서울시립대학교
UNIVERSITY OF SEOUL

학기말 보고서

다변량 통계를 통한
아파트 경매가격 분석

과목명	다변량 통계학
교수명	김규성 교수님
학 과	통계학과
학 번	
이 름	이유민
제출일	2020. 12. 14

차례

1. 서론

1.1 연구 목적

한국의 부동산 시장은 IMF 위기 이후 선진국과 같은 면모로 성장했다. 하지만 최근 급변하는 부동산 정책에 따라, 최근에는 단순히 적당한 부동산을 고르거나 아파트 분양을 받는 방법에서 조금 벗어난 '부동산 경매'에 눈길이 더욱 쏠리고 있다.

부동산 경매 시장에서 부동산 투자자들과 물건의 소유자인 개인 및 기관 너나할 것 없이 가장 관심을 가지는 것은, 경매 물건의 '최종 낙찰률'이라 할 수 있다. 이와 관련하여 국내 여러 기관에서는 부동산 경매와 관련된 각종 지표들을 제공하고 있다. 그러나 현재 제공되고 있는 대부분의 부동산 경매 관련 지표 및 정보들은 법원에서 제공하고 있는 경매 물건에 대한 기본적인 사항을 정리해 둔 것에 불과하며, 특히 가장 관심사인 낙찰률은 관련된 정보로는 지역별, 시기별 혹은 사용 용도별 평균만 제공하고 있다.

또한 부동산 경매와 관련된 우리나라의 연구 결과들은 보통 실제로 부동산 경매에 참여한 당사자들간의 법적 관계와 관련된 연구들이 주를 이루고 있다. 실제로 부동산 경매 자료를 이용한 실증적인 연구는 활발하지 않다. 실증적 차원에서 진행된 연구들을 살펴 본 결과, 부동산 경매에 있어서 낙찰률에 영향을 미칠 수 있는 여러 요인들을 조사하여 중요 요인을 찾고자 하는 연구, 상관분석을 이용하여 낙찰률과 여러 외적 요인들 간의 관계를 분석한 연구 정도가 있다. 하지만 거의 10년 전에 진행된 연구이기 때문에 최근의 실정을 제대로 반영하고 있는지는 미지수이다.

따라서 많은 사람들은 부동산 경매는 어려울 것이라고 생각하여 시작할 엄두조차 내지 못하는 경우가 많다. 권리 분석과 관련한 법을 잘 모르는 상태에서 어설프게 나섰다간 오히려 손해만 막심하게 입을 수 있다는 걱정 때문이다.

이번 연구에서는 부동산 경매에 참여하는 사람들과 기관들이 가장 관심있게 생각하는 지표(변수)들에 대해 다양한 방법으로 살펴보려고 한다. 또한 여러 변수들에 대해 평균벡터를 비교하고, 주성분 분석과 인자분석, 정준상관분석을 통해 데이터 변수들 사이에서의 유의미한 관계를 찾아내려 한다.

1.2 문헌 연구

1.2.1 부동산 경매 Big Data를 활용한 Chaos 분석

해당 연구에서는 빅데이터 분석 및 활용이 확대되는 현재, 국내 부동산 영역에서의 연구는 다른 영역에 비해서 미진하다고 한다. 하지만 부동산 경매 분야는 데이터 축척이 계속되고 있고, 그 형태도 정형화되어 충분한 연구가 주목되는 분야이다. 하지만 국내에서 부동산 경매 낙찰가율 데이터를 활용한 Chaos 분석 연구는 찾아볼 수 없었다. 따라서 해당 연구에서는 Hurst 지수, correlation dimension, maximum Lyapunov 지수와 같은 3가지 Chaos 분석기법을 활용하여 낙찰가율의 비선형 결정론적 동역학계적 특성을 확인하고, Chaos 분석을 통하여 얻은 결과와 실무 데이터를 비교하여 그 의미들을 해석했다. 높은 Hurst 지수에 따르는 추세와, maximum Lyapunov 지수의 측정을 통한 지속성, 그리고 correlation dimension 분석의 결과에 따라 time lag가 개시결정일에서 낙찰일, 배당요구종기일에서 낙찰일까지와 일치하는 점으로부터, Chaos 분석이 낙찰가율의 움직임 예측에 유용함을 확인하였다. 해당 논문을 통해 앞으로 진행할 분석에서 관심을 가지고 생성해야 할 변수들에 대한 인사이트를 얻을 수 있었다.

1.2.2 낙찰률 예측 모형에 관한 연구

해당 연구는 월별 평균 낙찰률을 예측하기 위하여 단순한 지역별, 기간별 평균값을 보완하고 의사결정나무 분석을 이용하여 예측오차를 보정하는 방법을 제안하였고, 선형회귀모형을 이용하여 개별 경매 물건별 낙찰률을 예측하는 모형을 제안했다. 구축된 모형은 전국 아파트 경매 물건에 적용하여 예측 모형을 구현 하였으며 그 응용방법으로 예측결과에 대한 등급화를 함께 수행하였는데, 이번 분석을 수행한 후 평가지표를 보완하는 개념으로 사용하고자 하였다.

1.2.3 모델의 불확실성을 반영한 아파트가격지수 예측 모형 연구

해당 연구는 베이지안 방법론을 이용해 모델의 복잡성 문제와 예측력을 개선할 수 있는 연구 모델을 도출했다. 베이지안 이론을 기반으로 하는 Bayesian Model Selection(BMS)과 Bayesian Model Averaging (BMA)은 모델의 불확실성을 반영한 분석이 가능할 뿐만 아니라, 기존의 방법론보다 많은 정보를 활용하기 때문에 예측력 개선에 있어 강점이 있다. 베이지안 분석 방법론을 이용해 추정한 아파트 실거

래가지수와 기존의 분석에서 주로 사용되는 AR모형을 통해 추정한 아파트 실거래가지수를 실측치와 상대적으로 비교하여 보다 나은 예측 모형을 선정했다. 또한 서로 다른 지역을 분석 대상으로 선정함으로써 지역적 특색을 반영한 분석 결과를 도출했다. 연구에서 분석 대상의 다변화를 위한 여러 방법을 제시하고 있으므로, 수업 시간에 다루었던 내용과 함께 분석 방법을 결정하는 데에 도움을 줄 것이다.

1.3 데이터 설명

1.3.1 데이터명

: 아파트 경매가격 데이터 (DACON 데이터)

1.3.2 출처

: Dacon 3회 아파트 경매가격 예측 모델링 대회 데이터

(데이터 Link: https://dacon.io/competitions/official/17801/overview/)

1.3.3 description

데이터는 데이콘 3회 아파트 경매가격 예측에 사용하는 데이터이며, 원데이터는 41개의 열(변수)로 구성되어 있다. 한국의 서울과 부산 지역의 최근 2년간 아파트 경매물의 등기부. dlaclk, 감정과, 유찰 횟수, 낙찰가의 정보를 담고 있다. 해당 대회에서의 목적은 아파트 낙찰가를 예측하는 것이였다.

1.3.4 이상치 확인 및 데이터 가공

raw data의 변수가 16개나 되었지만, 큰 의미를 가지는 변수가 많이 없었다. 따라서 여러 전처리 과정을 거쳐 새로운 피처(변수)를 생성하고 분석을 진행했다. 데이터 EDA와 전처리는 모두 파이썬으로 진행했으며, 이후 csv 파일로 내보내어 sas에서의 분석으로 이어나갔다.

이상치로 의심되는 점 또한 함께 확인해주었다. Hammer_price(최종경매가격)가 지나치게 큰 점 두 곳이 발견되어 이상치인지에 대해 생각해보았다. 분석 결과 건물의 매우 넓어 가격이 비정상적으로 커 보이는, 다시 말해 정상치였다. 이상치가 아님을 발견했으나 sas에서 분석하는 과정에서 시각화가 용이하지 않기 때문에 이번 분석에서는 해당 두 개의 이상치를 제외하였다.

따라서, 생성한 변수(피처)는 다음과 같다. 새로운 변수를 생성하는 데에 사용한 raw data의 변수와 생성하는 과정도 함께 기술하였다.

1. **hpl(Hammer Price Log)** : 낙찰가에 로그를 취해 준 값이다. raw data에서 타겟 변수인 Hammer_price의 분포가 정규분포를 따르지 않기 때문에 log를 취한 hpl 변수를 생성함으로써 정규분포를 따르게 하였다.
2. **HPR(Hammer Price Ratio)** : EDA 과정에서, 상관관계 히트맵을 통해 Total_building_auction_area와 Hammer_price의 매우 높은 상관성(1.0)을 확인할 수 있었다. 낙찰가율은 실제 경매에서도 많이 사용되고 있으므로 이를 도입하고, 이를 지역구별 평균 낙찰가율로 환산하여 낙찰가율이 높은 지역과 낮은 지역에 대한 Feature 또한 생성하였다. 원래 데이터의 정보를 이용하여, Hammer_price/Total_appraisal_price 로 생성해주었다.
3. **Claim_price** : 경매 신청인의 청구 금액을 의미한다.
4. **Total_appraisal_price** : 총감정가를 의미한다.
5. **Total_appraisal_price_log** : Hammer Price와 마찬가지고 총감정가의 분포도 정규분포를 따르지 않았다. 따라서 정규분포를 따르도록 로그변환 해주었다.
6. **Total_land_gross_area** : 총토지 전체 면적으로, 제곱미터 단위이다.
7. **Total_land_real_area** : 총토지 실면적으로, 단위는 제곱미터이다.
8. **Total_land_auction_area** : 총토지경매면적으로, 제곱미터 단위이다.
9. **Total_building_area** : 총건물 면적으로, 단위는 제곱미터이다.
10. **Total_building_auction_area_log** : 총토지 경매면적의 분포가 정규분포를 따르지 않았기 때문에 로그변환을 통해 분포를 고르게 한다. 추후 평당 가격 생성을 위한 보조변수로 사용하였다.
11. **PPP (Price per pyung)** : 제곱미터 단위가 아닌 '평'의 의미를 담을 수 있도록 3을 나눈 면적 변수를 생성했다. Total_appraisal_price/(Total_building_auction_area/3) 와 같은 방식을 사용하였다.
12. **Current_floor_type** : 저층:0, 중층:1, 고층:2 로 층을 분류하는 피처를 생성한다. 층분류 별 row 수는 0:689, 1:553, 2:691 로 고르게 분포한다.

13. **Final_auction_weekday** : 경매 요일 정보를 담은 피처를 생성한다. 날짜 변수를 매칭시켜 요일을 부여한 후 int로 인코딩한다. (0:월요일, 6:일요일)

14. **Auction_period** : 경매기간 정보를 담은 변수이다. Datetime 형식으로 저장되어 있기 때문에 days를 문자열로 바꾸고, 숫자만 추출해 '일수'로 단위를 맞춘다.

	Claim_price	Total_appraisal_price	Total_appraisal_price_log	HPR
count	1.931000e+03	1.931000e+03	1931.000000	1931.000000
mean	3.655668e+08	4.806061e+08	19.664711	0.970028
std	1.324106e+09	4.706126e+08	0.819217	0.128217
min	0.000000e+00	4.285000e+06	15.270631	0.181727
25%	7.727625e+07	2.090000e+08	19.157845	0.910000
50%	1.720438e+08	3.600000e+08	19.701615	0.983325
75%	3.554527e+08	5.715000e+08	20.163775	1.030385
max	2.286481e+10	5.810000e+09	22.482846	1.868845

Total_land_gross_area	Total_land_real_area	Total_land_auction_area	Total_building_area	Total_building_auction_area_log
1.931000e+03	1931.000000	1931.000000	1931.000000	1931.000000
3.458806e+04	40.974076	39.949990	94.157877	4.399550
9.446369e+04	26.402675	26.625119	48.255855	0.522595
0.000000e+00	0.000000	0.000000	9.390000	0.405465
2.997000e+03	25.870000	24.545000	61.415000	4.093761
1.424140e+04	37.510000	36.710000	84.900000	4.440885
4.140310e+04	51.770000	51.295000	114.935000	4.743540
3.511936e+06	603.200000	603.200000	1203.760000	7.093205

PPP	Current_floor_type	Final_auction_weekday	Auction_period	hpl
1.931000e+03	1931.000000	1931.000000	1931.000000	1931.000000
1.475505e+07	1.001036	1.456758	67.394614	19.624016
8.538547e+06	0.844978	1.265144	155.920624	0.834662
1.650165e+06	0.000000	0.000000	0.000000	15.656536
9.064364e+06	0.000000	0.000000	0.000000	19.100121
1.315068e+07	1.000000	1.000000	35.000000	19.683394
1.775841e+07	2.000000	2.000000	49.000000	20.137015
7.227488e+07	2.000000	4.000000	1939.000000	22.039671

총 14 개의 변수로 분석을 진행한다.

1.4 분석 방법

다변량통계학 과목에서 다루었던 방법들을 이용하여 여러 방식으로 분석하려 한다. 크게 네 가지의 분석으로 나뉜다.

1.4.1 주성분 분석

비슷한 단위(제곱미터, 원 등)를 가지는 집단이 뚜렷하게 존재하고, 또한 이 변수들 사이에서의 공분산-분산의 구조를 주성분 분석으로 알아본다. 원데이터에 대한 정보 손실을 최소화하기 위해 적절한 수의 주성분을 선택한 후 주성분을 통해 변수들의 관계를 파악한다.

1.4.2 인자 분석

주성분 분석을 기반으로 하여 인자분석을 수행한다. 이를 통해 변수들 간에 내재하고 있는 공통의 구조를 파악하고 데이터의 특성을 2-3가지의 인자로 축약하여 설명하고자 한다. 또한 varimax 회전으로 해석이 용이해진 인자들을 해석한다.

1.4.3 평균벡터 추론

Current_floor_type 변수의 저층, 고층 여부에 따라 데이터를 분리한다. 두 경우의 아파트 경매가 변수, 낙찰률 등의 변수들의 평균벡터가 동일한지에 대해 검정을 수행한다.

1.4.4 정준상관분석

면적과 가격으로 정준상관분석을 진행한다. 즉, 반응변수를 Claim_price, Total_appraisal_price_log, Total_appraisal_price로 하고, 설명변수를 Total_land_gross_area, Total_land_real_area, Total_land_auction_area, Total_building_area 로 했을 때의 정준상관분석 결과를 살펴본다.

1.5 결과 활용 및 기대 효과

부동산 투자에 막연한 두려움을 가지고 있는 사람들에게 경매 데이터의 특징을 잘 담은 분석을 제공함으로써 부동산 경매로의 진입장벽을 낮출 수 있을 것이다. 또한 현실적으로 사람들이 관심있어 하는 분야를 다루는 것이니만큼, 보다 현실적으로 쓰임새가 많은 분석으로의 시작이 될 것이다.

2. 본론

2.1 분석 방법 소개

주성분 분석, 인자분석, 정준상관분석, 모집단 평균벡터에 대한 추론을 이용한 분석을 진행하려 한다. 여러 분석을 진행하므로, 분석 방법에 대한 소개는 각 분석을 진행하기 전에 해당 방법에 대한 설명을 함께 작성하였다.

2.1.1 변수 기초통계량 확인, 상관계수 확인

본격적인 데이터 분석에 앞서 데이터의 특징을 살펴보았다.

proc means 프로시저를 통해 변수들 사이의 단위가 동일하지 않음을 알 수 있다. 따라서 공분산행렬 대신 상관계수행렬을 사용한다. 또한 변수들 간의 상관계수를 살펴보았을 때, HPL은 total_appraisal_price_log등과 같은 가격 변수와의 상관계수가 높아 보인다.

변수	N	평균	표준편차	최솟값	최댓값
VAR1	1931	965.00	557.58	0.00	1930.00
Claim_price	1931	365566786.33	1324106235.4	0.00	22864813241
Total_appraisal_price	1931	480606104.91	470612584.00	4285000.00	5810000000.0
Total_appraisal_price_log	1931	19.66	0.82	15.27	22.48
HPR	1931	0.97	0.13	0.18	1.87
Total_land_gross_area	1931	34588.06	94463.69	0.00	3511936.00
Total_land_real_area	1931	40.97	26.40	0.00	603.20
Total_land_auction_area	1931	39.95	26.63	0.00	603.20
Total_building_area	1931	94.16	48.26	9.39	1203.76
Total_building_auction_area_log	1931	4.40	0.52	0.41	7.09
PPP	1931	14755048.54	8538547.09	1650165.02	72274881.52
Current_floor_type	1931	1.00	0.84	0.00	2.00
Final_auction_weekday	1931	1.46	1.27	0.00	4.00
Auction_period	1931	67.39	155.92	0.00	1939.00
hpl	1931	19.62	0.83	15.66	22.04

피어슨 상관 계수, N = 1931
H0: Rho=0 가정하에서 Prob > |r|

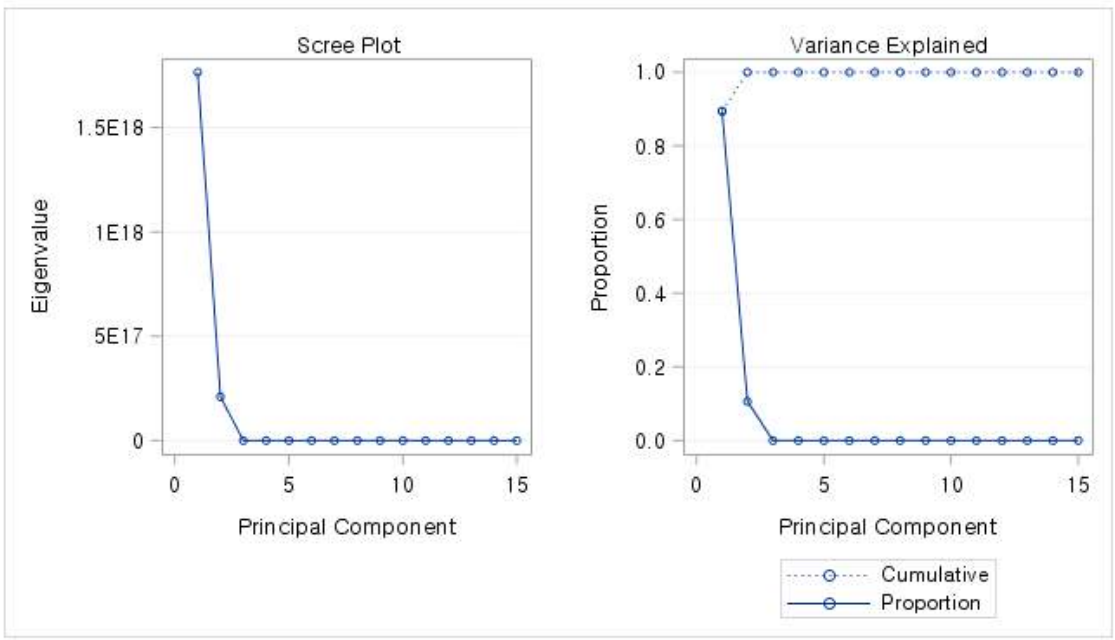
	VAR1	Claim_price	Total_appraisal_price	Total_appraisal_price_log	HPR	Total_land_gross_area	Total_land_real_area	Total_land_auction_area	Total_building_area	Total_building_auction_area_log	PPP	Current_floor_type	Final_auction_weekday	Auction_period	hpi
VAR1	1.00000	0.11020 <.0001	0.31667 <.0001	0.42791 <.0001	0.05536 0.0150	0.03117 0.1709	0.11442 <.0001	0.10236 <.0001	0.10430 <.0001	0.07658 0.0008	0.47228 <.0001	-0.05826 0.0104	-0.44538 <.0001	0.02733 0.2301	0.42652 <.0001
Claim_price	0.11020 <.0001	1.00000	0.22318 <.0001	0.17888 <.0001	0.10236 <.0001	0.04494 0.0281	0.11712 <.0001	0.12120 <.0001	0.16010 <.0001	0.14049 <.0001	0.14571 <.0001	0.05504 0.1238	-0.00297 0.8963	0.03015 0.1854	0.18424 <.0001
Total_appraisal_price	0.31667 <.0001	0.22318 <.0001	1.00000	0.81463 <.0001	-0.05814 0.0136	0.07919 0.0005	0.50102 <.0001	0.51369 <.0001	0.70676 <.0001	0.58223 <.0001	0.74433 <.0001	0.10882 <.0001	-0.04482 0.0489	0.04038 0.0761	0.78868 <.0001
Total_appraisal_price_log	0.42791 <.0001	0.17688 <.0001	0.81463 <.0001	1.00000	0.00466 0.8579	0.12316 <.0001	0.42837 <.0001	0.46130 <.0001	0.64042 <.0001	0.76720 <.0001	0.72009 <.0001	0.16400 <.0001	-0.15767 <.0001	0.05395 0.0190	0.98321 <.0001
HPR	0.05536 0.0150	0.10236 <.0001	-0.05814 0.0136	0.00466 0.8579	1.00000	0.08674 0.0033	-0.02388 0.2940	-0.02168 0.3410	-0.16817 <.0001	-0.11991 <.0001	0.10469 <.0001	0.01786 0.4328	-0.12852 <.0001	-0.32519 <.0001	0.18151 <.0001
Total_land_gross_area	0.03117 0.1709	0.04494 0.0281	0.07919 0.0005	0.12316 <.0001	0.08674 0.0033	1.00000	0.14201 <.0001	0.13718 <.0001	0.05395 0.0177	0.06791 0.0028	0.12019 <.0001	0.08802 <.0001	-0.02482 0.2756	-0.02823 0.2150	0.13209 <.0001
Total_land_real_area	0.11442 <.0001	0.11712 <.0001	0.50102 <.0001	0.42837 <.0001	-0.02388 0.2940	0.14201 <.0001	1.00000	0.97708 <.0001	0.68815 <.0001	0.48354 <.0001	0.20566 <.0001	-0.08525 0.0002	-0.04541 0.0460	-0.00252 0.9084	0.41502 <.0001
Total_land_auction_area	0.10236 <.0001	0.12120 <.0001	0.51369 <.0001	0.46130 <.0001	-0.02168 0.3410	0.13718 <.0001	0.97708 <.0001	1.00000	0.67728 <.0001	0.55666 <.0001	0.18310 <.0001	-0.08510 0.0002	-0.04207 0.0645	-0.00153 0.9465	0.44979 <.0001
Total_building_area	0.10430 <.0001	0.16010 <.0001	0.70676 <.0001	0.64042 <.0001	-0.16817 <.0001	0.05395 0.0177	0.68815 <.0001	0.67728 <.0001	1.00000	0.76511 <.0001	0.23194 <.0001	0.09021 <.0001	0.01425 0.5314	0.09419 0.0002	0.60083 <.0001
Total_building_auction_area_log	0.07658 0.0008	0.14049 <.0001	0.58223 <.0001	0.76720 <.0001	-0.11991 <.0001	0.06791 0.0028	0.48354 <.0001	0.55666 <.0001	0.76511 <.0001	1.00000	0.17921 <.0001	0.12080 <.0001	0.02200 0.3340	0.06155 <.0001	0.73530 <.0001
PPP	0.47228 <.0001	0.14571 <.0001	0.74433 <.0001	0.72009 <.0001	0.10469 <.0001	0.12019 <.0001	0.20566 <.0001	0.18310 <.0001	0.23194 <.0001	0.17921 <.0001	1.00000	0.09900 <.0001	-0.14500 <.0001	-0.00882 0.7049	0.72277 <.0001
Current_floor_type	-0.05826 0.0104	0.05504 0.1238	0.10882 <.0001	0.16400 <.0001	0.01786 0.4328	0.08802 <.0001	-0.08525 0.0002	-0.08510 0.0002	0.09021 <.0001	0.12080 <.0001	0.09900 <.0001	1.00000	0.04899 0.0313	0.02454 0.2810	0.16812 <.0001
Final_auction_weekday	-0.44538 <.0001	-0.00297 0.8963	-0.04482 0.0489	-0.15767 <.0001	-0.12852 <.0001	-0.02482 0.2756	-0.04541 0.0460	-0.04207 0.0645	0.01425 0.5314	0.02200 0.3340	-0.14500 <.0001	0.04899 0.0313	1.00000	0.06973 0.0022	-0.17795 <.0001
Auction_period	0.02733 0.2301	0.03015 0.1854	0.04038 0.0761	0.05395 0.0190	-0.32519 <.0001	-0.02823 0.2150	-0.00252 0.9084	-0.00153 0.9465	0.09419 0.0002	0.06155 <.0001	-0.00882 0.7049	0.02454 0.2810	0.06973 0.0022	1.00000	-0.01314 0.5639
hpi	0.42652 <.0001	0.18424 <.0001	0.78868 <.0001	0.98321 <.0001	0.18151 <.0001	0.13209 <.0001	0.41502 <.0001	0.44979 <.0001	0.60083 <.0001	0.73530 <.0001	0.72277 <.0001	0.16812 <.0001	-0.17795 <.0001	-0.01314 0.5639	1.00000

2.2 데이터 분석 및 결과 설명

2.2.1 주성분 분석

주성분 분석은 여러 개의 반응변수로 얻어진 다변량 데이터가 있을 때, 분산-공분산 구조를 변수들의 선형결합식, 즉 주성분으로 설명하고자 하는 접근방법이다. 주성분 분석에서는 차원을 축소하고, 변동이 큰 축을 탐색하며, 주성분을 통한 데이터의 해석 등 크게 세 가지의 목적을 가지고 진행한다.

변수들의 단위가 다른 경우, 상관계수 행렬을 이용하여 주성분 분석을 진행하는 것이 해석하기 편리하다. 따라서 상관행렬을 이용하여 분석을 진행하였다.



Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.76578E18	1.55679E18	0.8942	0.8942
2	2.08991E17	2.08958E17	0.1058	1.0000
3	3.24769E13	3.24681E13	0.0000	1.0000
4	8791437606	8791197899	0.0000	1.0000
5	239706.972	239706.972	0.0000	1.0000

proportion을 통해, 첫 번째 주성분에서 이미 전체 분산의 약 89%가 설명되었음을 알 수 있다. 보통 주성분이 설명하는 누적 분산의 비율이 70~80%정도가 되는 주성분 개수를 선택하지만, 이번 분석에서는 단위에 차이가 커서인지 하나의 주성분에 쏠려 있는 것을 볼 수 있다. 표본 주성분을 두 번째까지 구해 보면 다음과 같이 나타난다.

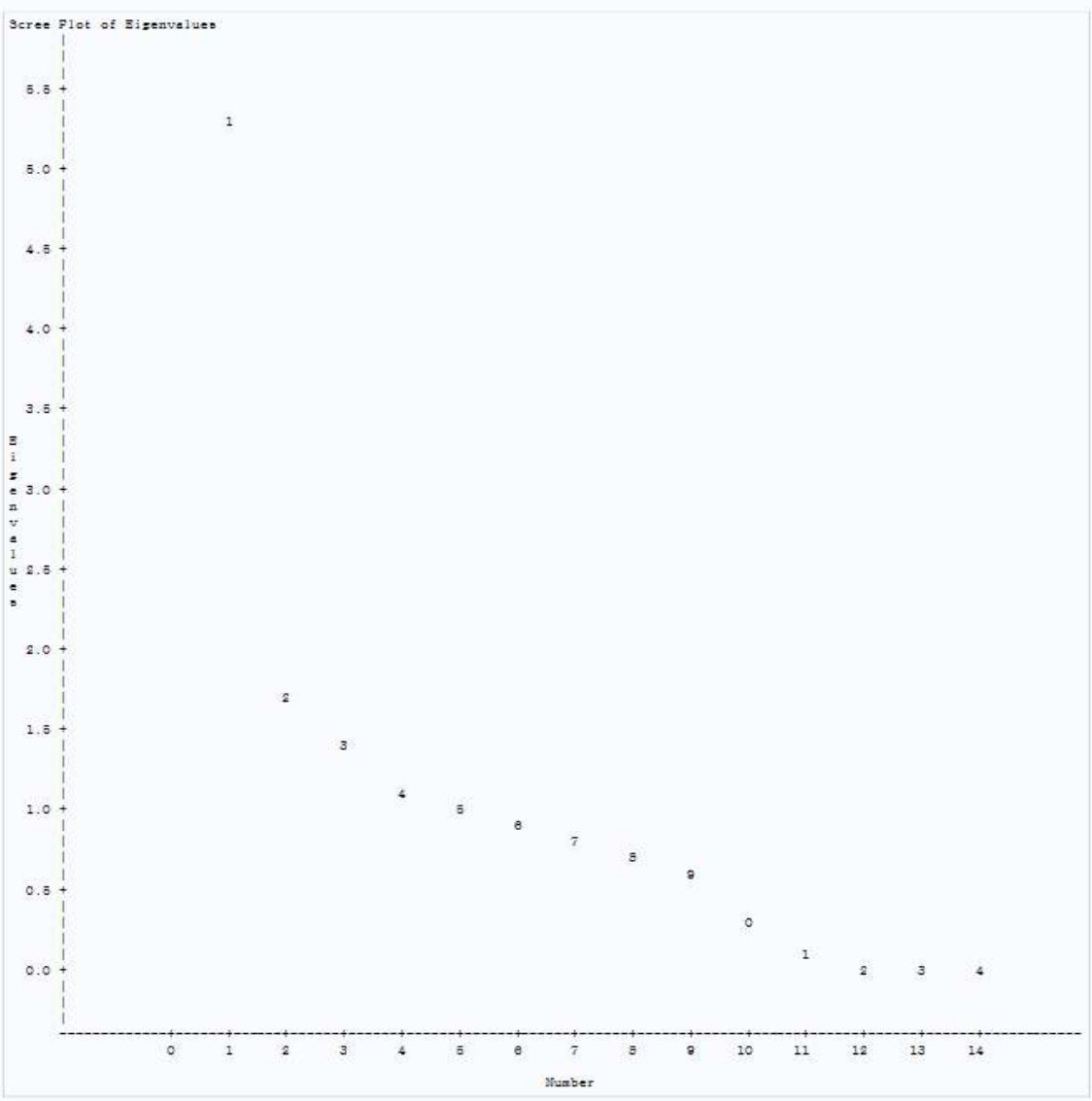
$$Y_1 = 0.9959 * ClaimPrice + 0.001 * PPP$$
$$Y_2 = -0.08970 * ClaimPrice + 0.995877 * TotalAppraisalPrice$$

하나의 주성분이 분산의 89%를 설명하고 있으므로, 유의미한 분석이 될 것으로 기대하기는 힘들어 보인다. 이는 데이터들 사이의 공분산의 영향으로 추정된다. 따라서 인자분석으로 넘어가 분석을 진행한다.

2.2.2 인자분석

인자분석은 공분산 구조를 몇 개의 관측 불가능한 인자로 설명하려는 것으로, 변수들 간에 내재하고 있는 공통의 구조를 파악하고 데이터의 특성을 몇 개의 인자로 축약하여 설명하고자 한다. 인자 분석은 여러 개의 변수들 사이의 상관성 구조를 나타내는 몇 개의 인자로 분석하는 것인데, 이를 통해 구성된 인자를 통해 인자 변수의 의미를 파악할 수 있다. 주성분 분석과의 차이는, 주성분 분석의 주성분들은 관측된 변수들의 선형결합식으로 정의되는 반면 인자분석의 변수들은 공통인자의 선형결합식으로 정의된다는 것이다. 앞선 주성분 분석에서는 큰 의미를 가지는 결

과를 찾지 못하였기 때문에 이번 절에서 주성분법을 이용한 인자분석을 진행한다. 주축인자법으로 인자분석을 시행하는 경우 데이터에 따라 반복적인 계산에서 공통성이 수렴하지 않는 상황이 발생할 수 있는데, 정제되지 않은 실제 데이터를 쓰고 있기 때문에 공통성이 s^ii값보다 커질 수 있다. 또한 최대우도법으로 인자분석을 시행하는 경우 다른 방법보다 Heywood상황이 발생할 수 있기 때문이다. 물론 이런 경우는 모두 1로 보는 등의 해결방안이 존재하기는 한다.
주성분법에 따른 인자분석의 스크리플롯은 다음과 같이 나타난다. 주성분법에서도 발견했듯, 역시 첫 번째 인자에서 그래프가 매우 급격하다. 세 번째 인자부터는 완만해지는 것을 보아 factor2까지로 분석을 진행한다.



또한 14개 인자의 고유값의 일부는 아래와 같이 나타난다.

Eigenvalues of the Correlation Matrix: Total = 14 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.29654081	3.56679871	0.3783	0.3783
2	1.72974210	0.33028745	0.1236	0.5019
3	1.39945465	0.32927108	0.1000	0.6018
4	1.07018356	0.06098101	0.0764	0.6783
5	1.00920256	0.07786282	0.0721	0.7504
6	0.93133973	0.08172354	0.0665	0.8169

인자분석 결과는 베리맥스 회전한 후의 인자분석 결과를 첨부하였다. 베리맥스 회전을 한 후 보다 인자들에 대한 해석이 용이해지는 장점이 있기 때문이다.
먼저, 직교변환행렬 T는 아래와 같다.

Orthogonal Transformation Matrix			
	1	2	3
1	0.74501	0.66574	0.04190
2	0.53033	-0.62924	0.56816
3	0.40462	-0.40106	-0.82185

직교변환 후의 인자들과 각 인자들에 대한 해석은 다음과 같다.
인자1(factor1)에서는 주로 price(가격)과 관련한 변수, 인자2에서는 area(땅)과 관련한 변수, 인자3에서는 period(기간)과 관련한 변수들이 묶여 있음을 볼 수 있다. 또한 앞에서 살펴봤듯 인자1이 설명하는 분산 비율이 현저히 높으므로, 해당 데이터는 가격과 관련한 인자들이 중요해 보인다. SAS 결과는 다음 장에 첨부하였다.

Rotated Factor Pattern			
	Factor1	Factor2	Factor3
Claim_price	0.23597	0.09319	0.09086
Total_appraisal_price	0.79257	0.44093	0.00489
Total_appraisal_price_log	0.89171	0.37607	0.04751
HPR	0.04115	-0.11229	0.77388
Total_land_gross_area	0.10400	0.11193	0.20644
Total_land_real_area	0.11484	0.93238	0.09936
Total_land_auction_area	0.13556	0.94130	0.08789
Total_building_area	0.44316	0.75988	-0.18662
Total_building_auction_area_log	0.54402	0.60839	-0.19558
PPP	0.82001	-0.00098	0.23420
Current_floor_type	0.39336	-0.24607	-0.16355
Final_auction_weekday	-0.13911	0.02134	-0.42828
Auction_period	0.12453	-0.02208	-0.68597
hpl	0.88202	0.35119	0.18604

2.2.3 모집단 평균벡터에 관한 추론

current floor에 따른 경매가격에 차이가 있는지를 알아 보려 한다.

current floor변수는 중 0:저층, 1:중층, 2:고층에 해당하는 세 집단으로 분류된다. 여기서 저층과 중층의 집단을 추출하여 해당 두 집단 사이의 모집단 벡터에 대한 추론을 시행한다. 모집단의 분산이 같을 때와 다를 때 수치상의 차이만 있을 뿐, 같은 분석 결과를 보였기 때문에 이번 보고서에서는 합동 공분산 행렬을 추정하여 진행한 분석 결과만 첨부하였다.

하단은 검정을 시행하기 위해 생성한 데이터셋으로, COV, MEAN을 사용하여 검정

을 수행하였다.

OBS	_TYPE_	_NAME_	HPR	Total_land_auction_area	PPP	hpl
1	COV	HPR	0.02	-0.19	188808.86	0.03
2	COV	Total_land_auction_area	-0.19	1100.99	68240261.63	15.54
3	COV	PPP	188808.86	68240261.63	6.9689508E13	5288994.46
4	COV	hpl	0.03	15.54	5288994.46	0.84
5	MEAN		0.97	42.21	13436924.03	19.43
6	STD		0.14	33.18	8348024.20	0.92
7	N		688.00	688.00	688.00	688.00
8	CORR	HPR	1.00	-0.04	0.16	0.23
9	CORR	Total_land_auction_area	-0.04	1.00	0.25	0.51
10	CORR	PPP	0.16	0.25	1.00	0.69
11	CORR	hpl	0.23	0.51	0.69	1.00
12	COV	HPR	0.01	0.28	-6093.46	0.01
13	COV	Total_land_auction_area	0.28	457.31	24315747.91	6.06
14	COV	PPP	-6093.46	24315747.91	7.6989785E13	5085471.91
15	COV	hpl	0.01	6.06	5085471.91	0.57
16	MEAN		0.97	36.85	15438668.28	19.76
17	STD		0.11	21.38	8774382.30	0.76
18	N		690.00	690.00	690.00	690.00
19	CORR	HPR	1.00	0.12	-0.01	0.11
20	CORR	Total_land_auction_area	0.12	1.00	0.13	0.37
21	CORR	PPP	-0.01	0.13	1.00	0.77
22	CORR	hpl	0.11	0.37	0.77	1.00

추정된 합동 공분산 행렬과 검정통계량은 다음과 같다.
귀무가설을 매우 크게 기각하므로 두 집단의 평균에는 차이가 있다고 할 수 있다.
단, 해석에서 유의해야 하는 점이 있다. 귀무가설이 기각할 수 있음이 모든 평균벡터가 차이가 있다는 의미가 아니라, 적어도 하나의 평균벡터가 동일하지 않음을 의미한다.

S			
0.0159755	0.044297	91216.056	0.0194398
0.044297	778.6797	46246083	10.797839
91216.056	46246083	7.3345E13	5187085.3
0.0194398	10.797839	5187085.3	0.7085047

T
124.12444

q
9.5135827

2.2.4 정준상관분석

해당 데이터에는 area와 price를 포함하는 변수들이 대거 포함되어 있다. 해당 분석에서는 가격 집단과 면적 집단간에서의 정준상관분석을 진행한다. 이 때 설명변수는 면적, 반응변수는 가격과 관련한 변수로 둔다. 정준상관분석은 단순상관계수와 다중상관계수의 개념의 일반화로 이해할 수 있다. 정준상관분석을 통해 종속변수들과 독립변수들의 두 집단 간의 정준상관계수를 계산할 수 있으며, 이는 두 변수 집단 간의 관계를 파악하는 데 유용하다.

반응변수를 Claim_price, Total_appraisal_price_log, Total_appraisal_price로 하고, 설명변수를 Total_land_gross_area, Total_land_real_area, Total_land_auction_area, Total_building_area 로 했을 때의 정준상관분석 결과는 다음과 같다.

Raw Canonical Coefficients for the VAR Variables			
	u1	u2	u3
Claim_price	3.58424E-12	-3.90516E-12	7.747786E-10
Total_appraisal_price_log	0.43740845	2.0586216687	0.0266897027
Total_appraisal_price	1.4559682E-9	-3.358267E-9	-5.36189E-10

v1, v2, v3은 뒷장에 첨부하였다.

Raw Canonical Coefficients for the WITH Variables			
	v1	v2	v3
Total_land_gross_area	8.8594967E-7	4.832429E-6	-8.508798E-6
Total_land_real_area	-0.032647912	-0.16098	-0.069879123
Total_land_auction_area	0.0331102484	0.1452992436	0.0931312556
Total_building_area	0.0201099617	0.0018107408	-0.008356748

첫 번째 정준변수는 다음과 같다.

$u_1 = 3.58E-12^* X_1 + 0.4374^* X_2 + 1.456E-9^* X_3$
$v_1 = 8.8595E-7^* Y_1 - 0.0326^* Y_2 + 0.0331^* Y_3 + 0.0201^* Y_4$

편의상 X1=Claim_price, x2=Total_appraisal_price_log, x3=Total_appraisal_price으로, Y1=Total_land_gross_area, Y2=Total_land_real_area, Y3=Total_land_auction_area, Y4=Total_building_area로 나타내었다. 두 번째, 세 번째 정준변수도 같은 방식으로 구할 수 있다.

또한 정준변수의 정준상관계수는 차례로 0.7318, 0.2125, 0.0163으로 나타난다.
이 때, 가장 큰 정준변수를 갖도록 X1, X2, X3의 선형결합으로 U를, Y1, Y2, Y3, Y4의 선형결합으로 V를 만들었을 때의 상관계수가 첫 번째 정준상관계수 0.7318이다.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenva
1	0.731374	0.730624	0.010587	0.534908	1.1
2	0.212484	0.209130	0.021735	0.045149	0.0
3	0.016349	0.000700	0.022756	0.000267	0.0

2.5 결과 설명

아파트 경매와 관련한 데이터를 여러 분석법을 통해 알아보았다.

먼저 주성분 분석에서는, 변수들의 단위가 다른 경우, 상관계수 행렬을 이용하여 주성분 분석을 진행하는 것이 해석하기 편리하다. 따라서 상관행렬을 이용하여 분석을 진행하였다. 하지만 첫 주성분이 분산의 89%를 설명하고 있으므로, 유의미한 분석으로 보이지 않았다.

따라서 인자분석을 진행하였다. varimax 회전 후 인자분석 결과를 해석하면, 인자1(factor1)에서는 주로 price(가격)과 관련한 변수, 인자2에서는 area(땅)과 관련한 변수, 인자3에서는 period(기간)과 관련한 변수들이 묶여 있음을 볼 수 있다. 주성분법으로 분석했기 때문에 인자분석에서도 인자1이 설명하는 분산 비율이 현저히 높게 나타났다.

모집단 평균벡터에 대한 추론에서는 저층과 고층에 따라 변수들에 차이가 있는지를 검정하였다. 귀무가설을 매우 크게 기각하므로 두 집단의 평균에는 차이가 있다고 할 수 있다. 비교한 평균으로는 HPR, Total_land_auction_area, PPP, hpI이 있다. 다만 귀무가설이 기각할 수 있음이 모든 평균벡터가 차이가 있다는 의미가 아니라, 적어도 하나의 평균벡터가 동일하지 않음을 의미하므로 추가적인 분석이 필요하다.

마지막으로 정준상관분석에서는 반응변수를 Claim_price, Total_appraisal_price_log, Total_appraisal_price로 하고,

설명변수를 Total_land_gross_area, Total_land_real_area, Total_land_auction_area, Total_building_area으로 두고 분석을 진행하였다. 다만 표준화하지 않았기 때문에 표준화를 하는 경우 결과가 달라질 것으로 보인다. 또한 정준상관계수도 함께 나타내었다.

2.6 분석의 타당성 설명

주성분 분석과 인자분석에서는 스크리 플롯과 분산 설명비율을 확인하며 적절한 주성분과 인자의 수를 결정하였다. 모집단 평균벡터에 대한 추론에서는 검정통계량을 통해 귀무가설이 매우 유의하지 않아 기각할 수 있었다. 또한 분석 과정에서 여러 옵션들에 대해 가장 타당한 옵션들을 설정하여 진행하였는데, 예로 인자분석에서 주성분법을 선택한 것이 그러하다. 따라서 이번 분석에서 사용했던 방법들은 어느 정도 유의하다고 할 수 있다.

3. 결론

3.1 분석 결과 요약

아파트 경매와 관련한 데이터를 여러 분석법을 통해 알아보았다.
주성분 분석의 경우 첫 주성분이 분산의 89%를 설명하고 있으므로, 유의미한 분석으로 보이지 않았다. 따라서 인자분석을 진행하여 arimax 회전 후 인자분석 결과를 해석 결과 인자1(factor1)에서는 주로 price(가격)과 관련한 변수, 인자2에서는 area(땅)과 관련한 변수, 인자3에서는 period(기간)과 관련한 변수들임을 볼 수 있었다. 이 때도 역시, 주성분법으로 분석했기 때문에 인자분석에서도 인자1이 설명하는 분산 비율이 현저히 높게 나타났다. 모집단 평균벡터에 대한 추론에서는 저층과 고층에 따라 변수들에 차이가 있는지를 검정하였다. 귀무가설을 매우 크게 기각하므로 두 집단의 HPR, Total_land_auction_area, PPP, hpl 평균에는 차이가 있다고 할 수 있다. 정준상관분석을 통해 면적과 가격 사이의 관계도 확인할 수 있었다.

3.2 분석의 장점 및 한계점 설명

우선 최근 사람들의 관심을 받는 경매에 대해 보다 현실적인 접근을 시도한 점에서 의미가 있다. 또한 주성분 분석, 인자 분석, 정준상관분석 등을 통해 변수들 사이의 공분산-분산 관계를 파악하고 그 상관관계들을 파악하여 분석에 녹여내었다는 점에서도 의미를 가진다. 다만 평균벡터 추론에서 언급했듯, 4개의 평균벡터 중 적어도 하나는 평균이 다르다는 의미를 가지기 때문에 어떤 변수에서의 평균벡터가 차이나는지를 알기 위해서는 추가적인 분석이 필요하다.

3.3 추가 연구사항 제안

이번 분석에서는 경매 데이터를 통해 알아보았다. 다만 이번 분석에서는 단순히 변수들 사이의 관계를 알아보는 데에 그쳤다. 또한 요인분석에서의 ㅇㅛ인점수를 추출하는 과정을 생략하여 추가 연구 사항으로 제안하려 한다. 일반적으로 요인분석은 단순히 요인을 추출하고 혹은 특정한 요인을 형성하고 있는 대표적인 변수들에 관한 정보만 얻고 분석을 끝내는 경우는 많지 않다. 요인분석을 하는 목적은 입력변수가 너무 많아 이들 모두를 독립적으로 이용해서 모형을 개발하든가 하는 것이 비효과적이기 때문에 연관성이 높은 변수들을 한데 묶어 요인으로 만든 다음 이들을 이용해서 추가적인 분석을 하는 데에 있다. 따라서 대부분의 경우 요인분석을

통해서 얻을 수 있는 요인점수를 활용하여 추가적인 분석을 하거나 모형을 개발한다. 추가적인 분석을 위한 요인점수를 추출하는 방법으로는 크게 2가지가 있다.

첫 번째 방법은 가장 일반적인 방법으로 요인분석 과정에서 직접 요인점수(factor score)를 산출하는 것이다. 이때 산출되는 요인점수는 요인점수 계수와 입력변수 값을 이용하여 계산된다. 둘째, 각각의 요인에 속한다고 판단되는 입력변수값들을 단순히 산술평균하여 사용하는 방법이 있다. 이 경우에는 요인에 속한 변수들이 서로 어느 정도 연관성이 강한지에 대한 신뢰성검사를 실시한 후에 산출된 평균값을 요인값으로 사용하는 것이 바람직하다.

하지만 각 방법은 장단점을 가진다. 먼저 첫 번째 방법은 요인분석 과정에서 직접 요인점수(factor score)를 산출하는 방법은 가장 일반적이며 정확한 요인점수를 구하는 방법임에는 분명하나 각각의 요인과 높은 상관관계가 있는 입력변수뿐 아니라 연관이 적은 입력변수들도 포함됨으로써 모든 변수들이 요인점수 값에 영향을 미치기 때문에 요인의 대표성에 대한 명확한 해석이 어려울 수도 있다는 단점이 있다.

두 번째 방법은 각각의 요인에 속한다고 판단되는 입력변수값들을 단순히 산술평균하여 사용하는 방법은 요인값이 요인분석을 통하여 추출한 요인값과 정확하게 일치하지는 않으나 단순히 입력변수들의 산술 평균값을 사용함으로써 요인값에 영향을 미치는 입력변수가 무엇인지를 명확하게 파악할 수 있어 실무적으로 사용하기 편리하다는 장점이 있다.

참고 문헌

- **부동산 경매 Big Data를 활용한 Chaos 분석**

https://s3.amazonaws.com/media.guidebook.com/upload/TKcoBa16TXxQwJud7zuzotmfiOkLf7comSNoaQGJ/3CiW0PxoNvkPJFvxeh7T0VFMVJLEy4EwPrYD.pdf

- **경매 용어**

https://www.courtauction.go.kr/RetrieveAucTermInq.laf

- **낙찰률 예측 모형에 관한 연구**

https://www.kss.or.kr/jounalDown.php?IDX=2968 , 클릭시 바로 다운로드됩니다.

- **모델의 불확실성을 반영한 아파트가격지수 예측 모형 연구**

https://www.kab.co.kr/kab/home/common/download_cnt.jsp?sMenuIdx=036015015000032028&sBoardIdx=045005125004&sFileIdx=045005125002