北京市二手房房屋单价影响因素研究

安　彤
北京建筑大学　北京　102600

摘　要：随着我国二手房市场的逐渐放开，二手房数量不断增加，二手房交易规模也不断扩大。目前人们对于二手房的需求也比较大。本文主要以北京市二手房市场为研究对象，通过Python爬虫得到链家官网2021年11月的在售数据，对北京市二手房价格变化的影响因素进行定量分析，将爬取的数据进行整理，选取小区、地区、城区、装修情况、楼层、建筑类型、卧室数、厅数、建筑面积和房龄作为自变量，房屋单价作为因变量，进行了多元回归分析，通过探索城区与房龄及城区与建筑类型的交互作用修正回归模型进而提高模型的拟合度。最后通过一个实例来利用模型进行预测。

关键词：二手房房屋单价；多元回归分析

Study on the Influence of Unit Price of Second-hand Housing in Beijing

Tong An
Beijing University of Civil Engineering and Architecture, Beijing 102600

Abstract: With the gradual liberalization of our country's second-hand housing market, the number of second-hand houses has continued to increase, and the scale of second-hand housing transactions has also continued to expand. At present, people's demand for second-hand housing is also relatively large. This paper mainly takes the second-hand housing market in Beijing as the research object, obtains the sales data of Lianjia official website in November 2021 through Python crawler, quantitatively analyzes the factors affecting the price changes of second-hand housing in Beijing, sorts out the crawled data, and selects community, area, urban area, decoration situation, floor, building type, number of bedrooms, number of halls, building area and house age are used as independent variables, and the unit price of houses is used as a dependent variable, and a multiple regression analysis is carried out. The interaction of building types modifies the regression model and improves the fit of the model. Finally, an example is used to use the model to make predictions.

Keywords: Unit price of second-hand housing; Multiple regression analysis

1　数据来源及说明

本文所用数据来自2021年11月，通过Python爬取数据，删除信息不全、信息错误等冗余信息后，共计33992条数据，其中包括16个城区。网站所提供的信息包含了许多如位置、城区、房屋类型、建筑面积等多个变量。本文选取的数据共包含11个变量，其中3个连续型变量，8个离散型变量，详细变量说明见表1。将处理后数据导入SPSS进行数据分析。

2　多元线性回归模型构建

为了深入探索自变量和因变量房屋单价之间的具体定量关系，利用逐步筛选策略建立了多元线性回归模型。多元回归方程为：

$$y = 110745.837 - 7008.163x_1 - 1271.409x_2 - 33.701x_3 + 1688.388x_4 + 655.514x_5 + 916.062x_6 \quad (1)$$

如表2所示，调整后R方为0.636接近1，此时所包含的自变量有城区（x_1）、楼层（x_2）、建筑面积（x_3）、卧室数（x_4）、装修情况（x_5）、厅数（x_6）。方差分析结果中线性关系的检验（F检验）对应P值近似为0，小于显著水平 α=0.05，说明回归方程是显著的，模型拟合的程度较好。如表3所示，可以得到各变量前对应的系数，其中表格标准化系数这列说明了自变量对因变量影响的占比，由此说明其中影响最大的是城区，其次是建筑面积。

表1　数据变量说明表

变量类型		变量名	详细说明	取值范围	备注
因变量	连续变量	房屋单价	数值型变量（单位：元/平方米）	10020-18000000	
自变量	离散变量	小区	字符型变量		例：水上华城
		地区	字符型变量		例：崇文门
		城区	定性变量 共16个水平	1-6	1-16分别表示：西城、东城、海淀、朝阳、丰台、石景山、通州、大兴、延庆、顺义、昌平、房山、怀柔、密云、平谷、门头沟
		装修情况	定性变量 共4个水平	0-3	0-3分别表示：其他、毛坯、简装、精装
		楼层	定性变量 共5个水平	1-5	1-5分别表示底层、低楼层、中楼层、高楼层、顶层
		建筑类型	定性变量 共4个水平	1-4	1-4分别表示：板楼、板塔结合、塔楼、平房
		卧室数	数值型变量 整数（单位：室）	1-12	房屋卧室数
		厅数	数值型变量 整数（单位：厅）	0-9	房屋厅数
	连续变量	建筑面积	数值型变量（单位：平方米）	12.8-2323.87	房屋面积
		房龄	数值型变量（单位：年）	0-72	房龄=2021-建成时间

表2　关于房屋单价线性回归分析结果（一）

模型	R	R方	调整后R方	标准估算的错误	D-W
6	.797[f]	.636	.636	20857.680	1.093

3　变量间交互作用的分析

在上文中线性回归模型拟合度为0.636，同时DW值为1.093，说明因变量的规律没有被完全解释，同时拟合程度还可以提高。因此考虑到自变量之间可能存在交互作用，所以下面进行验证。

3.1城区与房龄

如表4所示，可以得到城区和房龄交互作用对应的概率P-值近似为0，小于显著性α水平，所以可以认为城区和房龄交互作用对于因变量的影响显著。

3.2城区与建筑类型

如表5所示，可以得到城区和建筑类型交互作用对应的概率P-值近似为0，小于显著性α水平，所以可以认为城区和建筑交互作用对于因变量的影响显著。

4　修正多元线性回归方程

在上一节中对城区和房龄的交互作用及城区和建筑类型的交互作用做了简单分析，下面进行建立带有交互项的回归方程。

多元回归方程为：

$$y = 88906.884 - 4044.232x_1 - 105.938x_1 \bullet x_7 + 752.854x_7 - 1049.404x_2 - 20.334x_3 + 1045.786x_4 + 820.336x_5 + 856.839x_6 + 4083.849x_8 - 739.459x_1 \bullet x_8 \quad (2)$$

表3　关于房屋单价线性回归分析结果（二）

模型	未标准化系数		标准化系数	t	显著性	共线性统计	
	B	标准错误	Beta			容差	VIF
（常量）	110745.837	569.373		194.505	.000		
城区	-7008.163	30.037	-.797	-233.320	.000	.946	1.057
楼层	-1271.409	99.378	-.043	-12.794	.000	.979	1.022
建筑面积（单位：平米）	-33.701	2.826	-.064	-11.926	.000	.384	2.606
卧室数（单位：室）	1688.388	179.446	.049	9.409	.000	.407	2.458
装修情况	655.514	123.310	.018	5.316	.000	.993	1.008
厅数（单位：厅）	916.062	268.343	.015	3.414	.001	.575	1.738

a.因变量：房屋单价（单位：元/平方米）

表4　房屋单价多因素方差分析结果

主体间效应检验					
因变量：　房屋单价（单位：元/平方米）					
源	III类平方和	自由度	均方	F	显著性
修正模型	32294031122563.582[a]	701	46068518006.510	208.914	.000
截距	2514125549949.514	1	2514125549949.514	11401.173	.000
城区	5156690916913.787	15	343779394460.919	1558.987	.000
房龄	139319616174.281	69	2019124872.091	9.156	.000
城区*房龄	910290399690.519	617	1475349108.088	6.690	.000
误差	7120417487672.190	32290	220514632.632		
总计	156968626644224.000	32992			
修正后总计	39414448610235.770	32991			

a.R 方 =.819（调整后 R 方 =.815）

表5　表房屋单价多因素方差分析结果

主体间效应检验					
因变量：　房屋单价（单位：元/平方米）					
源	III类平方和	自由度	均方	F	显著性
修正模型	31233594373771.270a	55	567883534068.569	2286.291	.000
截距	748600484101.862	1	748600484101.862	3013.855	.000
城区	12305251443451.656	15	820350096230.111	3302.718	.000
建筑类型	10398015101.657	3	3466005033.886	13.954	.000
城区*建筑类型	165514695766.598	37	4473370155.854	18.010	.000
误差	8180854236464.505	32936	248386392.897		
总计	156968626644224.000	32992			
修正后总计	39414448610235.770	32991			

a.R 方 =.792（调整后 R 方 =.792）

对于上式（4.3）进行解释：

（1）不同城区对应的不同房屋单价不同，在计算不同城区时，x_1对应的值应为初始赋值量（即西城区=1，海淀区=2等）。

（2）对于房龄这一变量，房龄每增加一年，房屋单价增加752.85元。

（3）对于城区和房龄这一交互变量，当房屋所处城区相同时，房龄每增加一年，房屋单价减少105.94元。而当房屋的房龄相同时，在海淀区的房屋比在西城区的房屋对应的房屋单价低105.94元。

（4）不同楼层类型对应不同的房屋单价，在计算不同楼层时，x_2对应的值应为初始赋值变量（即底层=1，低楼层=2，等等）。

（5）对应建筑面积这一变量来说，在此模型中，每增加一平米，房屋单价减少20.34元，由此看出该变量对于房屋单价的影响力极小，几乎可以忽略不计。

（6）对于卧室数这一变量来说，每增加一间卧室，房屋单价上涨1045.79元。

（7）对于装修情况这一变量来说，在计算不同楼层时，x_5对应的值应为初始赋值变量（即毛坯房=1，简装=2，等等）。

（8）对于厅数数这一变量来说，每增加一间客厅，房屋单价上涨856.84元。

（9）对于建筑类型这一变量来说，在计算不同建筑类型时，x_8对应的值应为初始赋值变量（即板楼=1，板塔结合=2，等等）。

（10）对于城区和建筑类型这一交互变量，当房屋所处城区相同时，板楼的房屋单价比板塔结合的房屋单价高739.46元。而当房屋的建筑类型相同时，如两个房屋均为板楼，则在海淀区的房屋比在西城区的房屋对应的房屋单价低739.46元。

表6　带有交互项线性回归分析结果（一）

模型摘要[k]					
模型	R	R 方	调整后 R 方	标准估算的错误	德宾－沃森
10	.814[j]	.662	.662	20103.414	1.209

如表6所示，调整后R方为0.662接近1，并且高于

表7　关于房屋单价线带城区和建筑类型性回归分析结果（二）

系数[a]								
模型		未标准化系数		标准化系数	t	显著性	共线性统计	
		B	标准错误	Beta			容差	VIF
10	（常量）	88906.884	919.838		96.655	.000		
	城区	−4044.232	81.936	−.460	−49.359	.000	.118	8.465
	城区和房龄	−105.938	2.177	−.279	−48.655	.000	.312	3.210
	房龄（单位：年）	752.854	19.741	.226	38.136	.000	.292	3.424
	楼层	−1049.404	96.111	−.035	−10.919	.000	.972	1.029
	城区和建筑类型	−739.459	44.273	−.132	−16.702	.000	.164	6.082
	建筑类型	4083.849	289.641	.087	14.100	.000	.270	3.707
	装修情况	820.336	119.810	.022	6.847	.000	.977	1.024
	卧室数（单位：室）	1045.786	176.369	.030	5.930	.000	.391	2.556

之前的模型，拟合度变高了，因此接受当前的模型。此时所包含的自变量有城区（x_1）、楼层（x_2）、建筑面积（x_3）、卧室数（x_4）、装修情况（x_5）、厅数（x_6）、房龄（x_7）、建筑类型（x_8）。方差分析结果中线性关系的检验（F检验）对应P值近似为0，小于显著水平 $\alpha=0.05$，说明回归方程是显著的，模型拟合的程度较好。如表7所示，可以得到各变量前对应的系数，其中表格标准化系数这列说明了自变量对因变量影响的占比，由此说明其中影响最大的是城区，其次是城区和房龄的交互作用。

5　利用多元线性回归方程预测房价

实例预测：客户要购买一间在海淀区，中楼层，面积在85平米左右两室一厅的房子，房龄在6年左右，房型为板塔结合型，根据预测方程可以得到，这样的一间房屋对应的房屋均价在76913.20元，总价在653.76万元。

参考文献：

[1]贾俊平，何晓群，金勇进.统计学[M].北京：中国人民大学出版社，2012.

[2]沈孝玲.北京二手房市场的统计分析[D].北京理工大学，2017.

[3]薛薇.基于SPSS的数据分析[M].中国人民大学出版社，2014.