基于多元回归模型对住房价格的预测与影响因素分析

李钦 2020012872 未央-水木 02

李泽宇 2020012879 未央-水木 02

陈奕玮 2020012881 未央-水木 02

指导教师: 吴🄵

Engineering Economy

Weiyang College

Tsinghua University

2022 年 1 月 16 日

摘要

房价问题一直以来都是百姓关注的热门话题. 作为国家的支柱产业之一, 房价的走势与区域经济的发展常常有着密不可分的联系. 在城市元素日益多样化, 商品住房因金融属性的增加而呈现较大的不确定性的今天, 影响房价的相关因素的分析显得尤为重要. 本研究着眼于住房周边的地点, 探究住房周边设施, 场所等区位因素及其距离远近对于住房价格的影响, 并根据建立的简单模型进行拟合, 取得较为良好的估测效果.

本项目已在 https://github.com/godvix/housing-price-model 开源.

关键词: 住房价格; 多元回归分析; POI

目录

插图

表格

1 引言

房价问题一直以来都是百姓关注的热门话题. 作为国家的支柱产业之一, 房地产的走势常常能够影响到区域经济的发展. 近年来不断走高的房价, 不断膨胀的房地产泡沫令人担忧. 但是, 住房的真实价值却一直是一个未知数. 在不同人眼中, 住房的价值可能完全不同. 但对于消费市场而言, 住房的真实价值应当是相对确定的. 通过分析住房价格的影响因素能够一定程度上确定商品房的市场价格中金融属性所占的比重.

现有住房价格影响因素的分析往往基于居民收入, 税收政策等宏观因素, 或基于交通设施等单一因素. 住房价格通常与周边诸多环境, 基础设施等高度相关, 而不仅仅只与单一因素相关. 如果能够量化这种多元相关性, 住房价格的空间分布能够在一定程度上表征城市居民对住房周边设施所带来的效益的支付意愿, 这将能够作为评价支付意愿的重要依据之一, 有利于计算难以量化的外部性的影子价格. 此外, 通过引入更加全面的影响因素, 可能可以识别出城市的特征, 发现不同城市间的偏好差异.

在课堂上, 老师介绍了市场分析中的回归分析法. 回归分析法可以通过输入数据, 得到一系列自变量对于因变量的影响因子, 给出不同自变量对于因变量的影响程度大小; 同时, 这种影响因子的结果也有助于预测一定自变量条件下因变量的值.

这样的模型可以通过考虑交通便捷性对于房价的影响来估算人们对于通勤时间的支付意愿, 从而计算地铁的社会总效益. 经过思考后与阅读文献后, 本文拟采用回归分析法来对房价的影响因素进行分析, 得到因素对于房价的作用因子; 得到影响因子后, 本文借助该模型对于房价进行预测.

2 相关工作

在研究某一地区住房价格影响因素的过程中, 许多研究者采用回归分析法, GWR 模型法等方法进行研究. 由于本文的模型是考虑区位对房价的影响情况, 本部分主要关注那些采用区位的视角来考察房价的研究. 西南大学的刘青霞, 王方民两人在其硕士论文中都利用了回归分析法对某个特定的因素对于房价的影响做了估计. [1, 2]

阅读文章, 可以发现刘青霞建立的模型在一些方面合理且细致, 但不难发现, 研究的结果并不如人意. 例如, 我们发现刘青霞的回归分析模型 R^2 只有 0.5415, 可以发现, 这一方面显然来源于其模型的过度简化.

在刘的研究中, 对于所有的影响因素, 都统一取了该居民小区到距离其最近的该类建筑的距离, 而这种过度的简化其实是不合理的. 首先, "取最小值" 的方法本身对于某些影响因素是不合理的. 例如, 在研究中, 作者对于居民建筑与所有火车站之间的距离取最小值. 然而, 无论是在大家购买房子的时候, 还是在实际使用的过程中, 居民实际前往哪一个火车站是基于特定列车的始发车站的, 与 "最近的火车站距离" 其实没有直接关系. 其次, "取最小值" 的处理就导致在其研究中, 所有的模型在组内都是不分排名, 不加权重的. 比如, 在模型中所有的轨道交通站都是一致的. 然而, 在现实生活中, 多条线路交汇的轨道交通站与只有一条线路的轨道交通站对于周围房价的影响一定是有不同的.

对于以上两个方面, 如果模型如果能改成对于一定区域内的轨交车站, 火车站按照其重要性取加权平均, 那么就会显得更加合理.

3 模型选择

受限于专业知识的不足, 本研究选用较为简单的模型作为示例进行分析.

假设住房价格受到工业, 教育, 商业, 医疗, 交通, 旅游等周边因素的影响, 影响程度与住房与周边重要区位中心的距离有关. 简单起见, 本研究选用直线距离作为权重, 选取若干企业, 学校, 商业中心, 医院, 交通枢纽, 作为 POI (Point of Interest), 可以得到住房与这些 POI 之间的直线距离.

对于不同的 POI, 其影响力必然不同. 假设选取的所有 POI 的影响范围 ($Scope$) 从 e^3, 按排名, 指数递减至 e^{-3}. 考虑某个 poi, 其对房价的影响为

$$\delta\left(Price\right) = \exp\left\{-\frac{Dist\left[poi\right]}{Scope\left[poi\right]}\right\}$$

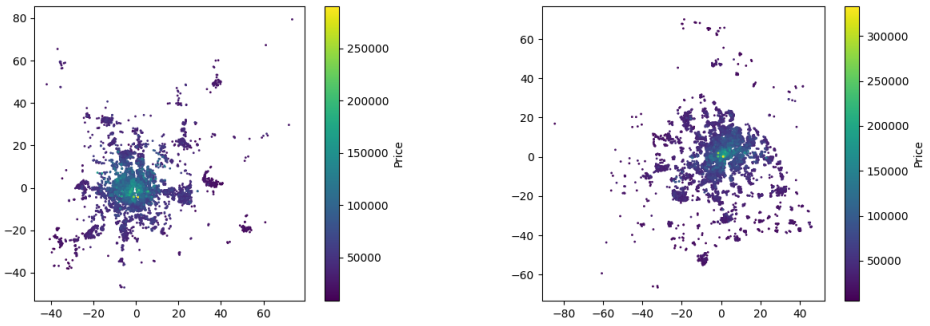
其中 $Dist\left[poi\right]$ 表示住房与该 POI 之间的直线距离. 住房的价格近似满足

$$Price = \sum_{category}\left(Impact\left[category\right] * \sum_{poi}\exp\left\{-\frac{Dist\left[poi\right]}{Scope\left[poi\right]}\right\}\right)$$

通过回归计算可以得到相应的系数 $Impact\left[category\right]$, 这表征着该类 POI 对于房价的影响程度.

4 数据来源

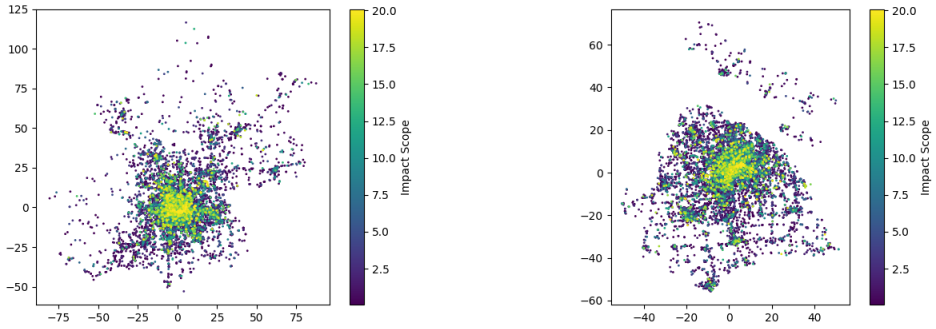
为了对房价的区位因素进行分析, 本研究需要北京与上海两地各个小区房价, 各类设施位置与排名两方面数据. 对于两地各小区房价, 采用网络上的两地 2018 年的各小区房价信息, 如图 1 所示.



(a) 北京市　　(b) 上海市

图 1: 住房价格样本分布

对于两地各类设施位置, 利用高德地图的 POI 搜索功能, 能够获取两地地图上的全部设施的信息, 如图 2 所示.



(a) 北京市　　(b) 上海市

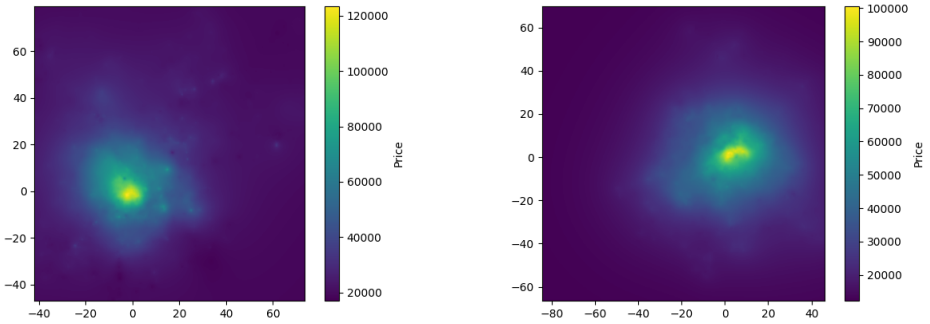
图 2: POI 分布

令人惊喜的是, 所有的设施已经被高德开放平台划入了一定的类别. 基于本文研究目标, 本文对高德开放平台分类进行了一些调整: 对一些类别进行了合并, 对一些与研究目的无关的设施信息进行了删除.

5 数据分析

5.1 价格热力图

通过对搜集到的数据进行回归分析与预测, 我们可以得到北京市与上海市的房价预测热力图, 如图 3 所示.



(a) 北京市 (b) 上海市

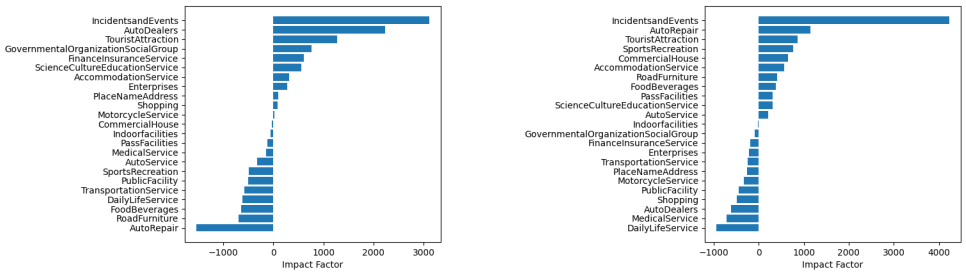
图 3: 模型预测

从预测图 (图 3) 和样本分布 (图 1) 的对比可以主观地发现二者的房价分布之间呈现出一致的趋势, 在北京市与上海市的中心均有显著的 10 万以上高房价的区域, 由内而外大致呈现递减的变化趋势.

值得注意的是, 在整张地图上, 市中心外存在多个高房价的区域, 离市中心的远近不是房价唯一的影响因素, 一个区域拥有商场, 医院, 公园等重要区位条件也是高房价的成因.

5.2 区位分析

在生成房价预测的同时, 通过回归分析同时可以得出各个区位因素的影响因子, 如图 4 所示.



(a) 北京市 (b) 上海市

图 4: 影响因子

在北京与上海的对比中可以发现, 高居第一的都为 Incidents and Events. 此外, Tourist Attraction 的重要性也较大. 其他区位的影响因子具有一定的随机成分, 这是由于这些区位间并不独立: 例如, 由于市中心拥有良好的地理位置能够吸引足够大的人流量, 上海的景点 (豫园, 外滩等), 商业区 (淮海路, 南京路等) 同时多位于市中心区域, 那么这两个因素必然具有很大的相关性, 那么我们便难以排除这二者之间潜在的联系, 从而间接导致了区位排行的混乱性.

这是我们的模型的一大局限之处 — 由于文化, 历史, 经济等多方面综合因素, 我们难以兼顾区位选择的完整性和独立性.

5.3 模型评价

5.3.1 2018 年住房价格估测

使用前文所述模型对北京市和上海市 2018 年的住房价格进行回归计算, 结果的判定系数 R^2 分别为 0.732 与 0.669. 同时, 回归计算中也出现了异常偏大的条件数, 这意味着回归因子之间可能存在强相关性. 更加详细的计算结果见 A 部分.

我们用得到的模型对已有的数据回测, 得到误差分布如图 5.

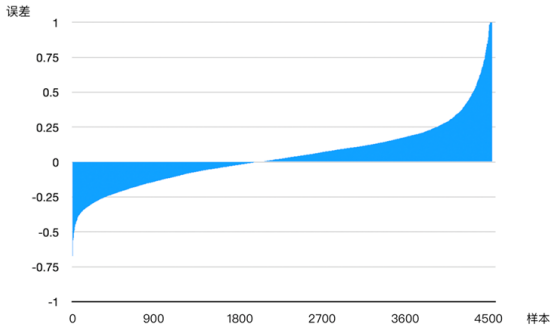


图 5: 误差分布

5.3.2 2022 年住房价格估测

在北京市随机选取灯市口, 西四, 五道口, 杨庄, 百子湾, 五个地点进行当前房价预测, 结果与出现了较大的误差, 在调查国家统计局数据后, 发现 2020 年的北京房价较 2018 年有 14 % 的上涨, 因此基于全北京房价均匀上涨的假设, 对数据进行了线性的校准, 得到结果如表 1 所示, 可以发现在预测 2022 年房价上也是较为可观的.

表 1: 2022 年住房价格估测

地区	2018 年房价 (估测)	2022 年房价 (估测)	2022 年房价 (样本)	误差
灯市口	111 743.1228	127 460.9104	125 000	1.97 %
西四	119 677.2265	136 511.0252	151 000	−9.60 %
五道口	82 661.687 14	94 288.880 05	113 000	−16.56 %
杨庄	57 558.578 81	65 654.768 51	53 000	23.88 %
百子湾	53 822.890 11	61 393.617 83	61 000	0.65 %

6 应用场景

6.1 消除信息不对称

在房产市场中可能发生的一种情况是: 由于买方卖方的信息不对称而产生价格差额, 从而损害交易的健康性. 在本文提出的模型中, 只需给出目标房产的经纬度, 就可获得该房产的客观估值, 从而促进交易的公平性.

6.2 房产估值

已建成的房产价格在市场平衡中往往趋于稳定, 而在建工程的楼盘往往没有一个明确的报价, 本文的模型可用于在建工程的报价预测, 为广大消费者提供未来财产分配的合理建议.

7 未来工作

7.1 模型选择

受限于专业知识的不足, 笔者无法建立更加符合现实情况的模型进行统计. 目前人工智能发展迅速, 使用柔性更强的神经网络也许能够取得更好的结果.

此外, 由于北京, 上海的住房价格分布呈现明显的单中心性, 某地的住房价格很大程度上由其与市中心的距离决定, 这使得细部特征的研究较为困难. 如果能够使用统计学方法更加精确地提取细部特征, 则能够取得更好的效果.

除地理位置以外, 物业水平, 地方政策, 住房建成年限, 装修情况, 居民收入等也是影响房价的重要因素. 受限于数据来源的不足, 本研究暂未将其它因素纳入考虑范围.

7.2 区位之间的相关性

区位之间的相关性引起了本文主要的挑战, 在消除相关性后进行实验, 便能得到明确的区位比较. 在这里提出一个思路: 对区位进行进一步的细分, 到市中心的距离是一个重要因素, 那么与之相关的诸多因素, 例如商业, 餐饮, 可分为一类; 与到市中心距离关联不大的因素分为另一类: 例如景观等.

7.3 密集数据

本文的算法对世界各地的城市是通用的, 在获取一个城市的各主要区位分布, 房价分布后, 便能模拟出全城的房价分布, 因此可以尝试获取国外城市的类似数据, 例如可以从 Google Map 上收集城市的相关信息, 构建一个全球的房价分布. 在这个基础上, 我们可以进一步进行城市之间的比较——对比城市的系数组, 从而得到各个城市影响房价的主要区位条件, 这是研究世界文化, 经济差异的一个新思路.

8 总结

在工程经济学的学习过程中, 我们在回归分析法的自学过程中产生了从区位视角, 运用模型分析房价的这样一种想法.

从眼就开始到结束的过程中, 我们的模型也发生了很大的变化. 一开始, 我们的研究方法是考虑一系列小区到某几个重要的区位点的距离 — 如万象汇, 故宫博物院, 在拟合的过程中, 我们发现数据结果并不理想, 于是我们从前面提到的两篇硕士论文中寻找思路, 运用高德开放平台爬取大量 POI 点, 再利用模型进行分析. 一开始, 我们的模型算法也仅仅停留在最近点或是同类中心点距离的和, 由于 R^2 较低, 我们不断调整算法, 最终使得 R^2 达到比较好的水平, 同时这种算法对应的现实意义也是优于前两种的.

经过分析, 我们得到了价格热力图, 各个区位的影响因子, 并用这样的模型来估测 2018 与 2022 年的住房价格, 得到了比较好的结果.

其实, 这样的研究也是具有一定现实意义的. 正如老师在课堂上讲授的一样, 许多公共投资项目需要考虑其产生的社会总效益, 而房价正是反映人们对于某一建筑的支付意愿的一个窗口; 同时, 这样的研究也可以为我们辅助我们预测房价.

土木学科叫做 Civil Engineering, 也就是民用工程. 这是一个与人的生活息息相关的学科, 实际地服务于社会, 服务于人们的日常生活, 是这个学科的终极意义, 这也是我们在工程经济学课上所体悟到的价值内核.

参考文献

[1] 刘青霞. 重庆市中心城区景观可达性对住宅价格的影响研究. 硕士, 2021.

[2] 王方民. 基于生活圈的公共服务设施配置对房价的影响研究. 硕士, 2021.

A 回归结果

A.1 北京市回归结果

Dep. Variable:	price	**R-squared:**	0.732
Model:	OLS	**Adj. R-squared:**	0.731
Method:	Least Squares	**F-statistic:**	863.3
Date:	Sun, 16 Jan 2022	**Prob (F-statistic):**	0.00
Time:	23:05:17	**Log-Likelihood:**	-80268.
No. Observations:	7284	**AIC:**	1.606e+05
Df Residuals:	7260	**BIC:**	1.607e+05
Df Model:	23		
Covariance Type:	nonrobust		

表 2: Beijing OLS Regression Results

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	1.815×10^4	1144.337	15.862	0.000	1.59×10^4	2.04×10^4
0	−318.2477	180.670	−1.761	0.078	−672.413	35.918
1	2235.4543	229.650	9.734	0.000	1785.273	2685.636
2	−1534.4204	236.678	−6.483	0.000	−1998.378	−1070.463
3	23.2936	73.607	0.316	0.752	−120.997	167.584
4	−644.9779	125.628	−5.134	0.000	−891.244	−398.711
5	82.6199	172.075	0.480	0.631	−254.698	419.938
6	−614.9914	160.017	−3.843	0.000	−928.671	−301.312
7	−487.5924	158.244	−3.081	0.002	−797.797	−177.388
8	−142.9470	161.561	−0.885	0.376	−459.654	173.760
9	312.4724	116.548	2.681	0.007	84.003	540.941
10	1270.1086	116.237	10.927	0.000	1042.251	1497.967

表 2: Beijing OLS Regression Results (续)

	coef	std err	t	P> |t|	[0.025	0.975]
11	−26.1004	103.495	−0.252	0.801	−228.980	176.779
12	763.8384	98.091	7.787	0.000	571.551	956.126
13	563.2973	99.021	5.689	0.000	369.187	757.408
14	−583.0906	113.205	−5.151	0.000	−805.006	−361.175
15	605.8885	121.843	4.973	0.000	367.040	844.737
16	279.9137	95.020	2.946	0.003	93.647	466.181
17	−697.0469	132.489	−5.261	0.000	−956.763	−437.330
18	103.6793	149.566	0.693	0.488	−189.514	396.872
19	−499.3558	139.606	−3.577	0.000	−773.024	−225.688
20	3109.6958	489.380	6.354	0.000	2150.368	4069.024
21	−55.7886	19.411	−2.874	0.004	−93.841	−17.737
22	−114.4076	120.922	−0.946	0.344	−351.449	122.634

Omnibus:	2542.377	**Durbin-Watson:**	1.210
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	31068.689
Skew:	1.319	**Prob(JB):**	0.00
Kurtosis:	12.768	**Cond. No.**	2.02e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.02e+03. This might indicate that there are strong multicollinearity or other numerical problems.

A.2 上海市回归结果

Dep. Variable:	price	**R-squared:**	0.669
Model:	OLS	**Adj. R-squared:**	0.669
Method:	Least Squares	**F-statistic:**	1026.
Date:	Sun, 16 Jan 2022	**Prob (F-statistic):**	0.00
Time:	23:12:51	**Log-Likelihood:**	-1.2861e+05
No. Observations:	11681	**AIC:**	2.573e+05
Df Residuals:	11657	**BIC:**	2.574e+05
Df Model:	23		
Covariance Type:	nonrobust		

表 3: Shanghai OLS Regression Results

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	1.213×10^4	1104.216	10.982	0.000	9962.140	1.43×10^4
0	208.1793	137.317	1.516	0.130	−60.984	477.343
1	−620.6409	160.637	−3.864	0.000	−935.517	−305.765
2	1154.0144	166.506	6.931	0.000	827.634	1480.395
3	−328.8537	110.295	−2.982	0.003	−545.050	−112.657
4	377.6965	116.365	3.246	0.001	149.601	605.792
5	−489.5849	118.257	−4.140	0.000	−721.388	−257.782
6	−941.0288	143.450	−6.560	0.000	−1222.215	−659.843
7	769.1201	129.818	5.925	0.000	514.655	1023.585
8	−711.9080	92.832	−7.669	0.000	−893.875	−529.941
9	560.3915	129.874	4.315	0.000	305.818	814.965
10	869.9834	127.214	6.839	0.000	620.622	1119.344
11	645.5831	92.588	6.973	0.000	464.095	827.071
12	−95.8444	136.095	−0.704	0.481	−362.613	170.924
13	301.4136	98.198	3.069	0.002	108.929	493.899
14	−253.1105	69.496	−3.642	0.000	−389.335	−116.886
15	−183.6645	99.961	−1.837	0.066	−379.604	12.275
16	−215.7938	82.412	−2.618	0.009	−377.335	−54.253
17	400.7241	124.958	3.207	0.001	155.785	645.663
18	−262.7602	30.338	−8.661	0.000	−322.228	−203.292
19	−448.9212	75.897	−5.915	0.000	−597.692	−300.150
20	4231.6910	3216.305	1.316	0.188	−2072.805	1.05×10^4
21	−15.9718	12.171	−1.312	0.189	−39.830	7.886
22	308.3106	138.158	2.232	0.026	37.498	579.123

Omnibus:	5219.570	**Durbin-Watson:**	1.042
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	144950.970
Skew:	1.561	**Prob(JB):**	0.00
Kurtosis:	19.973	**Cond. No.**	8.12e+03

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 8.12e+03. This might indicate that there are strong multicollinearity or other numerical problems.

B 成员分工

李钦 摘要, 引言, 模型选择, POI 数据爬取, 回归计算与样本估测

李泽宇 数据分析, 模型评价, 应用场景, 未来工作

陈奕玮 引言, 相关工作, 数据来源, 总结