2021-1 서강대학교 통계자료분석 학기말 프로젝트

-회귀분석을 이용한 부동산 매매가격 예측-

20191245 수학과 노유정

2021.6.23

서론-1, 상황 설명

* Reyem Affiar씨는 마음에 드는 콘도를 발견하여 구매하려고 한다. 호가는 $169000이나, 지난 5년간의 매사추세츠 콘도 판매 현황을 이용하여 분석 후 최종 구매가격을 낮추려고 한다.

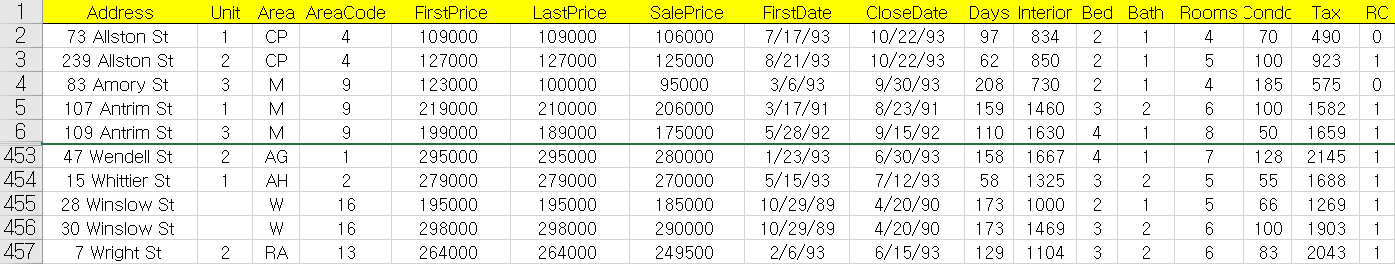
\* Affiar가 마음에 들어하는 콘도의 특징:

|  |  |
| --- | --- |
| 방 개수(ROOMS) | 5개(침실 2, 욕실 1) |
| 지역(AREA,AREACODE) | Mid-Cambridge(M,9) |
| 크기(INTERIOR) | 1040 평방피트 |
| 월세(CONDO) | $175 |
| 연 재산세(TAX) | $1121 |
| 가격(FIRST PRICE) | $169000 |
| Put on market | 1994년 4월 11일 |
| 임대 관리 상태(RC) | Rent control(1) |

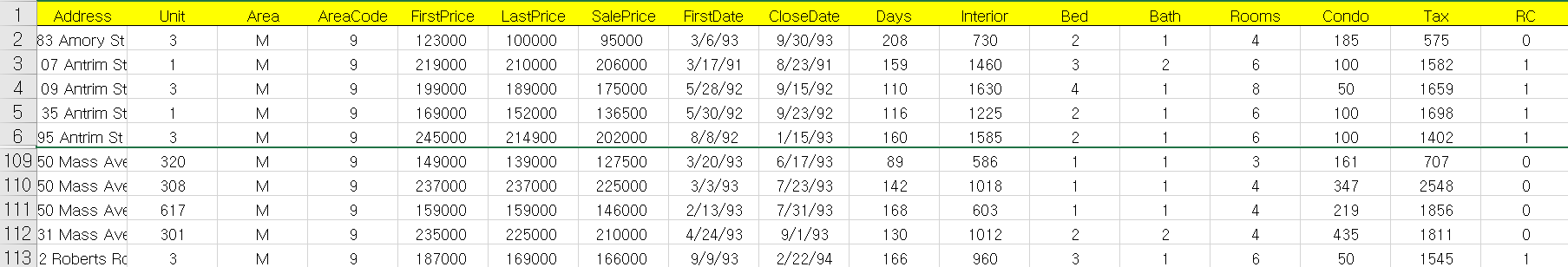
서론-2, .Raw data 및 데이터 가공 & 분석 용어 정리

\* 데이터 가공

raw data는 아래와 같이 456개의 데이터로 이루어져 있음



Mid-Cambridge의 데이터를 가지고 분석을 진행하기 위해 필터링, 112개의 데이터가 남음



\* 분석에 사용된 용어 개념 정리

상관계수: 두 변수 사의의 상관관계의 정도를 나타내는 지표, 두 변수의 공분산 / 각 변수의 표준편차의 곱 으로 계산함.

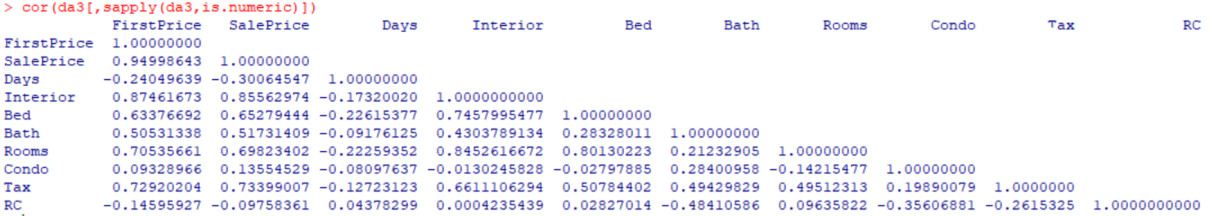
다중공선성: 독립변수들 사이의 상호의존도, 강한 상호의존도가 존재하는 경우 회귀계수의 분산을 크게 만들어 분석에 부정적인 영향을 끼치게 됨.

VIF: 분산확대인자, 변수의 다중공선성을 평가하는 지표이며 통상적으로 10을 넘는 경우 그 변수를 분석에서 제외한다.

본론, 데이터 분석

우선 종속변수는 최종 가격인 SalesPrice, 독립변수는 FirstPrice, Days, Interior, Bed, Bath, Rooms, Condo, Tax, RC(더미->메트릭)의 9가지가 될 수 있다.

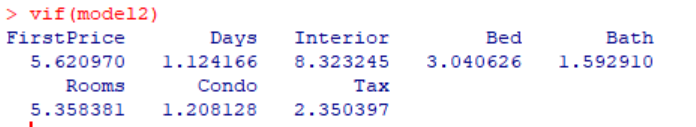
각 독립변수간의 상관계수를 살펴보면 아래의 결과가 나온다.



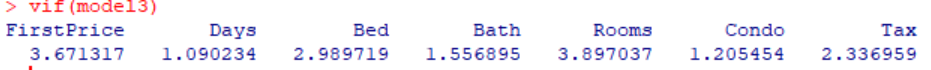
종속변수인 SalesPrice와 상관관계가 낮은 Condo와 RC를 분석에서 제외할 수 있다.

또한 Interior과 FirstPrice, Interior과 Rooms의 상관관계가 서로 높으므로 다중공선성을 없애기 위해 Interior을 분석에서 제외하였다.

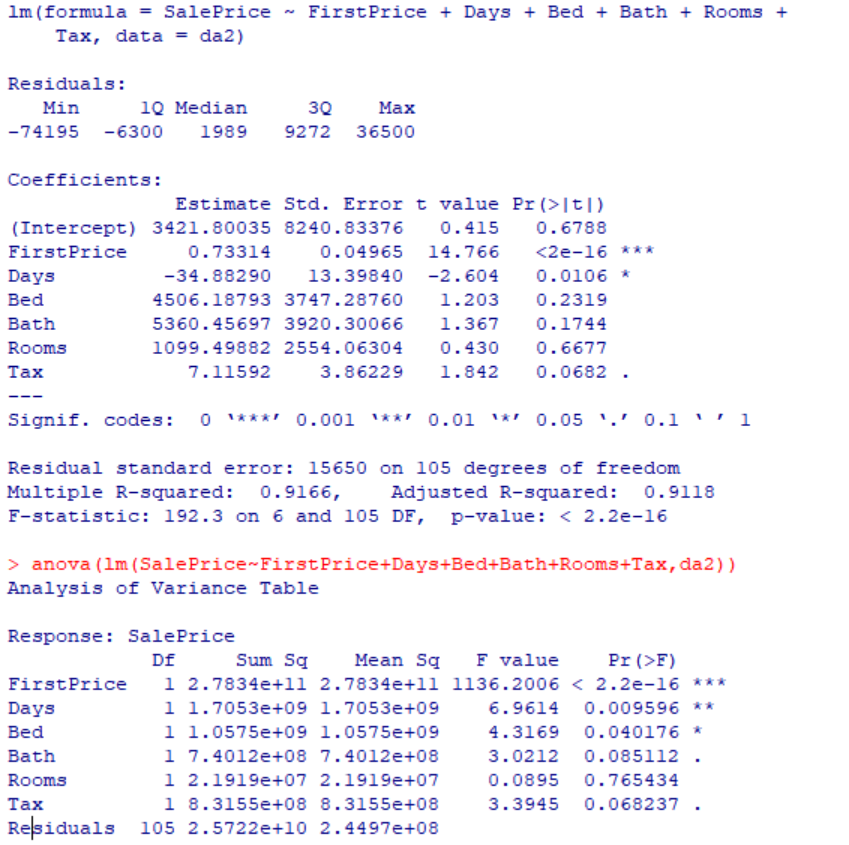
실제로 vif값을 계산한 결과 interior의 vif가 비교적 높은 편에 속하였고,



interior을 제외한 결과 vif가 모두 5 이하로 떨어졌다.

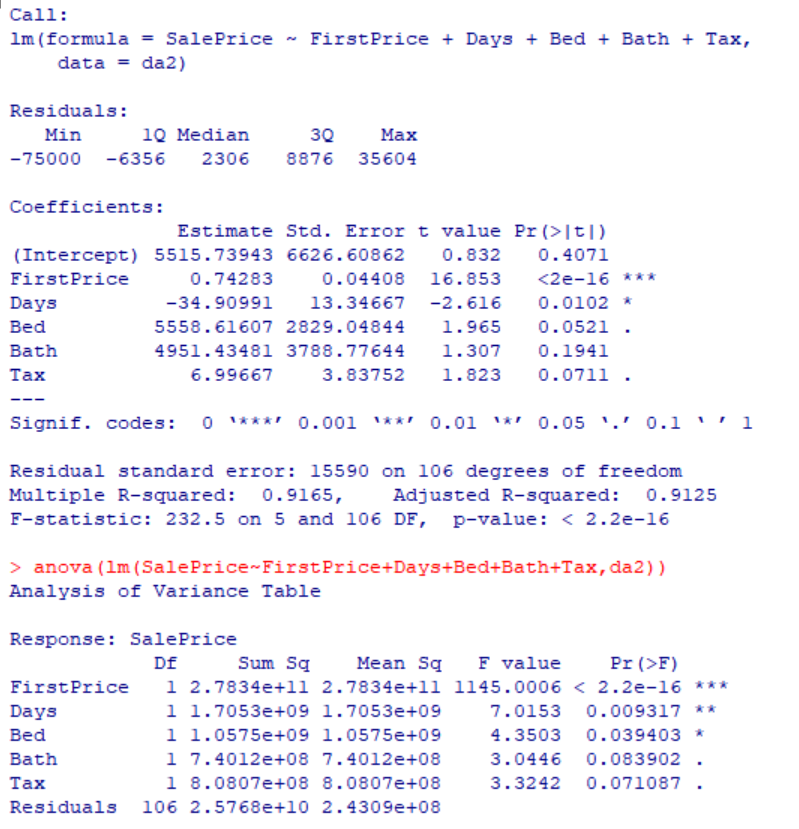


따라서 세 변수를 제외한 FirstPrice, Days, Bed, Bath, Rooms, Tax의 여섯 가지 변수를 사용하여 분석을 진행하였다.



하지만 Rooms에서의 p값이 너무 크게 나와 Rooms가 유효하다는 귀무가설을 기각하게 된다.  
(앞서 언급된 상관계수에서도 Rooms와 Bed의 상관관계가 높은 것으로 판명되었음)

따라서 Rooms를 제외한 다섯가지 변수를 이용하여 계산한 결과,



위와 같이 비교적 전체적으로 안정적인 p값을 가진 모델을 찾아낼 수 있었고,

따라서 종속변수 SalesPrice / 독립변수 FirstPrice, Days, Bed, Bath, Tax로 이루어진 본 모델을 최종 모델로 채택하였음.

결론, 분석결과

최종 모델의 분석 결과에 따라

SalePrice = 5515.74 + 0.75\*FirstPrice - 34.91\*Days + 5558.62\*Bed + 4951.43\*Bath + 7.00\*Tax

라는 식을 도출해낼 수 있다.

따라서 Reyem씨가 원하는 집의 경우는,

최종가 = 5515.74 + 0.75\*169000 – 34.91\*23 + 5558.62\*2 + 4951.43\*1 + 7.00\*1121 = 149820

로 169000 – 149820 = 19180달러를 내려 계약할 수 있다.

(Days는 94년 4월 11일에서 94년 5월 4일까지의 날짜 수, 23일)

<reference>

이군희 2014, 연구방법론의 이해, 북넷, 419-420