



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。

1

## 背景介绍

- 北京市房地产市场是我国房地产市场最为发达，也是最具有代表性的房地产市场之一。
- 什么样的因素影响着北京市商品房的销售价格？
- 价格上的巨大差异又是怎样产生的？
- 本案例收集了北京城内六区16210套在售二手房相关数据，试图揭开影响房价因素的神秘面纱！

2

## 目录

- 数据来源和说明
- 描述性分析
  - 区位因素：城区、地铁、学区
  - 内部因素：卧室数、是否有客厅、面积、楼层
- 数据建模
  - 简单线性模型：解读→诊断→因变量取对数
  - 对数线性模型：解读→关心位置因素→城区与学区交互
  - 带有交叉项的模型：解读→关于学区的有趣现象
  - 预测
- 结论和建议

3

## 数据来源和说明

- 数据来自某二手房中介网站
- 2016年5月，北京在售二手房的相关数据
- 样本量 $n=16210$

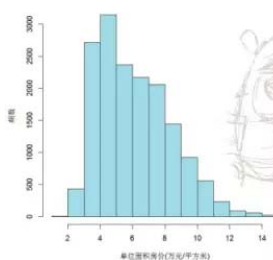
4

## 数据来源和说明

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	
	房屋面积	单位：平方米	30.06~299.00	
自变量	卧室数	单位：个	1~5	
	厅数	单位：个	0~3	建模处理成是否有客厅
	所属楼层	定性变量共3个水平	低楼层、中楼层、高楼层	相对楼层
区位因素	所属城区	定性变量共6个水平	朝阳、东城、丰台、海淀、石景山、西城	
	是否邻近地铁	定性变量共2个水平	1代表邻近地铁 0代表不邻近地铁	82.78%邻近地铁
	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	30.31%是学区房

5

## 因变量：单位面积房价

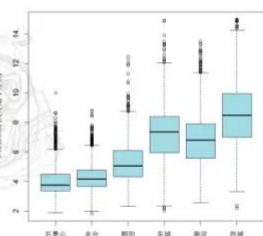


- 均值：6.12万元/平方米
- 中位数：5.74万元/平方米
- 最小值：1.83万元/平方米
- 丰台区东山坡三里的一间两居室
- 总面积100.83平米
- 最大值：14.99万元/平方米
- 西城区金融街的一套三室一厅
- 总面积77.40平米

6

## 自变量：区位因素 2/1

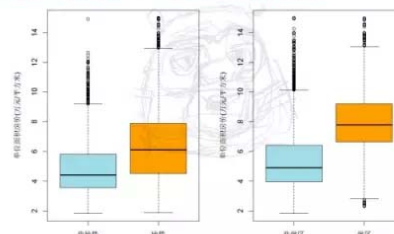
- 各城区的单位面积房价差异大
- 第一梯队：西城、海淀、东城；
- 第二梯队：朝阳、丰台、石景山
  - 猜想：与各城区的资源禀赋有关
  - 例如：学区资源、交通便利程度等



7

## 自变量：区位因素 2/2

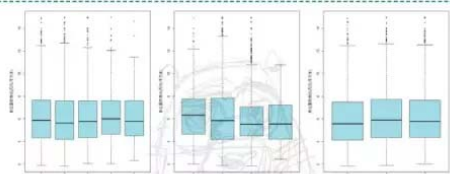
- 学区房、地铁房 → 单位面积房价高



8

## 自变量：内部因素 2/1

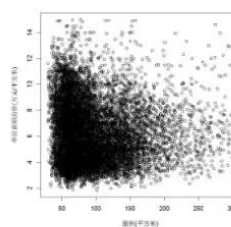
- 卧室数：不同的卧室数量，单位面积房价差异并不明显
- 厅数：厅数越少，单位面积房价越高
  - 0厅的样本比例很少（5%），不少是“房价逆天”的胡同四合院
- 楼层：不同楼层的单位面积房价差异并不明显



9

## 自变量：内部因素 2/2

- 房屋面积与单位面积房价存在一定的负相关（相关系数=-0.07且显著）
- 同等面积房屋的单位面积房价波动较大，尤其是100平方米以下的房屋
- 猜测：受到其他因素的影响（比如：是否为学区房、是否为地铁房等）



10

## 小结

可能影响单位面积房价的因素：

- 区位因素：城区、地铁、学区
- 内部因素：卧室数、是否有客厅、面积、楼层

11

## 线性回归模型（因变量：单位面积房价）

变量	回归系数	p值	备注
截距项	3.315	<.0001	
城区-丰台	0.131	0.001	
城区-朝阳	0.875	<.0001	
城区-东城	2.443	<.0001	基准组：石景山区
城区-海淀	2.191	<.0001	
城区-西城	3.705	<.0001	
学区房	1.183	<.0001	
地铁房	0.672	0.001	
楼层-中层	0.152	<.0001	基准组：高层
楼层-低层	0.198	<.0001	
有客厅	0.163	<.0001	
卧室数	0.111	<.0001	
房间面积	-0.002	<.0001	
F检验	p值<.0001	调整的R2	0.5901

12

## 线性回归模型：结果解读

控制其他因素不变时

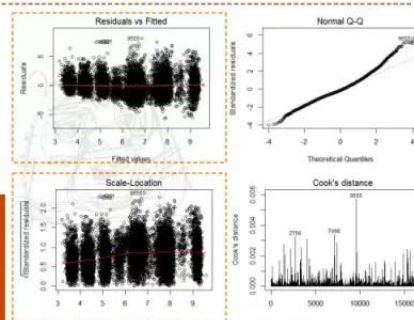
- ▶ **城区**：石景山区单位面积房价最低，西城区单位面积房价最高，比石景山区每平米平均高出**3.70**万元
- ▶ **学区房**比非学区房单位面积房价平均高出**1.18**万元
- ▶ **地铁房**比非地铁房单位面积房价平均高出**6720**元
- ▶ **高层房屋**单位面积房价最低，其次是中层，低层房屋单位面积房价最高
- ▶ **有客厅**的房子单位面积房价更高
- ▶ **卧室数**每增加一间，单位面积房价平均增加**1110**元
- ▶ **房屋面积**的增加会带来单位面积房价的降低

13

## 线性回归模型：模型诊断

随着预测值的增大，  
残差的波动也随之增大  
→ “异方差”现象

解决方案：  
对因变量（单位面积  
房价）取对数



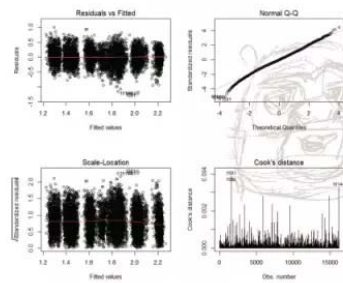
14

## 对数线性模型（因变量：对数房价）

变量	回归系数 ( $\times 10^{-4}$ )	p值	备注
截距项	12.360	<.0001	
城区-丰台	0.441	<.0001	基准组：石景山区
城区-朝阳	2.057	<.0001	
城区-东城	4.577	<.0001	
城区-海淀	4.320	<.0001	
城区-西城	6.270	<.0001	
学区房	1.719	<.0001	基准组：高层
地铁房	1.282	<.0001	
楼层-中层	0.152	<.0001	
楼层-低层	0.198	<.0001	
有客厅	0.275	0.001	
卧室数	0.140	<.0001	
房间面积	-0.003	<.0001	
F检验	p值 <.0001	调整的R2	0.6079

15

## 对数线性模型：模型诊断



- ▶ “异方差”现象得到极大改善
- ▶ Cook距离表现正常，没有异常点

16

## 对数线性模型：结果解读

- ▶ 与线性模型不同，对数线性模型的系数估计解读为“增长率”
- ▶ 控制其他因素不变时
- ▶ **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%**
- ▶ **学区房**比非学区房单位面积房价平均贵**17.19%**
- ▶ **地铁房**比非地铁房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

17

## 交互模型（对数线性模型）

- ▶ 根据描述分析结果发现，不同的城区、是否为学区房的房价有非常明显的差异
- ▶ 因此，在对数线性模型的基础上，考虑更为复杂的情况：城区和学区的交互作用
- ▶ 交互效应：城区\*学区

18



变量	回归系数 ( $\times 10^{-3}$ )	p值	备注
截距项	12.410	<.0001	
城区-丰台	0.429	<.0001	基准组：石景山区
城区-朝阳	2.184	<.0001	
城区-东城	4.467	<.0001	
城区-海淀	4.121	<.0001	
城区-西城	6.177	<.0001	
学区房	-1.800	<.0001	
地铁房	1.257	<.0001	
楼层-中层	0.263	<.0001	基准组：高层
楼层-低层	0.343	<.0001	
有客厅	0.270	0.001	
卧室数	0.140	<.0001	
房间面积	-0.003	<.0001	
丰台*学区	2.948	<.0001	
朝阳*学区	2.780	<.0001	
东城*学区	3.706	<.0001	
海淀*学区	3.876	<.0001	
西城*学区	3.638	<.0001	
F检验	p值<.0001	调整的R2	0.6108

19

### 交互模型：结果解读 2/1

学区系数估计变成负数 → 石景山区：学区房比非学区房的单位面积房价低！

数据原因：

来自石景山区的样本中，学区房比例极低

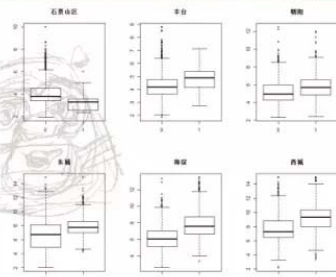
城区	样本量	学区房占比 (%)
石景山	1947	0.92
丰台	2947	3.18
朝阳	2864	20.84
东城	2783	45.81
海淀	2919	47.48
西城	2750	56.10

20

### 分城区的学区房单位面积房价对比

客观原因：

石景山区的学区资源在6个城区中相对较差



21

### 交互模型：结果解读 2/2

整体而言（不考虑学区因素），西城区与海淀区的单位面积房价之比

对数线性模型： $e^{0.627-0.432} = e^{0.195} = 1.215$

学区房房价哪家强？

学区房：西城区与海淀区的单位面积房价之比

交互模型： $e^{0.617+0.363-0.412-0.387} = e^{0.181} = 1.198$

非学区房：西城区与海淀区的单位面积房价之比

交互模型： $e^{0.617-0.412} = e^{0.205} = 1.228$

结论：交互模型能发现更多信息！

22

### 预测

- 假设：一家三口，为了孩子在西城区上学，想买一套邻近地铁的两居室，面积是85平方米，低层楼层
- 根据交互模型，预测的单位面积房价是 9.29 万元/平方米，总价 789.78 万元

23

### 结论和建议

- 影响北京市二手房单位面积房价的主要因素有：
  - 区位因素：城区、地铁、学区
  - 内部因素：卧室数、是否有客厅、面积、楼层
- 使用对数线性模型，可以克服数据中存在的异方差问题
- 使用交互模型，能带来更好的模型解读
  - “学区优势”对各城区单位面积房价的影响有所区别

24

未来研究方向

- 房价的影响因素诸多，应在模型中加入更多因素：
  - 小区位置（如：地处几环？）
  - 小区环境（如：绿化情况）
  - 周边配套（如：商圈、医院，等等）
  - .....
- 若要推广到其他城市，还可进一步考虑城市特有因素（如：是否为海景房等）

- 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）。