北京二手房房价影响因素分析

《Python数据科学：全栈技术详解》
第七章 线性回归

Ben

背景介绍

在对房价的影响因素进行模型研究之前，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，进而建立房价预测模型。

步骤如下：

（一） 因变量分析：单位面积房价分析

（二） 自变量分析：

2.1 自变量自身分布分析

2.2 自变量对因变量影响分析

（三）建立房价预测模型

3.1 线性回归模型

3.2 对因变量取对数的线性模型

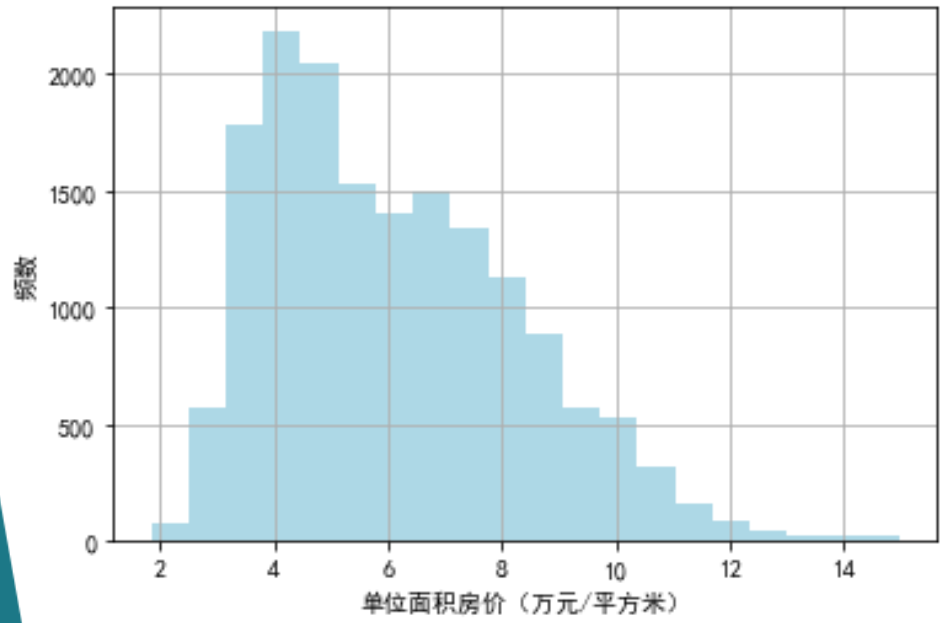
3.3 考虑交互项的对数线性

（四）预测： 假设有一家三口，父母为了能让孩子在东城区上学，想买一套邻近地铁的两居室，面积是70平方米，中层楼层，那么房价大约是多少呢?

❶ 描述统计

2 线性模型

dist	roomnum	halls	AREA	floor	subway	school	price
chaoyang	1	0	46.06	middle	1	0	48850
chaoyang	1	1	59.09	middle	1	0	46540
haidian	5	2	278.95	high	1	1	71662



平均值	6.11518
中位数	5.7473
标准差	2.22934

1 描述统计

2 线性模型

dist	丰台	海淀	朝阳	东城	西城	石景山
频数	2947	2919	2864	2783	2750	1947

roomnum	2	3	1	4	5
频数	7971	4250	3212	675	102

halls	1	2	0	3
频数	11082	4231	812	85

floor	middle	high	low
频数	5580	5552	5078

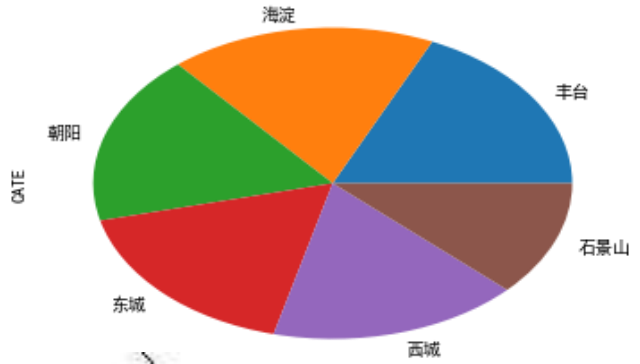
subway	1	0
频数	13419	2791

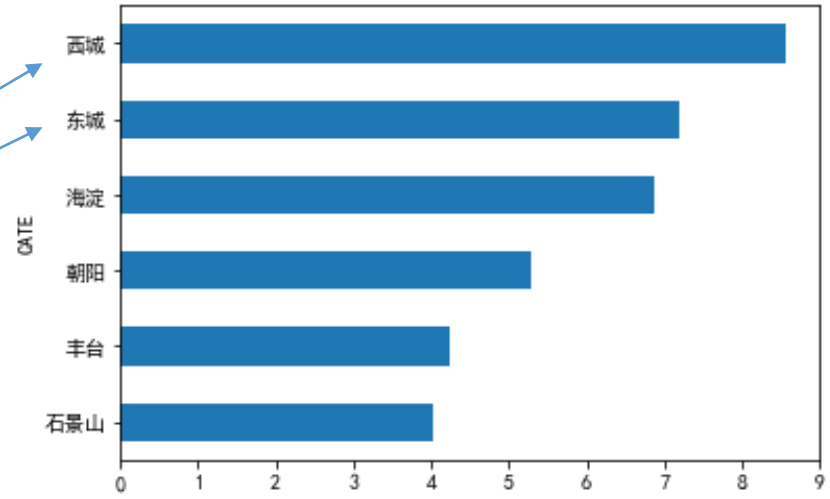
school	0	1
频数	11297	4913

AREA	最小值	平均值	众数	最大值	标准差
	30.06	91.7466	78.83	299	44.00077

❶ 描述统计

2 线性模型

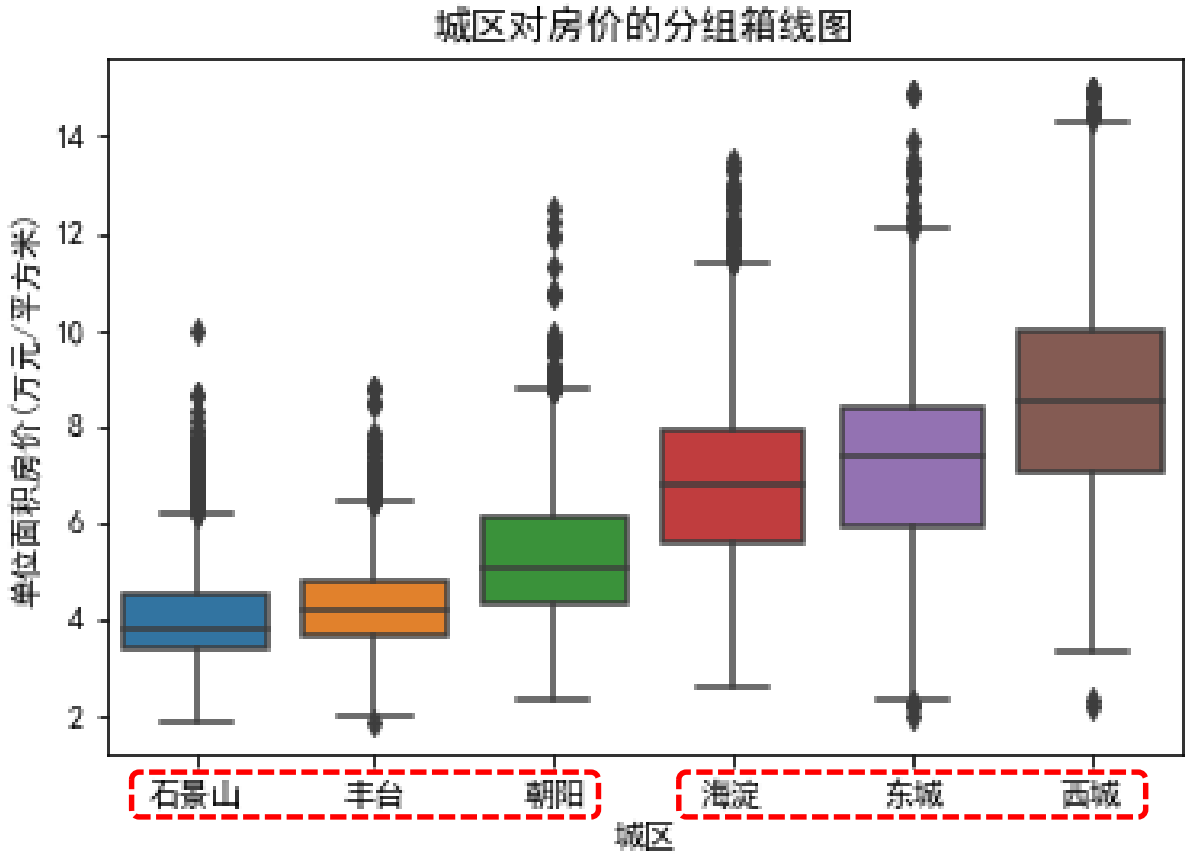




dist	丰台	海淀	朝阳	东城	西城	石景山
频数	2947	2919	2864	2783	2750	1947

❶ 描述统计

2 线性模型



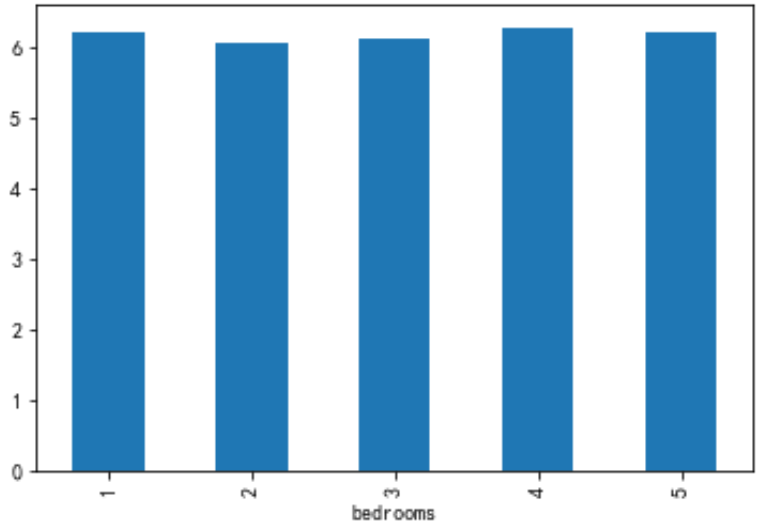
平均值	6.11518
中位数	5.7473
标准差	2.22934

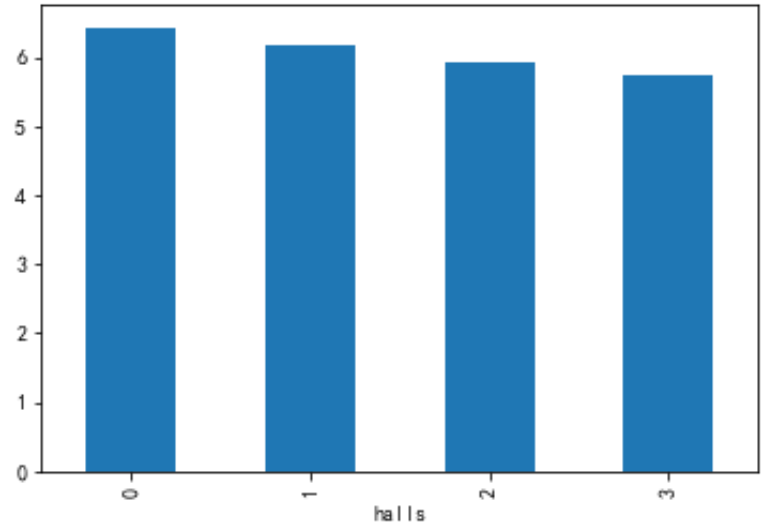
dist	丰台	海淀	朝阳	东城	西城	石景山
频数	2947	2919	2864	2783	2750	1947

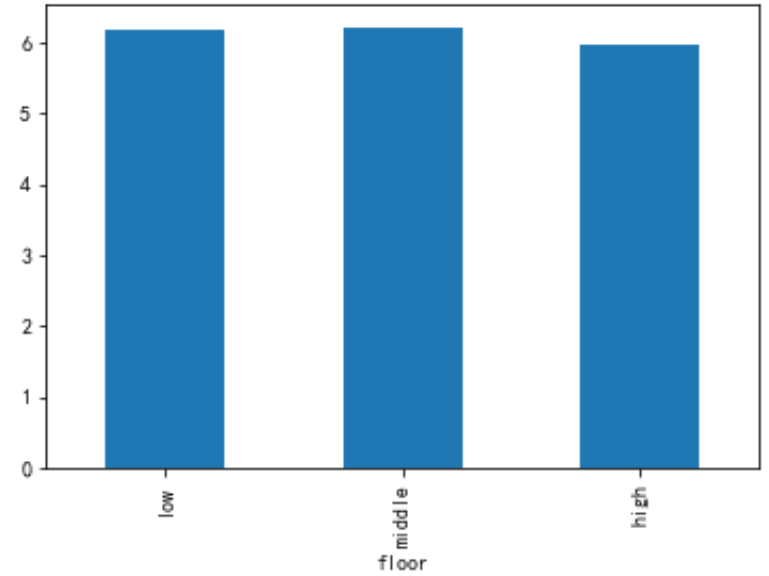
1 描述统计

2 线性模型

自变量：卧室数、厅数和楼层





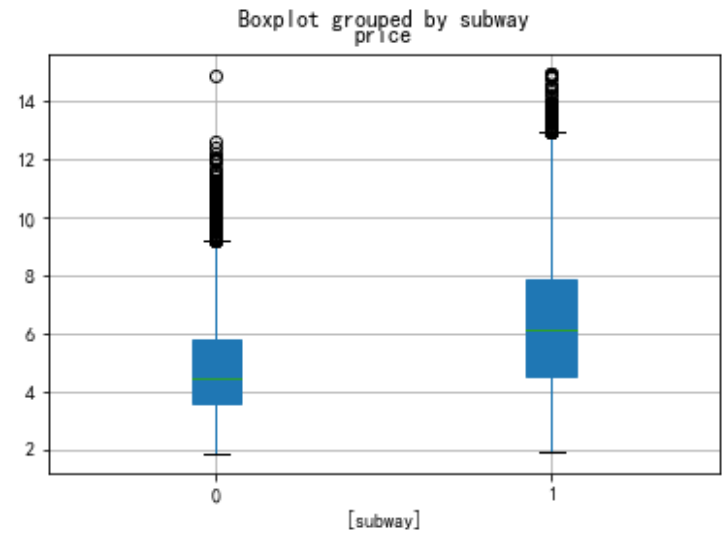


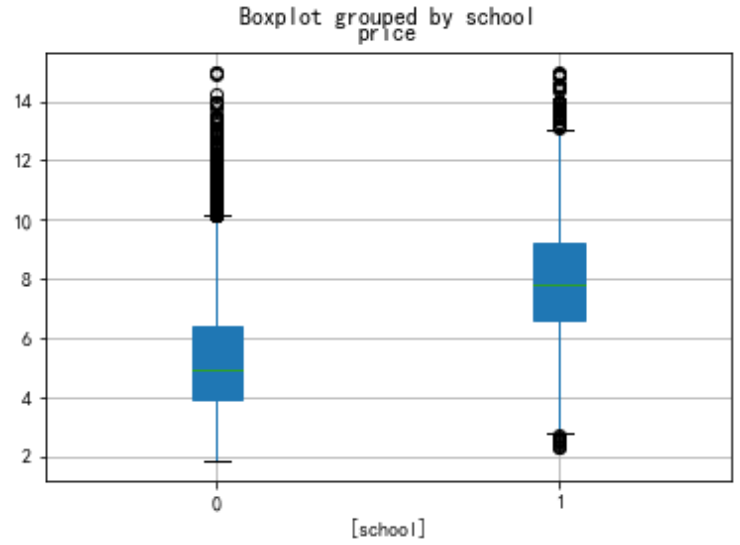
- 不同**卧室数**的房价差别不大
- **厅数**对单位面积房价有轻微影响
- 不同**楼层**的单位面积房价差异不明显

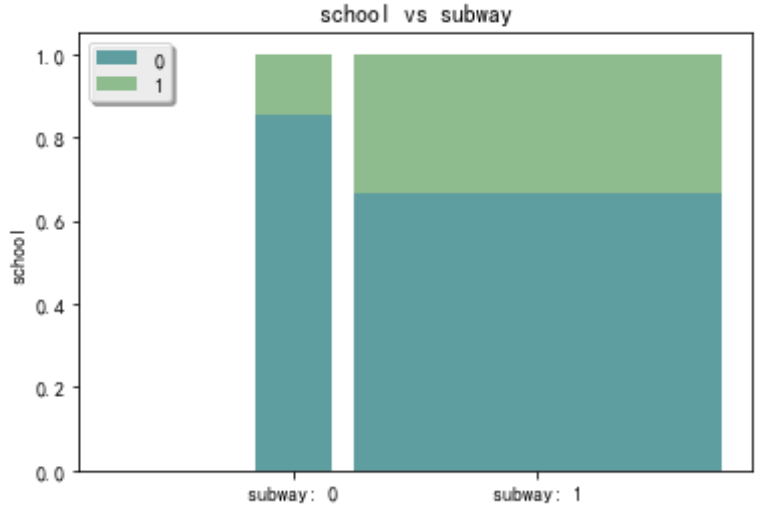
《Python数据科学：全栈技术详解》

❶ 描述统计

2 线性模型



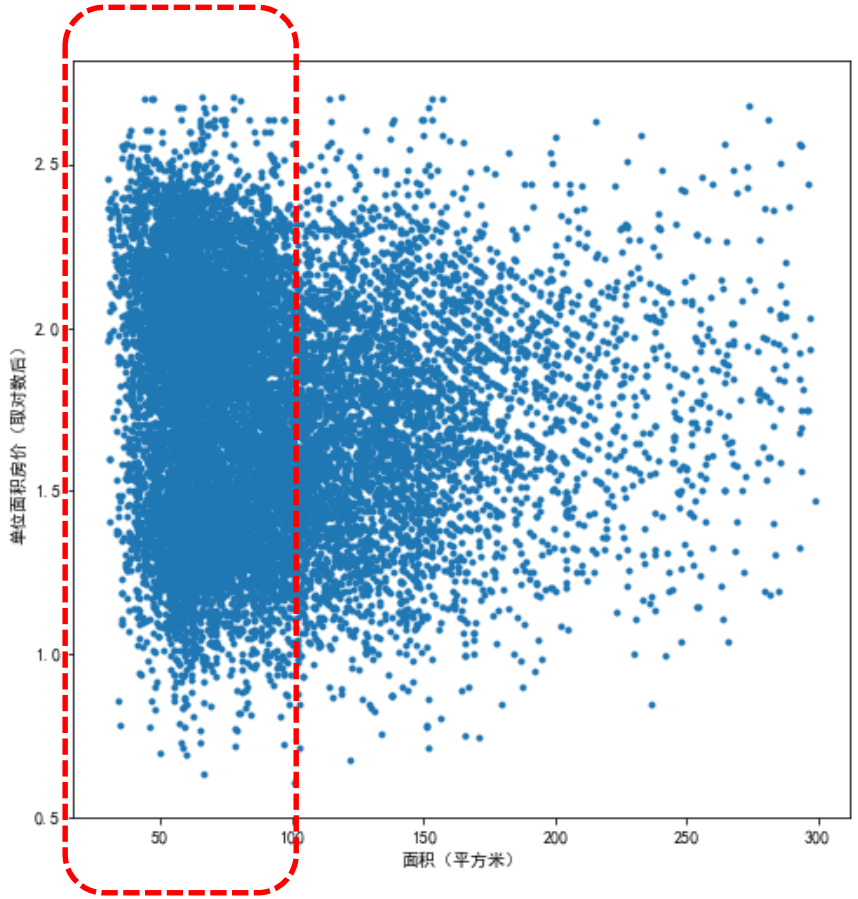




- **学区房、地铁房**的单位面积房价高
- 样本中，**地铁房中的学区房**比重比非地铁房中的学区比重更大

❶ 描述统计

2 线性模型



- **对数化的**房屋面积与**对数化的**单位面积房价存在一定的**负相关**（相关系数=-0.09）
- 同等面积房屋的**单位面积房价波动较大**，尤其100平米以下的房屋

1 描述统计

2 线性模型

按照区分层，每个区抽取400个样本：

```
dat0=get_sample(datall, sampling="stratified", k=400, stratified_col=['dist'])
```

提出假设

区的单位面积房价影响大

~~卧室数对单位面积房价影响不大~~

客厅数与单位面积房价有轻微影响

不同楼层的单位面积房价差异不明显

地铁房单价高

学区房单价高

```
dist的P值为:0.0000
roomnum的P值为:0.8794
halls的P值为:0.0480
floor的P值为:0.0029
subway的P值为:0.0000
school的P值为:0.0000
```

支持结论

*注意：由于是随机抽样，每次的结果肯定一样，但是多次的结论是一致的。

n<100 alfa取值[5%,20%]之间
100<n<500 alfa取值[1%,10%]之间
500<n<3000 alfa取值[0.1%,5%]之间

1 描述统计

2 线性模型

OLS Regression Results

Dep. Variable:	price	R-squared:	0.600
Model:	OLS	Adj. R-squared:	0.596
Method:	Least Squares	F-statistic:	178.2
Date:	Sat, 09 Jun 2018	Prob (F-statistic):	2.46e-228
Time:	23:50:25	Log-Likelihood:	-2136.6
No. Observations:	1200	AIC:	4295.
Df Residuals:	1189	BIC:	4351.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.8683	0.161	24.075	0.000	3.553	4.184
dist_丰台	0.1067	0.146	0.732	0.464	-0.179	0.393
dist_朝阳	0.7846	0.151	5.185	0.000	0.488	1.081
dist_东城	2.4079	0.157	15.330	0.000	2.100	2.716
dist_海淀	2.2490	0.155	14.477	0.000	1.944	2.554
dist_西城	3.6773	0.160	23.030	0.000	3.364	3.991
school	1.2051	0.106	11.374	0.000	0.997	1.413
subway	0.5633	0.114	4.940	0.000	0.340	0.787
floor_middle	0.2379	0.101	2.352	0.019	0.039	0.436
floor_low	0.2228	0.103	2.162	0.031	0.021	0.425
AREA	-0.0040	0.001	-4.044	0.000	-0.006	-0.002

1 描述统计

2 线性模型

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3.67万元

学区房比非学区房单位面积房价平均高出1.20万元

地铁房比非地铁房单位面积房价平均高出5633元

高层房屋单位面积房价最低，其次是中层，低层房屋面积房价最高

房屋面积的增加会带来单位面积房价的降低

1 描述统计

2 线性模型

OLS Regression Results

Dep. Variable:	price_ln	R-squared:	0.623
Model:	OLS	Adj. R-squared:	0.620
Method:	Least Squares	F-statistic:	196.5
Date:	Sat, 09 Jun 2018	Prob (F-statistic):	1.16e-243
Time:	23:49:32	Log-Likelihood:	73.489
No. Observations:	1200	AIC:	-125.0
Df Residuals:	1189	BIC:	-68.99
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.4789	0.073	20.173	0.000	1.335	1.623
dist_丰台	0.0358	0.023	1.549	0.122	-0.010	0.081
dist_朝阳	0.1825	0.024	7.611	0.000	0.135	0.230
dist_东城	0.4499	0.025	18.086	0.000	0.401	0.499
dist_海淀	0.4355	0.025	17.705	0.000	0.387	0.484
dist_西城	0.6097	0.025	24.122	0.000	0.560	0.659
school	0.1759	0.017	10.475	0.000	0.143	0.209
subway	0.1211	0.018	6.700	0.000	0.086	0.157
floor_middle	0.0413	0.016	2.577	0.010	0.010	0.073
floor_low	0.0333	0.016	2.038	0.042	0.001	0.065
AREA_ln	-0.0487	0.016	-3.046	0.002	-0.080	-0.017

1 描述统计

2 线性模型

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60.97%

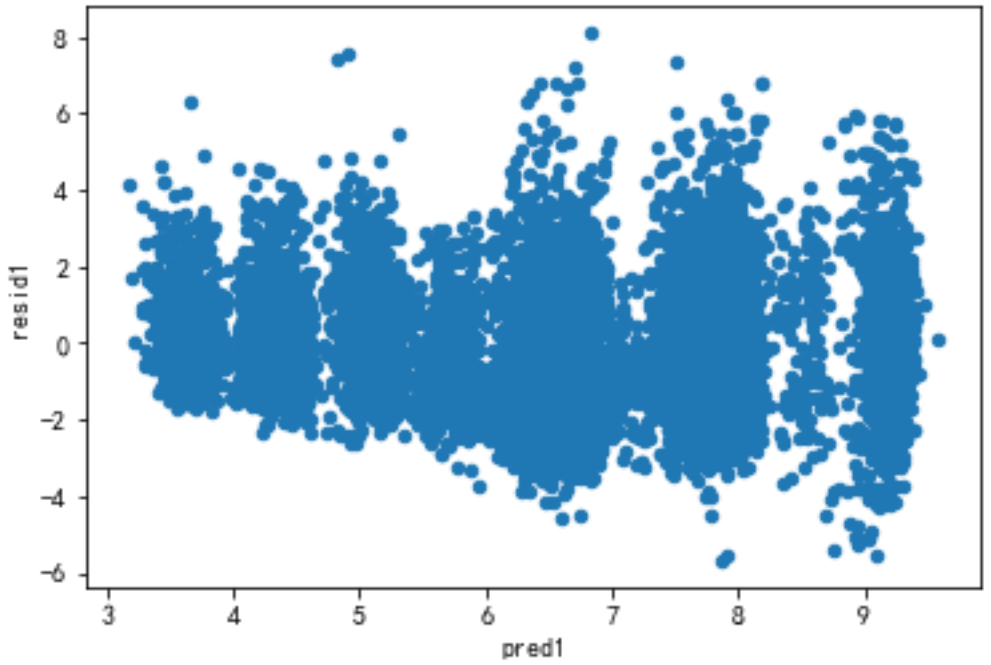
学区房比非学区房单位面积房价平均贵17.59%

地铁房比非地铁房单位面积房价平均贵12.11%

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

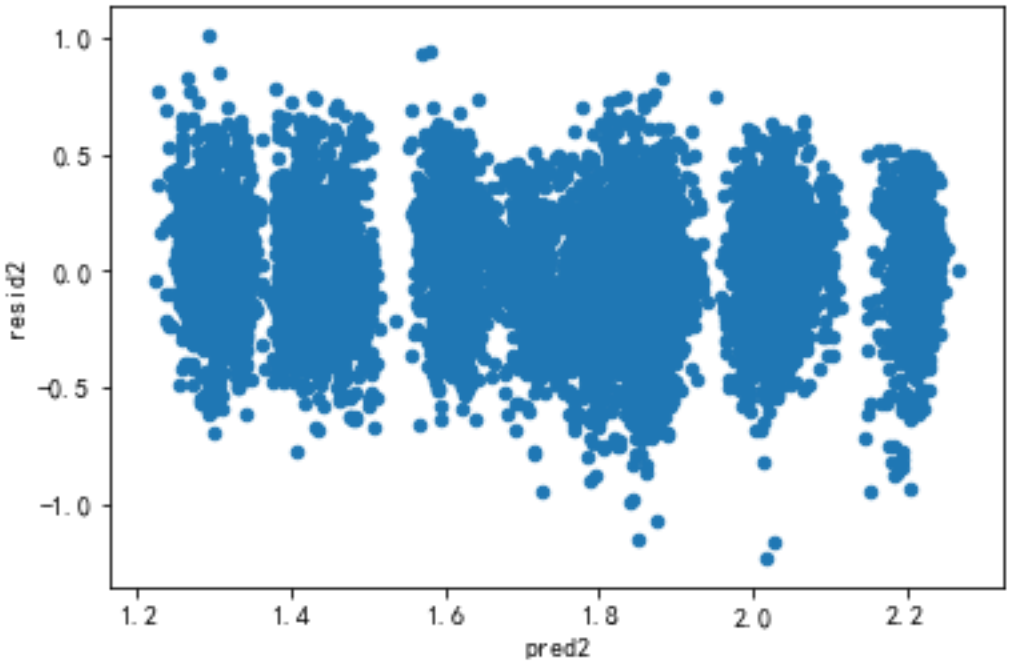
房屋面积的增加会带来单位面积房价的降低

为什么要做对数化处理?



对数化前的残差分析:

随着预测值的增大，残差的波动也随之增大，存在“**异方差**”现象

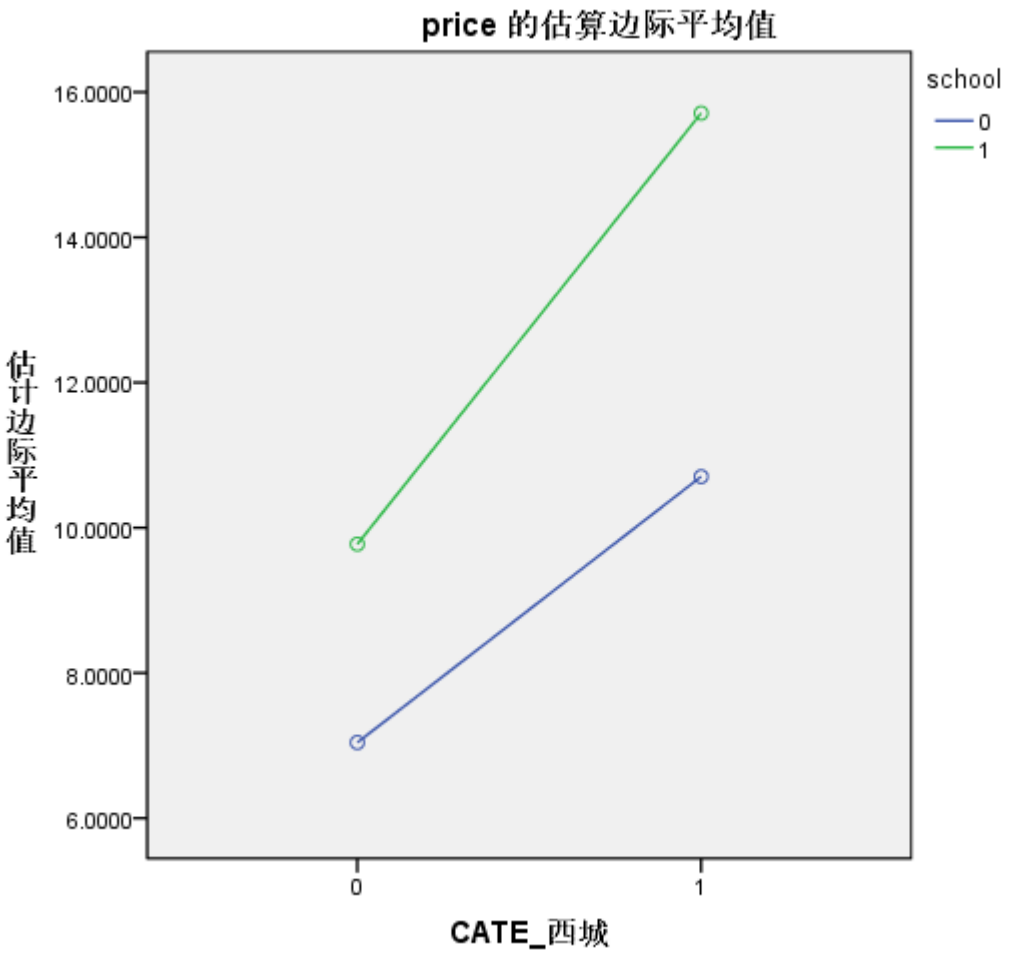


对数化后的残差分析:

“**异方差**”现象得到极大改善

《Python数据科学：全栈技术详解》

什么是交互作用?



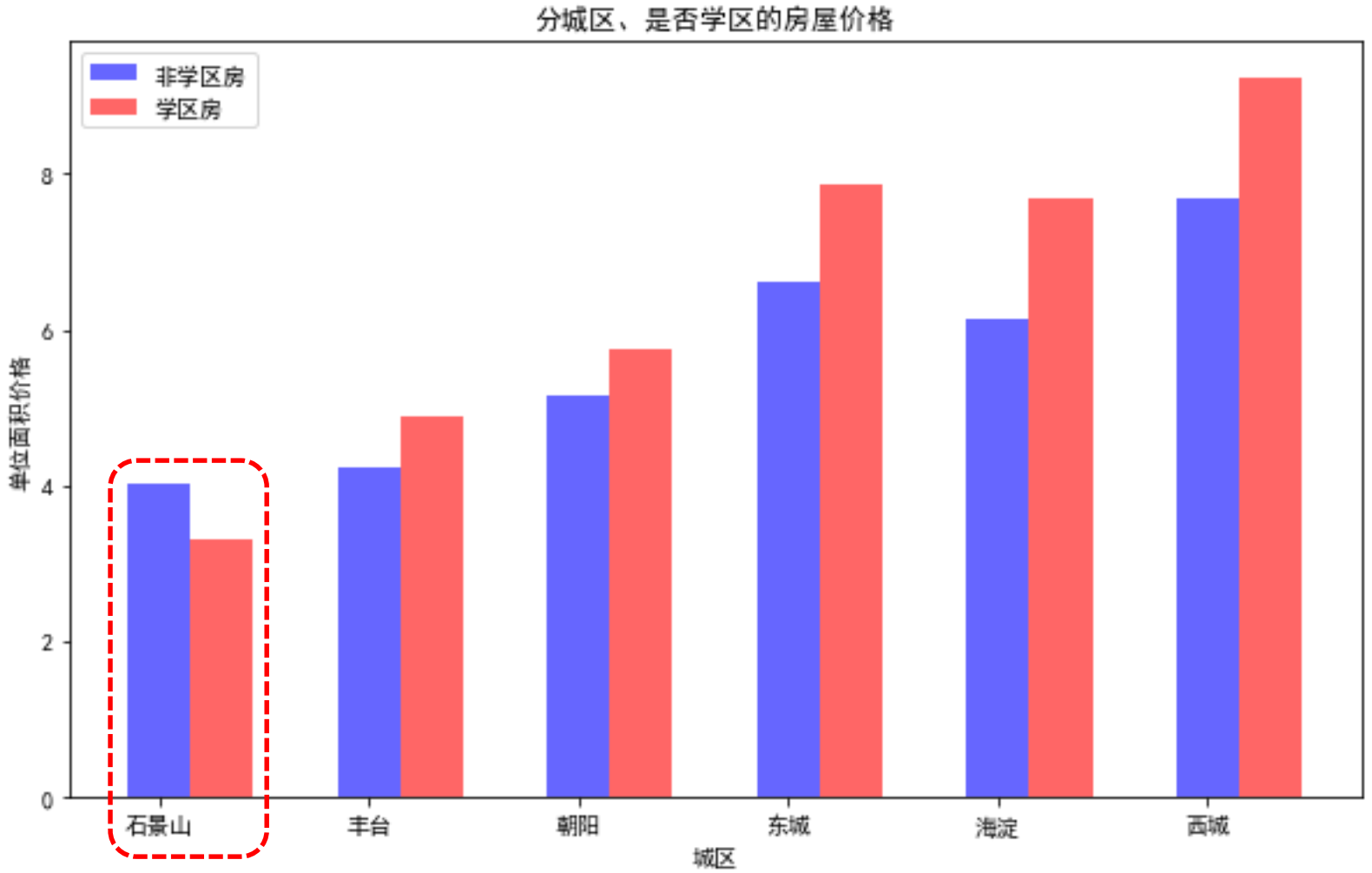
$$y=a+b*x1+c*x2+d*x1x2$$

x1x2被称为交互项

当系数d显著的时候，交互作用存在

《Python数据科学：全栈技术详解》

交互作用能发现更多信息?



6个城区只有**石景山的学区房**

价格比非学区房低

《Python数据科学：全栈技术详解》

1 描述统计

2 线性模型

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3936	0.071	19.640	0.000	1.254	1.533
dist_丰台	0.0190	0.023	0.831	0.406	-0.026	0.064
dist_朝阳	0.1869	0.024	7.689	0.000	0.139	0.235
dist_东城	0.4241	0.027	15.536	0.000	0.371	0.478
dist_海淀	0.4100	0.028	14.654	0.000	0.355	0.465
dist_西城	0.6351	0.029	21.762	0.000	0.578	0.692
school	-0.1494	0.224	-0.666	0.505	-0.589	0.291
dist_丰台:school	0.2733	0.243	1.127	0.260	-0.203	0.749
dist_朝阳:school	0.2588	0.228	1.135	0.257	-0.189	0.706
dist_东城:school	0.3744	0.227	1.653	0.099	-0.070	0.819
dist_海淀:school	0.3550	0.227	1.567	0.117	-0.090	0.800
dist_西城:school	0.2554	0.227	1.127	0.260	-0.189	0.700
subway	0.1099	0.018	6.026	0.000	0.074	0.146
floor_middle	0.0433	0.016	2.768	0.006	0.013	0.074
floor_low	0.0370	0.016	2.292	0.022	0.005	0.069
AREA_ln	-0.0262	0.016	-1.690	0.091	-0.057	0.004

1 描述统计

2 线性模型

- 假想：一家三口，为了孩子在**东城区上学**，想买一套**邻近地铁**的**两居**室，面积是**70平方米**，**中层**楼层

- 根据交互模型，预测的单位面积房价是**7.86万元/平方米**，总价**550.37万元**

1 描述统计

2 线性模型

- 影响北京市二手房单位面积房价的主要因素
  - ✓ 区位因素：城区、地铁、学区
  - ✓ 内部因素：房屋面积、楼层
- 使用对数线性模型，可以克服数据中存在的异方差
- 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

Thanks