Reyeme Affiar

2023-2학기 통계자료분석 1조

2023-12-06

## 1. 데이터 정리

### 1-1. 결측값

주어진 데이터는 456개의 관측치와 17개의 변수로 구성되어 있다. 이 중 결측값 9개를 확인했다.

결측값 중 한개를 제외하곤 모두 UNIT 열 값이었다. UNIT는 건물내 특정 주거공간을 나타내는 호수(ex 101호, 102호)이다. Reyem Affiar가 구매하려는 건물의 UNIT이 알려져 있지 않았기 때문에 Reyem Affiar가 구매하려는 건물 가격을 추정하는데에 UNIT 변수는 사용하지 않기로 결정했다.

나머지 한개의 결측값은 First Date에서 나타났으며, 254행에 있다. FirstDate열은 차후 HousePriceIndex 변수의 각 행의 값를 정하는 기준으로 사용한다. 때문에 First Date 값이 누락된 254행을 제거하기로 결정했다.

### 1-2. 변수 선택

#### 1) Area

건물 가격을 예측함에 있어서 지역변수는 매우 중요하다고 판단한다. 같은 도시라고 하여도 그 지역, 입지에 따라 건물 가격의 평균이 달라진다. 예를 들어, 강남구의 건물과 금천구의 건물은 같은 서울일지라도 크게 다르다. Cambridge 역시 마찬가지 행태를 보일 것이라 예상한다.

#### 2) FirstDate

FirstDate는 SalePrice와 높은 상관관계를 갖는 Price1이 정해진 시기이다. 건물 가격을 결정하는데 시점은 중요한 의미를 갖는다. 시점에 따라 부동산 시장의 상황은 크게 변화할 수 있다. 때문에, 건물 가격을 예측하는데 시점변수는 매우 중요하다. 그러나, FirstDate를 독립변수로 하고, SalePrice를 종속변수로 하여 분산분석 결과를 사용하는 것보다는, FRED에서 제공하는 1989년에서 1994년 Cambridge 지역 주택가격 지수를 사용하는 것이 건물가격을 추정하는데 더 나은 정확도를 제공할 것이라고 판단했다.

#### 3) HousePriceIndex

FirstDate는 1993-07-17 와 같은 형식으로 된 데이터이다. 여기에서 연월만 추출하여 다시 분기로 대체한 데이터를 사용했다. 예를 들어, 1993-07-17 데이터는 1993-3Q 와 같이 ‘연 - 분기’ 형식으로 대체했다.

이와 같이 분기별로 나눈 이유는 FRED에서 제공하는 “All-Transactions House Price Index for Cambridge-Newton-Framingham, MA”(https://fred.stlouisfed.org/series/ATNHPIUS15764Q) 자료가 분기별로 값을 제공하기 때문이다.

위 자료는 1995년 1분기의 가격 지수를 100 기준으로 한다. 자료에 따르면, 1989년 1분기(데이터에서 1920년을 제외한 가장 오래된 관측치)에서 1994년 2분기(Reyme Affiar가 건물 매수를 고려하는 시점) 사이 중, 1989년 4분기의 가격지수가 108.92로 가장 높았으며, 1992년 2분기의 가격지수가 96.83으로 가장 낮다.

FRED에서 제공하는 House Price Index 분기별 자료와 각 관측치의 ’연-분기’를 대조하여 HousePriceIndex 변수를 만들었다. 예를 들어, 1993-3Q 변수값을 갖는 관측치는 HousePriceIndex 변수값으로 98.70을 갖게 된다.

#### 4) Days

첫 호가를 올린 시점부터 판매 시점까지의 기간이 길어질수록 SalePrice가 떨어질 것이라 판단했다. 부동산 시장은 보통 매수자가 있기 마련이라 대부분의 경우 적절한 기간이 지나면 거래가 체결된다. 그러나, 건물의 호가가 시장이 생각하는 가격에 비해 높거나, 부동산 시장 상황 악화로 매수자가 줄어드는 경우 거래완료까지 기간이 늘어날 수 있다. 이 경우 매도자는 호가를 낮춰 거래를 성사시키려 하기 때문에, Days가 커질수록, SalePrice는 떨어진다.

#### 5) Interior

내부면적으로 공용면적을 제외한 수치이다. 내부면적은 SalePrice를 결정하는데 중요한 변수라 판단한다. 내부면적이 커질 수록 SalePrice가 커질 것이다. 그러나, Interior 변수는 공용면적을 포함하고 있지 않기 때문에, 공용면적을 나타내는 다른 변수가 필요하다.

#### 6) Condo

매달 관리비를 나타내는 변수다. 관리비는 utility(가스비, 수도세, 전기세 등)와 공용공간 관리비를 포함한다. 예를 들어, 대부분의 오피스텔 관리비에 포함된 복도, 엘리베이터, 분리수거장 등 공용공간에 대한 청소용역비가 Condo에 해당할 것이다. Condo는 공용면적이 늘어날 수록 비례해서 늘어날 것이라 추측할 수 있다.

Condo가 클수록 헬스장, 수영장등을 포함한 공용면적이 넓은 고급 주거공간일 수 있기 때문에, SalePrice가 커질 것이라 판단했다.

#### 7) Rooms

방의 수는 Interior(내부 면적)과 높은 상관관계를 갖고 있기 때문에 두 변수를 사용할 경우, 다중공선성이 일어날 수 있다. 하지만 내부면적이 같은 경우에도 방의 수가 너무 많다면, 선호도가 떨어져 SalePrice가 낮아지는 경우가 생길 수 있다. 즉, Rooms는 내부면적과 관련이 있지만, 건물의 퀄리티(설계 수준)과 관련이 있기 때문에 SalePrice를 결정하는 요인으로 판단했다.

#### 8) Bed

침실의 수는 방의 수와 높은 상관관계를 갖기 때문에 역시 다중공선성을 조심해야 하는 변수이다. 그러나, 방의 수로만 설명할 수 없는 지점을 침실의 수 변수와 함께 설명할 수 있다. 예를 들어, 방의 수가 동일한 건물에서 침실의 수가 너무 많다면, 작은 침실이 많은 구조도를 갖고 있을 것이다. 작은 침실은 보통 선호도가 떨어져 SalePrice가 낮아지는 경우가 생긴다. 즉, Bed 역시 건물의 퀄리티(설계 수준)와도 관계가 되기 때문에 SalePrice를 결정하는 요인으로 판단했다.

#### 9) Bath

화장실의 수는 방의 수와 높은 상관관계를 갖기 때문에 역시 다중공선성을 조심해야 하는 변수이다. 그러나, 침실의 수로만 설명할 수 없는 지점을 화장실의 수 변수와 함께 설명할 수 있다. 예를 들어, 침실의 수가 3개 이상이어서 4인 이상 가구가 거주할 것으로 예상하는 건물에, 화장실의 수가 1개인 건물과 2개인 건물에 대한 선호도는 다를 것이다.(보통 화장실이 2개인 건물을 선호할 것이다.) 이로 인해 SalePrice에 차이가 생길 수 있다. 즉, Bath 역시 건물의 퀄리티(설계 수준)와도 관계가 되기 때문에 SalePrice를 결정하는 요인으로 판단했다.

#### 10) Tax

Cambridge가 위치한 메사추세츠 주에서는 부동산에 부과하는 세금(재산세)를 공시가(Assesed value)를 기준으로 부과한다. 공시가는 감정평가사(Professional appraiser)에 의해 정해지는데, 감정평가사는 아래와 같은 기준으로 공시가를 정한다.

1.건물 퀄리티2. 방 상태3. 동네4. 평수5. 침실수 6.화장실 수

Tax는 SalePrice를 결정하는 많은 요인들이 반영되어 있는 변수이기 때문에, 다른 변수들과의 다중공선성을 조심해야 한다. 그러나 건물 퀄리티, 방 상태와 같이 다른 변수들이 충분히 설명하지 못하는 부분을 반영하고 있기 때문에 사용하기로 판단했다.

#### 11) RC

RC(Rent Control Status)는 임대료 규제가 있는지에 대한 여부를 보여주는 변수이다. 값이 “1”이면 건물주가 임대를 할 때, 규제에 의해 임대료를 일정 수준 이상 올릴 수 없다. 때문에 임대료 규제가 존재한다면, SalePrice가 낮아질 것이라 추측한다.

## outlier 제거

#### 1. 1920년 데이터 제거

주어진 데이터는 주로 1988년 ~ 1994년에 이뤄진 매매에 대한 자료이다. 그런데, 1920년에 이뤄진 거래에 대한 관측치가 1개 존재한다. 이 자료로 인해 회귀분석 결과가 크게 왜곡될 수 있기 때문에, 제거하였다.

#### 2. 너무 크거나 작은 내부면적(Interior)을 갖는 관측치 제거

Interior 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 8개의 관측치를 제거했다.

#### 3. 너무 크거나 작은 Tax를 갖는 관측치 제거

Tax 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 20개의 관측치를 제거했다.

#### 4. 너무 크거나 작은 Condo를 갖는 관측치 제거

Tax 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 19개의 관측치를 제거했다.

#### 5. 너무 많거나 적은 Rooms를 갖는 관측치 제거

Rooms 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 20개의 관측치를 제거했다.

#### 6. 너무 많거나 적은 Bath/Rooms (= BathPerR변수)를 갖는 관측치 제거

방의 수 대비 너무 많거나 적은 화장실을 갖고 있는 건물은 극단값에 해당한다고 보았다. BathPerR 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 1개의 관측치를 제거했다. 해당 관측치는 Bath/Rooms = 0.6으로 방 5개 중 화장실이 3개, 침실이 2개인 구조를 갖고있는 건물이다.

#### 7. Days가 너무 긴 혹은 짧은 관측치 제거

Days 변수값에서 IQR 1.5배를 상한선, 하한선으로 하여 밖에 위치한 데이터는 극단값으로 보고 제거했다. 그 결과, 21개의 관측치를 제거했다.

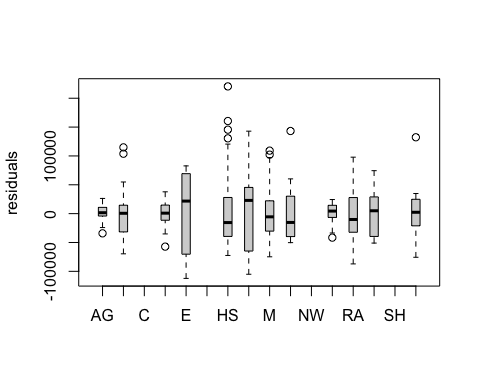
#### 8. Area 수준 중 관측치 10개 미만인 수준 제거

Area 수준 중 관측치가 10개 미만인 수준이 있으면 SalesPrice를 종속변수로 하고 Area를 독립변수로 하는 분산분석의 왜곡된 결과를 초래할 수 있다. 때문에, 10개 미만의 관측치를 갖는 수준은 제거한다.

C,FP,NW,SH 수준의 관측치가 10개 미만으로 나타났기 때문에, 해당 관측치는 모두 제거하고, C,FP,NW,SH 지역은 분산분석 및 회귀분석에서 고려하지 않는다. 총 14개의 관측치가 제거되었다.

#### 11. SalePrice ~ Area 분산분석 모형 극단값 제거

SalePrice ~ Area 분산분석 모형(M1)에서 각 Area 수준별 잔차가 IQR 1.5배 상한선, 하한선 밖에 위치한 극단값을 갖는 관측치를 제거한다. 그 결과 13개 관측치가 제거 되었다.



### outlier 제거 결과

총 338개 관측치가 남았다. 이는 주어진 456개 관측치 중 118개의 관측치를 극단값으로 판단하여 제거한 결과이며, 약 25%의 관측치를 극단값으로 판단한 것이다.

## SalePrice ~ Area 분산분석

# 분산분석(SalePrice ~ Area)

Analysis of Variance Table  
  
Response: SalePrice  
 Df Sum Sq Mean Sq F value Pr(>F)   
Area 11 5.7672e+10 5242927264 2.7916 0.001736 \*\*  
Residuals 326 6.1225e+11 1878077747   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p값이 0.001736 \*\*으로 Area에 따른 SalePrice의 평균이 같다는 귀무가설을 기각했다.

### 등분산성 확인

그래프상에서 Area별로 SalePrice의 분포가 크게 다른 것으로 보인다. 때문에, 등분산성을 만족하지 않는 것으로 예상할 수 있다.

잔차의 분포는 상당히 정규성을 띄는 분포로 보인다. 그러므로 bartlett test를 이용하여 등분산성을 검정한다.

Bartlett test of homogeneity of variances  
  
data: SalePrice by Area  
Bartlett's K-squared = 80.965, df = 11, p-value = 9.6e-13

p값이(9.6e-13)으로 ’등분산성을 만족한다’는 귀무가설을 기각하고 있기 때문에 등분산성을 갖고 있다고 볼 수 없다. 때문에 비모수검정을 진행한다.

### 비모수검정

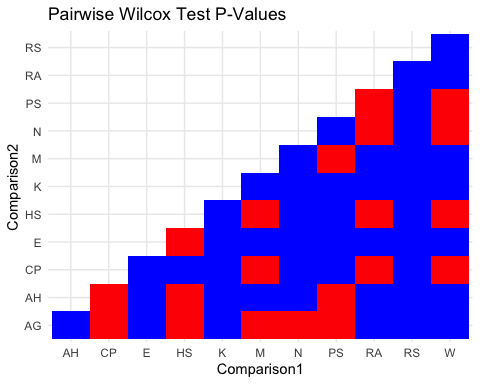
Kruskal-Wallis rank sum test  
  
data: SalePrice by Area  
Kruskal-Wallis chi-squared = 25.193, df = 11, p-value = 0.008542

p-value(p-value= 0.008542)가 매우 작으므로 그룹간 차이가 없다는 귀무가설을 기각하고 있다. 때문에 그룹간 차이가 존재하는 것으로 판단할 수 있다.

#### 사후검정, 다중비교

작은 차이라도 찾아내기 위해 fisher lsd 검정과 같이 개별 유의수준을 통제하는 방식으로 wilcoxon 검정을 진행했다.

wilcoxon 검정 결과를 기반으로 두 변수간 p값이 0.1 이하인 경우 붉은 타일이 나타나는 타일플롯을 그렸다.

A graph of different numbers

Description automatically generated with medium confidence

### Area 분산분석 결과

위 분산분석결과를 통해 SalePrice 평균 차이가 나지 않는 지역을 묶고, SalePrice 평균 차이가 나는 지역을 나눈 새로운 Zone 변수를 만들었다.

1Zone = Area의 변수값 “E”,“K”,“RA”,“AG”  
2Zone = Area의 변수값 “W”,“AH”, “M”,“RS”  
3Zone = Area의 변수값 “HS”,“N”,“CP”,“PS”

## 다중회귀분석

### 다중 회귀 모형 - 단계별 선택법

모든 변수(Interior,Zone,HousePriceIndex,Days,Rooms,Bed,Bath,Condo,Tax,RC)를 독립변수로 하고 모든 상호작용을 고려한 Full model(SalePrice ~ Interior\*Zone\*HousePrice\*Index\*Days\*Rooms\*Bed\*Bath\*Condo\*Tax\*RC)은 다중공선성 문제가 발생하는 것으로 판단하여, 상호작용을 고려하지 않는 model에서 시작하여 단계별 선택법을 통해 다중회귀 모형을 결정하기로 했다.

# 상호작용을 고려하지 않은 다중회귀 모형  
LM2 = lm(SalePrice ~ Interior + Zone + HousePriceIndex + Days + Rooms + Bed + Bath + Condo + Tax + RC, dataG)  
F2 = anova(LM2)  
F2[11,3] # MSE : 715,208,196

[1] 715208196

단계별 선택법의 결과는 아래와 같다.

# 다중회귀 모형의 model selection 단계별 선택법 결과  
Final Model:  
SalePrice ~ Interior + Zone + HousePriceIndex + Days + Rooms +   
 Bed + Bath + Condo + Tax + RC + Zone:Tax + Rooms:Bath + Interior:Zone +   
 Zone:Bath + Condo:RC + Days:Bath + Bed:RC + Interior:Rooms +   
 Interior:Tax + Rooms:Bed + Interior:Zone:Tax

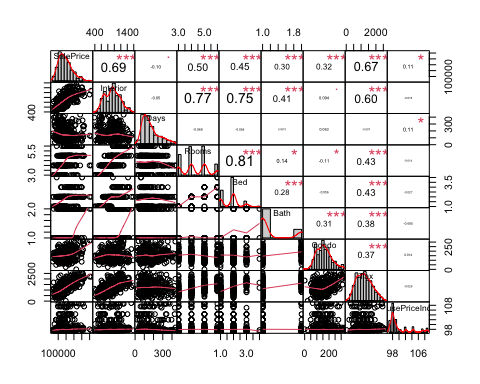
단계별 선택법에 의한 다중회귀 모형의 MSE는 아래와 같다.

[1] 571449668

상호작용항을 고려하지 않은 모형의 MSE(715,208,196)에 비해 크게 감소한 것을 확인할 수 있다.

## 다중공선성 및 영향점 확인

### 다중회귀모형에 사용한 변수간 상관관계 확인(수치형 변수만)



Interior – Rooms, Interior – Bed, Interior – Bath, Interior – Tax, Rooms - Bed변수간에 상관관계가 특히 높다.

다중회귀 모형의 회귀계수와 분산분석을 확인해보면, Bed와 Bath 변수의 회귀계수 부호, 통계적 유의성에 문제가 보인다. 그러나, Bed 혹은 Bath와 Bed 혹은 Bath가 속한 상호작용을 제거할 경우, MSE가 지속적으로 증가하여 예측력이 떨어지는 것을 확인하였다. 때문에 다중공선성 문제보단 영향점을 제거하는데 집중했다.

### 잔차 분석

gvlma(LM3)  
  
Call:  
 gvlma(x = LM3)   
 Value p-value Decision  
Global Stat 367.8211 0.00000 Assumptions NOT satisfied!  
Skewness 7.5930 0.00586 Assumptions NOT satisfied!  
Kurtosis 356.2798 0.00000 Assumptions NOT satisfied!  
Link Function 0.2561 0.61283 Assumptions acceptable.  
Heteroscedasticity 3.6923 0.05466 Assumptions acceptable.

gvlma 결과 영향점 있을 수 있다고 판단했다.

### 영향점 확인

외적 스튜던트화된 잔차가 1.5이상인 경우를 영향점으로 봤다. 총 28개의 관측치가 해당되었으며 모두 제거했다.

### 영향점 제거 후 다중회귀모형 성능 확인

영향점 제거 후 동일한 모형의 MSE를 확인할 결과 259,626,691로, 영향점 제거 전 MSE의 45% 수준으로 감소했다.

# 영향점 제거 후 MSE 확인   
F3 = anova(LM3)  
F3[22,3] # MSE = 259,626,691 (이전보다 반으로 줄어듦 571,449,668)

[1] 259626691

영향점 제거 후 각 회귀계수의 통계적 유의성을 검정하는 F값이 증가하여 p값이 크게 감소하였다. 이는 앞서 다중회귀모형의 회귀계수 부호와 통계적 유의성 문제의 원인이 다중공선성이 아닌 영향점이라는 것을 알 수 있다.

# 영향점 제거 후 다중회귀모형 분산분석 확인  
Analysis of Variance Table  
  
Response: SalePrice  
 Df Sum Sq Mean Sq F value Pr(>F)   
Interior 1 2.8640e+11 2.8640e+11 1103.1051 < 2.2e-16 \*\*\*  
Zone 2 4.1369e+10 2.0684e+10 79.6694 < 2.2e-16 \*\*\*  
HousePriceIndex 1 1.1726e+10 1.1726e+10 45.1667 9.946e-11 \*\*\*  
Days 1 4.3986e+09 4.3986e+09 16.9421 5.060e-05 \*\*\*  
Rooms 1 1.8155e+09 1.8155e+09 6.9926 0.0086415 \*\*   
Bed 1 3.9586e+09 3.9586e+09 15.2474 0.0001180 \*\*\*  
Bath 1 1.3430e+07 1.3430e+07 0.0517 0.8202441   
Condo 1 1.9153e+10 1.9153e+10 73.7718 5.981e-16 \*\*\*  
Tax 1 3.6231e+10 3.6231e+10 139.5489 < 2.2e-16 \*\*\*  
RC 1 9.7797e+08 9.7797e+08 3.7668 0.0532716 .   
Zone:Tax 2 1.7380e+10 8.6900e+09 33.4713 8.958e-14 \*\*\*  
Rooms:Bath 1 9.1333e+09 9.1333e+09 35.1787 8.754e-09 \*\*\*  
Interior:Zone 2 6.7598e+09 3.3799e+09 13.0183 3.903e-06 \*\*\*  
Zone:Bath 2 3.7627e+09 1.8814e+09 7.2465 0.0008528 \*\*\*  
Condo:RC 1 3.8838e+09 3.8838e+09 14.9593 0.0001364 \*\*\*  
Days:Bath 1 4.0836e+09 4.0836e+09 15.7289 9.267e-05 \*\*\*  
Bed:RC 1 4.3710e+09 4.3710e+09 16.8357 5.335e-05 \*\*\*  
Interior:Rooms 1 1.9870e+09 1.9870e+09 7.6534 0.0060395 \*\*   
Interior:Tax 1 4.9622e+09 4.9622e+09 19.1127 1.733e-05 \*\*\*  
Rooms:Bed 1 4.2460e+08 4.2460e+08 1.6354 0.2020025   
Interior:Zone:Tax 2 7.0722e+09 3.5361e+09 13.6200 2.252e-06 \*\*\*  
Residuals 283 7.3474e+10 2.5963e+08   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### 회귀계수 확인

회귀계수 확인시 더미변수를 모두 확인하려면 항이 너무 많이지기 때문에 Reyeme Affiar가 고려하는 건물의 Area(M)에 해당하는 2Zone=1, RC=1 인 경우만 고려하여 확인했다. 전체 다중회귀 모형은 다음과 같을 것이다.

##### 1) Interior의 회귀계수

는 2zone에 속한 모든 집에서 0보다 클 것이다. Rooms가 7이상일 경우 이 되는데, 데이터상에서는 Rooms가 7이상인 관측치가 존재하지 않는다. 때문에 Interior가 클수록 SalePrice가 증가한다는 예측과 부합하는 회귀계수이다. 또한 내부면적이 같은 경우에도 방의 수가 너무 많다면, 선호도가 떨어져 SalePrice가 낮아지는 경우가 생길 수 있을것이라는 예측과 관련하여, Rooms와의 상호작용에 의한 항도 예측과 부합한다고 판단했다.

##### 2) Rooms의 회귀계수

침실의 수는 모든집에서 1개 이상이므로, 총합을 계산해 보면 이다. 합당한 회귀계수 부호를 갖고 있다고 판단한다.

##### 3) Bed의 회귀계수

침실의 수가 너무 많으면 SalePrice가 감소할 것이라는 예측에 부합하는 회귀계수 부호를 갖고 있다고 판단했다.

##### 4) Bath의 회귀계수

화장실의 수가 너무 많으면 SalePrice가 감소할 것이라는 예측에 부합하는 회귀계수 부호를 갖고 있다고 판단했다.

##### 5) Condo의 회귀계수

Condo가 커지면, SalePrice도 클 것이라는 예상과 부합한다.

##### 6) Tax의 회귀계수

Interior가 보통 1000을 상회하기 때문에 총합을 계산하면, 이다. Tax가 커지면, SalePrice도 클 것이라는 예상과 부합한다.

##### 7) Days의 회귀계수

Days의 회귀 계수는 해석할 수 없다. 다중공선성 문제나 영향점 제거가 안되었을 가능성이 있다.

##### 8) HousePriceIndex의 회귀계수

HousePriceIndex가 클수록, SalePrice는 클 것이라는 예상과 부합하는 회귀계수 이다..

##### 9) (Zone이 2Zone일 경우)의 회귀계수

Reference Group인 1Zone에 2Zone은 가격이 낮은 지역이므로, 2Zone일 경우, SalePrice가 떨어질 것이라는 예상과 부합한다.

##### 10) (RC= 1 일 경우)의 회귀계수

임대료 규제가 있다면(RC=1) 없는 경우보다 SalePrice가 떨어질 것이다.

### 영향점 제거 후 다중회귀모형을 통한 Reyeme Affiar의 건물 가격 예측

다중회귀모형을 이용한 매매가격에 대한 예측 결과는 아래와 같다.(단, 매매가 5월 4일에 이뤄졌다고 가정)

# 매매가격에 대한 예측  
predict(LM3, newdata = data.frame(Interior = 1040, Zone = "2Zone",HousePriceIndex = 99.38, Days = 24, Rooms = 5, Bed = 2, Bath = 1, Condo = 175 , Tax = 1121 , RC = "1"), interval = "prediction")

fit lwr upr  
1 168540.1 136353.8 200726.4

이 건물의 매매가격은 $168,540 로 추정할 수 있으며, $200,726 이상으로 매매가격이 나타나는 것은 이례적으로 비싼 가격이라 할 수 있다. $136,353이하로 매매가 이루어지기는 어려울 것이다.