파키스탄 내륙지방 주택 가격 분석

2023학년도 2학기 응용계량경제학 과제

경북대학교 식품자원경제학과 2022111224 최문석

1. 서론

해당 분석의 목적은 파키스탄의 내륙지방 주택 가격을 분석하는 것입니다. 2014년과 2015년에 수집된 데이터를 사용하여 분석하였으며, 이를 통해 주택 구매자 및 판매자를 위한 데이터 기반 통찰을 얻을 수 있습니다.

2. 파기스탄 내륙지방 주택 가격 분석

2.1 분석환경

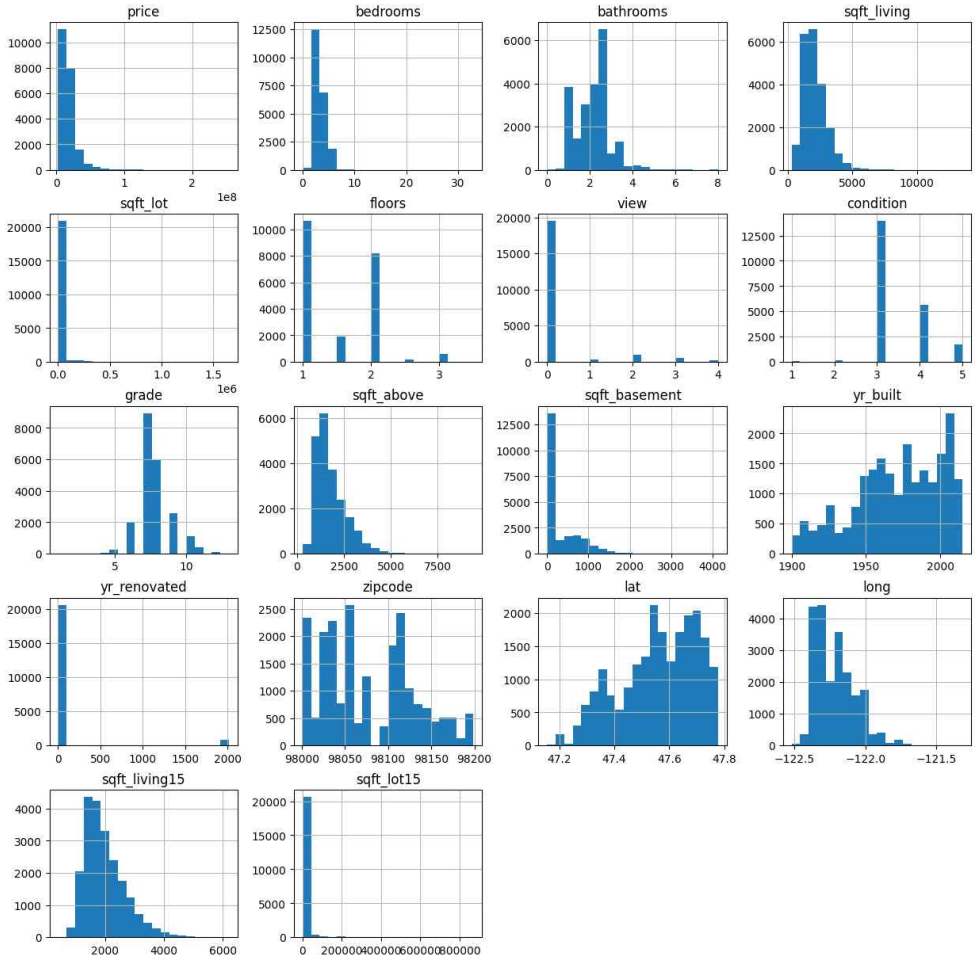
본 분석에서는 로컬 환경에서 Python 3.11.4을 활용하여 분석하였습니다. 데이터를 정규화를 위해 'sklearn'라이브러리의 StandardScaler 클래스, 통계 분석을 위해 'statsmodels'라이브러리 그리고 시각화 및 행렬 연산 처리를 위해 'matplotlib', 'seaborn', 'pandas'라이브러리를 사용하였습니다.

2.2 데이터셋

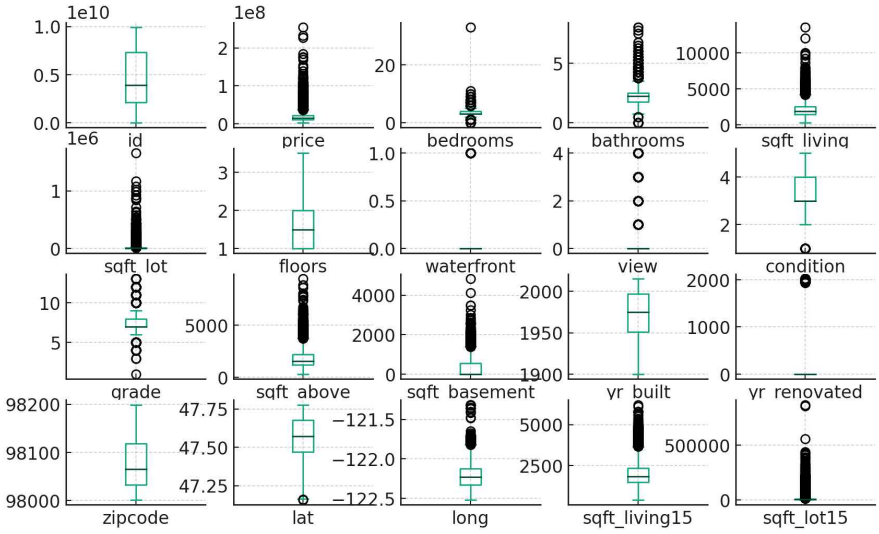
본 분석에 사용될 데이터 셋은 Kaggle의 'Comprehensive Dataset of House Prices in Pakistan'데이터를 사용했습니다. 해당 데이터셋은 파키스탄의 주거용 부동산 가격에 대한 포괄적인 내용을 담고 있으며, 다양한 출처에서 웹 스크래핑을 통해 수집되었습니다. 해당 데이터셋은 표1과 같이 총 21개의 칼럼을 포함하고 있습니다. 본 분석의 목적에 맞게 'waterfront'칼럼의 값이 0인, 즉 해안가 부동산이 아닌 데이터만을 전처리하여 사용하였습니다. 또한 'id'를 제거하였고, 해당 데이터셋이 2014년에서 2015년의 데이터로 만들어진 것을 감안해 시간에 따른 큰 변화가 없을것이라고 가정하여 'date'를 제거해 사용하였습니다. 그림1은 각 칼럼의 히스토그램을 시각화하였습니다. 그림2는 각 칼럼의 박스플랏입니다. 이를 통해서 각 칼럼의 분포를 확인할 수 있습니다. 그림3은 칼럼간 상관관계 히트맵인데, 이를 통해서 각 변수간 상관관계를 시각적으로 확인할 수 있습니다. 해당 데이터에서 변수간 상관관계가 어느정도 존재하는 것으로 보입니다. 가격에 대한 분석이므로 종속변수는 'price'입니다.

칼럼명	설명
id	각 주택 기록에 대한 고유 식별자.
date	주택 거래 또는 기록 날짜.
price	부동산의 상장 가격.
bedrooms	부동산 내의 침실 수.
bathrooms	부동산 내의 화장실 수.
sqft_living	총 거주 공간의 제곱피트.
sqft_lot	총 대지 면적의 제곱피트.
floors	부동산 내의 층 수.
waterfront	해안가 부동산 여부를 나타내는 이진 지시자 (1은 해당, 0은 비해당).
view	부동산에서의 전망 수준을 나타내는 지수.
condition	부동산의 전반적인 상태를 나타내는 지수.
grade	부동산의 전반적인 등급을 나타내는 지수.
sqft_above	지상 수준 이상의 내부 거주 공간 제곱피트.
sqft_basement	지하 수준 이하의 내부 거주 공간 제곱피트.
yr_built	부동산이 건설된 연도.
yr_renovated	부동산이 마지막으로 리노베이션된 연도.
zipcode	부동산의 우편 번호.
lat	부동산 위치의 위도.
long	부동산 위치의 경도.
sqft_living15	가장 가까운 15명의 이웃에 대한 거주 공간 제곱피트.
sqft_lot15	가장 가까운 15명의 이웃에 대한 대지 면적 제곱피트.

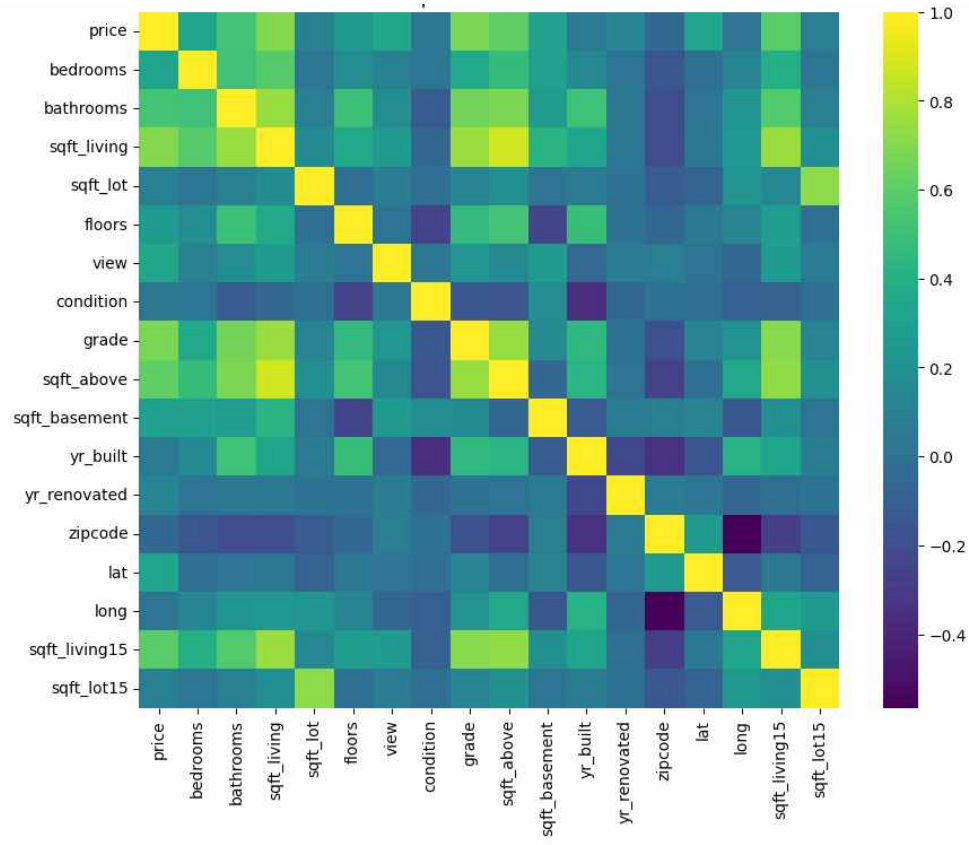
<표1. 칼럼명과 그 의미>



<그림1. 각 칼럼의 히스토그램>



<그림2. 각 칼럼의 박스플랏>



<그림3. 상관관계 히트맵>

2.3 분석 모형

2.3.1 데이터 전처리

선형회귀분석을 진행하기에 앞서, 각 변수의 단위가 모두 다르므로 정규화를 진행하였습니다.

2.3.2 다중공선성 문제

앞서 그림3을 통해 변수간 상관관계가 있다는 것을 확인하였습니다. 그중에서는 어느정도 강한 상관관계가 있는 변수 쌍도 확인되었는데, 이를 해결하기 위해서 독립변수에 대해 VIF 검정을 통해 다중공선성 문제를 해결합니다.

	Feature	VIF
16	sqft_living	inf
15	sqft_basement	inf
14	sqft_above	inf
6	grade	3.389677
1	bathrooms	3.341101
12	sqft_living15	2.980869
7	yr_built	2.434021
13	sqft_lot15	2.154606
2	sqft_lot	2.121415
3	floors	2.013460
11	long	1.832017
9	zipcode	1.666941
17	bedrooms	1.656483
5	condition	1.249916
4	view	1.207541
10	lat	1.181523
8	yr_renovated	1.140228
0	const	1.000000

<표2. VIF 검정결과>

	Feature	VIF
14	sqft_above	4.042020
6	grade	3.276277
12	sqft_living15	2.864945
1	bathrooms	2.758665
7	yr_built	2.403312
13	sqft_lot15	2.152554
2	sqft_lot	2.120454
3	floors	1.829843
11	long	1.820720
9	zipcode	1.663200
15	bedrooms	1.534972
5	condition	1.244868
4	view	1.178763
10	lat	1.178497
8	yr_renovated	1.140109
0	const	1.000000

<표3. 변수 제거후 VIF 검정결과>

표2는 데이터에 대한 VIF검정 결과를 나타냅니다. VIF가 10 이상인 변수를 차례로 제거하면서 VIF가 10 이하가 될 때까지 진행하였으며, 'sqft_basement'와 'sqft_living'변수를 제거하였습니다.

2.3.3 이분산성 문제

이분산성을 파악하기 위해, 종속변수인 'price'칼럼에 대해서 OLS(Ordinary Least Squares) 회귀 모델을 적합하고 White's test를 진행합니다. 테스트 결과, p_value 0.0으로 이분산성이 존재함을 확인할 수 있습니다.

테스트 결과	값
테스트 통계량	8997.31
p_value	0.0
f_value	114.07
F통계량의 p-value	0.0

<표4. white's test 결과((소수 셋째자리에서 반올림)>

2.3.4 모델 적합

변수간 상관관계가 존재하고 이분산성 또한 존재하므로 GLS(Generalized Least Squares)회귀를 진행합니다.

2.4 분석결과 및 해석

GLS Regression Results

Dep. Variable:	price	R-squared:	0.673
Model:	GLS	Adj. R-squared:	0.672
Method:	Least Squares	F-statistic:	2935.
Date:	Sun, 17 Dec 2023	Prob (F-statistic):	0.00
Time:	06:29:08	Log-Likelihood:	-18462.
No. Observations:	21450	AIC:	3.696e+04
Df Residuals:	21434	BIC:	3.708e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.78e-15	0.004	1.22e-12	1.000	-0.008	0.008
bathrooms	0.1859	0.006	28.636	0.000	0.173	0.199
sqft_lot	0.0214	0.006	3.755	0.000	0.010	0.033
floors	-0.0364	0.005	-6.890	0.000	-0.047	-0.026
view	0.1336	0.004	31.476	0.000	0.125	0.142
condition	0.0608	0.004	13.936	0.000	0.052	0.069
grade	0.3688	0.007	52.124	0.000	0.355	0.383
yr_built	-0.2443	0.006	-40.321	0.000	-0.256	-0.232
yr_renovated	0.0292	0.004	6.998	0.000	0.021	0.037
zipcode	-0.0781	0.005	-15.488	0.000	-0.088	-0.068
lat	0.2444	0.004	57.597	0.000	0.236	0.253
long	-0.1052	0.005	-19.955	0.000	-0.116	-0.095
sqft_living15	0.0959	0.007	14.494	0.000	0.083	0.109
sqft_lot15	-0.0195	0.006	-3.403	0.001	-0.031	-0.008
sqft_above	0.2859	0.008	36.385	0.000	0.271	0.301
bedrooms	-0.0475	0.005	-9.818	0.000	-0.057	-0.038

Omnibus:	19496.363	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2458532.711
Skew:	3.916	Prob(JB):	0.00
Kurtosis:	54.860	Cond. No.	5.06

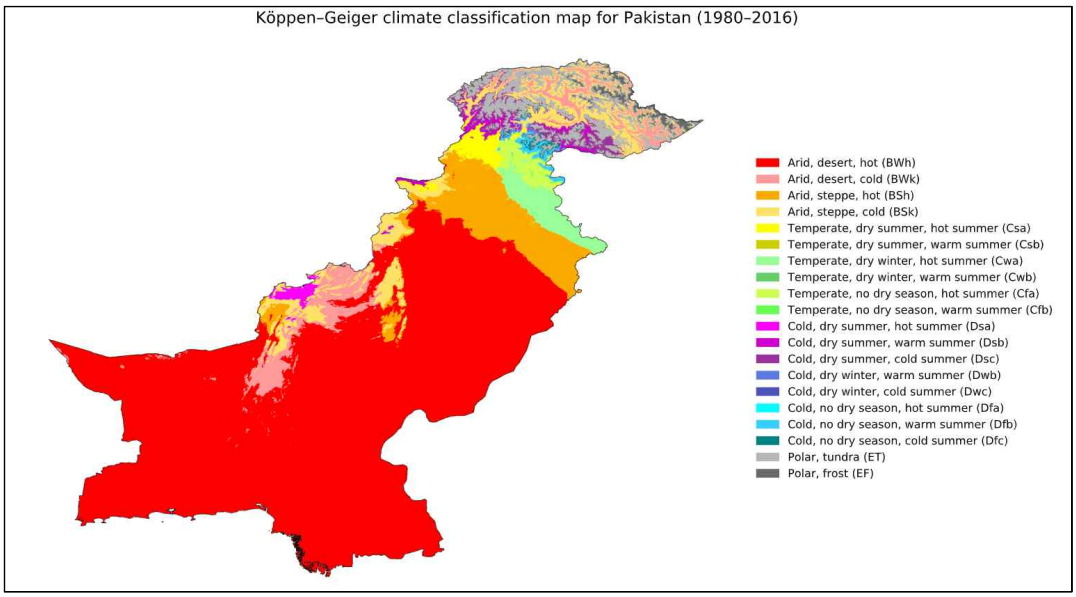
<표5. GLS 회귀 결과 요약>

표5에는 GLS회귀 결과가 잘 나타나있습니다. 이 GLS(Generalized Least Squares) 회귀분석 결과의 해석은 다음과 같습니다:

2.4.1 결과 해석
1) 종속 변수(Dep. Variable): 'price'. 분석의 대상이 되는 변수입니다.
2) R-squared: 0.673 : 이 모델이 데이터의 68.4%를 설명한다는 것을 의미합니다.
3) Adj. R-squared: 0.672 : 독립 변수의 수를 고려한 후 조정된 R-squared 값입니다.
4) F-statistic: 2935. F-statistic이 2935인 경우, 이는 회귀 모델이 전체적으로 통계적으로 매우 유의미하며, 모델이 종속 변수를 설명하는 데 유의미한 예측력을 가지고 있다는 것을 나타냅니다. F-statistic 값이 매우 크므로 모델의 적합도가 높다고 해석할 수 있습니다."
5) Prob (F-statistic) : 0.00. 모델이 통계적으로 유의하다는 것을 나타냅니다.

2.4.1 계수 해석
1) `const`: 모델의 상수항. 여기서는 4.78e-15로 거의 0에 가깝습니다.
2) 각 계수의 'std err'은 계수의 표준 오차를, 't'와 'P>|t|'는 t-통계량과 그 유의성을 나타냅니다. 유의수준 0.05에서 독립변수 'sqft_lot15'의 계수는 p-value 0.01로 유의하며 나머지 모든 독립변수의 계수 p-value 0.00으로 독립변수들의 계수가 적합하다는 것을 나타냅니다.
3) '[0.025 0.975]': 각 계수의 95% 신뢰 구간입니다.

2.4.2 모델 해석:
'grade'는 계수의 절댓값 기준 0.3688로 가장 큽니다. 이는 부동산의 등급이 양의 방향으로 주택 가격에 가장 큰 영향을 미친다는 의미입니다. 나머지는 차례로 'sqft_above', 'lat', 'yr_built'등이 차례로 절댓값 기준 계수의 값이 컸습니다. 해당 변수들이 주택 가격에 미치는 정도는 어느정도 당연한 부분이지만, 단계적으로 어떤 변수의 중요도가 큰지를 수치적으로 모델링 했다는 것에 의의가 있습니다. 주의깊게 봐야 할 부분은 절댓값 기준 계수가 세 번째로 큰 'lat'변수에 관한 것인데, 그림4와 같이 북쪽과 남쪽의 기후 차가 심한 파키스탄의 특성상 위도가 집값에 중요한 영향을 미친 것을 확인할 수 있습니다.



<그림4. 파키스탄 기후(1980~2016), 출처 : 위키피디아>

2.4.3 추가정보:
1) Prob(Omnibus) : Omnibus 테스트의 p-value입니다. 일반적으로 0.05 미만이면 잔차가 정규분포를 따른다고 볼 수 있으나, 여기서는 0.000으로 매우 낮아 잔차가 정규분포를 따르지 않음을 나타냅니다.

2) Durbin-Watson: 잔차의 독립성을 검정하는 테스트로, 값이 2에 가까우면 잔차가 독립적이라고 볼 수 있습니다. 여기서는 1.981로, 잔차 간에 상당히 독립적입니다.

3) Cond. No.: 조건수로, 다중공선성 문제를 나타냅니다. 값이 높을수록 다중공선성의 문제가 있을 가능성이 높습니다. 여기서는 5.06으로, 다중공선성이 큰 문제는 아닐 수 있습니다.

3. 알림
본 분석의 코드는 다음 링크에 공개되어 있습니다.
(https://github.com/chlanstjr/assignment)