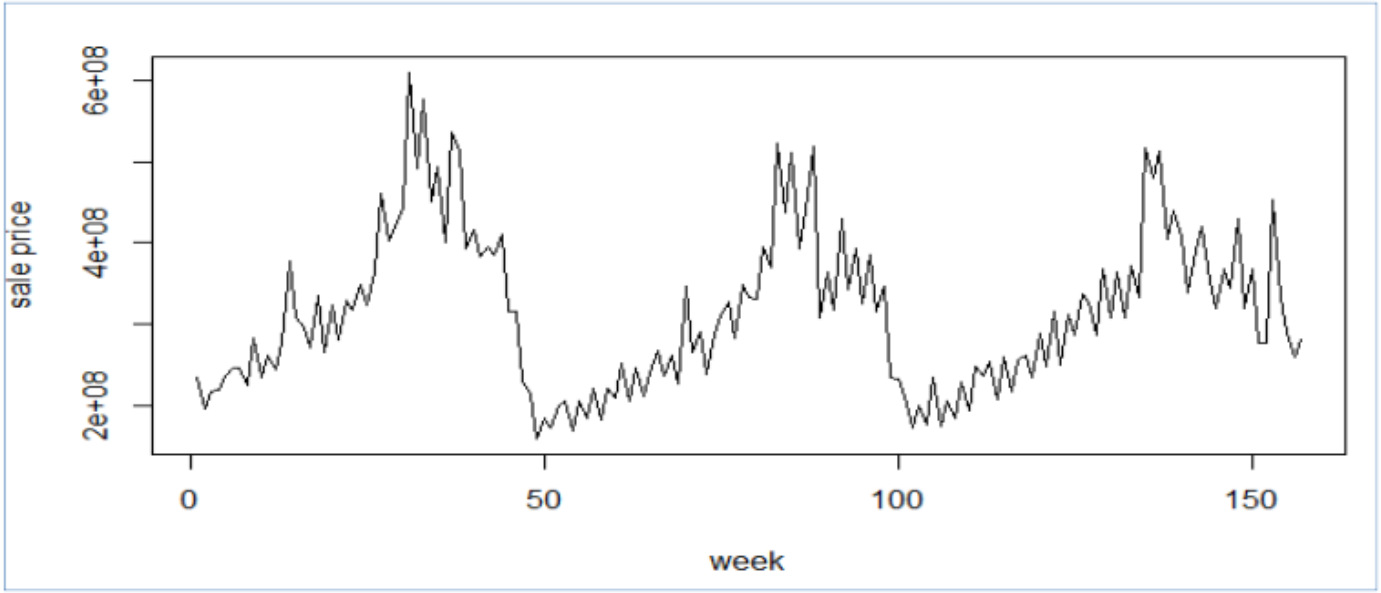
우선은 수량의 단위가 명확하지 않으므로 판매량의 규모는 매출액으로 잡자. 우선은 대형마트금액으로 시계열 분석을 해보는 것이 좋을 것 같다. 컬럼명은 우선 영문으로 모두 바꾸었다.

```
> colnames(kimchi) <- c("YYWW" , "LAST_WK" , "BIG_CNT" , "BIG_SALE" , "DEPT_CNT" , "DEPT_SALE" ,"SUPER_CNT" , "SU
PER_SALE" , "CONV_CNT" , "CONV_SALE")
> head(kimchi)
  YYWW  LAST_WK BIG_CNT  BIG_SALE DEPT_CNT DEPT_SALE SUPER_CNT SUPER_SALE CONV_CNT CONV_SALE
1 1301 20130106   27916 233968900    11971  99796735      1795   11561690     1603   2264200
2 1302 20130113   23057 194593960    11678 103106940      1832   11493710     2149   3073450
3 1303 20130120   25153 216322950    11634 106922870      1783   11541050     2277   3368720
4 1304 20130127   24645 218053670    12102  95765955      1949   11773330     2169   3172410
5 1305 20130203   26603 236095130    11404 102729615      1988   12808900     2085   3147500
6 1306 20130210   25509 245607800    11689 117911615      1829   12558230     2185   3178160
```

1. 생활에서 만나는 문제

3) 대형마트 김치매출 시각화

```
> sale <- kimchi$BIG_SALE
> plot(sale , type = "l" , xlab = "week" , ylab = "sale price")
```



대충 그림을 그려보니 50 주정도마다 일정한 패턴이 반복되고 있음을 알 수 있다. 이것이 시계열이 가지고 있는 '주기성'이라고 한다. 2013 년에서 2015 년까지의 3 년치의 데이터가 이런 식으로 거의 비슷한 패턴을 그린다면 앞으로도 그럴 가능성이 많다고 볼 수 있을 것이다. 자! 그럼 본격적으로 모델을 한번 만들어 보자.

2. 본격적인 분석작업

1) 추세(Trend)

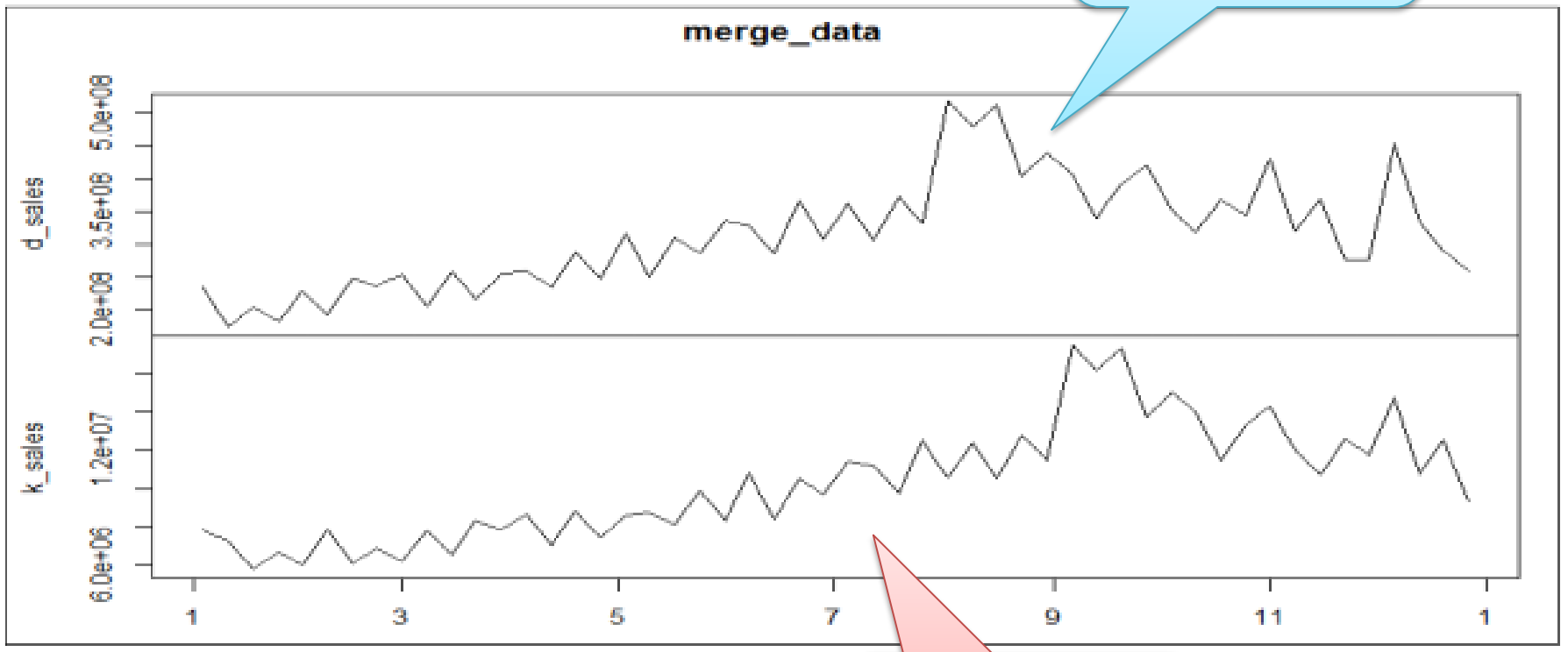
관측계열이 1 년이나 또는 계절마다 또는 월마다 주기적으로 같은 패턴을 그리는 것을 '계절성(seasonality)'을 갖는다고 한다. 보통 기온이나 강수량 등의 기후데이터나 산업생산 , 수입 및 수출과 관련된 데이터는 이런 주기적인 패턴을 갖는 경우가 많다.
시계열에서 '계절성'말고도 다른 특징을 엿볼 수 있는데 전반적으로 주기적이 아닌 어떤 패턴을 그리면서 시간에 따라 움직이는 경우가 있는 데 이것을 '추세(trend)'라고 한다. 그래서 우리가 예측하기 힘든 잡음을 a_t라고 하고 추세를 T_t , 계절성을 S_t라고 한다면 시계열 x_t는 다음과 같은 모형을 만족하는 식이 된다.

$$x_t = T_t + S_t + a_t \quad , a_t \sim N(0,\sigma)$$

2. 본격적인 분석작업

2) 대형마트 vs 효성이네 반찬가게

대형마트
- 김치 매출금액



2015년 1월 ~ 12월

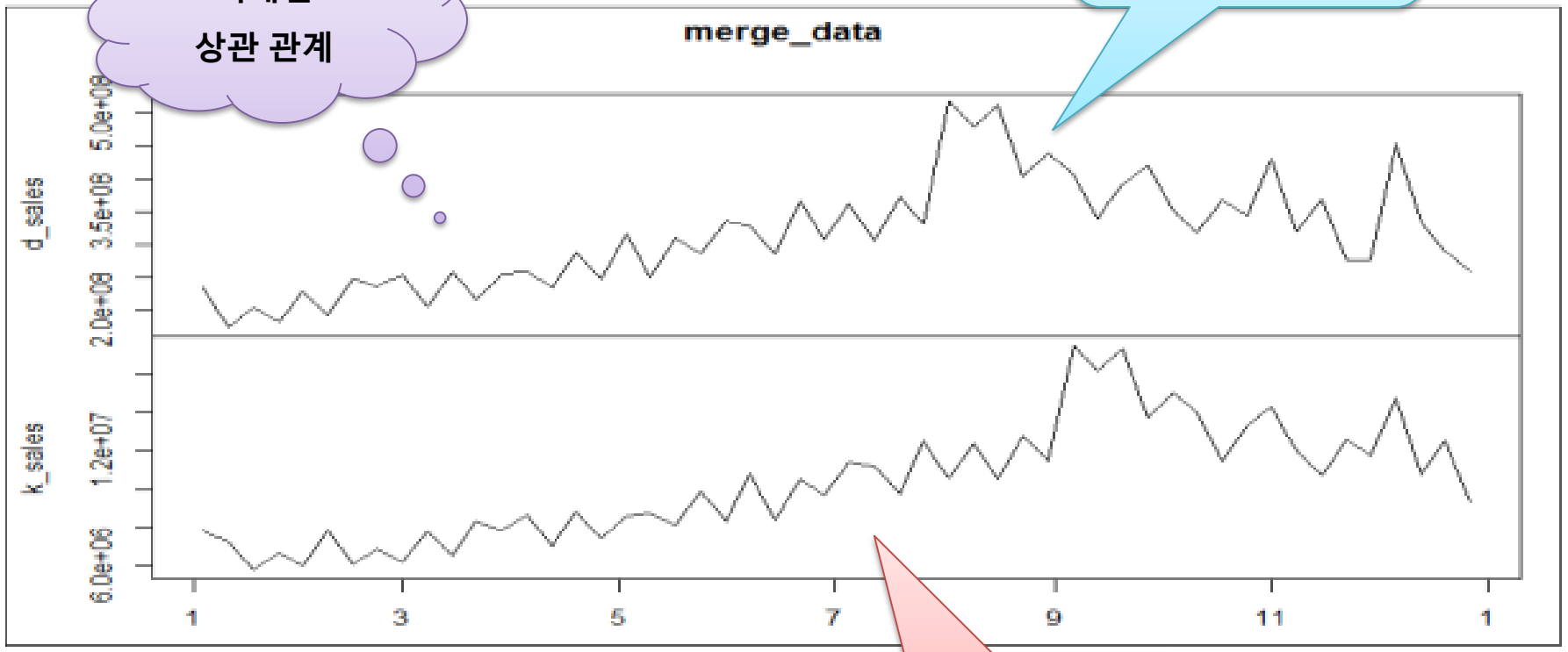
반찬가게
- 김치 매출금액

2. 본격적인 분석작업

2) 대형마트 vs 효성이네 반찬가게

시계열
상관 관계

참조 모델

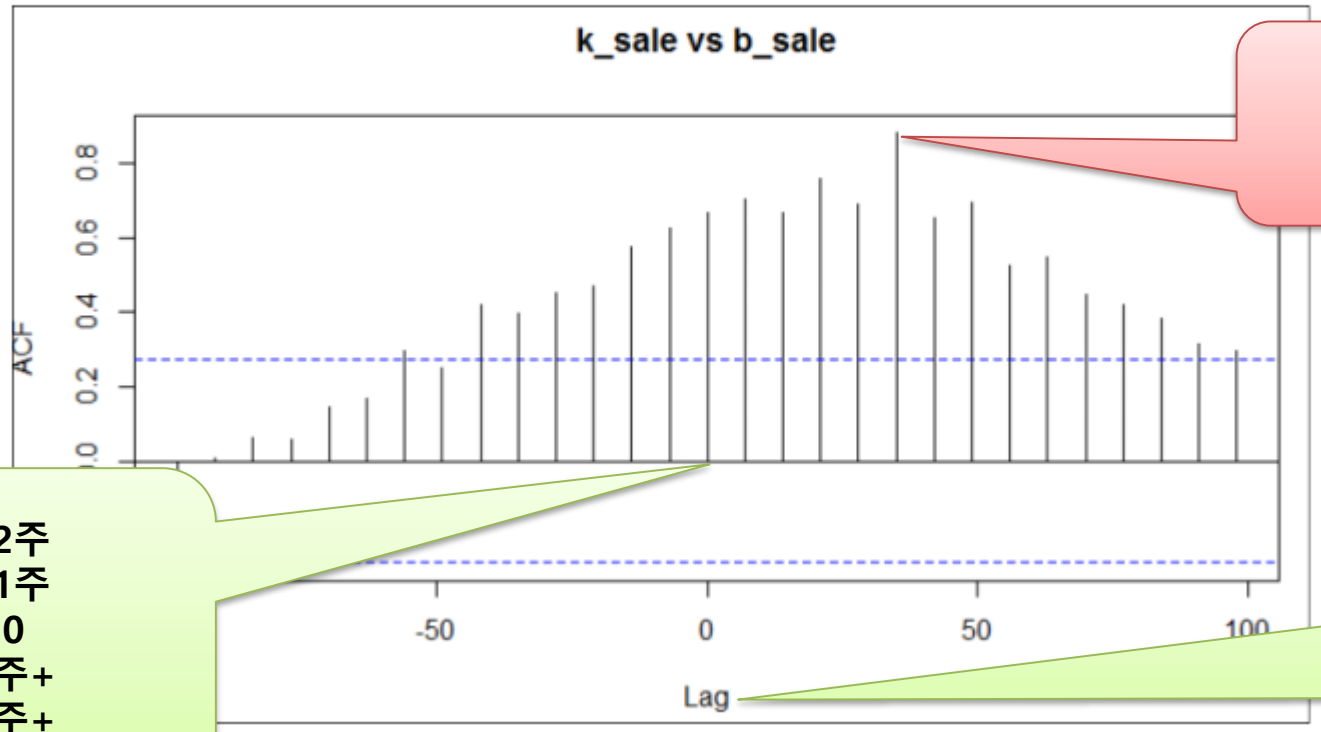


우리가 궁금한 것?

2. 본격적인 분석작업

3) 시계열 상관 관계(시차에 따른 상관 분석)

```
>ccf(k_sales , window(b_sales , start = "2015-01-01" , end = "2015-12-31") , main =
"k_sale vs b_sale")
```



5주 : 35일

-2주
-1주
0
1주+
2주+

[Lag]

시차(일)

2. 본격적인 분석작업

4) 대형마트 매출 예측

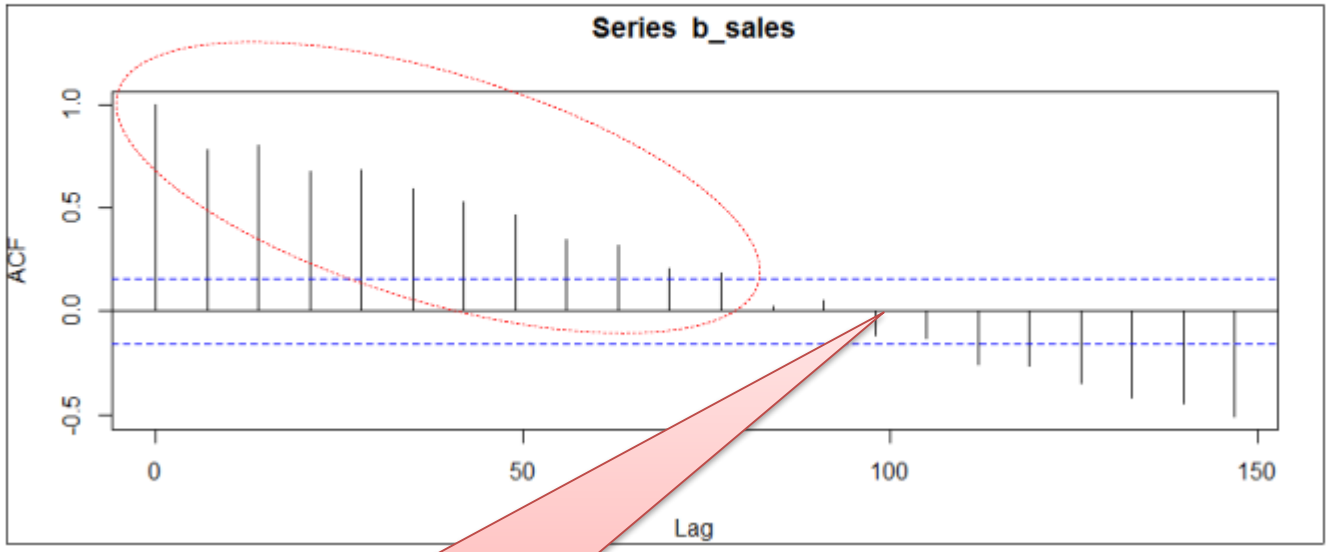
이제부터는

대형마트의 매출을 예측해보자

2. 본격적인 분석작업

4-1) 대형마트 : 자기상관함수(Autocorrelation Function)

```
> acf(b_sales)
```



11주

⇒ MA(Moving Average) = 11

2. 본격적인 분석작업

4-2) 자기상관이 있는가? -> Box-Pierce 검정, Ljung-Box 검정

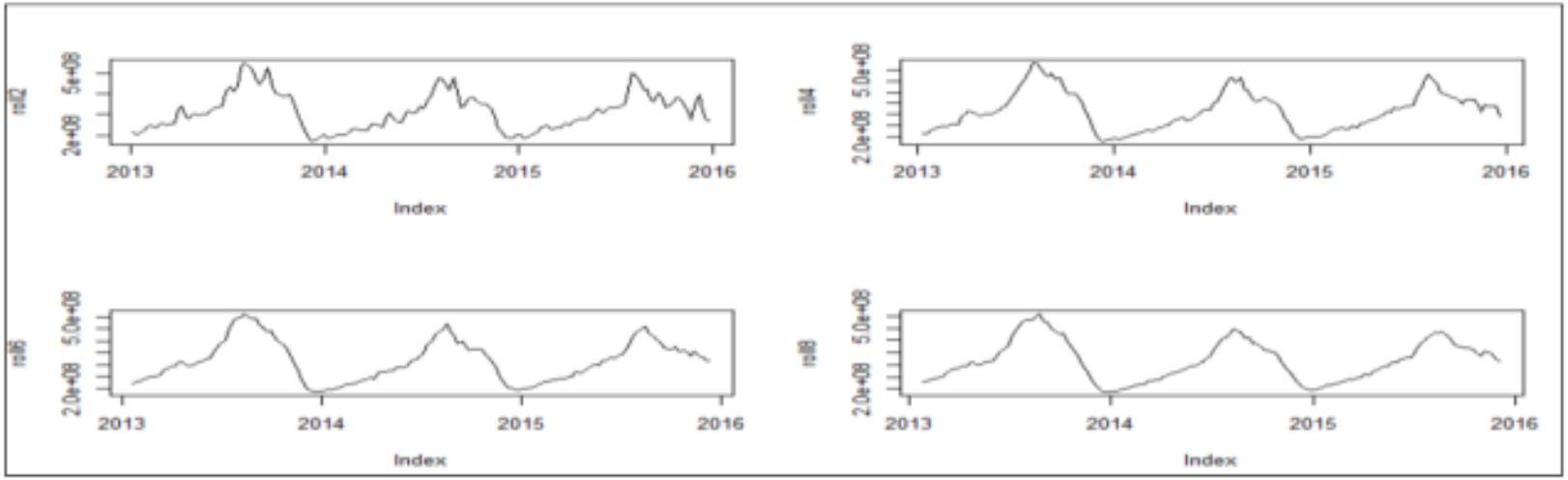
```
> Box.test(b_sales)

        Box-Pierce test

data:  b_sales
X-squared = 95.295, df = 1, p-value < 2.2e-16

> Box.test(b_sales , type = "Ljung-Box")

        Box-Ljung test

data:  b_sales
X-squared = 97.128, df = 1, p-value < 2.2e-16
```

P-value가
0.05 이하

2. 본격적인 분석작업

4-3) 이동 평균 그래프 그리기

```
> par(mfrow=c(2,2))
> roll2 <- rollapply(b_sales , 2 , mean , aligh = "right")
> roll4 <- rollapply(b_sales , 4 , mean , aligh = "right")
> roll6 <- rollapply(b_sales , 6 , mean , aligh = "right")
> roll8 <- rollapply(b_sales , 8 , mean , aligh = "right")
> plot(roll2);plot(roll4);plot(roll6);plot(roll8)
```



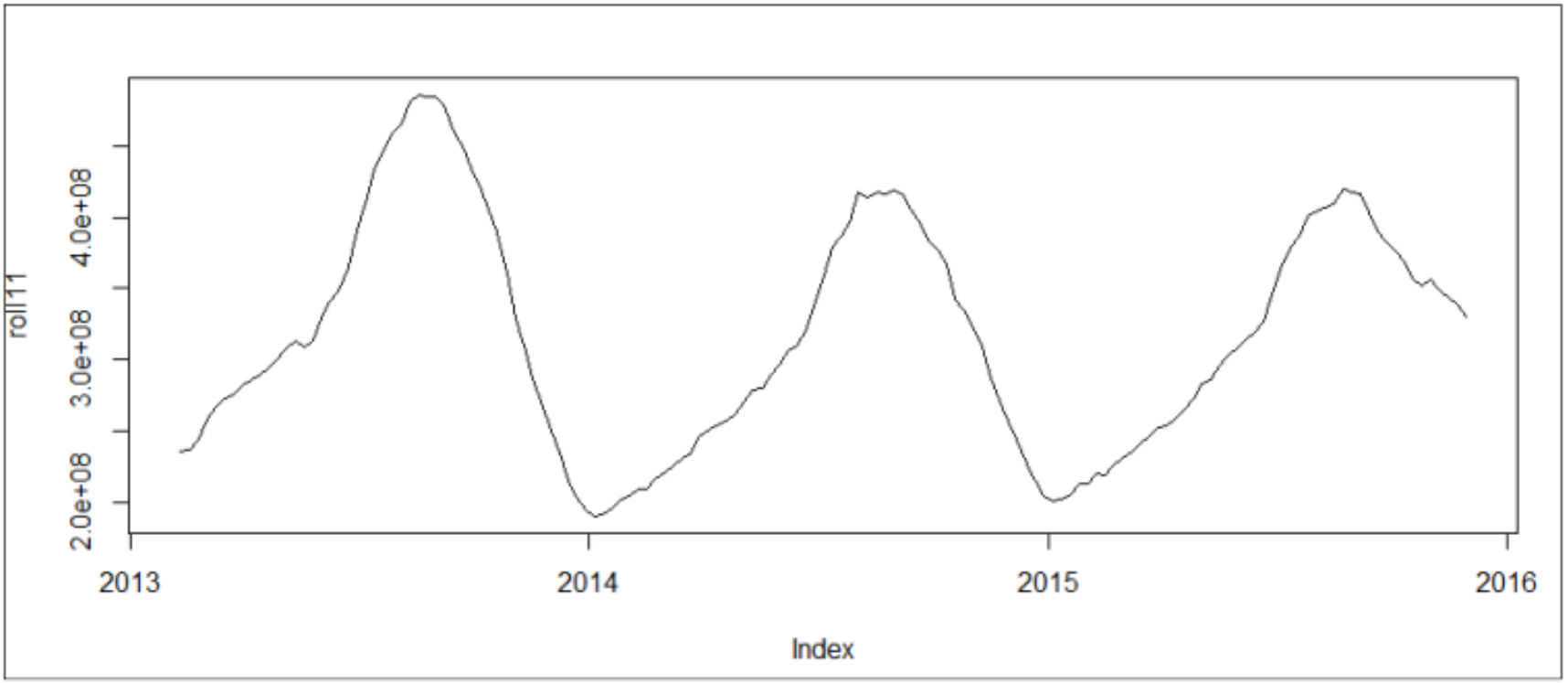
2. 본격적인 분석작업

4-3) 이동 평균 그래프 그리기

11주

⇒ MA(Moving Average) = 11

```
> par(mfrow=c(1,1))
> roll11 <- rollapply(b_sales , 11 , mean , aligh = "right");plot(roll11)
```



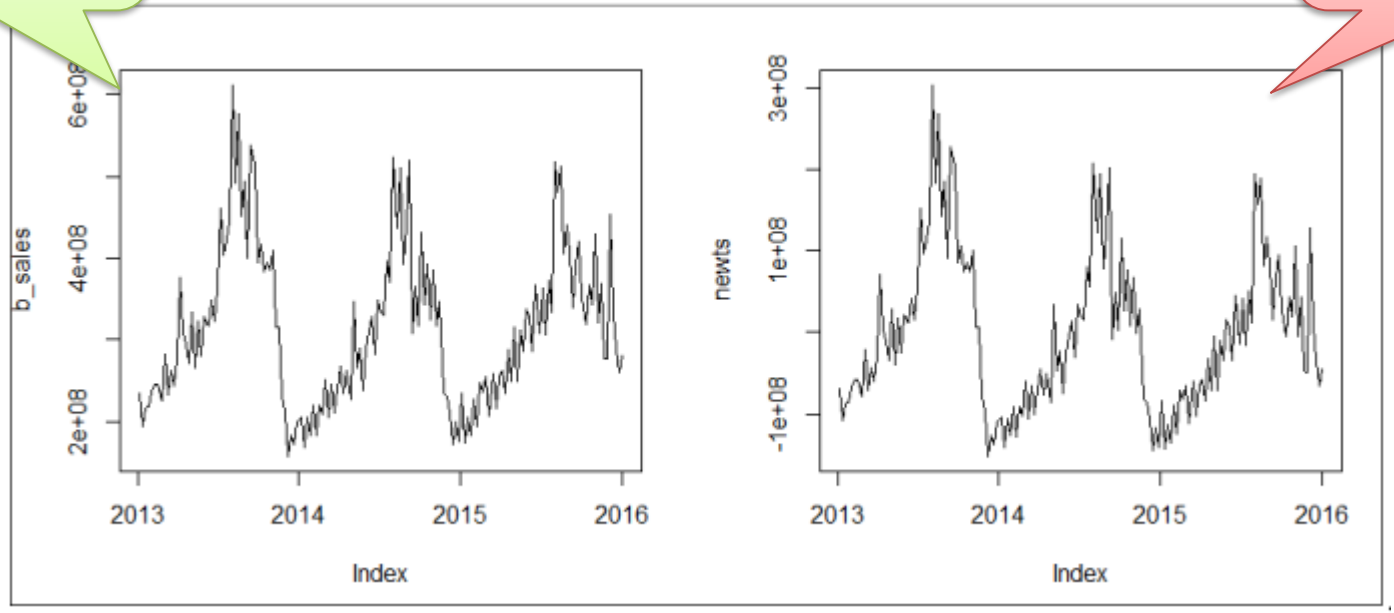
2. 본격적인 분석작업

4-4) 추세 찾기

```
> #추세 구하고 없애기
> par(mfrow=c(1,2))
> m1 <- lm(coredata(b_sales) ~ index(b_sales))
> newts <- zoo(resid(m1) , index(b_sales))
> plot(b_sales);plot(newts)
```

원래
데이터

원본 - 추세



거의 동일
- 추세가 없기 때문

2. 본격적인 분석작업

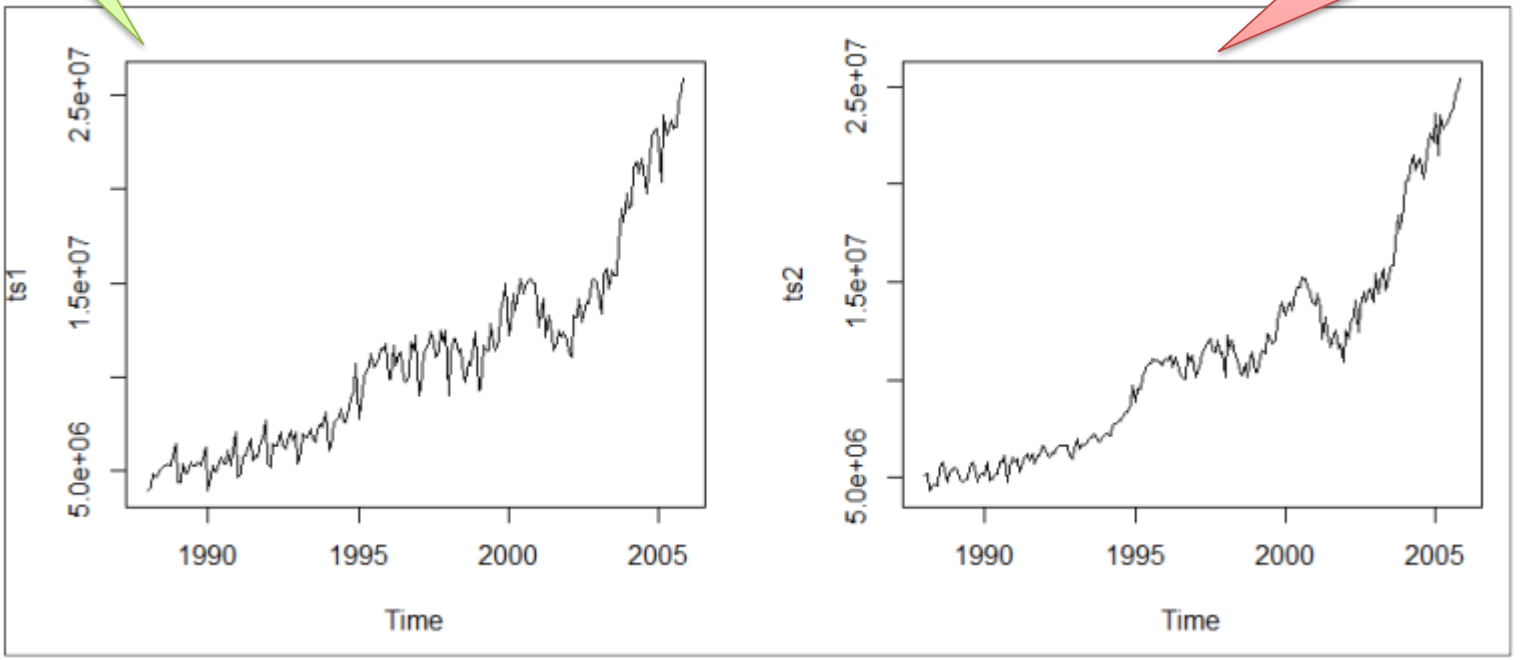
월별 수출액 데이터로 추세와 계절적 성향 분석

```
> #decompose
> par(mfrow=c(1,2))
> export <- read.table("Export_1988.txt" , header = T)
> ts1 <- ts(export$Series , start = c(1988,1) , frequency = 12)
> plot(ts1)
> decomp.result <- decompose(ts1)
> ts2 <- ts1 - decomp.result$seasonal
> plot(ts2)
```

원래 데이터

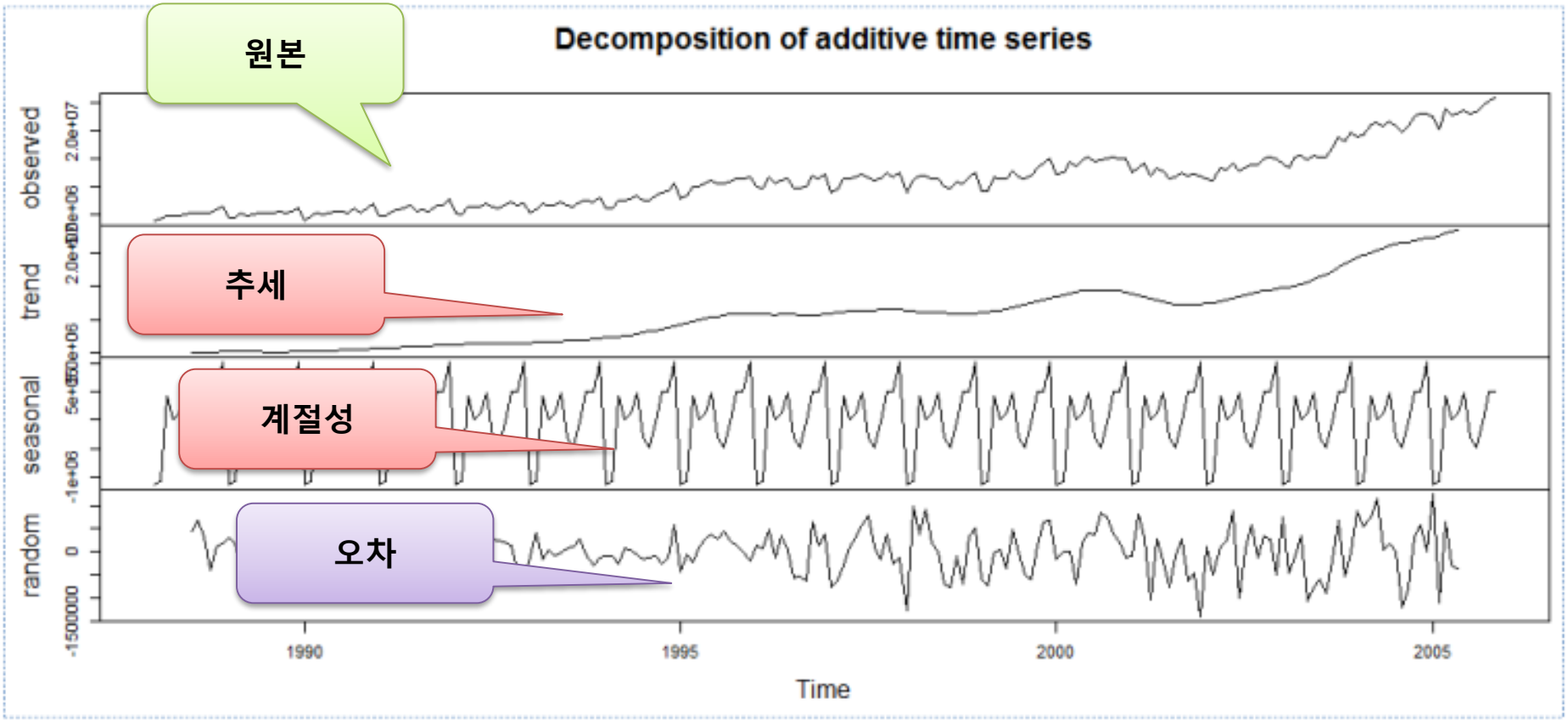
(원본 – 계절성)

⇒ 추이



2. 본격적인 분석작업

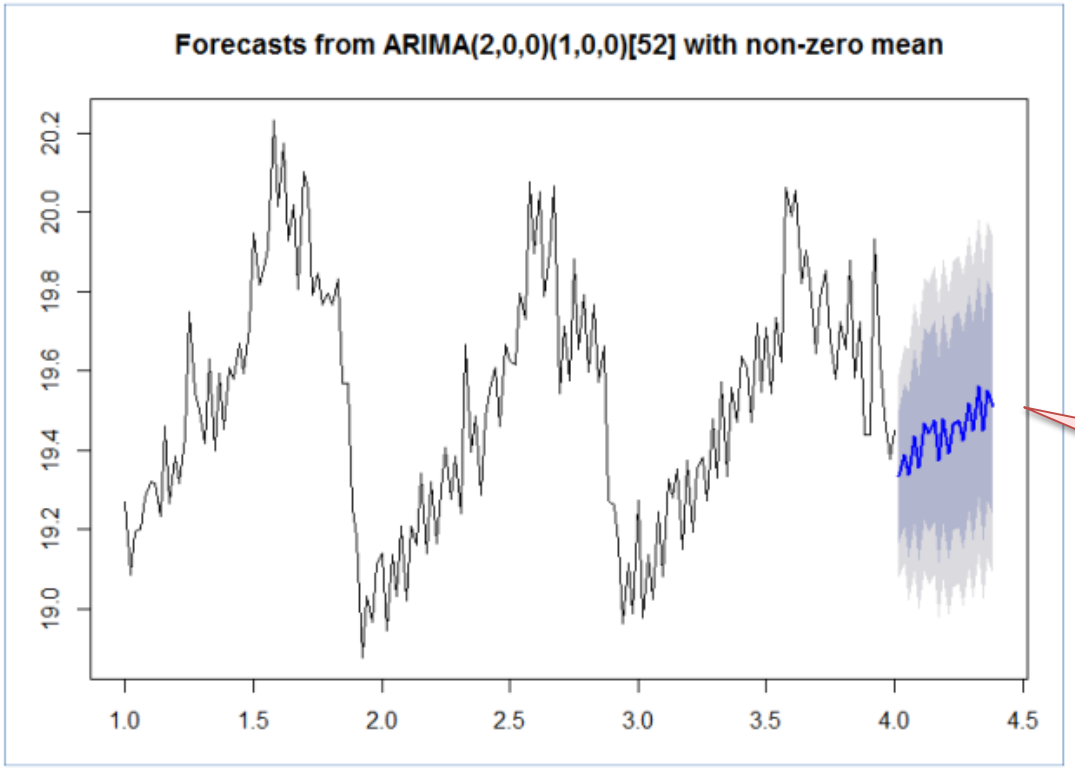
월별 수출액 데이터로 추세와 계절적 성향 분석



2. 본격적인 분석작업

4-5) 예측하기 : ARIMA(p,d,q) -> AR계수, 차분계수, MA계수

```
> #예측값 및 신뢰구간
> library(forecast)
> fit <- auto.arima(ts(log(BIG_sales) , frequency = 52) , seasonal = TRUE)
> plot(forecast(fit,h=20))
```



계절성 : 52주

20개 예측

2. 본격적인 분석작업

4-5) 예측하기 : 향후 10주간의 대형마트 예상 매출액 생성

```
> p <- predict(fit , n.ahead=10)
> p
$pred
Time Series:
Start = c(4, 2)
End = c(4, 11)
Frequency = 52
 [1] 19.33269 19.38859 19.33828 19.43415 19.35490 19.46747 19.44192 19.47386 19.37433 19.47860

$se
Time Series:
Start = c(4, 2)
End = c(4, 11)
Frequency = 52
 [1] 0.1305193 0.1405464 0.1621457 0.1716342 0.1817108 0.1883901 0.1942385 0.1986777 0.2023623 0.2052929
```

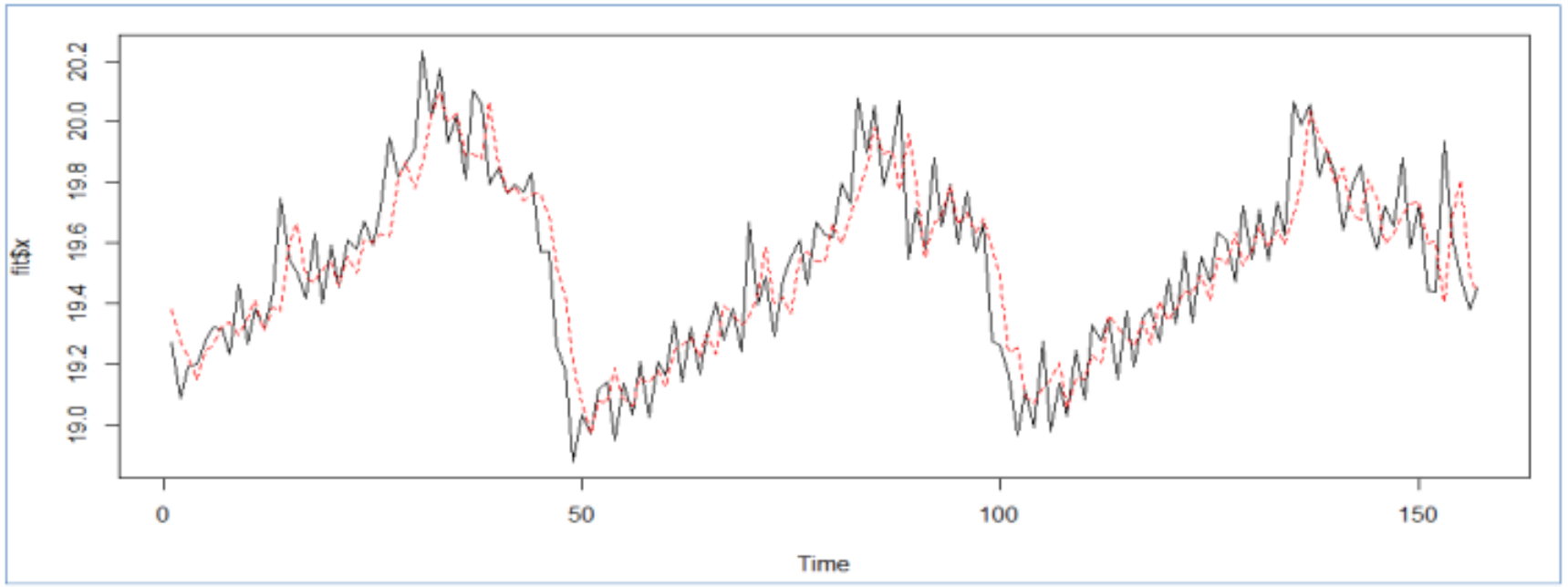
예측값

표준 오차

2. 본격적인 분석작업

4-6) 반찬가게 매출 예측

```
> par(mfrow=c(1,1))
> plot(fit$x , lty = 1)
> lines(fitted(fit) , lty = 2 , lwd = 1 , col = "red")
```



3. 알고리즘과 수학적 정의

1) 시계열

시계열 데이터는 크게 신호와 잡음으로 분리한다. 간단하게 표현하면 다음과 같다.

$$x_t = s_t + \varepsilon_t \ , \ \varepsilon_t \sim N(0,\sigma)$$

s_t는 신호이고 ε_t는 잡음이다. 여기서 신호 s_t는 추세성 T_t 와 계절성 S_t이 포함되어 있다. 따라서, 다음과 같이 나타낼 수 있다.

$$x_t = T_t + S_t + a_t \ \ , a_t \sim N(0,\sigma)$$

감사합니다.