**서울시 아파트 실거래 가격 지수 예측**

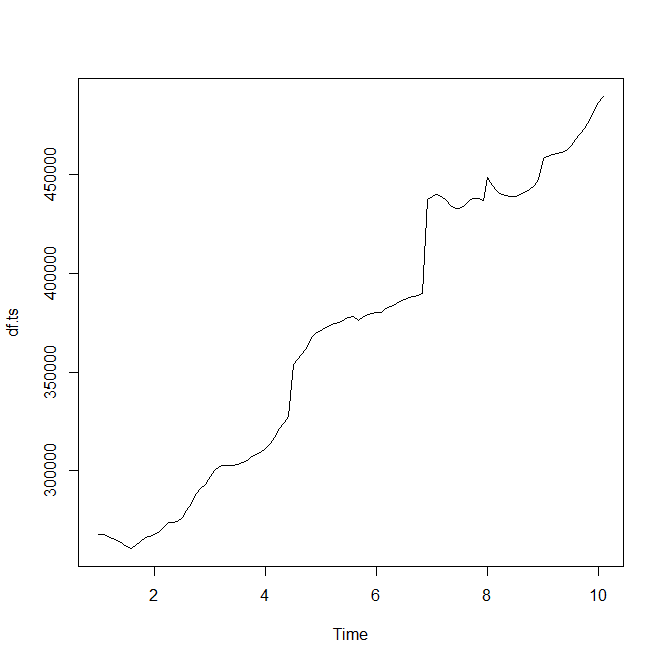
고영희

1. **Box-Jenkins의 모형 식별**
2. **시계열 자료 특징 파악**

: 시계열 그림을 통해 시계열 패턴과 추세성, 계절성, 이상점, 정상성 여부 등 판단.

|  |
| --- |
| > 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))  > plot(df.ts) |

Output]

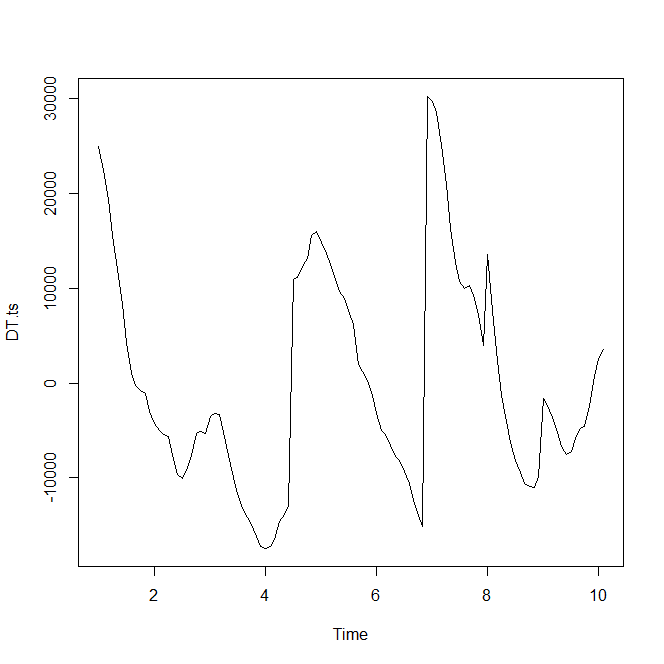


* 시간에 따라 점차 증가하는 추세를 보이며 계절성은 보이지 않는다.
* 이상점은 존재하지 않으며 시간에 따라 평균값이 증가하므로 비정상 시계열이다.

1. **추세 제거**

|  |
| --- |
| > x<-time(df.ts)  > DT2<-lm(df.ts~x+I(x^2))$residuals  > lines(lowess(DT2),col="blue")  > DT.ts<-ts(DT2,start=start(df.ts),frequency = 12)  > plot(DT.ts) |

Output]

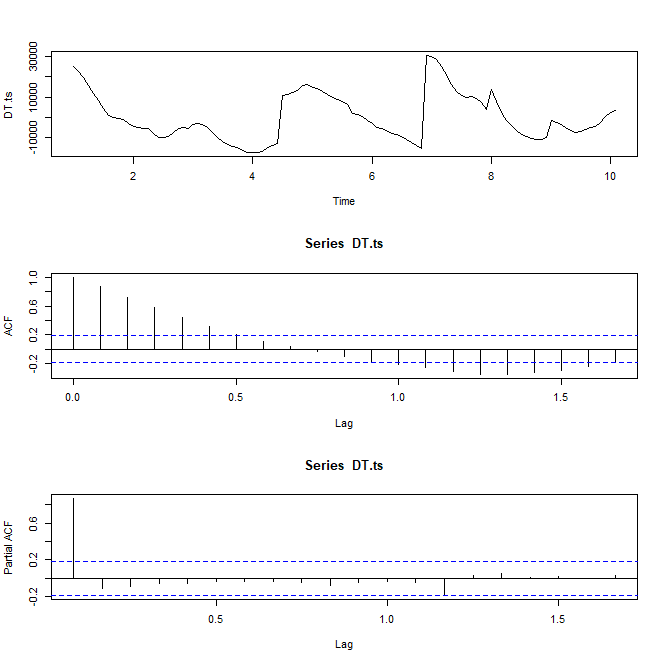


* 과제5에 따라 2차항까지 사용해 추세 제거를 진행한다.
* 회귀분석을 통해 추세제거를 시행했지만, 시간에 따른 진폭(분산)이 달라져 여전히 비정상 시계열이라고 할 수 있다.

1. **차분 차수 결정**

|  |
| --- |
| > par(mfrow=c(3,1))  > plot(DT.ts)  > acf(DT.ts)  > pacf(DT.ts)  > adf.test(DT.ts) |

Output]



|  |
| --- |
| Augmented Dickey-Fuller Test  data: DT.ts  Dickey-Fuller = -3.2946, Lag order = 4, p-value = 0.0757  alternative hypothesis: stationary |

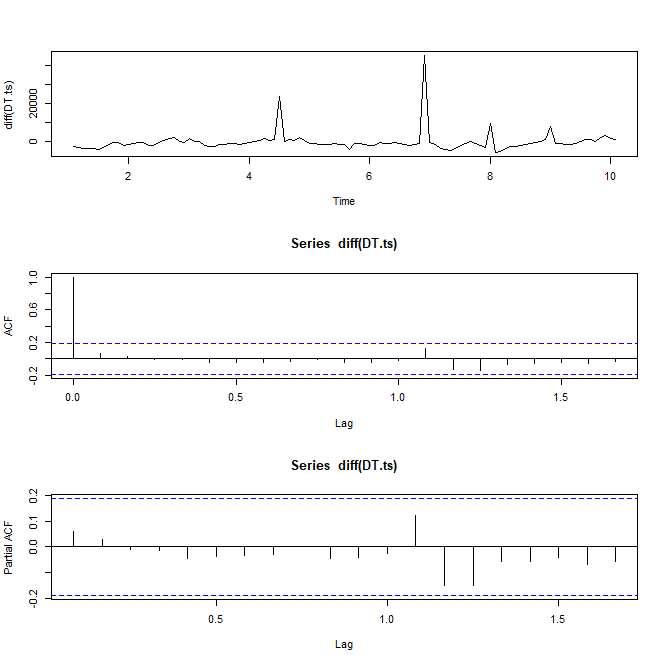
* SACF가 완만하게 줄어들어 차분이 필요한 비정상 시계열임을 알 수 있다.
* 추가적으로 adf test를 통해 객관적으로 차분이 필요한 지를 검정해보면, p-value가 0.08로 큰 값을 가져 유의수준 0.05 하에서 귀무가설을 채택한다.
* 귀무가설 (H0) : 단위근 존재 ⇔ 차분 필요 ⇔ 비정상 시계열
* 따라서 추세제거 한 데이터도 확률적 추세가 존재하는 비정상 시계열이므로 차분을 해야만 한다.

1. **차분시행**

|  |
| --- |
| > kpss.test(diff(DT.ts))  > par(mfrow=c(3,1))  > plot(diff(DT.ts))  > acf(diff(DT.ts))  > pacf(diff(DT.ts)) |

Output]

|  |
| --- |
| KPSS Test for Level Stationarity  data: diff(DT.ts)  KPSS Level = 0.10439, Truncation lag parameter = 4, p-value = 0.1 |



* 1차 차분한 데이터를 kpss 단위근 검정을 통해 보인 결과, p-value는 0.05보다 커 귀무가설을 채택한다. 따라서 차분차수는 1로 지정한다.
* 귀무가설(H0) : ⇔ 차분 필요 없음.
* 1차 차분한 데이터의 ACF 및 PACF 를 통해 파악한 결과, ACF와 PACF는 모두 지수적으로 감소하는 모습을 볼 수 있다.
* 따라서 ACF와 PACF가 모두 Tails off한 형태인 ARMA모형을 선택한다.

1. **모형의 추정**
2. **모수 추정**

|  |
| --- |
| > ma1<-Arima(DT.ts,order=c(0,1,1),include.mean = TRUE,include.drift=TRUE)  > ar1<-Arima(DT.ts,order=c(1,1,0),include.mean = TRUE,include.drift=TRUE)  > arma11<-Arima(DT.ts,order=c(1,1,1),include.mean = TRUE,include.drift=TRUE)  > arma11 |

Output]

|  |
| --- |
| Series: DT.ts **ARIMA(1,1,1)** with drift  Coefficients:  ar1 ma1 drift  0.9174 -1.0000 -82.3364  s.e. 0.0471 0.0244 150.8145  sigma^2 estimated as 28985695: log likelihood=-1090.45  AIC=2188.9 AICc=2189.28 BIC=2199.66 |

* 앞서 그린 ACF, PACF 그림을 통해 ARMA 모형과 함께 MA(1) 모형과 AR(1) 모형까지 적합 시킨다.
* 이때 적합 방법은 R의 디폴트 값인 “CSS-ML”을 사용하고 이는 최대 우도 추정을 이용하되, 초기값은 조건부 최소 제곱 값을 사용해 적합한다.
* Ma1의 AIC값은 2189.22, ar1은 2189.2, arma11은 2188.9 값이 도출되었기 때문에 AIC값이 가장 작은 arma11모형을 택한다.

1. **과적합 모형 파악하기**

|  |
| --- |
| > arma12<-Arima(DT.ts,order=c(1,1,2),include.mean = TRUE,include.drift=TRUE)  > arma21<-Arima(DT.ts,order=c(2,1,1),include.mean = TRUE,include.drift=TRUE)  > arma12 ; arma21 |

Output]

|  |
| --- |
| Series: DT.ts **ARIMA(1,1,2)** with drift  Coefficients:  ar1 ma1 ma2 drift  0.8958 -0.8975 -0.1025 -69.7302  s.e. 0.0554 0.1004 0.0973 137.9625  sigma^2 estimated as 28913306: log likelihood=-1089.9  AIC=2189.79 AICc=2190.37 BIC=2203.25 |

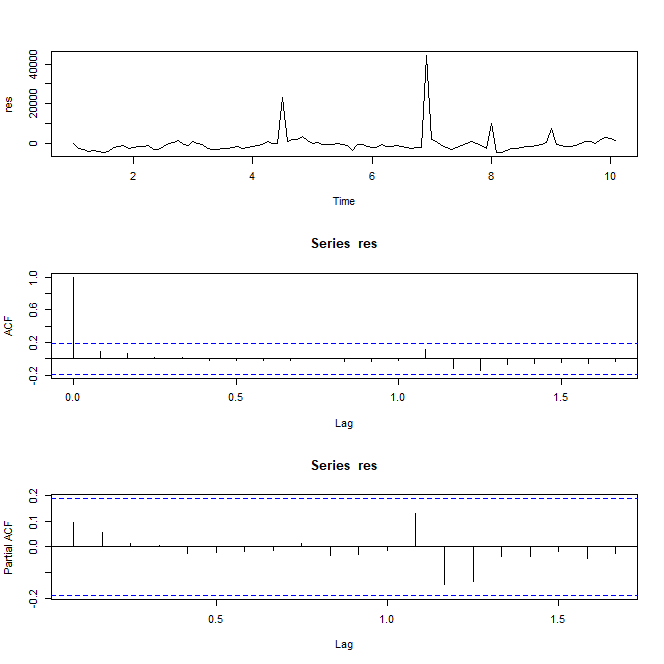
|  |
| --- |
| Series: DT.ts **ARIMA(2,1,1)** with drift  Coefficients:  ar1 ar2 ma1 drift  1.0129 -0.1116 -1.0000 -65.7928  s.e. 0.0950 0.0967 0.0252 134.1883  sigma^2 estimated as 28839256: log likelihood=-1089.79  AIC=2189.58 AICc=2190.16 BIC=2203.03 |

* ARIMA(1,1,1) 보다 MA차수가 하나 더 많은 ARIMA(1,1,2) 모형은 새롭게 추가된 ma2계수의 유의성을 파악한다.
* |-0.1025| < 0.0973\*2 (=0.1946) 이므로 새로운 ma2계수는 유의하지 않다.
* ARIMA(1,1,1) 보다 AR차수가 하나 더 많은 ARIMA(2,1,1) 모형은 새롭게 추가된 ar2계수의 유의성을 파악한다.
* |-0.1116| < 0.0967\*2 (=0.1934) 이므로 새로운 ar2계수는 유의하지 않다.
* 따라서 과적합을 방지하기 위해 ARIMA(1,1,1) 모형식을 선택한다.
* 최종 선택된 모형식은 이다.

1. **모형 진단**
2. **잔차 분석**

|  |
| --- |
| > res<-residuals(arma11)  > par(mfrow=c(3,1))  > plot(res)  > acf(res)  > pacf(res)  > Box.test(res,lag=10,type="Ljung-Box",fitdf=2) |

Output]



|  |
| --- |
| Box-Ljung test  data: res  X-squared = 1.9654, df = 8, p-value = 0.9821 |

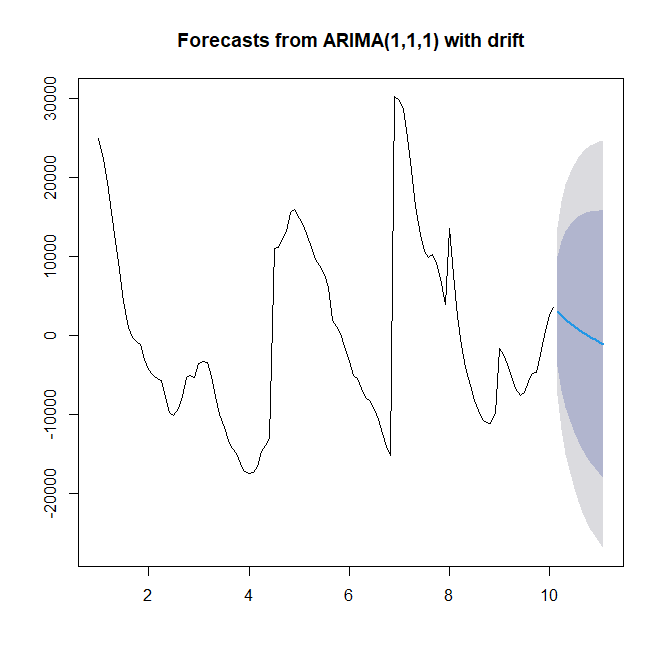
* 시계열 그림에서 몇 이상치를 제외하고 0에 가까운 값이므로 정상성 시계열임을 파악할 수 있다.
* ACF 그림에서 lag>=1 일 때 모든 acf값이 0이고, PACF 그림에서 모든 lag에 대해 값이 0이므로 적합된 ARIMA(1,1,1) 모형의 가정에는 문제가 없다.
* 객관적인 검정을 위해 Box.test를 통해 화이트 노이즈 검정을 진행한 결과, p-value는 0.9로 매우 커서 귀무가설 채택. 즉 , 모든 시차에서의 자기 상관함수는 0이고 잔차는 White Noise를 따른다고 볼 수 있다. 따라서 ARIMA(1,1,1) 모형은 잘 적합된 모형이라고 말할 수 있다.

1. **예측**
2. **잔차 분석**

|  |
| --- |
| > forecast(arma11,h=12)  > plot(forecast(arma11,h=12)) |

Output]

|  |
| --- |
| Point Forecast Lo 80 Hi 80 Lo 95 Hi 95  Mar 10 3039.7736 -3885.861 9965.408 -7552.07 13631.62  Apr 10 2526.7037 -6906.715 11960.122 -11900.46 16953.87  May 10 2049.2301 -9091.680 13190.141 -14989.32 19087.78  Jun 10 1604.4112 -10814.922 14023.745 -17389.32 20598.14  Jul 10 1189.5483 -12230.483 14609.579 -19334.62 21713.71  Aug 10 802.1658 -13421.945 15026.276 -20951.73 22556.07  Sep 10 439.9928 -14441.557 15321.543 -22319.37 23199.36  Oct 10 100.9458 -15324.924 15526.815 -23490.89 23692.78  Nov 10 -216.8863 -16097.742 15663.970 -24504.56 24070.79  Dec 10 -515.2567 -16779.329 15748.816 -25389.01 24358.50  Jan 11 -795.7738 -17384.643 15793.096 -26166.26 24574.71  Feb 11 -1059.9130 -17925.535 15805.709 -26853.66 24733.83 |



* Forecast 함수를 통해 예측한 값을 보면, 점차 감소하는 모습을 볼 수 있다. 이는 앞서 첫 과정에서 추세를 제거했기 때문에 가격이 하락하는 것을 볼 수 있다.
* 증가 추세를 제거하고 한 번 차분한 정상 시계열은 증가와 하락이 반복되는 순환성을 가졌고, 2021년 2월 이후 미래의 값에 대해서는 하락하는 주기를 맞을 것을 예측하였다.
* 하지만 이 때 예측한 자료는 원 데이터를 변형한 정상 시계열 자료이고, 실제 데이터는 정상 시계열과 증가 추세가 합쳐진 데이터이기 때문에 실제 서울시 아파트 실거래 가격 지수는 시간에 따라 증가할 것으로 예상할 수 있다.