데이터 마이닝 7조 모델링 중간 보고서

12132591 정동호

12131820 이건도

12161890 하나영

< 목 차 >

1. 주제

- **주제**

  인천광역시 아파트 단지 데이터를 이용한 아파트 제곱미터당 평균 매매 가격 예측 및 영향을 주는 변수 분석

- **분석 단위**

  최근 3년(16년 1월부터 18년 12월까지의 월 기준) 인천광역시 단지별 아파트 평균 매매 가격 기준

2. 변수 설명

분류	변수	변수 설명
외부변수		
	1.역세권점수	해당 읍면동을 지나는 지하철역의 개수
	2.교육시설 수	읍면동 초,중,고 합계
	3.문화시설 수	해당 읍면동의 영화관, 박물관, 공원, 테마파크 등 문화시설 수
	4.쇼핑시설 수	해당 읍면동의 백화점, 마트, 편의점 등 쇼핑시설 수
	5.공시지가변동률* (시계열)	1) 월별 : 읍면동의 매월 지가변동 상황 2) 누계 : [(당해월 지가지수 / 전년말 지가지수) - 1] * 100
	6.기준금리* (시계열)	금리 체계의 기준이 되는 금리
	7.서울집값	같은 시기의 강남 평균 매매가격(단위-만원)
	8.개발호재* (범주형)	해당 읍면동의 택지개발사업, 교통시설 예정지역 등의 존재 유무
	9.평균 인구수	읍면동 인구수의 3년 평균 (단위-건수)
	10.평균 세대수	읍면동 세대수의 3년 평균 (단위-건수)
	11.평균 혼인건수	구별 혼인건수의 3년 평균 (단위-건수)
	12.스타벅스	해당 읍면동 안의 스타벅스 매장 개수.(단위-건수)
	13.토지면적	해당 읍면동의 토지 면적(단위-km^2)
		* 제외하고 연속형 변수
내부변수		
	1.평균 전세가	아파트의 월별 m^2당 평균 전세가(단위-만원)
	2.분양면적	아파트의 전용면적+주거면적. 소위 집안 면적(단위-미터제곱)
	3.아파트 노후도	18 년 12 월 기준 아파트가 지어진 연도의 차(연식)
	4.단지세대수	아파트 전체 단지 내 입주하고 있는 세대의 수
	5.개별세대수	아파트 분양면적에 따라 입주하고 있는 세대의 수

3. Data Set 구분

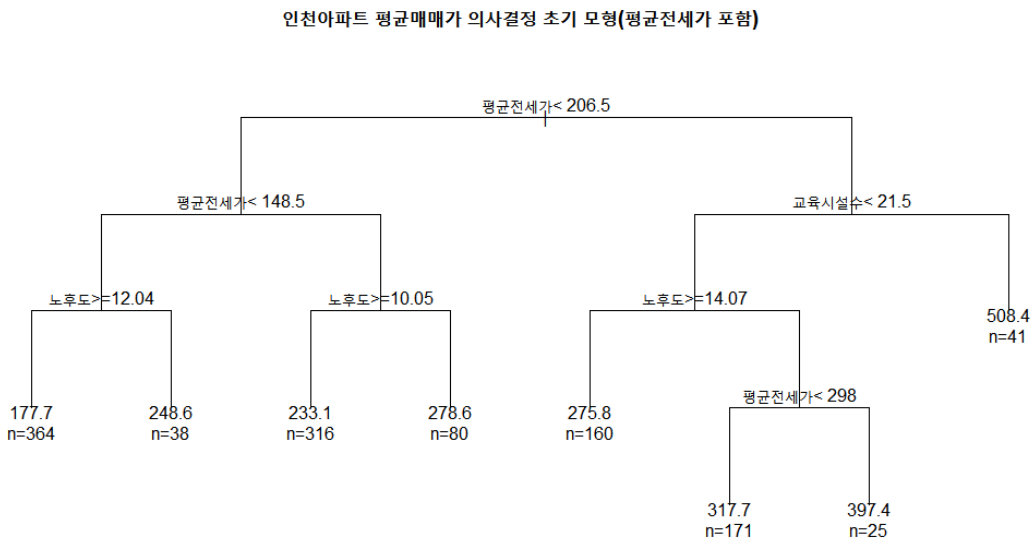
모델링을 학습하고 검증하고 테스트하기 위해서 저희는 Train, validation, test set을 4 : 3 : 3의 비율로 랜덤추출을 통해 구분하였습니다.

Data Set 종류	Data 수
Train set	1195
Validation set	895
Test set	895

4. 의사결정학습법

- 의사결정 학습법 package 중 rpart package 이용
- 목적변수 연속형에 따른 회귀 의사결정 학습법

4-1 초기모형 - Train set을 통한 의사결정학습법 결과



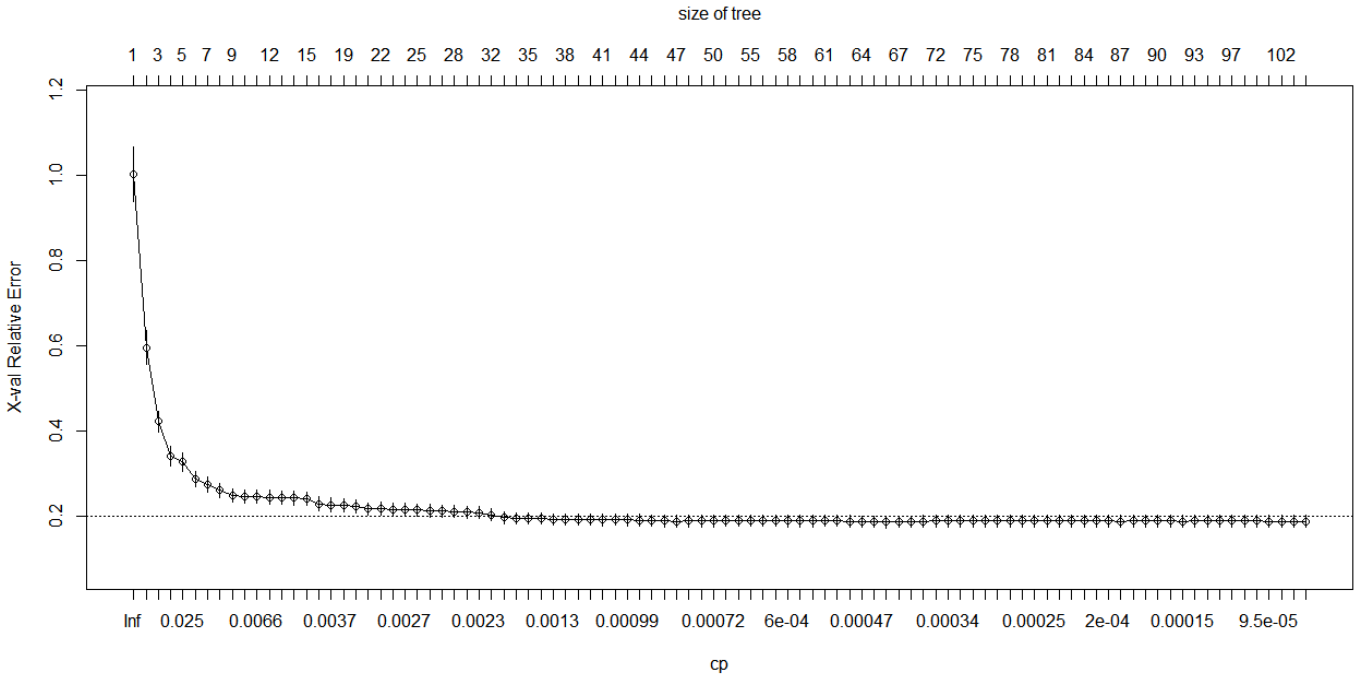
<그림 1>

초기 모형의 경우, Terminal node의 수는 8개이고, 노드 분리에 있어 가장 영향을 주는 변수가 "평균전세가"임을 알 수 있습니다. 그 다음으로는 "노후도"임을 알

수 있습니다. 그리고 가장 높은 값인 508.4의 경우, 평균 전세가가 206.5보다 크고 교육시설수가 21.5보다 클 때라는 것을 알 수 있습니다.

4-2 Pruning 적용

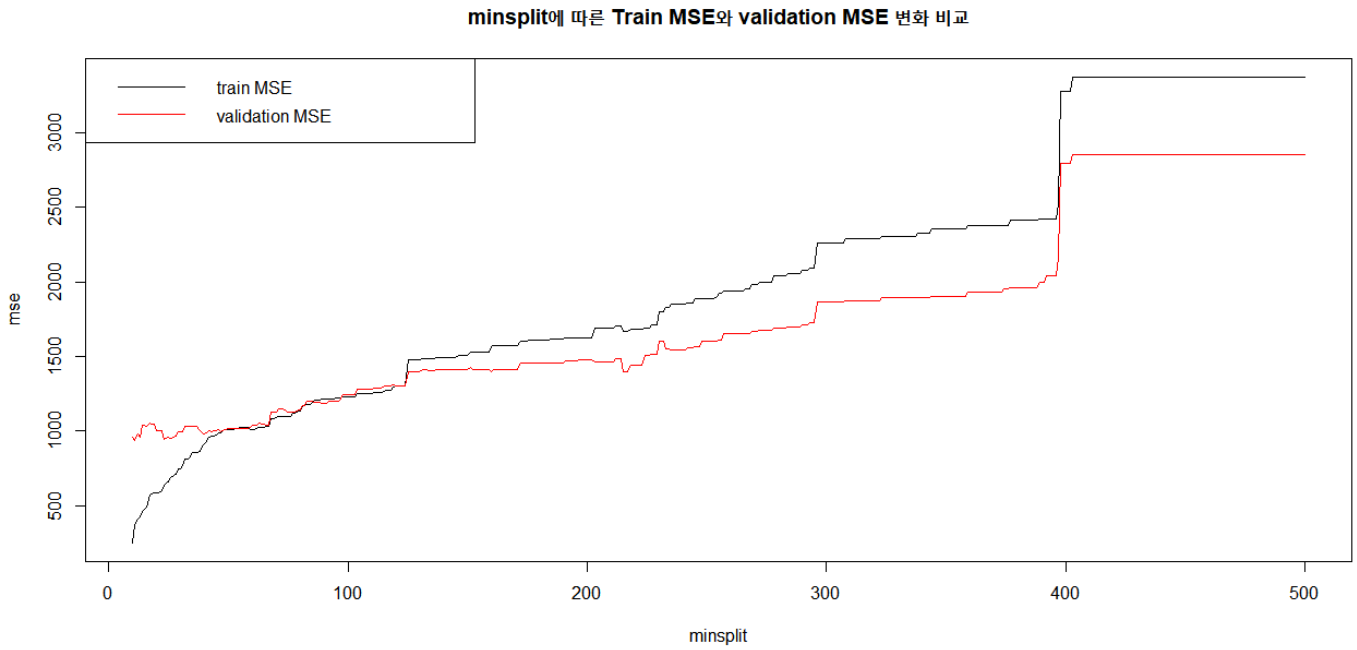
초기모형 그림1의 경우 rpart 함수의 적용 결과만을 판단한 것이기에 정확한 Pruning을 위해 Complexity parameter(오분류값)이 0이라는 조건을 두어 가장 낮은 오분율에서 적절한 terminal node 수가 몇 개인지 판단하기로 했습니다.



<그림 2>

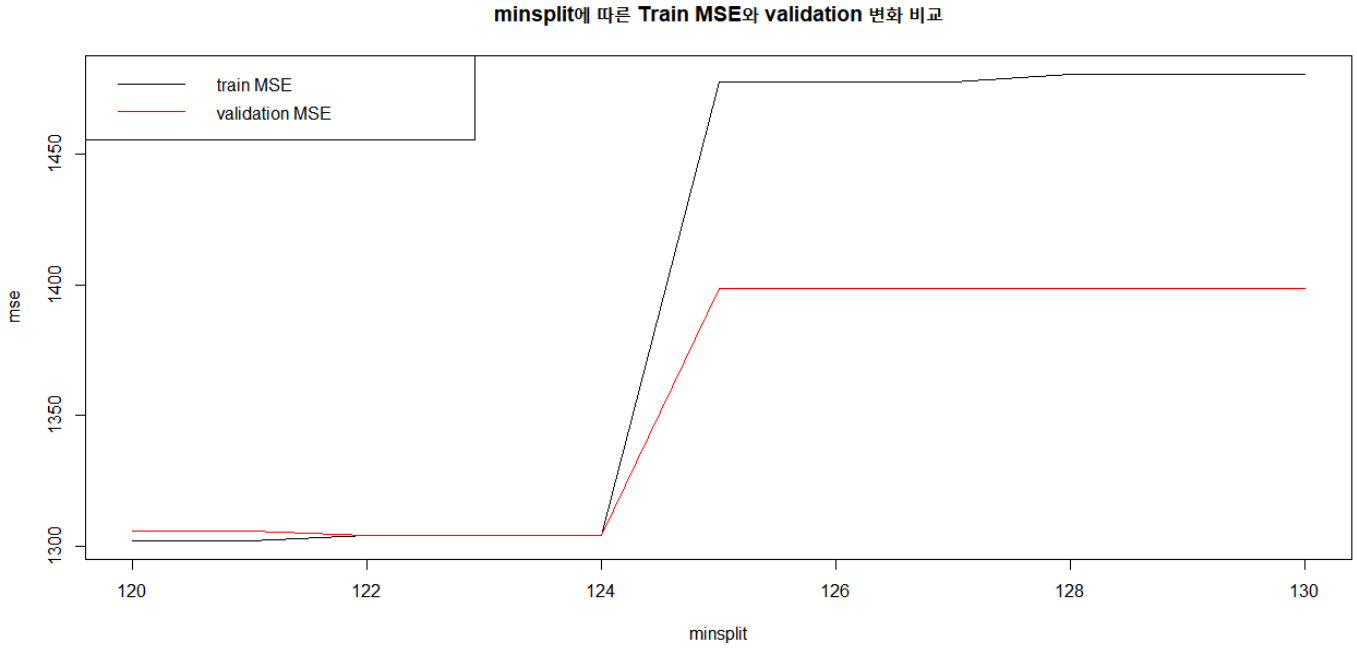
그림2를 보시면, terminal node의 수가 증가할수록 오분율 값이 계속 낮아지는 것을 확인할 수 있습니다.

적절한 Pruning을 하기전에 Validation set과 Train set을 통해 terminal node수에 따라 각 MSE의 변동을 알아보았습니다. 먼저 minsplit(노드 분할하기 위해 필요한 최소 데이터의 개수)변동은 결국 terminal node 수와 연관 있기 때문에 minsplit에 따라 validation set과 train set의 MSE을 측정했습니다.



<그림 3>

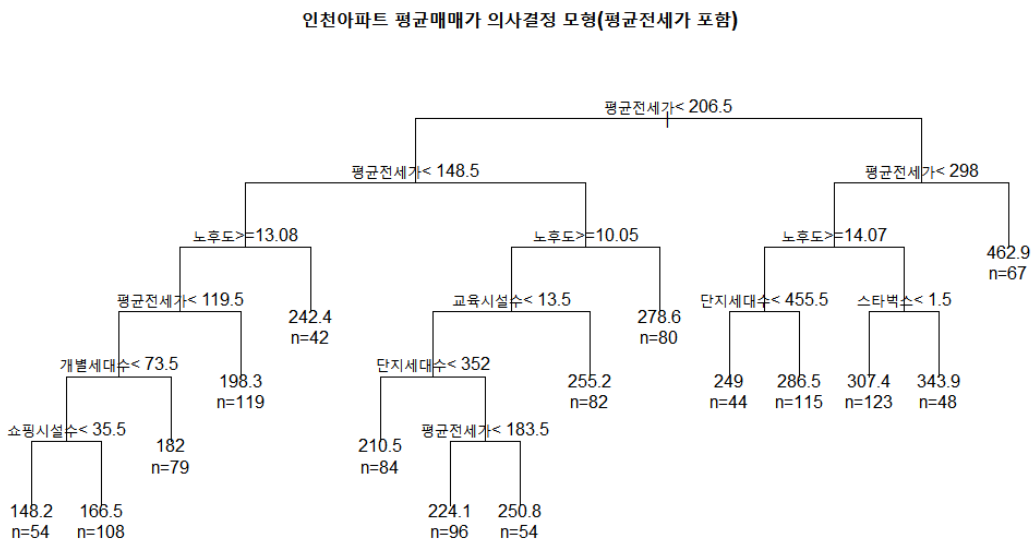
Minsplit이 낮을수록(terminal node 수가 많을수록) validation과 train의 MSE 또한 낮아지는 것을 확인할 수 있습니다.



<그림 4>

<그림3>과 <그림4>를 보면 minsplit이 500부터 125까지 validation MSE가 train MSE보다 계속 낮다는 것을 확인할 수 있습니다. 그러므로, <그림2>의 결과와 <그림4>의 분석에 따라 minsplit이 125일 때 지점을 선택하여 Pruning을 했습니다.

4-3 최종 모형 회귀 의사결정 학습법 결과



<그림 5>

<그림 5>에 대해 train MSE, validation MSE, test MSE을 비교한 결과

Train MSE	1477.903
Validation MSE	1398.396
Test MSE	1303.861

세 개의 MSE 값 중에서 Test MSE가 가장 낮은 것을 확인할 수 있습니다.

종합적인 판단 결과 terminal node수가 15개를 가진 회귀 의사결정 모형으로 선택했습니다. 또한, Root node을 나누는 결정적인 변수가 "평균전세가"인 것을 알 수 있으며, 또한 이후 노드 분할에 있어 가장 많이 기준이 되는 변수는 평균전세가"임을 알 수 있습니다.

변수명	분기 노드 수
평균전세가	5
노후도	3
단지세대수	2
교육시설수, 스타벅스, 개별세대수, 쇼핑시설수	1

위의 평균전세가, 노후도, 단지세대수가 인천아파트 평균매매가 형성에 많은 영

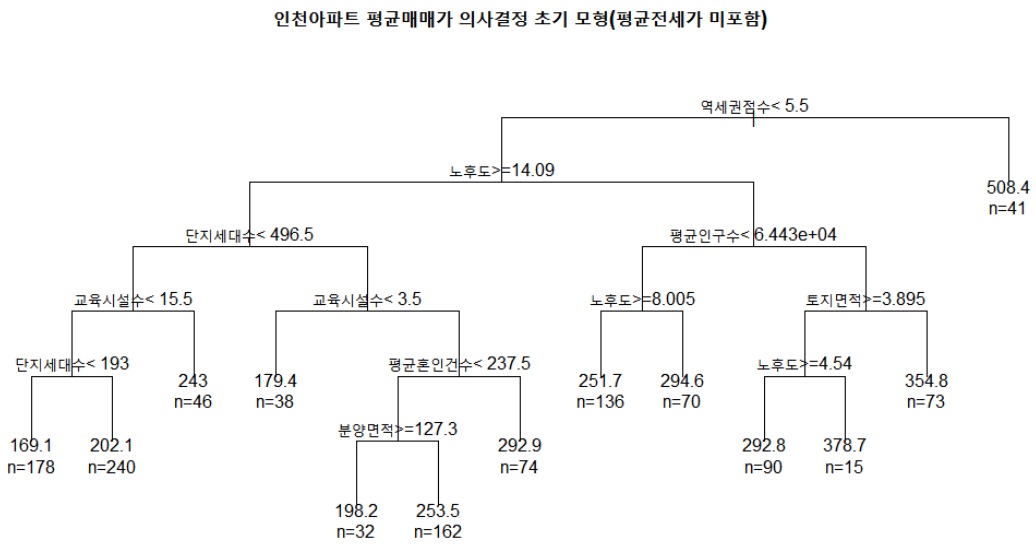
향을 주는 변수라는 것을 의미한다고 볼 수 있습니다.

4-4 결론

회귀 의사결정학습법을 통해서 인천아파트 평균매매가를 분석한 결과, 아파트와 관련된 내부변수, 즉 평균전세가, 노후도, 단지세대수가 다른 외부변수에 비해 인천 아파트 평균매매가에 많은 영향을 준다는 것을 확인할 수 있습니다. 또한 초기모형(그림1)에서 가장 높은 값을 형성하는데 영향을 주는 변수가 "평균전세가"와 "교육시설수" 변수였지만, 최종모형(그림5)에는 모두 "평균전세가"임을 확인할 수 있습니다.

4-5 추가) "평균전세가"를 제외한 의사결정 학습법

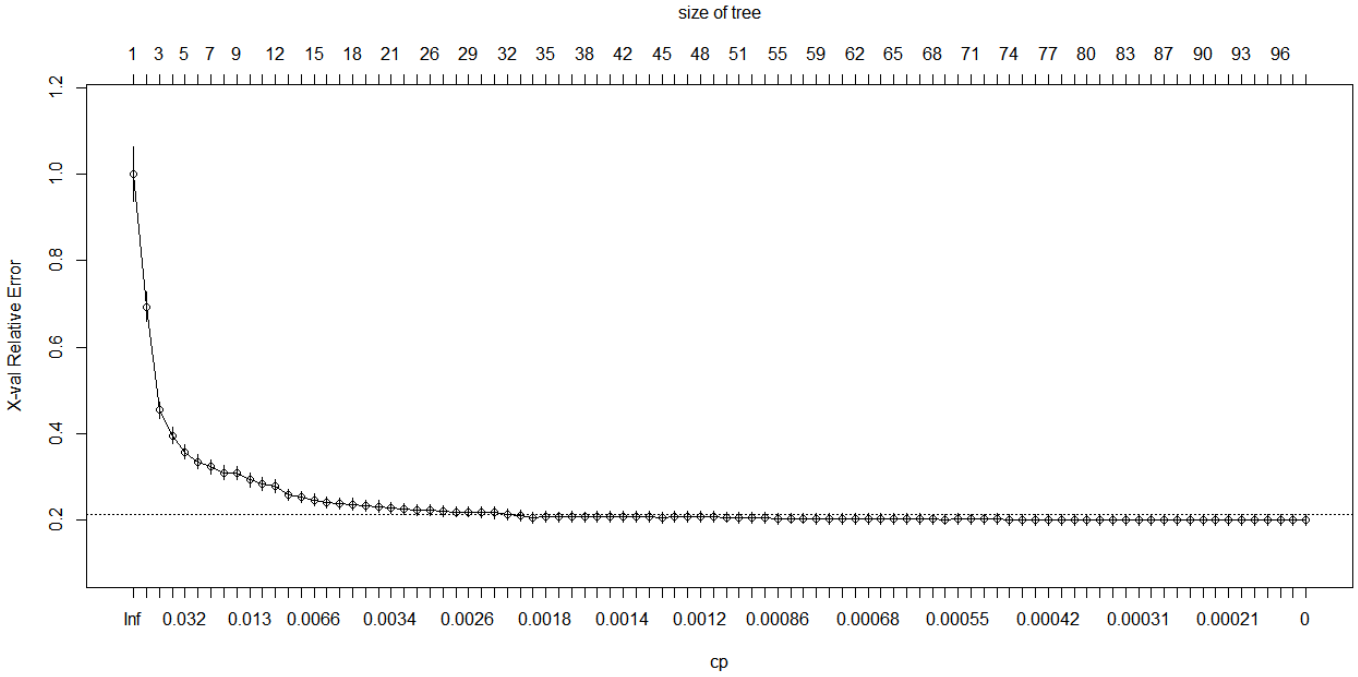
- **초기모형**



<그림 6>

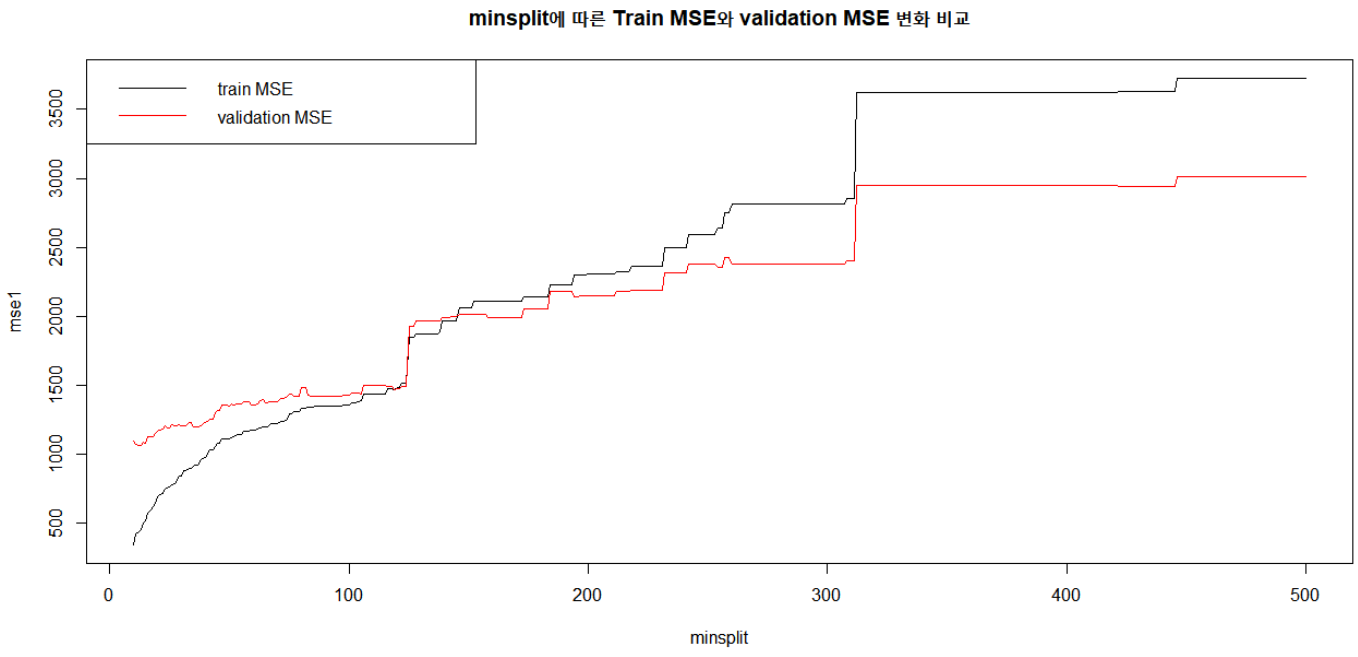
초기모형 그림6에 따르면, root node 분기 기준이 "역세권점수"임을 파악할 수 있습니다.

- **Pruning 적용**

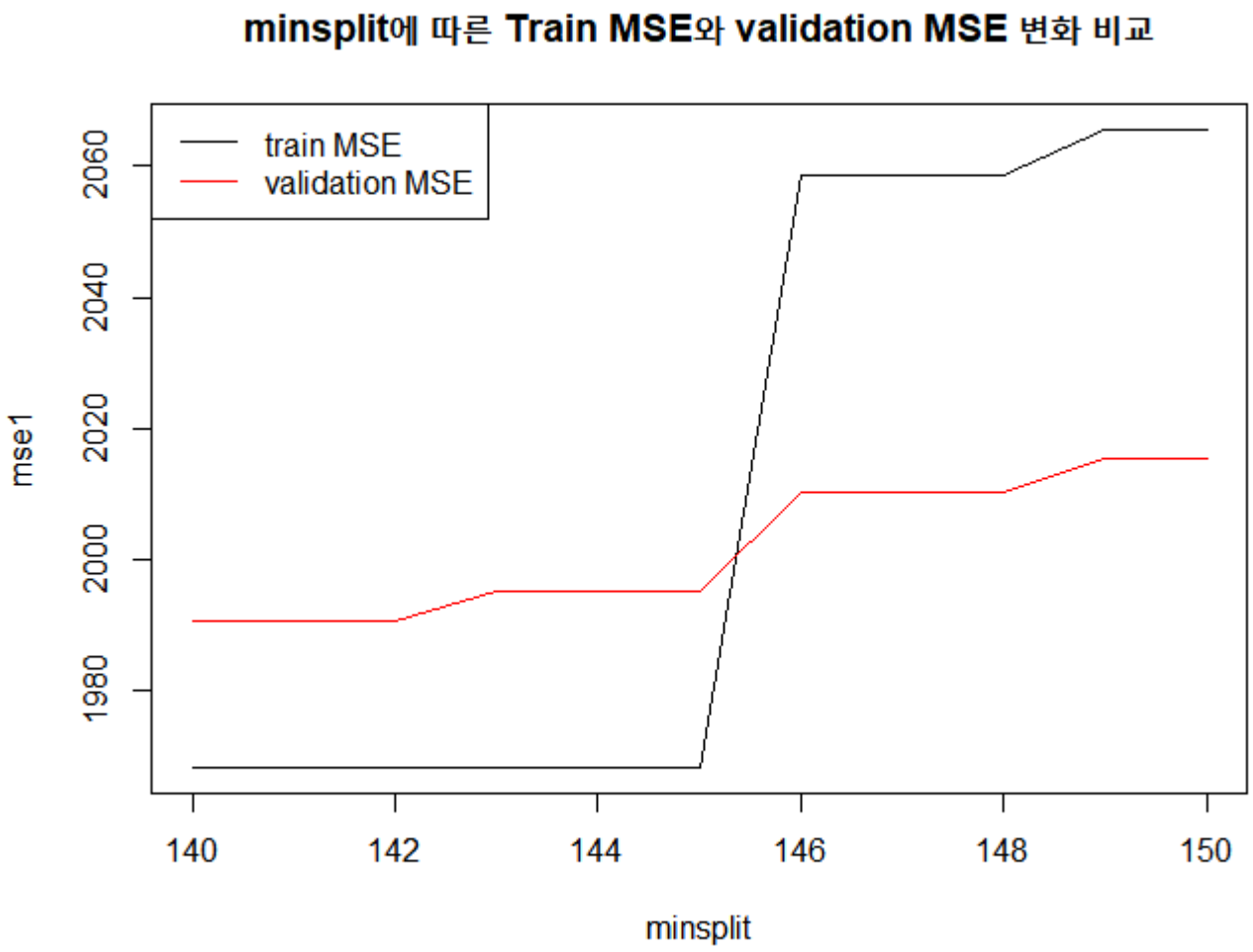


<그림 7>

그림7을 보면 terminal node 수 증가에 따라 오분율이 계속 낮아지는 것을 확인할 수 있습니다.



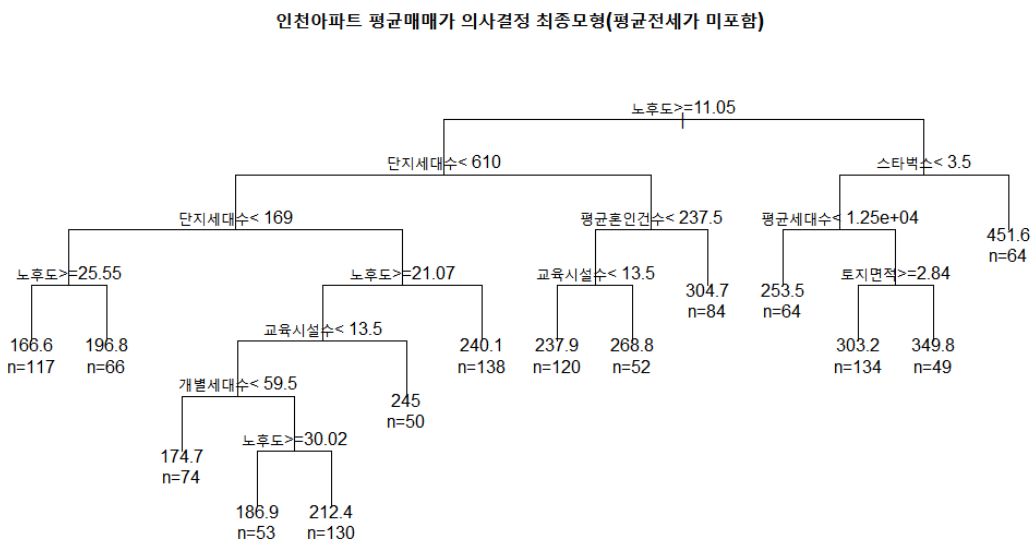
<그림 8>



<그림 9>

그림8과 그림9을 종합하여 판단하면, minsplit이 146에서 선택하는 것이 최적이라는 것을 고려할 수 있습니다.(terminal node 수는 14개)

- **최종모형 결과**



<그림 10>

최종모형 그림10에 따라 train MSE, validation MSE, test MSE을 비교해보면,

Train MSE	2058.659
Validation MSE	2010.161
Test MSE	2244.963

오히려 Train MSE, Validation MSE보다 Test MSE가 높다는 결과를 볼 수 있습니다.

변수에 따른 분기 노드의 수를 파악한다면

변수명	분기 노드 수
노후도	4
단지세대수, 교육시설수	2
스타벅스, 평균세대수, 개별세대수, 평균혼인건수, 토지면적	1

가장 영향을 주는 변수는 "노후도"로 내부 변수에 해당합니다.

- **결론**

평균전세가 미포함한 초기모형과 최종모형을 비교하면, 초기모형과는 다르게 최종모형에서의 root node가 바뀐 것을 확인할 수 있습니다.

평균전세가를 포함한 결과와 미포함한 결과를 분석하면, 전반적인 data set의

MSE의 변화를 통해 평균전세가가 인천아파트 평균매매가에 많은 영향을 준다는 것을 유추할 수 있습니다. 노후도 변수 또한 인천 아파트 평균매매가 형성에 상당하게 영향을 주는 것을 유추할 수 있습니다.

4-5 한계점

Pruning 단계에서 terminal node수가 많을수록 오분율이 계속 낮아지는 추세를 갖기에 이 부분에 있어 인터넷 검색을 통해 정보를 얻으려고 했으나 거의 전무했기에 적정한 수준의 terminal node를 선정하는데 어려움이 있었습니다. 추후 학습이 필요한 부분이라 생각이 됩니다.

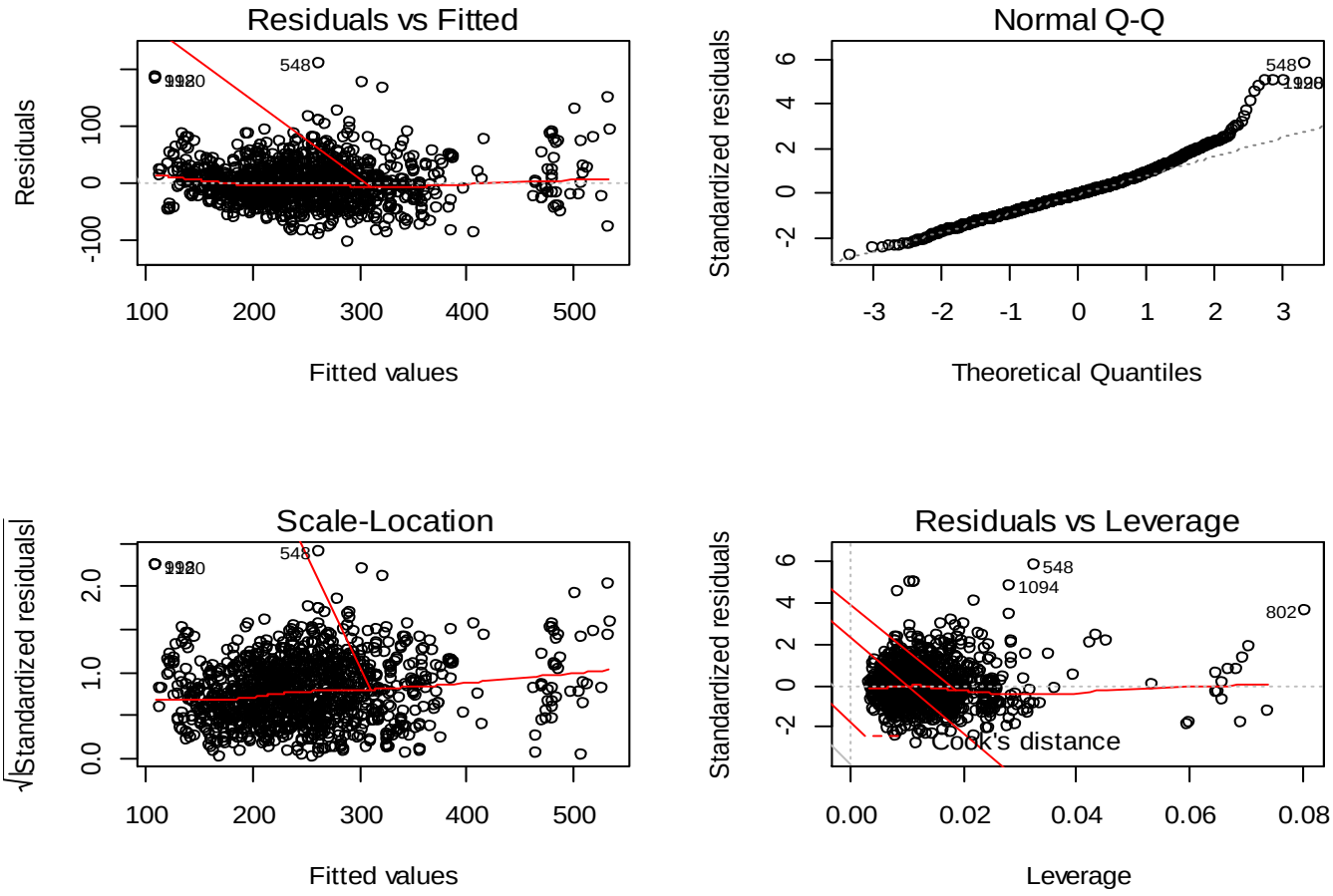
5. 회귀분석

- 회귀모형의 이상치 찾기
- 독립성(자기상관성) , 선형성, 등분산성, 정규성 만족하는지 검정
- 다중공선성 검정 후 3가지 모형선택법을 이용해서 설명변수 선택하기
- (1)Forward selection , (2)Backward selection, (3) Stepwise selection

5-1. 초기모형 설정 및 모형진단

```
# 변수 설정

Y = train$평균매매가

X1 = train$분양면적

X2 = train$단지세대수

X3 = train$개별세대수

X4 = train$노후도

X5 = train$역세권점수

X6 = train$평균전세가
```

```
X7 = train$토지면적

X8 = train$문화시설수

X9 = train$쇼핑시설수

X10 = train$스타벅스

X11 = train$평균인구수

X12 = train$평균세대수

X13 = train$교육시설수

X14 = train$개발호재

X15 = train$평균혼인건수


# 회귀 적합

reg.train= lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11

+X12+X13+factor(X14)+X15,data=train)
```



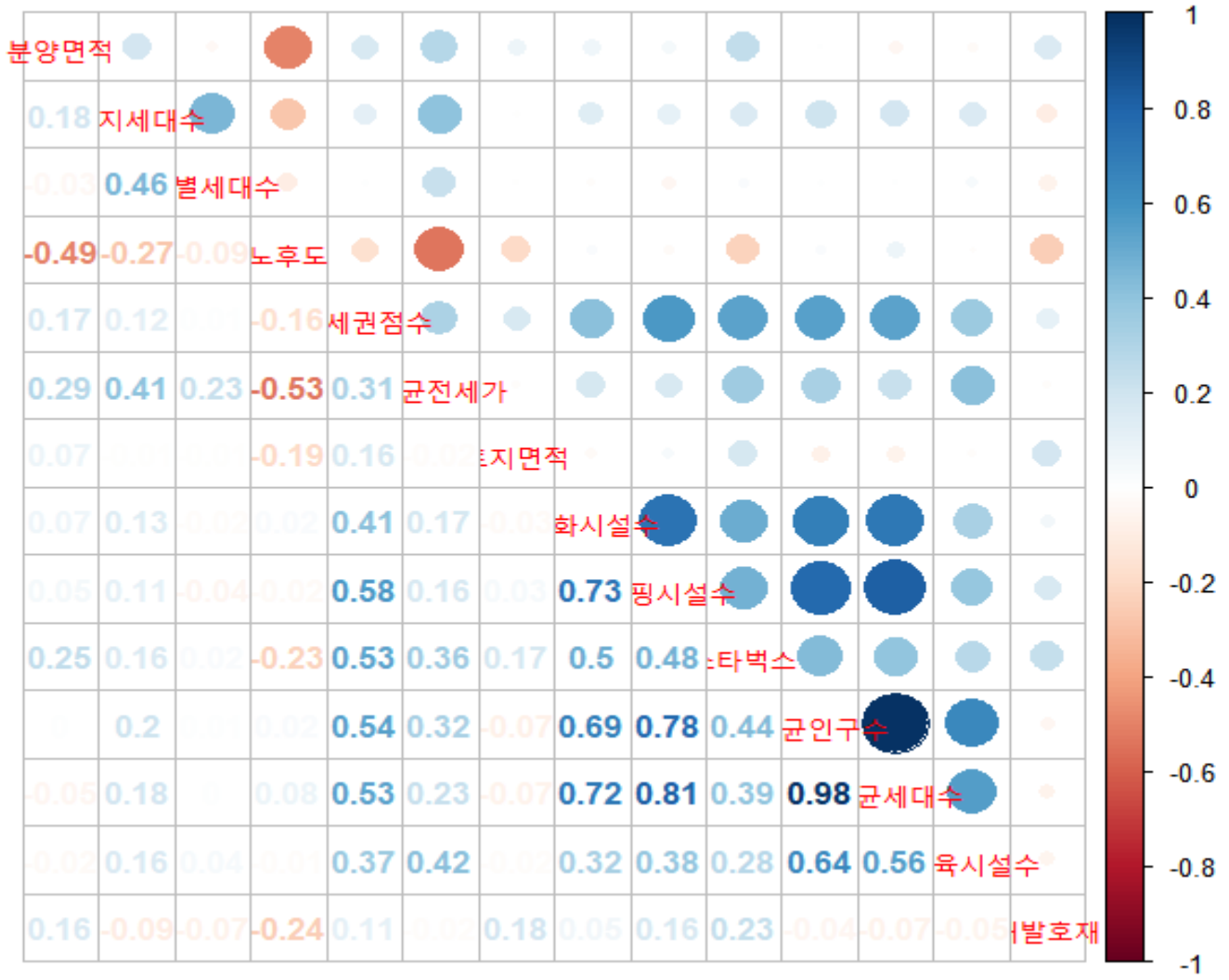
1) 정규성 : 정규확률그림에 따르면 직선을 보이다가 점점 직선을 벗어나는 것처럼 보인다. 하지만 표본수가 굉장히 크기 때문에 중심극한정리에 의해 큰 랜덤표본은 근사적으로 정규분포를 따른다.

2) 독립성 : 더빈왓슨 통계량에 의하면 2.045로 2보다 크므로 자기상관성이 없다고 볼 수 있다.

3) 선형성 : 잔차와 예측치 사이에 특별한 형태가 보이지 않으므로 만족한다.

4) 등분산성 : 잔차 플롯에 따르면 등분산성을 만족하지만 이상치가 존재한다. 이는 EDA에서 밝혔듯이 송도동의 데이터일 것이다.

5-2 다중공선성 확인



-다중공선성 VIF가 10보다 크면 다중공선성이 존재한다고 볼 수 있는데 평균인구수(X11)와 평균세대수(X12) 사이에서 나타난다.

- 두 변수간의 상관관계를 분석했을 때 0.9824로 매우 큰 상관관계가 있음을 알 수 있다.

-다중공선성은 추정된 회귀계수의 불안정성과 관련되어 있으므로 이를 해결하기 위해 변수선택법을 활용하겠다.

5-3. 변수선택법

1) Forward selection

: lm(formula = Y ~ X6 + X4 + X13 + X10 + factor(X14) + X5 + X9 + X11 + X12 + X2 + X15 + X7 + X8, data = train)

Multiple R-squared: 0.7989, Adjusted R-squared: 0.7967

2) Backward selection

: lm(formula = Y ~ X2 + X4 + X5 + X6 + X7 + X8 + X9 + X11 + X12 + X13 + factor(X14) + X15, data = train)

Multiple R-squared: 0.7988, Adjusted R-squared: 0.7968

3) Stepwise selection :

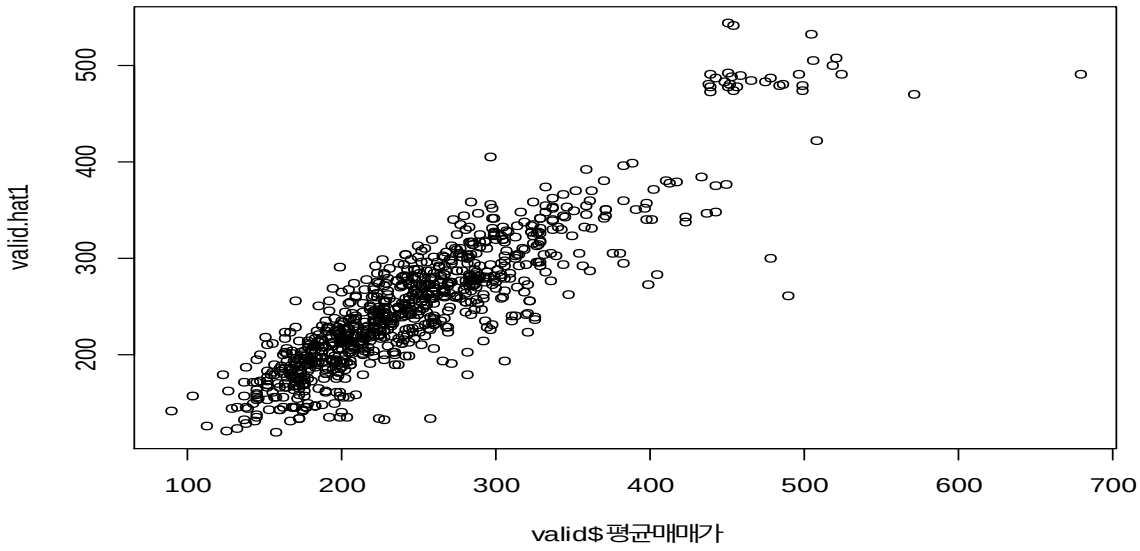
lm(Y~X2+X4+X5+X6+X7+X8+X9+X11+X12+X13+factor(X14)+X15

,data=train)

Multiple R-squared: 0.7988, Adjusted R-squared: 0.7968

-backward와 stepwise의 결과가 동일하게 나타나고 결정계수가 forward보다 크므로 lm(Y~X2+X4+X5+X6+X7+X8+X9+X11+X12+X13+factor(X14)+X15) 회귀모형을 채택하게 된다.

5-4 Validation 모형으로 예측해보기



-기존의 MSE가 1220.949인데 1217.055로 감소했다.

결정계수는 0.7888로 증가했으므로 설명력이 증가했음을 알 수 있다.

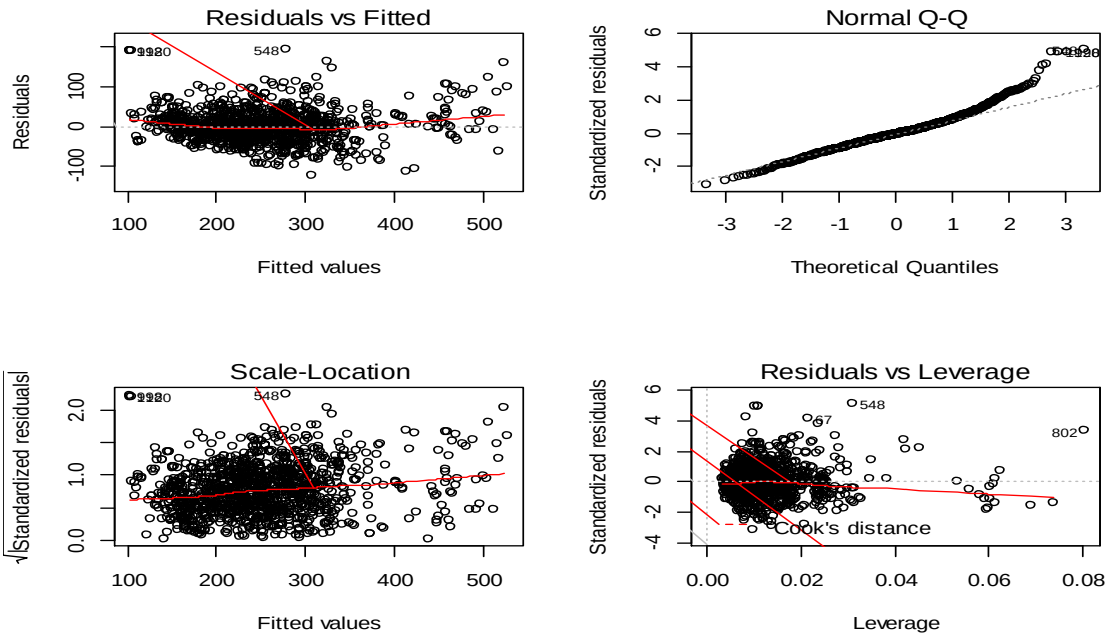
-하지만 여전히 X11과 X12사이에 공선성이 존재한다. 두 변수를 각각 제거하여 비교해보겠다.

1) X11을 제거 후 변수 선택

-validation set을 통한 MSE는 1348.721로 1220.949보다 높게 나왔다.

2)X12를 제거 후 변수 선택

- 위와 동일한 방법으로 MSE를 계산한 결과 1328.958로 X11 제거시(1347.754)보다 낮고, origin(1220.949)보다는 높다. 따라서 X12를 제거하기로 한다.



5-5. 변수 X12를 제거 후 다시 Stepwise selection 진행

lm(Y ~ X1+X2+X4+X5+X6+X7+X9+X10+X11+X13+factor(X14)+X15,data=train)

-X3과 X8이 제거 되었다.

Multiple R-squared: 0.7776, Adjusted R-squared: 0.7754

- 1334.331로 X11제거후 variable selection(1348.721)보다 낮지만, orign varable selection(1217.055)보다는 높다

<MSE 비교표>

모형	Train MSE	Validation MSE
초기 모형	1369.551	1220.949
초기 모형 stepwise	1367.26	1217.055
X11 제거 모형	1527.122	1347.754
X11 제거 stepwise	1525.889	1348.721
X12 제거 모형	1509.502	1328.958
X12 제거 stepwise	1511.268	1334.331

최종모형은 X12(평균세대수)변수를 제거하고 stepwise을 통한 모형이다.

-따라서 다중공선성을 해결하기위해 3가지 변수선택법을 이용하였다.

3가지 중 Stepwise 방법을 사용하여 변수 선택을 하였다. 그 모델을 validation에 적용시켰는데 여전히 공선성이 존재했다.

-X11과 X12 사이의 공선성을 해결하기 위해 두 변수를 각각 제거하여 MSE를 비교한 후 validation MSE가 더 작은 값이 나온 것을 선택하기로 했다. 그 결과, 원래 모델에서 X12변수(평균세대수)를 제거하고 다시 Stepwise방법을 이용하여 모델을 적합시켰다. 다중 공선성도 해결되었고, 결정계수도 조금 증가하였다.

이 모형을 통한 Test MSE값은 1474.591로 validation MSE보다 큰 것을 확인할 수 있다.

한계점은 다중공선성을 제거하고 변수선택을 했음에도 불구하고 Test MSE가 validation MSE보다 크다는 것이다.

모델링에 대한 자세한 해석은 최종보고서를 통해 더욱더 보강할 예정입니다.