서울시 아파트 실거래 가격 지수 예측

고영희

1. 데이터의 정상성 판단

1) 데이터 읽기

데이터를 불러오고, ts 함수를 통해 일변량 시계열 데이터로 변화시킨다.

```
> df<-read.csv('apartment_price.csv')

> colnames(df)<-c('year','month','price')

> head(df)

> df.ts<-ts(data=df$price,frequency = 12, start=c(2006,01))
```

Output]

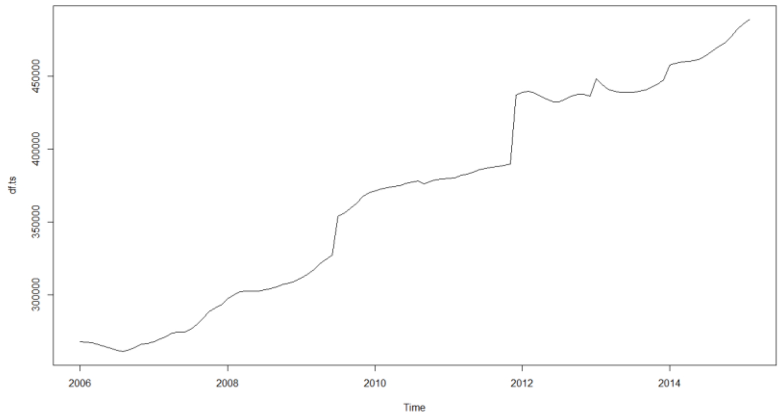
```
  year month   price
1 2012     1 267661
2 2012     2 267588
3 2012     3 267002
4 2012     4 265678
5 2012     5 264522
6 2012     6 263352
```

2) 데이터 시각화

시계열 데이터가 정상시계열인지 그림을 통해 판단한다. 이 때 정상 시계열이란, $E(X_t)$ 값이 t에 관계없이 일정하며, $Var(X_t)$ 가 유한한 값으로 존재하며 t의 값에 관계없이 일정하다는 것을 의미한다. 또한 $cov(X_t, X_{t+h})$ 가 t와 무관한 h값에만 의존하다는 것을 의미한다.

```
> plot(df.ts)
```

Output]



- ✓ 이 그래프를 통해 본 데이터는 시간 t에 따라 $E(X_t)$값이 변화하기 때문에 정상시계열이라고 말할 수 없다.
- ✓ 따라서 비정상 시계열이기 때문에 ARMA 모형에 바로 적합할 수 없으며, 비정상 시계열을 정상 시계열로 바꿔야 한다.

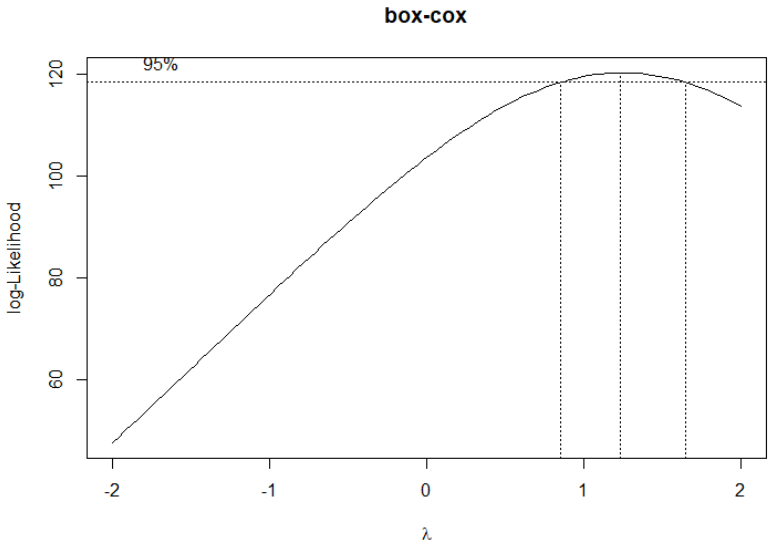
2. 분산의 정상화

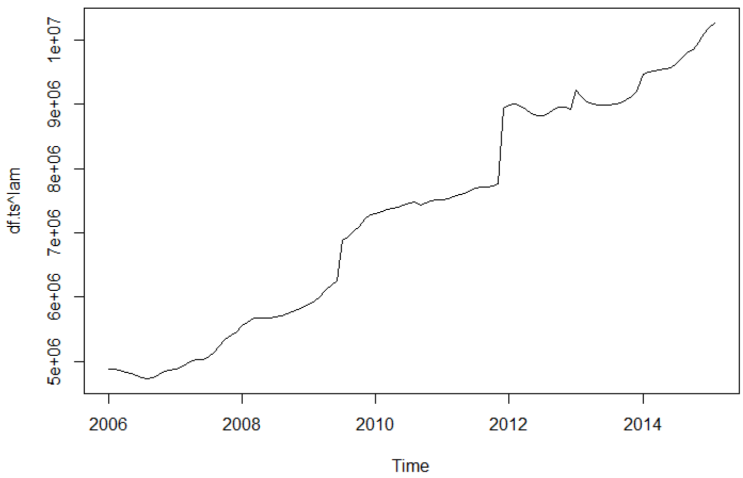
1) Box – cox 변환

Box-cox 변환은 분산을 안정화 하기 위한 변환법이다. $X_t^{(\lambda)} = \begin{cases} \frac{X_t^{(\lambda)}}{\lambda} & , \lambda \neq 0 \\ log(X_t) & , \lambda = 0 \end{cases}$ 변환을 사용하며 많이 사용하는 모수 λ는 +1,-1,+0.5,-0.5,0 을 사용한다. 선형회귀 모형에서 최대 가능도 추정법을 사용하여 모수를 추정하게 된다.

```
> library(MASS)

> bc<-boxcox(df.ts~time(df.ts))

> print(bc)

> lam<-bc$x[which.max(bc$y)]

> plot(df.ts^lam)
```

Output]





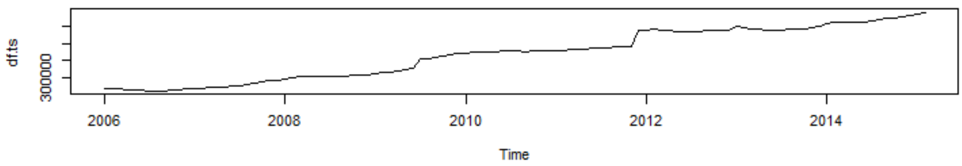
- ✓ Box-cox 변환을 통해 나온 모수(λ)는 약 1.2 였으며 모든 데이터 값에 약 1.2 승을 한 결과는 기존의 df.ts 선도표와 비슷하게 나왔다.
- ✓ 이는 원 데이터 df.ts 값의 분산은 시간 t 에 따라 변화하지 않았기 때문이라고 말할 수 있다.

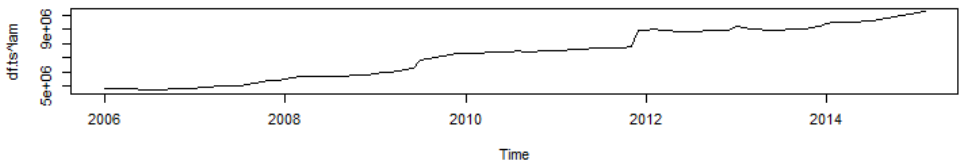
2) Log 변환

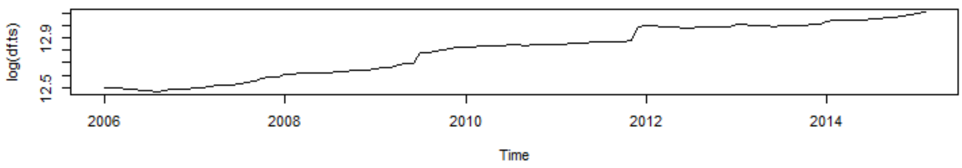
Box-cox 변환에서 모수 λ 가 0이 아님으로 앞선 box-cox 변환과 달리 λ가 0의 값을 가질 때 변환인 로그 변환을 사용해 분산 안정화를 시행한다.

```
> par(mfrow=c(3,1))

> plot(df.ts)

> plot(df.ts^lam)

> plot(log(df.ts))

> dev.off()
```

Output]







- ✓ 원 데이터의 선도표와 box-cox 변환을 시행했을 때, log 변환을 시행했을 때를 비교해보면, 모두 비슷한 그래프를 출력한다.
- ✓ 이는 시간에 따라 분산의 변화량이 적다는 것을 의미하고 box-cox 변환과 로그 변환 중, 데이터 값이 더 작아 계산을 쉽게 할 수 있는 log 변환을 사용한다.

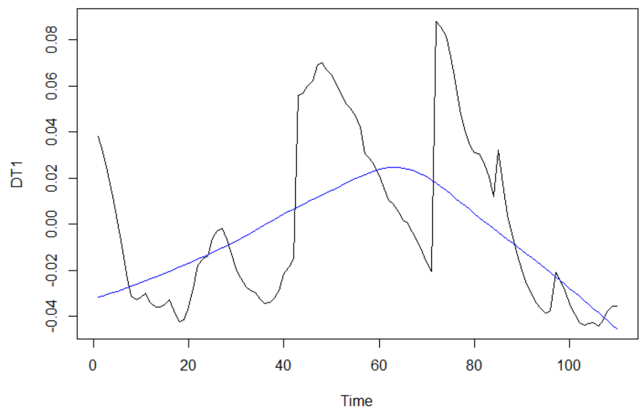
3. 평균의 정상화

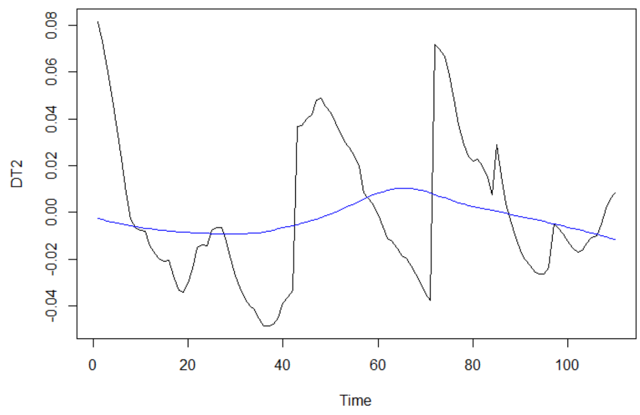
1) 추세제거(detrending)

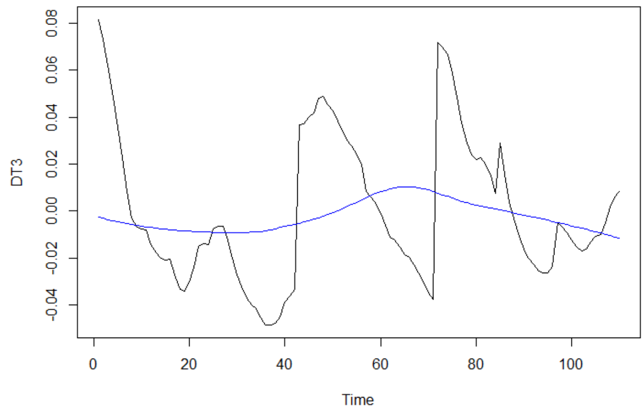
추세 제거를 시행할 때, 회귀분석기법을 사용하게 되는 데, 이 때 종속변수는 시계열 데이터이고 독립변수는 시간을 의미한다. 시간의 흐름에 따른 데이터가 비 선형적으로 증가하는 모습을 보이기 때문에 시간에 대한 독립변수는 1차 ~ 3차 까지 조정한다. 이 때 추세 제거된 데이터는 원 데이터값에서 회귀 적합값의 차이 이므로 회귀분석에서의 잔차를 통해 얻을 수 있다. 또한, 추세 제거가 잘 되었는지 판단하기 위해 국소적인 데이터의 평균을 나타내는 Lowess(국소회귀) 모형을 사용한다.

```
> DT1<-lm(log(df.ts)~time(df.ts))$residuals
> plot.ts(DT1)
> lines(lowess(DT1),col="blue")

> x<-time(df.ts)
> DT2<-lm(log(df.ts)~x+I(x^2))$residuals
> plot.ts(DT2)
> lines(lowess(DT2),col="blue")

> DT3<-lm(log(df.ts)~x+I(x^2)+I(x^3))$residuals
> plot.ts(DT3)
> lines(lowess(DT3),col="blue")
```

Output]







- ✓ DT1에서는 국소회귀 모형이 눈에 띄게 비선형이므로 사용하지 않는다.
- ✓ DT2와 DT3의 그래프에서는 추세 제거된 데이터와 국소회귀 모두 비슷하기 때문에 해석하기 쉬운 저 차원 DT2를 사용한다.

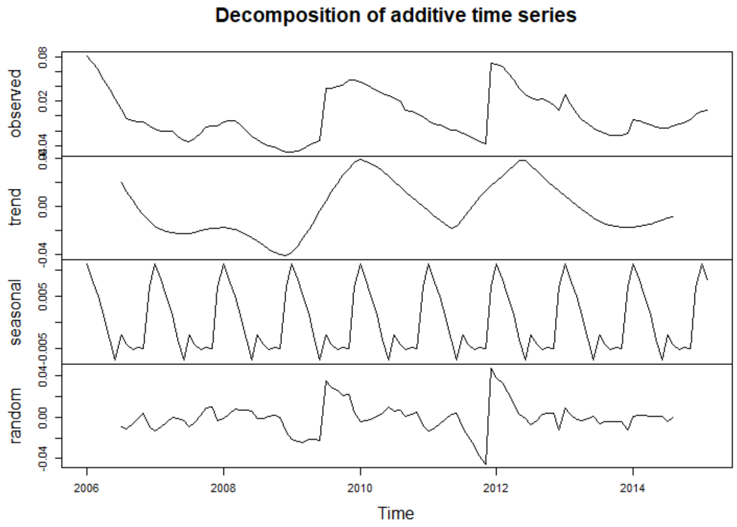
2) 계절조정 (Seasonal adjustment)

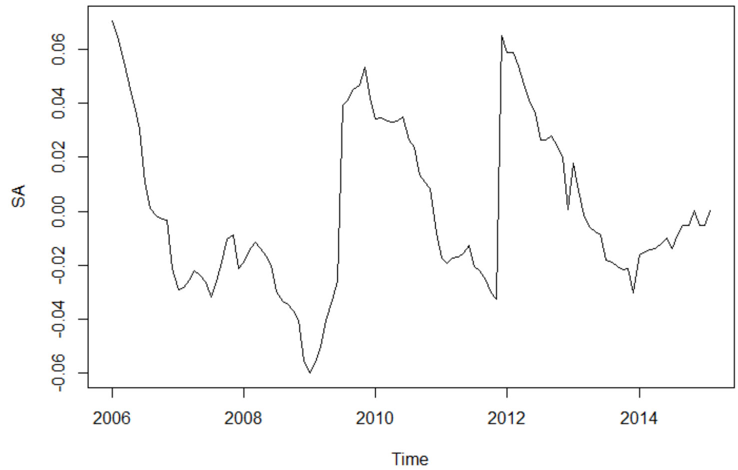
추세 조정된 데이터(DT2)를 이용하여 계절조정을 한다. 계절 조정은 회귀 분석 기법과 분해법을 이용해 데이터를 조정한 다음 두 결과 값을 비교한다.

① 분해법(Decompose) 이용

```
> start(df.ts)
> DT.ts<-ts(DT2,start=start(df.ts),frequency = 12)
> plot(decompose(DT.ts,type="additive"))

> dec<-decompose(DT.ts,type='additive')
> SA<-dec$x-dec$seasonal
> plot(SA)
```

Output]





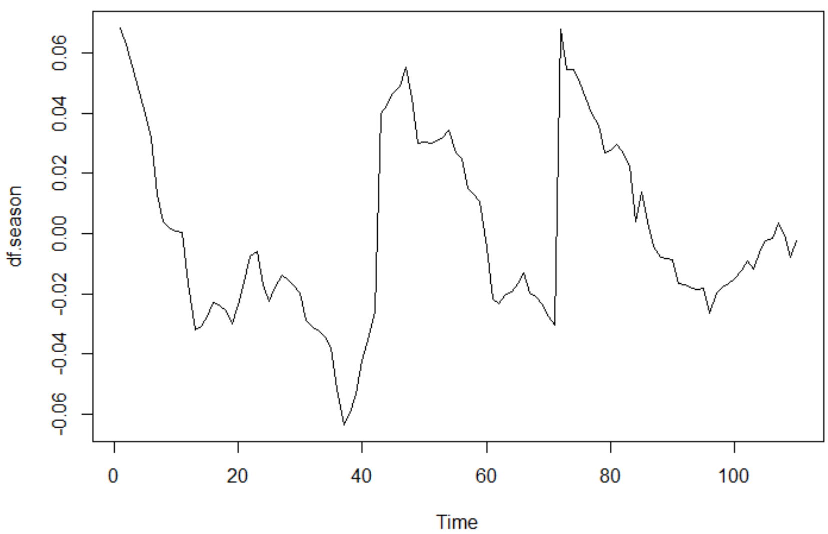
✓ 분해법을 이용해 계절 조정한 데이터는 평균이 0으로 일정하고 진폭(분산)이 시간의 흐름에 따라 일정하다는 것을 알 수 있다.

② 회귀분석 이용

회귀 분석을 이용할 때 앞서 추세분석을 통해 독립변수를 time^2로 파악했기 때문에 이를 이용해 종속변수는 시계열 데이터이고, 독립변수는 추세분석항과 계절성분 항으로 이루어진 회귀분석을 실시한다.

```
> xm<-1:length(df.ts)
> x1<-xm%%12==1 ; x2<-xm%%12==2 ; x3<-xm%%12==3
> x4<-xm%%12==4 ; x5<-xm%%12==5 ; x6<-xm%%12==6
> x7<-xm%%12==7 ; x8<-xm%%12==8 ; x9<-xm%%12==9
> x10<-xm%%12==10 ; x11<-xm%%12==11

> df.season<-lm(log(df.ts)~xm+I(xm^2)+x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11)$residuals
> plot.ts(df.season)
```

Output]



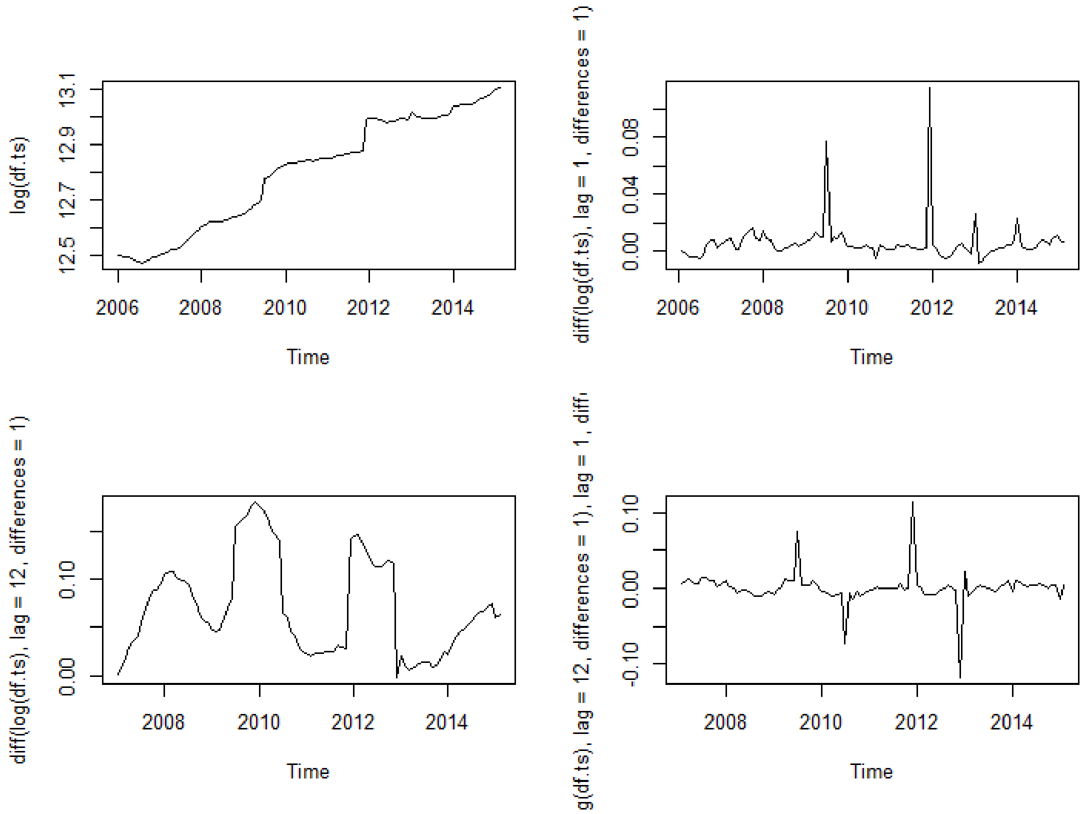
- ✓ 추세분석항 (xm+I(xm^2)) 과 계절성분항 (x1~x11) 모두 독립변수로 취급해 회귀분석을 시행한 결과, 잔차는 데이터 값과 적합값의 차이이므로 추세제거와 계절조정 모두 시행한 데이터임을 알 수 있다.
- ✓ 이는 평균이 0 이고 분산이 시간에 따라 일정하기 때문에 정상시계열이라 볼 수 있다.

4. 차분 (differencing)

현 시점의 데이터 값에서 이전 시점의 데이터 값을 빼는 것을 차분이라 한다. 원 시계열이 $X_t = \delta + X_{t-1} + a_t$, $a_t \sim WN(0, \sigma^2)$ 비정상 시계열을 따를 때, 1차 차분한 시계열은 $\nabla X_t = X_t - X_{t-1} = \delta + a_t$ 이며, 평균 $E(\nabla X_t) = \delta$, 분산 $var(\nabla X_t) = \sigma^2$, 자기 공분산 $cov(\nabla X_t, \nabla X_{t-1}) = 0, h \geq 1$이기 때문에 정상성을 만족한다.

```
> par(mfrow=c(2,2))
> plot(log(df.ts))
> plot(diff(log(df.ts),lag=1,differences = 1))

> plot(diff(log(df.ts),lag=12,differences = 1))

> plot(diff(diff(log(df.ts),lag=12,differences = 1),lag=1,differences
= 1))
> dev.off()
```

Output]



- ✓ 로그 변환된 데이터는 증가추세가 존재하기 때문에 비정상 시계열이다.
- ✓ 시차 (lag)를 1로 두고 1차 차분한 결과는 추세 조정한 데이터라고 볼 수 있고, 실제로 평균이 점차 증가하는 패턴이 사라짐을 확인할 수 있다.
- ✓ 시차 (lag)를 12로 두고 1번 차분한 그래프는 계절차분을 진행하였는데 반복 패턴이 제거되었다는 것을 알 수 있다.
- ✓ 1번 차분과 계절차분을 모두 시행한 가장 마지막 그래프에 대해서는 몇 개의 이상치 값만 제외하면 4개의 그래프 중 가장 정상시계열 데이터로 파악할 수 있다.