서울시 아파트 실거래 가격 지수 예측

고영희

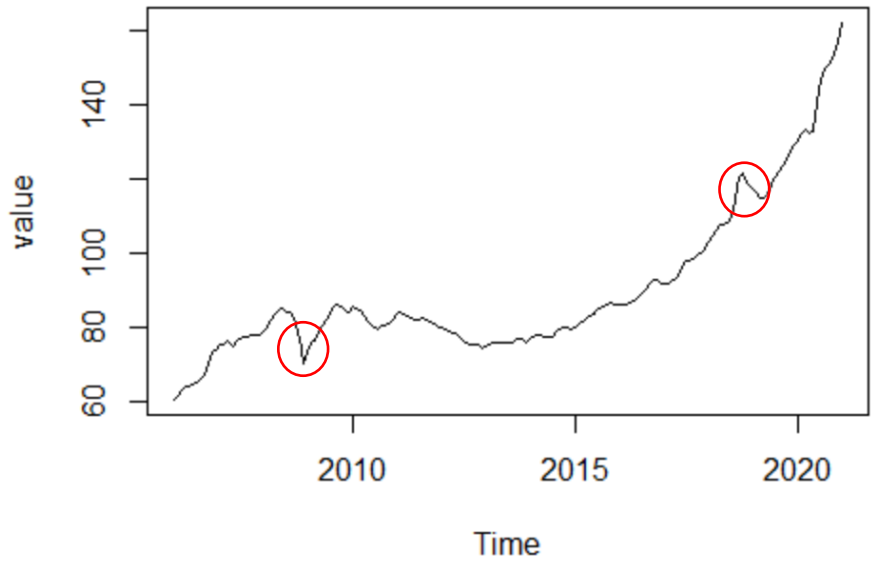
1. 분석 목적

최근 주식, 부동산 등에 대한 투자 열풍이 강하다. 케이비(KB)증권은 이러한 사회적 경향성은 사상 초유의 저금리가 지속되면서 근로소득으로 재산을 형성하기 어렵다는 판단이 확산되며 자산가격이 빠르게 오르는 주식과 아파트, 주택 등으로 재태크 하는 것으로 설명할 수 있다고 한다. 그 결과로 부동산 가격이 급상승하자, 현 정부는 출범이후 22개의 부동산 대책을 내놨지만 부동산 가격의 급격한 상승을 막지 못했다(BBC 코리아, 2020.07.16). 2021년 1월 18일 신년 기자회견에서 대통령이 직접 "부동산 시장 안정화에 성공하지 못했다."고 인정할 만큼 주택시장의 과열 시장을 잡을 수 있는 정책 결정은 쉽지 않다. KB 부동산 보고서에 따르면 2014년부터 집값이 상승세를 타기 시작해 올해 7년째 상승세를 이어가고 있다. 그렇다면 이러한 상승세는 올해도 지속될 것인지, 지속된다면 언제까지 상승세를 이어갈 수 있을지, 상승폭은 어느정도 될 것인지를 추정하는 것이 이번 프로젝트의 목적이다.

2. 데이터 수집

한국 부동산원에서 제공하는 지역별 매매(월별) 데이터 중, 서울시 지역을 대상으로 데이터 필터링을 진행하였다. 서울지역을 선택한 이유는 다른 시,도에 비해 주택 공급량이 매우 부족하고 학군에 대한 수요가 많기 때문에 타지역보다 좀 더 빠른 상승세를 겪은 도시이기 때문이다. 사용된 데이터는 2006년 1월부터 2021년 1월까지 서울시 아파트 평균 실거래 가격을 조사한 월별 데이터이다. 아파트 실거래가 지수란, 2017년 11월의 서울시 아파트 실거래 가격을 100으로 보았을 때 각 년도별, 월별 아파트 매매 평균 가격을 계산한 자료이다. 총 181개의 값이 관측되었고, 데이터를 구분하는 인덱스는 (년, 월)의 형태로 표현될 수 있다. 따라서 프로젝트에 사용된 데이터는 일변량 시계열 데이터라 할 수 있다.

3. 데이터 형태 파악



시간에 따른 서울시 아파트 실거래 가격지수는 위와 같이 선도표로 그려질 수 있다. 여기서 알 수 있는 점은 2008년 12월, 아파트 가격 변동률이 -5.53%로 폭락해 가격 지수 70.2까지 하락했다는 점이다. 2008년의 경우, 7월달을 기준으로 전년도 대비 가격 변동률이 0 이상인 달이 거의 없었다. 이는 2008년 금융위기 충격과 공급 과잉 현상이 맞물리면서 반영된 현상이라 볼 수 있다. 또한 10년 뒤인 2018년 8월부터 가격지수의 변동률은 4.07 ~ 5.53까지 상승하면서 10월에는 120.8이라는 지수를 달성했다. 집값 변동 폭이 크게 변했다는 것은 위 그림에서도 파악할 수 있다. 이 원인으로는 서울 연구 데이터 서비스에서 발간한 서울시 집값 상승과 대책 월간지에 따르면 부동산 투기와 주택 공급량 부족으로 인한 현상임을 알 수 있다. 2019년도 4월 이후 주택가격은 점점 상승하는 추세로, 이러한 추세를 따랐을 때, 미래의 주택가격은 상승할 것임을 기대할 수 있다.

4. 추세분석

1) 데이터 전처리

- 데이터 로드
- 시계열 데이터로 변환
- 시간에 따른 가격지수 산점도 형태로 시각화

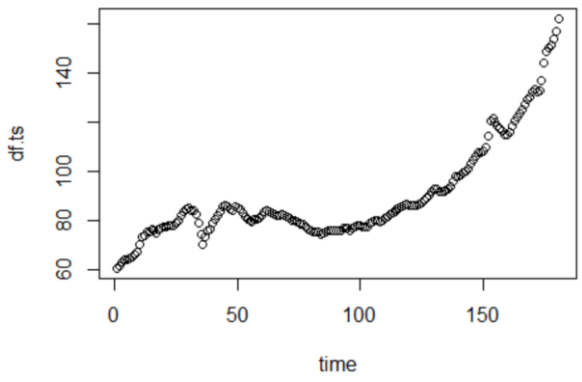
```
> df<-read.csv("apart_sales.csv")
> head(df)
> names(df)<-c('year','month','value')

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

> time<-1:nrow(df)
> plot(time,df.ts)
```

Output]

	year	month	value
1	2006	1	60.6
2	2006	2	61.7
3	2006	3	63.2
4	2006	4	64.1
5	2006	5	64.1
6	2006	6	64.4



2) 선형 모형 적합

- 회귀분석에 가장 기본이 되는 선형회귀식 적합
- 선형회귀 모형에서의 회귀계수 및 R^2값 파악
- 원 데이터의 산점도 위에 적합된 회귀식 시각화

```
> reg1<-lm(df.ts~time)

> summary(reg1)

> abline(reg1,col='red')
```

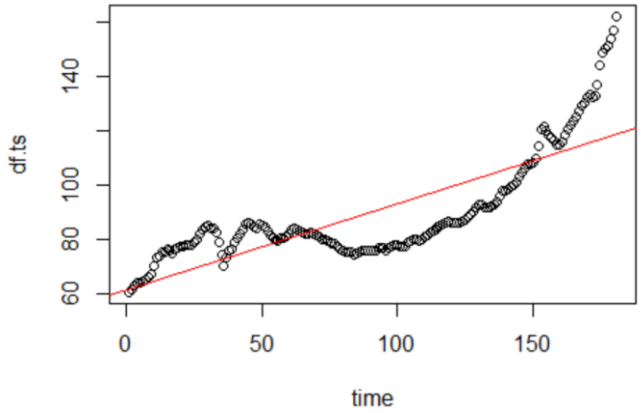
Output]

```
Call:
lm(formula = df.ts ~ time)

Residuals:
     Min       1Q   Median       3Q      Max
-16.239 -11.782   0.827   8.907  43.594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.1340     1.8363   33.29   <2e-16 ***
time          0.3175     0.0175   18.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.3 on 179 degrees of freedom
Multiple R-squared:  0.6478,    Adjusted R-squared:  0.6458
F-statistic: 329.2 on 1 and 179 DF,  p-value: < 2.2e-16
```



결과해석]

y=61.13 + 0.3175 * (time)이라는 회귀식을 도출하였다. y절편과 time의 회귀계수의 p-value가 0.05이하이므로 두 값 모두 유의한 것으로 파악되었다. 또한 모형의 결정계수 즉, R^2 값은 0.6478이라는 것을 알 수 있다. 실제 데이터 산점도 위 회귀식을 시각화 한 결과, 적합된 회귀식은 데이터의 분포 패턴을 제대로 표현하지 못한다고 볼 수 있다.

3) 2차 회귀식 적합

```
> reg2<-lm(df.ts~time+I(time^2))

> summary(reg2)

> lines(reg2$fitted.values,col='blue')
```

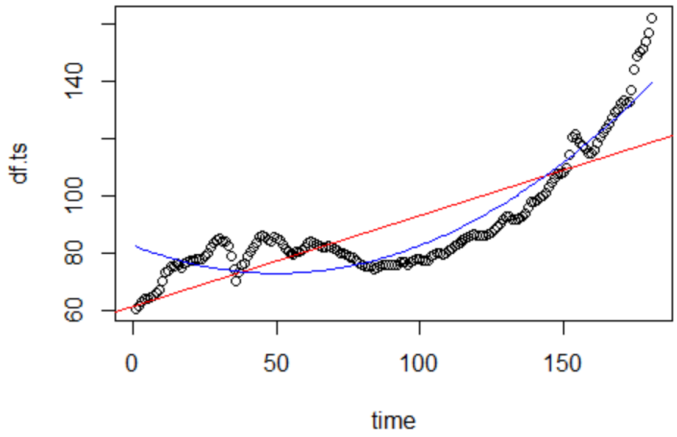
Output]

```
Call:
lm(formula = df.ts ~ time + I(time^2))

Residuals:
     Min       1Q   Median       3Q      Max
-21.742  -5.277  -1.777   6.539  22.703

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.7286466  1.7527231  47.200  < 2e-16 ***
time        -0.3904976  0.0444655  -8.782 1.31e-15 ***
I(time^2)    0.0038902  0.0002366  16.439  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.773 on 178 degrees of freedom
Multiple R-squared:  0.8601,    Adjusted R-squared:  0.8586
F-statistic: 547.4 on 2 and 178 DF,  p-value: < 2.2e-16
```



결과해석]

y=82.729 – 0.390 * (time) + 0.004 * (time^2) 회귀식을 도출하였다. y절편과 time, time^2의 회귀계수의 p-value 모두 0.05값 이하이므로 3개의 변수 모두 유의한 것으로 파악되었다. 또한 모형의 결정계수(R^2)는 0.8601이라는 것을 알 수 있다. 이 값은 앞선 선형모형의 결정계수 0.647보다 증가한 형태로 1차 회귀식보다 2차 회귀식에서 모형 설명력이 높아졌다는 것을 뜻한다. 또한 실제 데이터 산점도 위 2차 회귀식을 시각화 한 결과, 적합된 회귀식은 데이터 분포 패턴을 1차식보다 더 잘 설명하고 있다는 것을 알 수 있다.

4) 3차 회귀식 적합

```
> reg3<-lm(df.ts~time+I(time^2)+I(time^3))

> summary(reg3)

> lines(reg3$fitted.values,col='green')
```

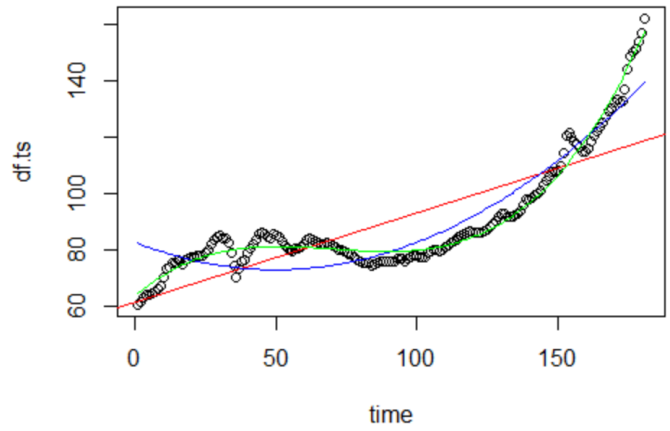
Output]

```
Call:
lm(formula = df.ts ~ time + I(time^2) + I(time^3))

Residuals:
      Min        1Q    Median        3Q       Max
-10.0134  -2.6731    0.4366    2.1014  10.9017

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.355e+01  9.838e-01   64.60   <2e-16 ***
time             8.566e-01  4.668e-02   18.35   <2e-16 ***
I(time^2)   -1.319e-02  5.951e-04  -22.17   <2e-16 ***
I(time^3)    6.258e-05  2.150e-06   29.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.241 on 177 degrees of freedom
Multiple R-squared:  0.9758,    Adjusted R-squared:  0.9754
F-statistic:  2382 on 3 and 177 DF,  p-value: < 2.2e-16
```



결과해석]

y=63.56 + 0.857 *(time) -0.013 * (time^2) + 0.000062 * (time^3) 회귀식을 도출하였다, y절편과 time, time^2, time^3 의 회귀계수의 p-value 모두 0.05값 이하이므로 4개 변수 모두 유의미하다는 것을 알 수 있다. 또한 모형의 결정계수는 0.976으로 앞선 2차 회귀식의 결정계수 0.8601보다 증가한 형태이므로 차원을 증가시켜 모형의 복잡성을 올리는 것이 고차원의 모형이 설명할 수 있는 편차가 훨씬 더 증가되었기 때문에 차원 증가의 충분한 이유가 된다. 이에 관해 적합된 3차 회귀식을 시각화하면 더 잘 파악할 수 있다. 기존의 1차, 2차 회귀식보다 3차 회귀식이 데이터의 패턴을 더 잘 표현하였고, 2008년도의 데이터에 대해서 실제 값보다 더 큰 값으로 적합되었고, 2018년도의 데이터에 대해서는 실제 값보다 약간 작은 값으로 적합되었다. 이는 예측불가능한 2008년 금융위기, 2018년 대량의 아파트 투기 등으로 인한 잔차로 생각할 수 있다.

5. 결론

```
> predict(reg3,newdata=data.frame(time=182), interval="confidence") # 2021.02

> predict(reg3,newdata=data.frame(time=193), interval="confidence") # 2022.01
```

적합된 3개의 모형중, 결정계수가 0.976으로 가장 큰 3차 회귀 모형을 최종모형으로 선택한다. y=63.56 + 0.857 *(time) -0.013 * (time^2) + 0.000062 * (time^3)으로 시간과 서울시 아파트 가격 지수의 관계를 표현할 수 있다. 미래에도 과거와 같은 데이터 분포 패턴을 따른다면, 한 달 뒤인 2021년 2월달의 아파트 가격 지수는 (157.752, 161.635) 신뢰구간 내 159.6932의 값을 가질 것을 알 수 있고, 1년뒤인 2022년 1월의 아파트 가격지수는 (184.32, 190.308) 신뢰구간 내 187.3139를 가질 것을 예측할 수 있다. 즉, 과거의 데이터 분포 패턴을 바꿀 수 있는 강력한 정부의 부동산 정책이 마련되지 않는다면 아파트 가격지수는 시간의 흐름에 따라 점점 더 증가할 것으로 예측할 수 있다.