



狗熊会精品案例系列

北京二手房房价影响因素分析

水妈，关蓉 - 2016 年 7 月



案例版权归狗熊会所有，如需转载，请联系小编

扫描二维码，获取更多精彩案例！

北京二手房房价影响因素分析

摘要：本案例以北京市二手房为研究对象，通过统计分析探究学区房、地铁房、房屋面积等相关因素对二手房房价的影响作用，建立了考虑“城区×学区”交互效应的对数线性模型来刻画各因素与单位面积房价的关联。结论表明区位因素和房屋的内部因素对于二手房房价都有显著影响，各城区的“学区优势”对单位面积房价的影响略有区别。

一、背景介绍

北京市房地产市场是我国最为发达、最具有代表性的房地产市场之一。近年来，北京市出台了一系列的已购公房上市出售政策，如《北京市已购公有住房上市出售实施办法》、《中央在京单位已购公有住房上市出售实施办法》等。诸多利好政策的出台让广大二手房业主能更加灵活地处理自己的房地产，如出售旧房或办理二手房抵押贷款。截至 2016 年 5 月 25 日的背景住宅年内交易数据显示，北京市二手房占市场住宅成交比例高达 86.2%，而这一比例在过去几年一直维持在 70% 左右，北京楼市已经全面进入二手房时代。

然而，广大业主往往因为强调某个或某些因素对价格的影响，错误估计自己的房屋价格，最终导致心理价格与现时的市场价格相距甚远。事实上，二手房的市场价格是多种因素综合作用的结果，到底有哪些因素在影响二手房的价格呢？房价的巨大差异又是如何产生的？本案例收集了北京城内六区 16210 套在售二手房相关数据，对二手房房价的相关影响因素展开研究。

二、数据来源与说明

本案例使用的是来自某二手房中介网站的北京在售二手房的相关数据，共 16210 条记录¹，数据采集时间为 2016 年 5 月。数据共包含 7 个变量，其中 4 个连续型变量，3 个离散型变量；因变量为单位面积房价，其他变量为自变量。具体的变量说明如表 2-1 所示。

¹ 本案例使用的 16210 条记录已经进行了数据预处理，具体处理方法见数据清洗文档。

表 2-1：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	单位面积房价	单位：万元/平方米	1.83~14.98	
自变量	房屋面积	单位：平方米	30.06~299.00	
	卧室数	单位：个	1~5	
	内部因素	厅数	单位：个	0~3
				建模时处理成是否有客厅
	楼层	定性变量：共 3 个水平	低楼层、中楼层、高楼层	相对楼层
	所属城区	定性变量：共 6 个水平	朝阳区、东城区、丰台区、海淀区、石景山区、西城区	
	区位因素	是否临近地铁	定性变量：共 2 个水平	1 代表临近地铁；0 代表不临近地铁
		是否学区房	定性变量：共 2 个水平	1 代表学区房；0 代表非学区房

三、描述性分析

在对房价的影响因素进行模型探究之前，首先对各变量进行描述性分析，以初步判断房价的影响因素，为后续研究做铺垫。

（一）因变量：单位面积房价

在本案例中，单位面积房价的最小值为 1.83 万元/平方米，所对应的房屋是丰台区东山坡三里的一间两居室，总面积 100.83 平米；最大值为 14.99 万元/平方米，所对应的房屋是西城区金融街的一套三室一厅，总面积 77.40 平米。

通过单位面积房价的直方图（图 3-1），可以看到：单位面积房价是呈现右偏分布的。具体来说，单位面积房价的均值为 6.12 万元/平方米、中位数为 5.74 万元/平方米。这一现象符合我们对于房价的基本认知，即存在少数天价房，从而拉高了房价的平均水平。

总的来讲，北京市城区的二手房价存在极大差距，有一半以上的二手房已经进入了“5 万元时代”。

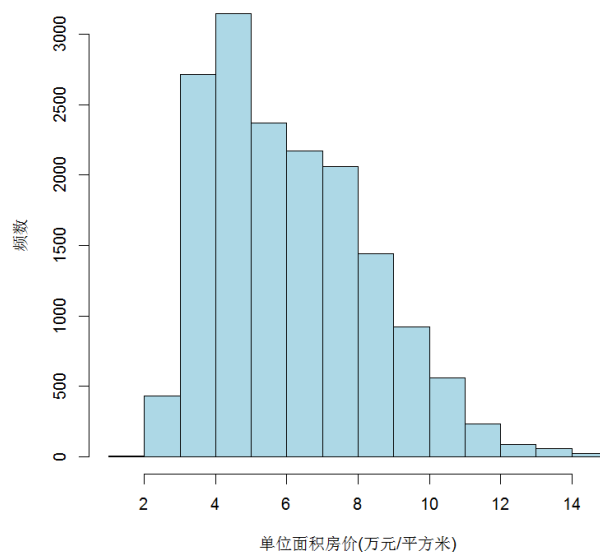


图 3-1：单位面积房价的分布情况

（二）自变量：外部因素

从图 3-2 可以看出，各城区的单位面积房价差异较大。总的来说，通过单位面积房价可以将六个城区分为两个梯队：第一梯队包括西城、海淀、东城；第二梯队为朝阳、丰台、石景山。显然，第一梯队的各区的单位面积房价明显高于第二梯队各城区的单位面积房价。初步猜想，这可能与各城区的资源禀赋有关，比如学区资源的多寡、优劣，交通的便利程度，等等。

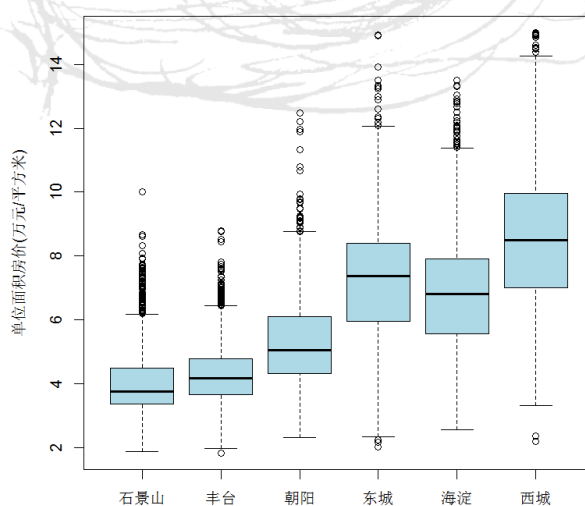


图 3-2：各城区的单位面积房价分布情况

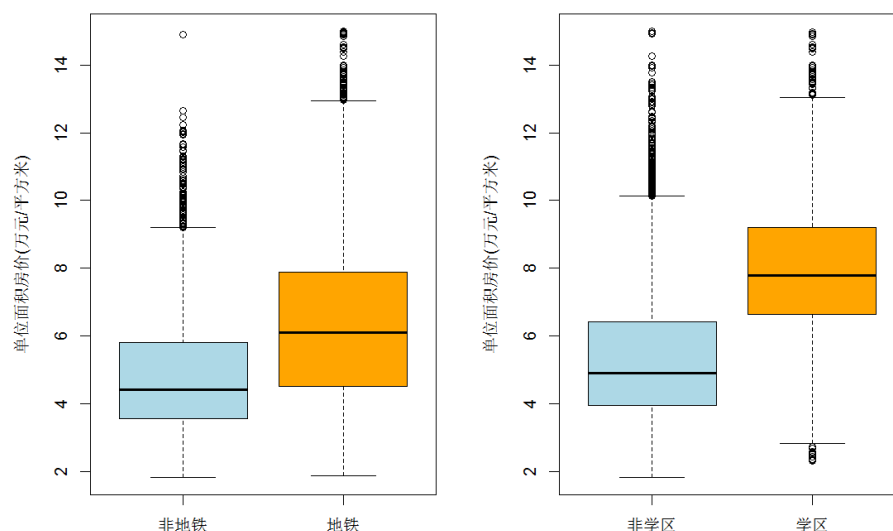


图 3-3：公共交通资源（地铁）、学区资源与单位面积房价的关联

图 3-3 进一步描述了公共交通资源（地铁）与学区资源对房价的影响。可以看出：地铁房比非地铁房的单位面积房价高，学区房比非学区房的单位面积房价高。这些初步结论都比较符合大众的预期。

（三）自变量：内部因素

从图 3-4 可以看出，对于不同的厅数而言（中图），厅数越少，单位面积房价越高，其中 0 厅的单位面积房价明显高于 1 厅、2 厅和 3 厅。仔细观察原始数据，可以发现 0 厅的样本比例较少（5%），而其中有不少是胡同四合院，房价普遍较高。因此，在后续的建模分析中，本案例将“厅数”这一变量处理为“是否有厅”，其中 0 表示无厅，1 表示有厅。

而对于不同的卧室数量（左图）、不同的楼层（右图），房屋的单位面积房价差异并不明显。值得注意的是，本案例中的楼层变量的取值在各小区样本内并不具有严格的可比性。尽管如此，在后续分析中，本案例依然将这两个因素放入模型，通过建模结果来判断其对单位面积房价是否存在影响。

图 3-5 的散点图描述了房屋面积（横轴）与单位面积房价（纵轴）之间的相关关系。容易看出，房屋面积与单位面积房价存在一定的负相关（经计算，相关系数 = -0.07 且显著），即：房屋面积越大，单位房屋面积房价越低。

此外，同等面积房屋的单位面积房价波动较大，尤其是 100 平米以下的房屋。

这充分说明，二手房的定价受到了除房屋面积以外的诸多因素的影响，比如前文提及的城区因素、学区因素、地铁因素，等等。

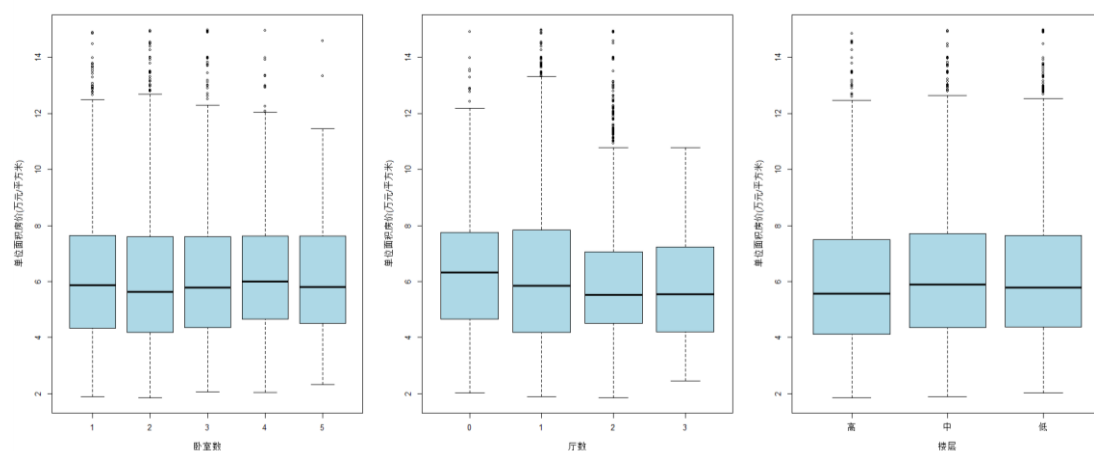


图 3-4：不同卧室数、不同厅数、不同楼层的单位面积房价分布情况

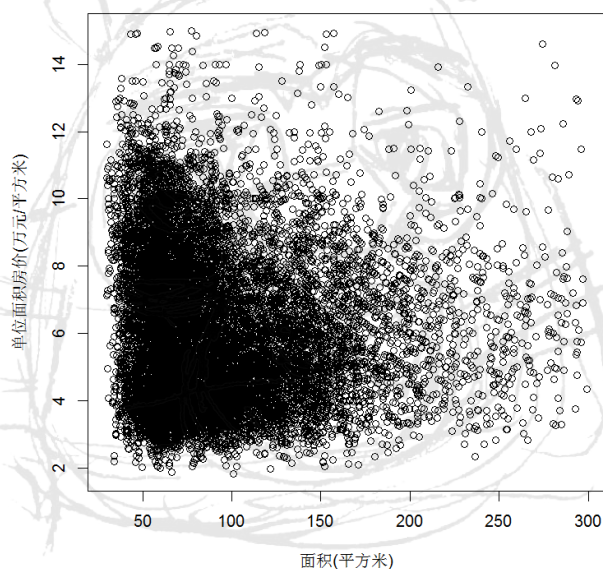


图 3-5：房屋面积与单位面积房价散点图

综上，通过对本案例数据的描述性分析，可以推测：对单位面积房价可能会产生影响的因素包括：区位因素（城区、地铁、学区）和内部因素（卧室数、是否有客厅、面积、楼层）；从影响作用来看，区位因素比内部因素更为明显一些。

四、模型建立

为了更深入地分析各因素对二手房房价的影响，本案例将建立单位面积房价关于区位因素和内部因素的回归模型，使用量化的方式更为精细地刻画两方面因素的影响作用大小，并且试图使用该模型来预测二手房房价。

（一）线性回归模型

首先，对数据建立简单的线性回归模型，结果如表 4-1 所示。

表 4-1：线性回归结果

变量	回归系数	p 值	备注
截距项	3.315	<0.001	
城区-丰台	0.131	0.001	
城区-朝阳	0.875	<0.001	
城区-东城	2.443	<0.001	基准组：石景山组
城区-海淀	2.191	<0.001	
城区-西城	3.705	<0.001	
学区房	1.183	<0.001	
地铁房	0.672	0.001	
楼层-中层	0.152	<0.001	基准组：高层
楼层-底层	0.198	<0.001	
有客厅	0.163	<0.001	
卧室数	0.111	<0.001	
房间面积	-0.002	<0.001	
F 检验	p 值<0.0001	调整的 R^2	0.5901

在控制其他因素不变时，可以得到如下结论：

- 对于城区这一变量，石景山区单位面积房价最低，西城区单位面积房价最高，比石景山区每平方米平均高出 3.70 万元；
- 对于学区这一变量，学区房比非学区房单位面积房价平均高出 1.18 万元；
- 对于地铁这一变量，地铁房比非地铁房单位面积房价平均高出 6720 元；
- 高层房屋单位面积房价最低，其次是中层，低层房屋单位面积房价最高；
- 有客厅的房子单位面积房价更高；
- 卧室数每增加一间，单位面积房价平均增加 1110 元；
- 房屋面积的增加会带来单位面积房价的降低。

这些结论与之前的猜想基本符合。而且模型的 F 检验拒绝原假设，说明建立的模型是显著的；调整的 R^2 为 0.59，模型的拟合程度尚可接受。

接下来进行模型诊断。如图 4-1 所示。从左上图可以看出，随着预测值的增

大，残差的波动也随之增大，说明模型可能存在异方差的问题。为了解决这一问题，考虑对因变量单位面积房价做对数处理。

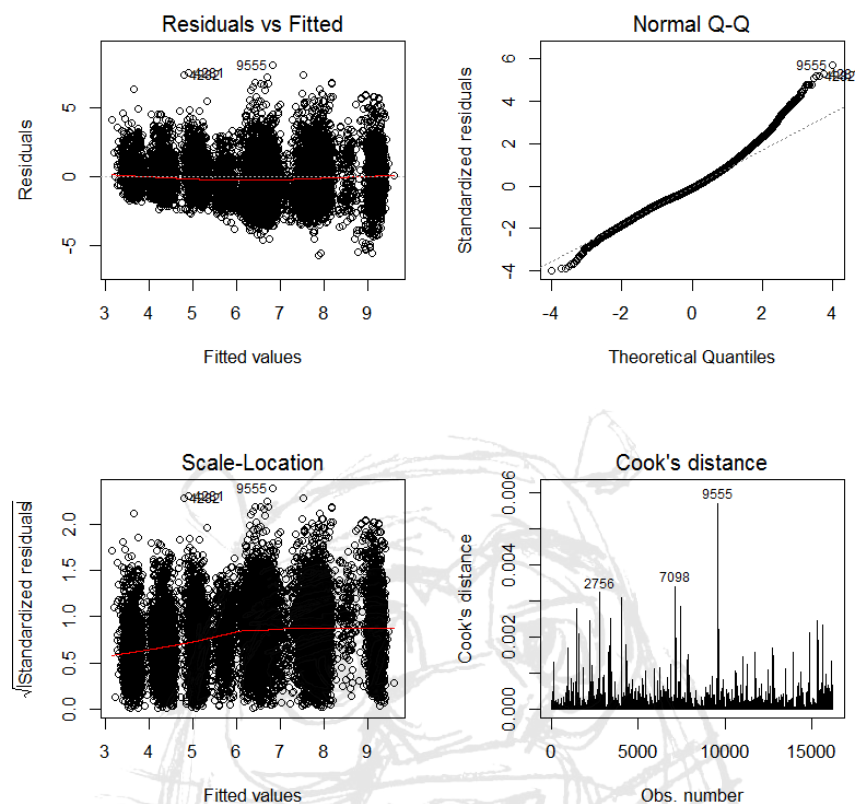


图 4-1：简单线性模型诊断图

（二）对数线性模型

对因变量做对数变换后，建立对数线性模型并进行诊断。估计结果如表 4-2 和所示，诊断结果如图 4-2 所示。

从图 4-2 可以看出，异方差问题得到极大改善；Cook 距离表现正常，表明没有异常点。因此，使用对数线性模型来刻画单位面积房价与各因素的关系是更为合理的。从表 4-2 可以看到，模型仍然是显著的（F 检验通过了），模型的拟合程度有略微上升（调整的 $R^2=0.6079$ ）。

需要特别注意的是，与线性模型不同，对数线性模型的系数估计值应解读为“增长率”，即某自变量变动一个单位，因变量变化的百分比。因此，通过表 4-2 可以得到如下结论：在控制其他因素不变的情况下，

表 4-2：对数线性回归结果

变量	回归系数（ $\times 10^{-1}$ ）	p 值	备注
截距项	12.360	<0.001	
城区-丰台	0.441	<0.001	基准组：石景山组
城区-朝阳	2.057	<0.001	
城区-东城	4.577	<0.001	
城区-海淀	4.320	<0.001	
城区-西城	6.270	<0.001	
学区房	1.719	<0.001	
地铁房	1.282	<0.001	
楼层-中层	0.152	<0.001	基准组：高层
楼层-底层	0.198	<0.001	
有客厅	0.275	0.001	
卧室数	0.140	<0.001	
房间面积	-0.003	<0.001	
F 检验	p 值<0.0001	调整的 R^2	0.6079

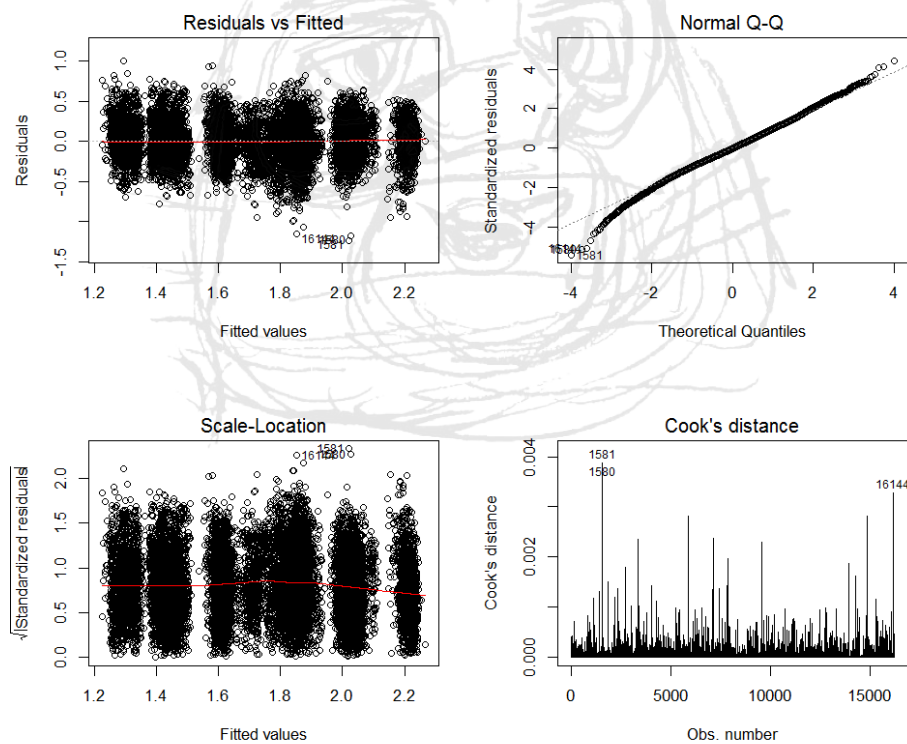


图 4-2：对数线性模型诊断图

- 对于不同城区，石景山区单位面积房价最低，西城区单位面积房价最高，比石景山区平均贵 62.70%；
- 对于学区这一变量，学区房比非学区房的单位面积房价平均贵 17.19%；

- 对于地铁这一变量，地铁房比非地铁房的单位面积房价平均贵 12.82%；
- 高层房屋单位面积房价最低，其次是中层，低层房屋单位面积房价最高；
- 有客厅的房子单位面积房价更高，平均贵 2.75%；
- 卧室数每增加一间，单位面积房价平均增加 1.40%；
- 房屋面积的增加会带来单位面积房价的降低。

（三）对数线性交互模型

根据描述性分析的结果发现，所属城区的不同以及是否为学区房会使房价存在非常明显的差异。因此，在对数线性模型的基础上，本案例考虑更为复杂的情况：城区和学区的交互作用，建立考虑“城区×学区”交互效应的对数线性模型，结果如表 4-3 所示。

可以看到，考虑交互效应后，模型仍然是显著的（F 检验的 p 值很小），且模型的拟合效果略有提高（调整的 $R^2=0.6108$ ）。

表 4-3：考虑交互效应的对数线性模型的回归结果

变量	回归系数 ($\times 10^{-1}$)	p 值	备注
截距项	12.410	<0.001	
城区-丰台	0.429	<0.001	基准组：石景山组
城区-朝阳	2.184	<0.001	
城区-东城	4.467	<0.001	
城区-海淀	4.121	<0.001	
城区-西城	6.177	<0.001	
学区房	-1.800	<0.001	
地铁房	1.257	<0.001	
楼层-中层	0.263	<0.001	基准组：高层
楼层-底层	0.343	<0.001	
有客厅	0.270	0.001	
卧室数	0.140	<0.001	
房间面积	-0.003	<0.001	
丰台×学区	2.948	<0.001	基准组： 石景山×学区
朝阳×学区	2.780	<0.001	
东城×学区	3.706	<0.001	
海淀×学区	3.876	<0.001	
西城×学区	3.638	<0.001	
F 检验	p 值<0.0001	调整的 R^2	0.6108

当考虑了“城区×学区”的交互效应后，一个明显的变化是：学区变量系数估计变成负数。由于“城区”、“城区×学区”两个因素的基准组均为石景山区，因此对应的结论是：在石景山区，学区房比非学区房的单位面积房价低，这一结论在图 4-3 中也可以体现。

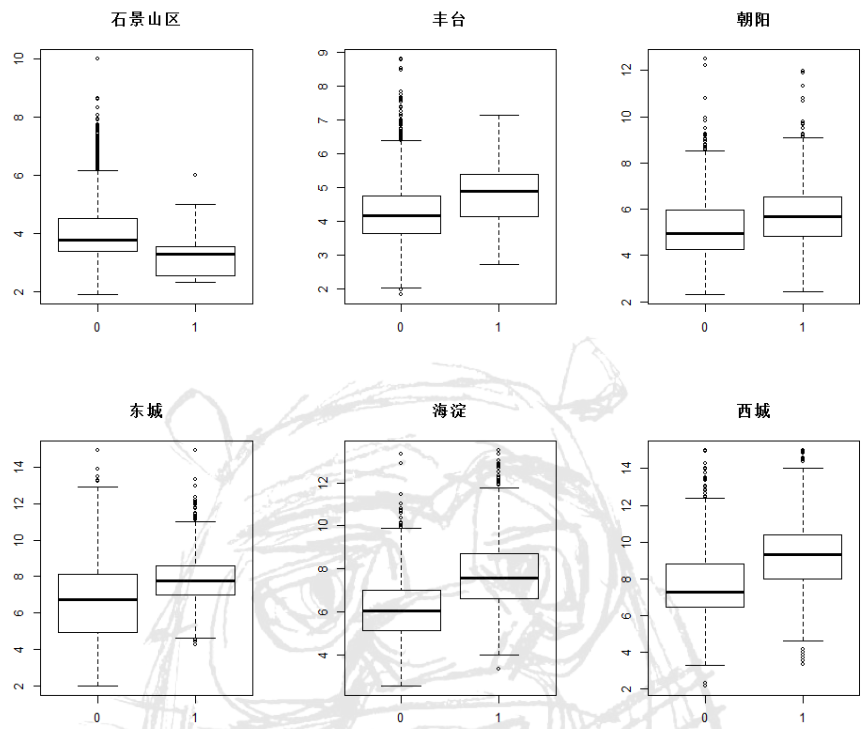


图 4-3：不同城区中，学区房与非学区房的单位面积房价分布情况

本案例推测，出现这一结果的原因有二。首先是数据原因，在来自石景山区的样本中，学区房比例极低（如表 4-4 所示）；其次是客观原因，与其他城区相比，石景山区的学区资源相对较差。因此，出现了“石景山区学区房比非学区房的单位面积房价低”的结果。

表 4-4：各城区样本分布

城区	样本量	学区房占比（%）
石景山	1947	0.92
丰台	2947	3.18
朝阳	2864	20.84
东城	2783	45.81
海淀	2919	47.48
西城	2750	56.10

若不考虑学区因素的区别，整体而言，西城区与海淀区的单位面积房价之比

可以通过对数线性模型计算，为： $e^{0.627-0.432} = e^{0.195} = 1.215$ 。若考虑“城区×学区”的交互效应，可以进一步分析各城区的学区房和非学区房的房价差异。此时，对于学区房，西城区与海淀区的单位面积房价之比为： $e^{0.617+0.363-0.412-0.387} = e^{0.181} = 1.198$ ；对于非学区房，西城区与海淀区的单位面积房价之比为： $e^{0.617-0.412} = e^{0.205} = 1.228$ 。由此可见，两个城区的房价差异主要体现为非学区的房价差异。类似地，还可以通过交互模型了解和比较其他城区的房价差异情况。

（四）预测

由于考虑交互效应的对数线性模型对本案例数据具有较强的解释力，因此以下将使用该模型开展预测。

假设有一家三口，父母为了能让孩子在西城区上学，想买一套邻近地铁的两居室，面积是 85 平方米，低层楼层，那么房价大约是多少呢？根据交互模型，预测的到的单位面积房价为 9.29 万元/平方米，总价高达 789.78 万元。

五、结论与建议

本案例对 2016 年 5 月北京市部分在售二手房的房价数据进行统计分析，得到如下结论：

- 影响北京市二手房单位面积房价的主要因素有：（1）区位因素：城区、地铁、学区；（2）内部因素：卧室数、是否有客厅、面积、楼层。
- “学区优势”对各城区单位面积房价的影响有所区别。海淀、西城、东城、丰台和朝阳区的学区房房价明显高于非学区房的房价；但石景山区的学区房房价比非学区房的房价低。

由于房价的影响因素有很多，因此再未来的研究中可以考虑在模型中加入更多因素，比如小区位置（地处几环）、小区环境（如绿化情况、容积率，等等）、周边配套设施（如商圈、医院，等等）等。另外，若要将模型推广到其他城市，还要进一步考虑城市特有因素（如：在旅游城市是否为海景房等）。