带惩罚项的回归方法

Sim

- 变量处理
- 数据的探索性分析
- 模型预测

变量处理

A	B	C	D	E	F	G	H	I
block	type	size	region	height	direction	price	built_date	price_unit
梅园六街坊	2室0厅	47.72	浦东	低区/6层	朝南	500	1992年建	104777
碧云新天地（一期）	3室2厅	108.93	浦东	低区/6层	朝南	735	2002年建	67474
博山小区	1室1厅	43.79	浦东	中区/6层	朝南	260	1988年建	59374
金桥新村四街坊（博兴路98(	1室1厅	41.66	浦东	中区/6层	朝南北	280	1997年建	67210
博山小区	1室0厅	39.77	浦东	高区/6层	朝南	235	1987年建	59089
伟莱家园	2室2厅	100.15	浦东	中区/6层	朝南北	515	2002年建	51422
羽北小区	2室2厅	69.88	浦东	低区/6层	朝南	560	1994年建	80137
证大家园（公寓）	3室2厅	122.75	浦东	低区/11层	朝南北	785	2002年建	63951
上南十村	1室1厅	40.17	浦东	低区/6层	朝南	240	1992年建	59746
鹏欣家园	1室1厅	59.42	浦东	中区/6层	朝南	410	1998年建	69000
恒大华城新华苑	2室1厅	68.9	浦东	高区/6层	朝南北	460	1997年建	66763
金橘新苑	1室1厅	61.65	浦东	低区/18层	朝南	370	2006年建	60016
陆家嘴花园（一期）	3室2厅	156.45	浦东	中区/11层	朝南北	1400	1999年建	89485

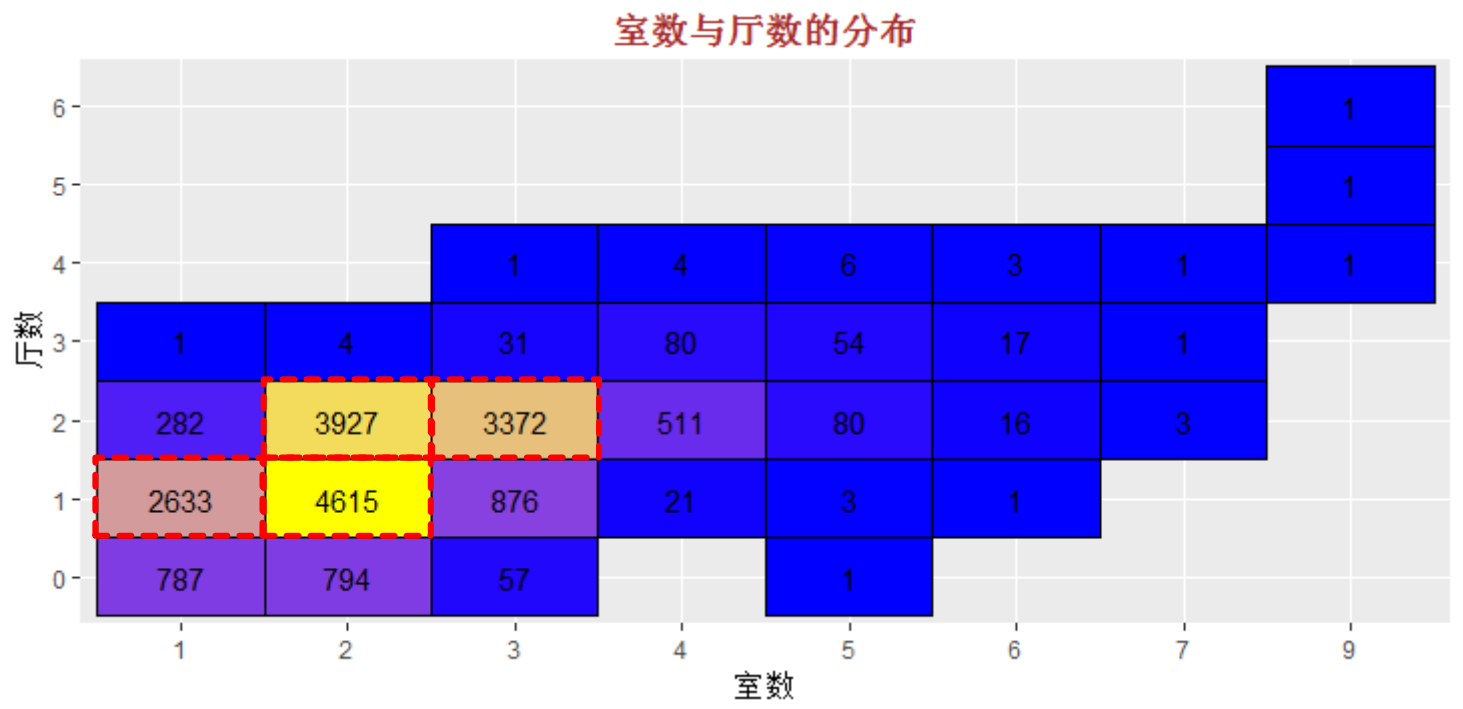
变量处理

A	B	C	D	E	F	G	H	I
size	region	direction	price	price_unit	rooms	halls	height_type	diff
47.72	浦东	南	500	104777	2	0	低区	25
108.93	浦东	南	735	67474	3	2	低区	15
43.79	浦东	南	260	59374	1	1	中区	29
41.66	浦东	南	280	67210	1	1	中区	20
39.77	浦东	南	235	59089	1	0	高区	30
100.15	浦东	南	515	51422	2	2	中区	15
69.88	浦东	南	560	80137	2	2	低区	23
122.75	浦东	南	785	63951	3	2	低区	15
40.17	浦东	南	240	59746	1	1	低区	25
59.42	浦东	南	410	69000	1	1	中区	19
68.9	浦东	南	460	66763	2	1	高区	20
61.65	浦东	南	370	60016	1	1	低区	11
156.45	浦东	南	1400	89485	3	2	中区	18

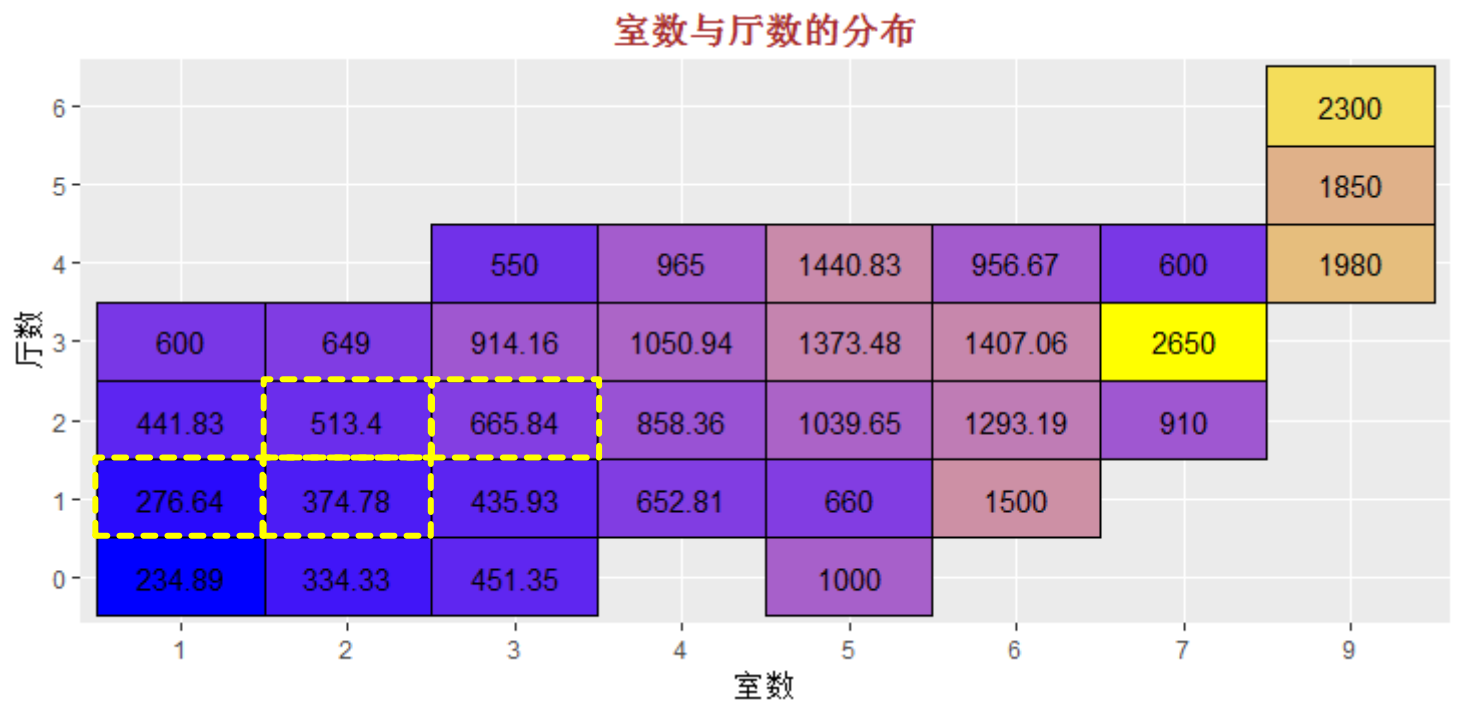
数据的探索性分析

1、从房型来看，什么样的格局比较常见？房价分布如何？
2、各行政区域下，二手房单价是否存在差异？
3、二手房价格是否为正态分布？
4、二手房楼层高低与价格之间是否存在差异？
5、二手房朝向对价格的影响是怎样的？
6、建筑时长与房价是否存在关系？

房型分析

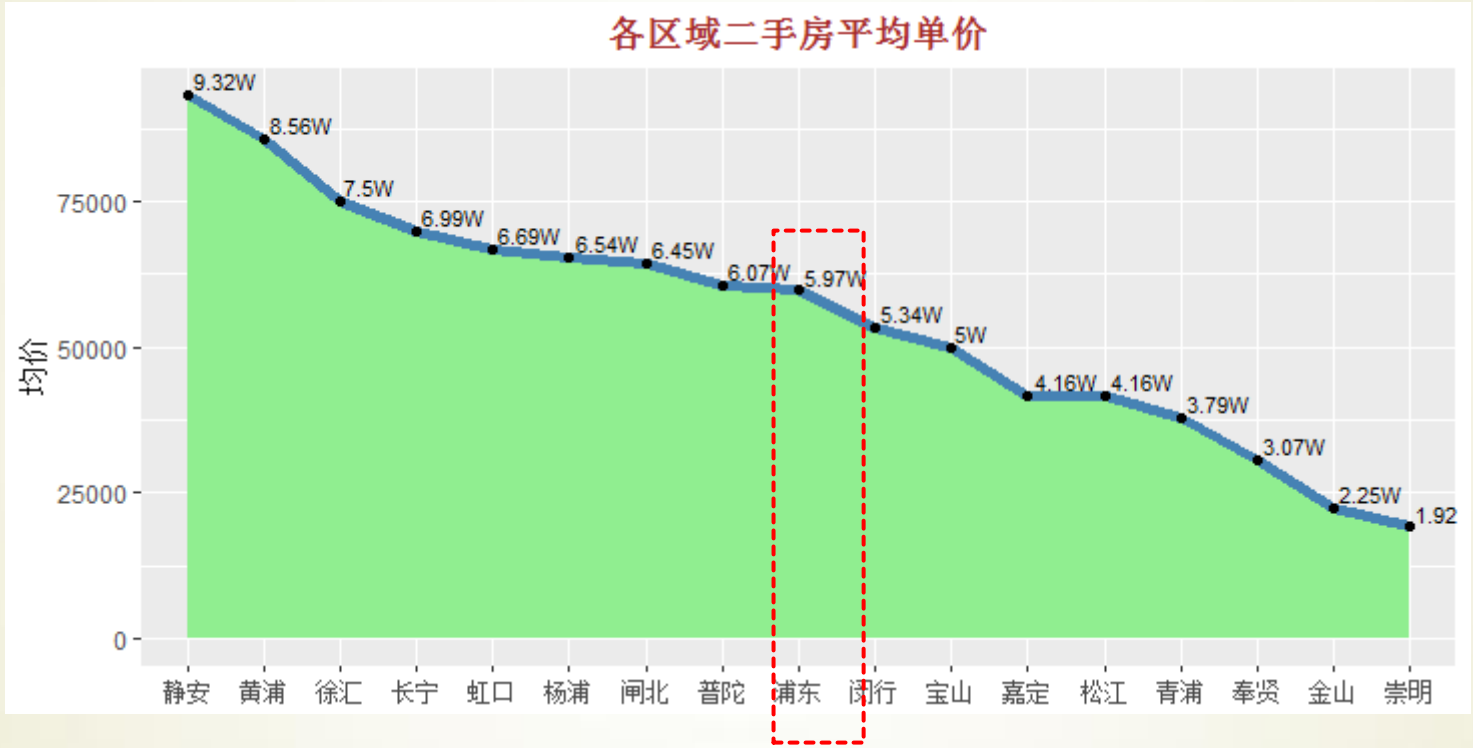

室数与厅数的分布
厅数
室数
6
5
4
3
2
1
0
1
2
3
4
5
6
7
9
1
1
1
4
6
3
1
1
4
31
80
54
17
1
282
3927
3372
511
80
16
3
2633
4615
876
21
3
1
787
794
57
1

房型分析

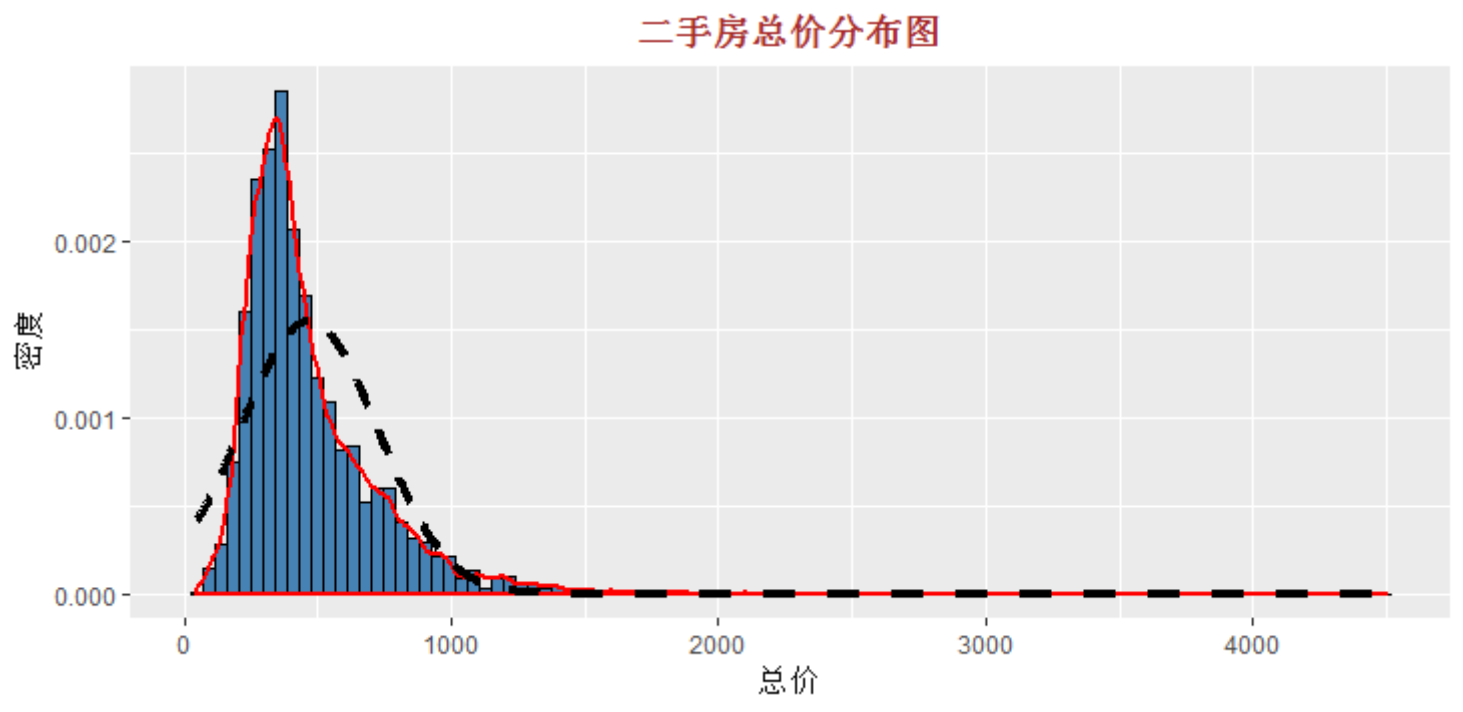


A	B	C	D	E
室数	厅数	二手房套数	平均总价	平均面积
1	1	2633	276.64	47.21
2	1	4615	374.78	66.53
2	2	3927	513.40	92.24
3	2	3372	665.84	124.41

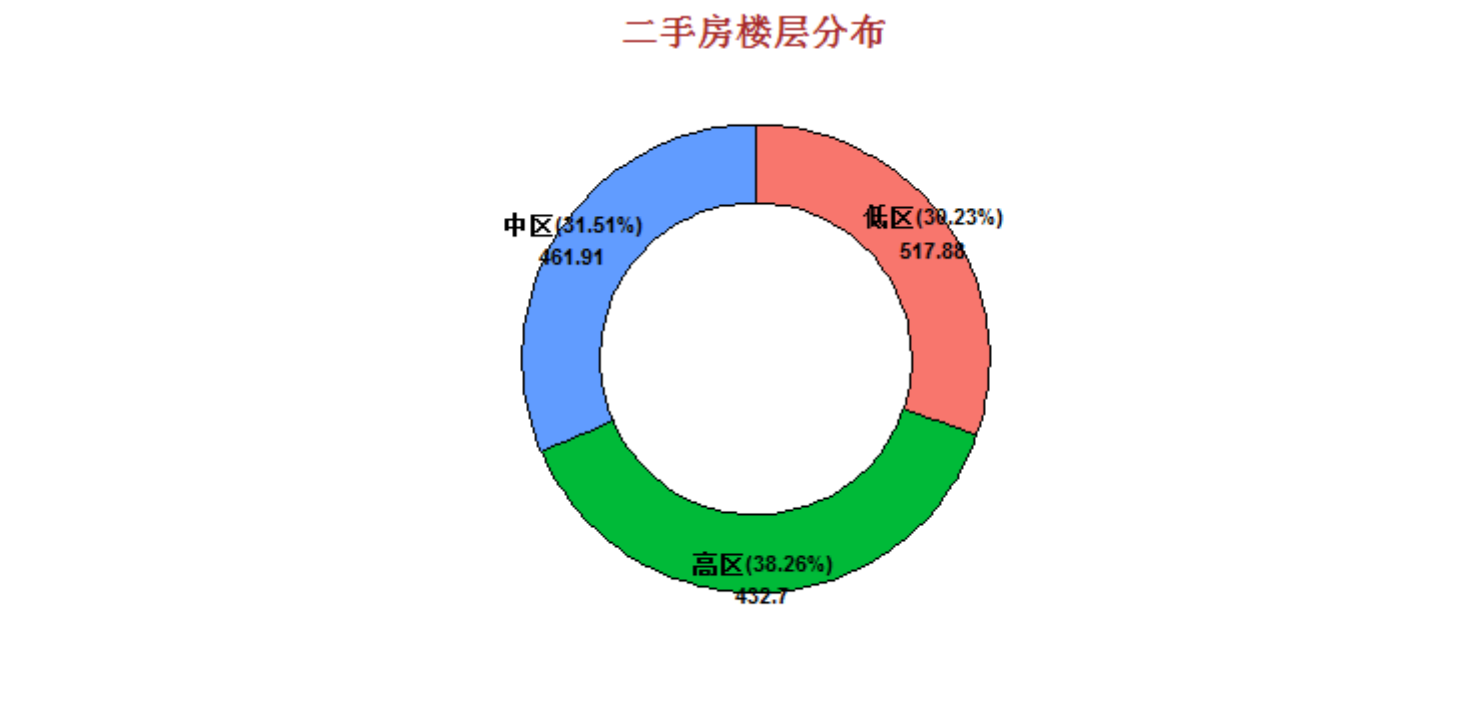
各区域单价分布


各区域二手房平均单价
均价
75000
50000
25000
0
9.32W
8.56W
7.5W
6.99W
6.69W
6.54W
6.45W
6.07W
5.97W
5.34W
5W
4.16W
4.16W
3.79W
3.07W
2.25W
1.92
静安
黄浦
徐汇
长宁
虹口
杨浦
闸北
普陀
浦东
闵行
宝山
嘉定
松江
青浦
奉贤
金山
崇明

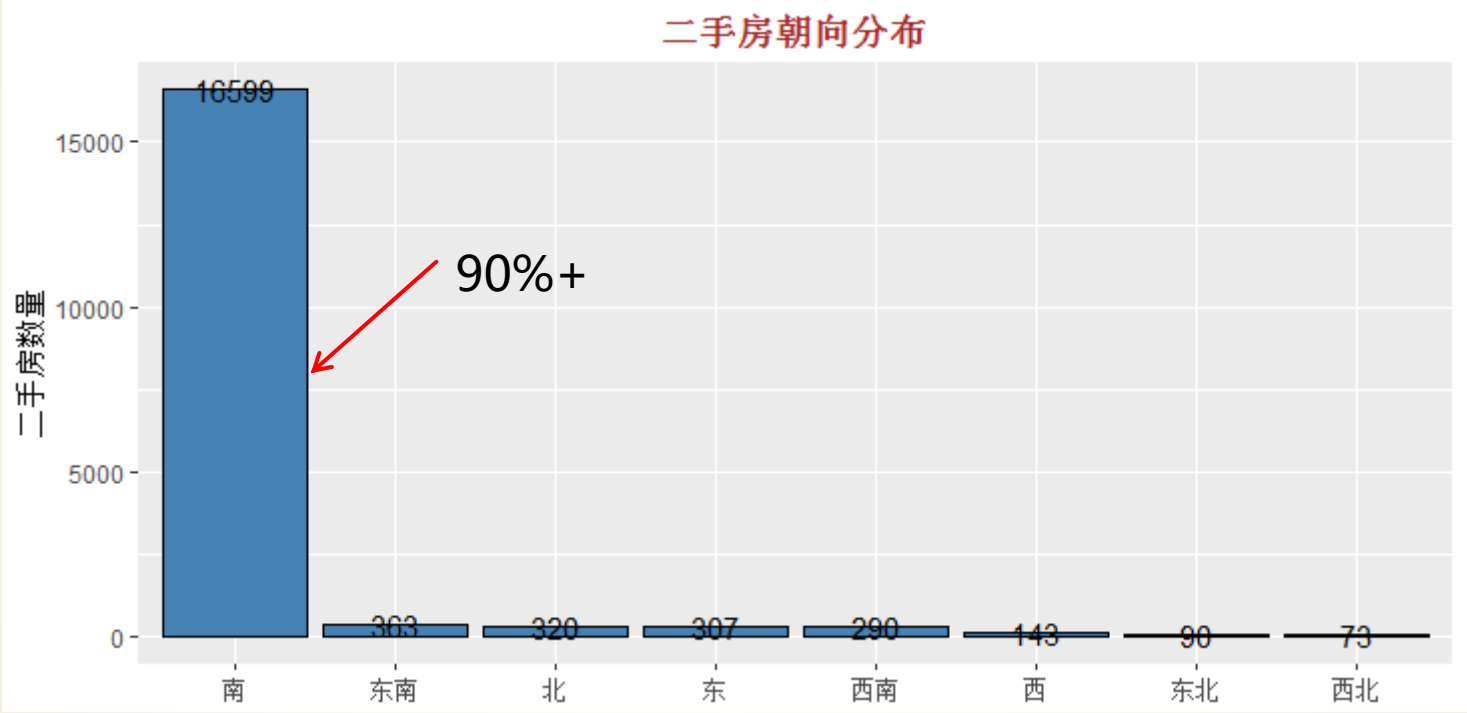
价格分布



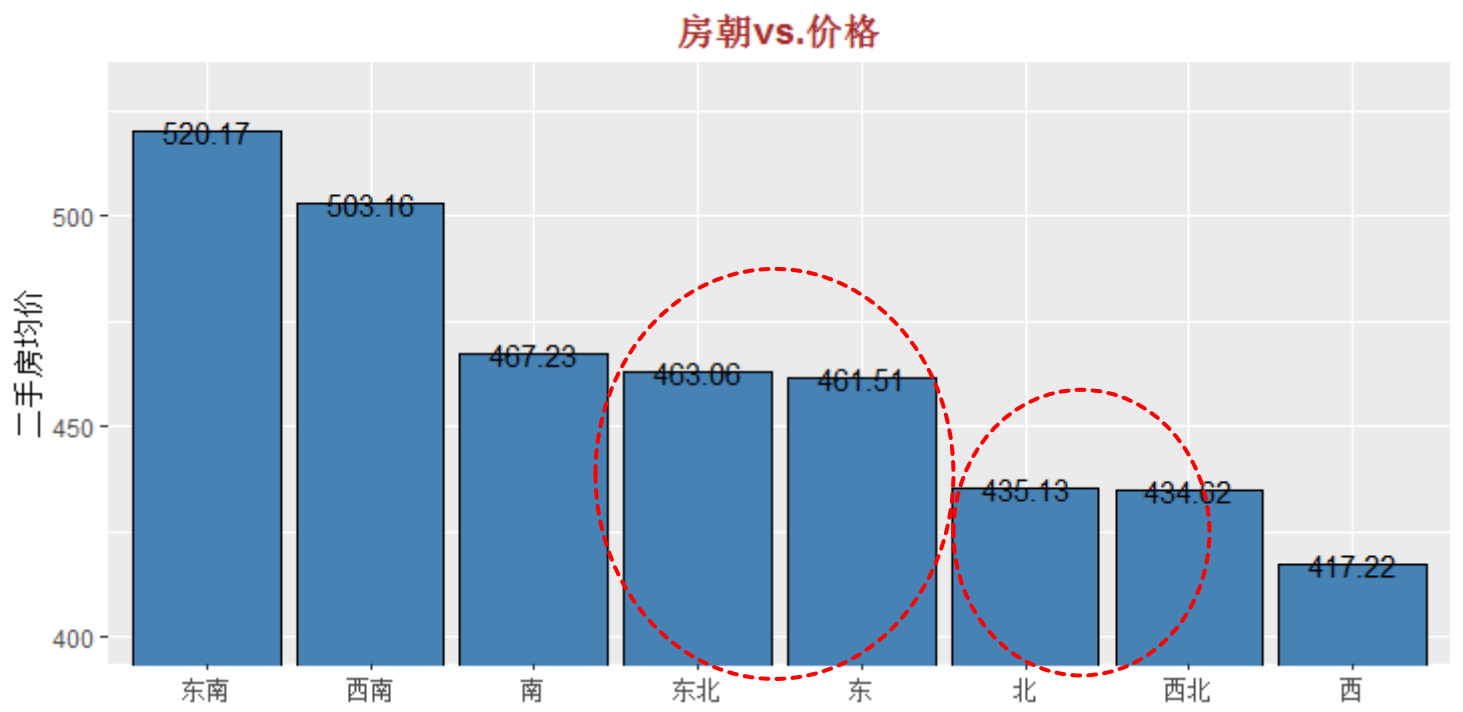
楼层分析



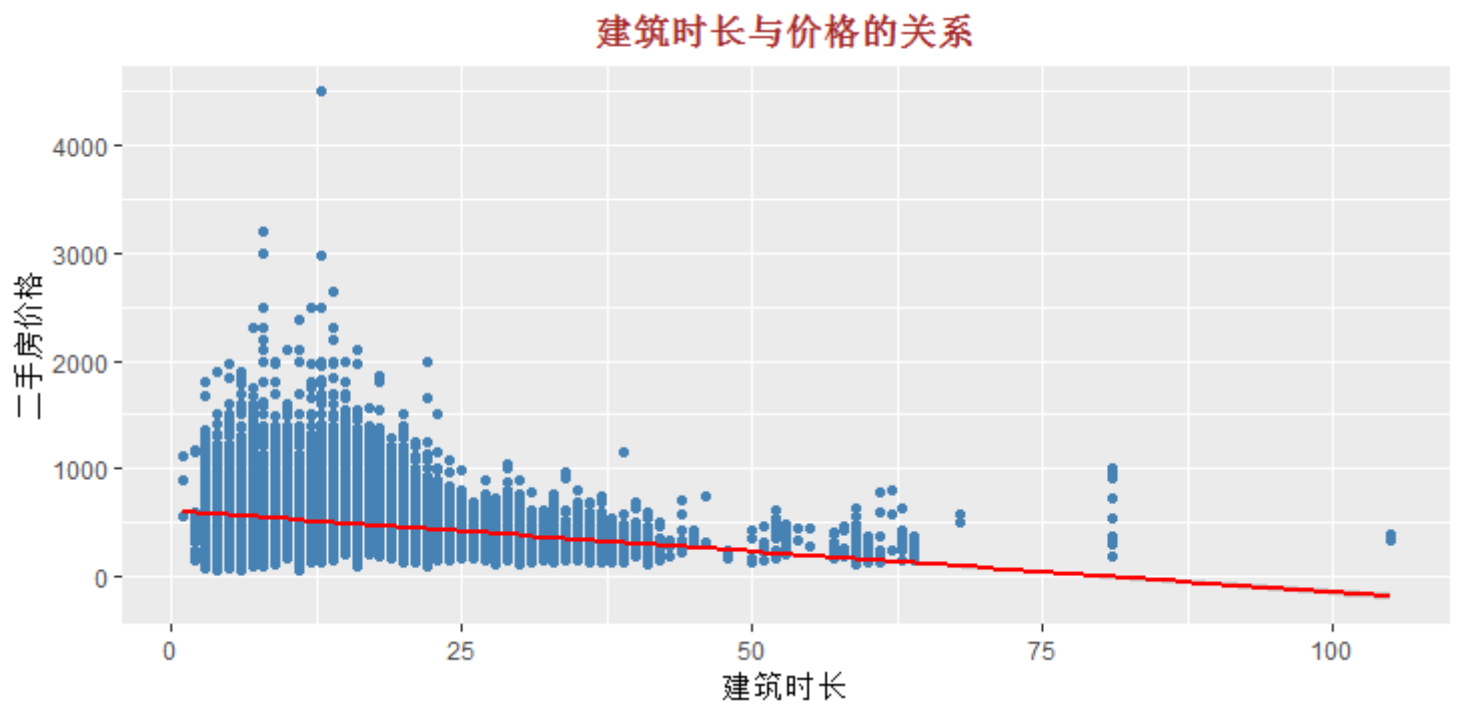
朝向分析



朝向分析



建筑时长与价格的关系



模型预测—线性回归

$$\underset{\arg\min}{J(\beta)} = \frac{1}{2}\sum(y - X\beta)^2 = \frac{1}{2}\sum\varepsilon^2$$

$$\underset{\arg\min}{J(\beta)} = \frac{\partial J(\beta)}{\partial \beta}$$

$$= \frac{1}{2}(0 - X'y - X'y + 2X'X\beta) = 0$$

$$\therefore \ 2X'X\beta = 2X'y$$

$$\beta = (X'X)^{-1}X'y$$

模型预测—线性回归

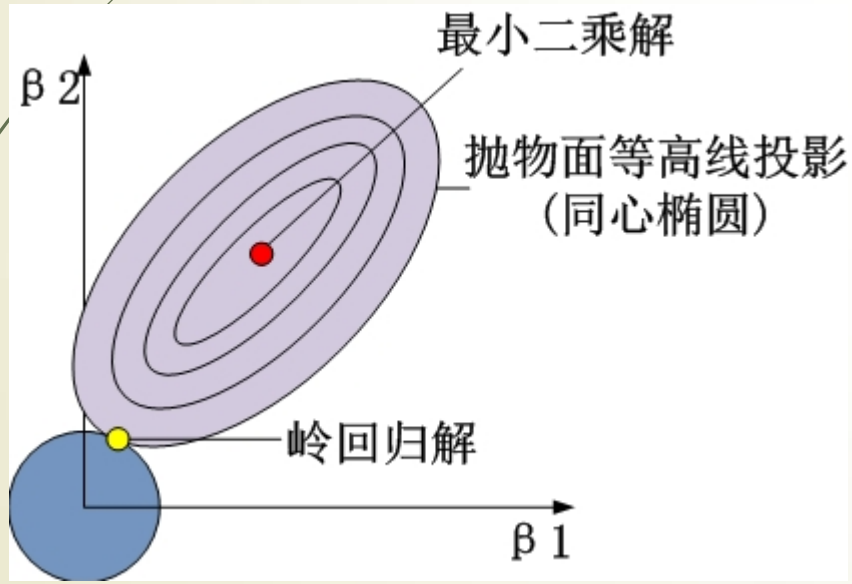
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.050e+01  1.663e-01  63.144  < 2e-16 ***
size             8.809e-02  8.065e-04 109.225  < 2e-16 ***
region闵行      -1.147e+00  8.682e-02 -13.206  < 2e-16 ***
region宝山      -1.951e+00  8.463e-02 -23.055  < 2e-16 ***
region徐汇       2.614e+00  9.288e-02  28.144  < 2e-16 ***
region普陀       4.027e-01  9.091e-02   4.429 9.53e-06 ***
region杨浦       1.255e+00  8.591e-02  14.605  < 2e-16 ***
region长宁       1.836e+00  9.672e-02  18.987  < 2e-16 ***
region松江      -4.307e+00  8.884e-02 -48.478  < 2e-16 ***
region嘉定      -4.087e+00  8.649e-02 -47.256  < 2e-16 ***
region黄浦       4.447e+00  1.204e-01  36.920  < 2e-16 ***
region虹口       1.687e+00  9.368e-02  18.007  < 2e-16 ***
region闸北       9.881e-01  8.797e-02  11.233  < 2e-16 ***
region静安       4.135e+00  7.133e-01   5.797 6.89e-09 ***
region青浦      -5.686e+00  1.043e-01 -54.512  < 2e-16 ***
region奉贤      -6.907e+00  1.035e-01 -66.745  < 2e-16 ***
region金山      -1.031e+01  7.129e-01 -14.455  < 2e-16 ***
region崇明      -1.040e+01  1.067e+00  -9.753  < 2e-16 ***
direction东      8.671e-01  1.887e-01   4.595 4.35e-06 ***
direction东北    8.570e-01  2.758e-01   3.108  0.00189 **
direction东南    1.967e+00  1.836e-01  10.715  < 2e-16 ***
direction南      1.984e+00  1.344e-01  14.761  < 2e-16 ***
direction西      3.961e-01  2.347e-01   1.688  0.09149 .
direction西北    1.248e+00  3.120e-01   4.001 6.34e-05 ***
direction西南    1.994e+00  1.947e-01  10.241  < 2e-16 ***
rooms            6.305e-01  3.503e-02  18.000  < 2e-16 ***
halls            1.299e+00  4.020e-02  32.320  < 2e-16 ***
height_type高区 -3.439e-01  4.372e-02  -7.866 3.93e-15 ***
height_type中区  1.236e-01  4.535e-02   2.725  0.00643 **
diff            -6.096e-02  2.440e-03 -24.982  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.13 on 14518 degrees of freedom
Multiple R-squared:  0.841,     Adjusted R-squared:  0.8407
F-statistic:  2649 on 29 and 14518 DF,  p-value: < 2.2e-16
```

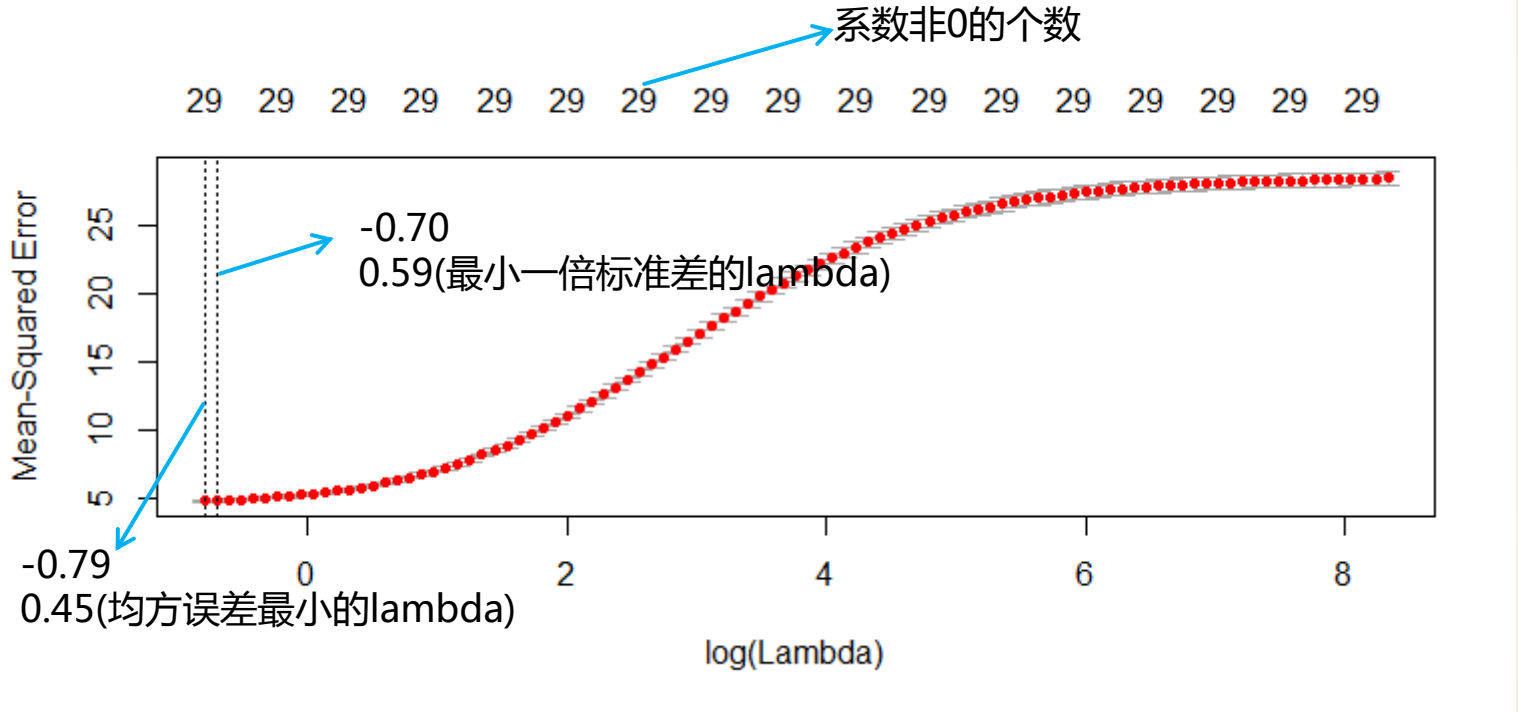
RMSE=2.0619

模型预测—岭回归

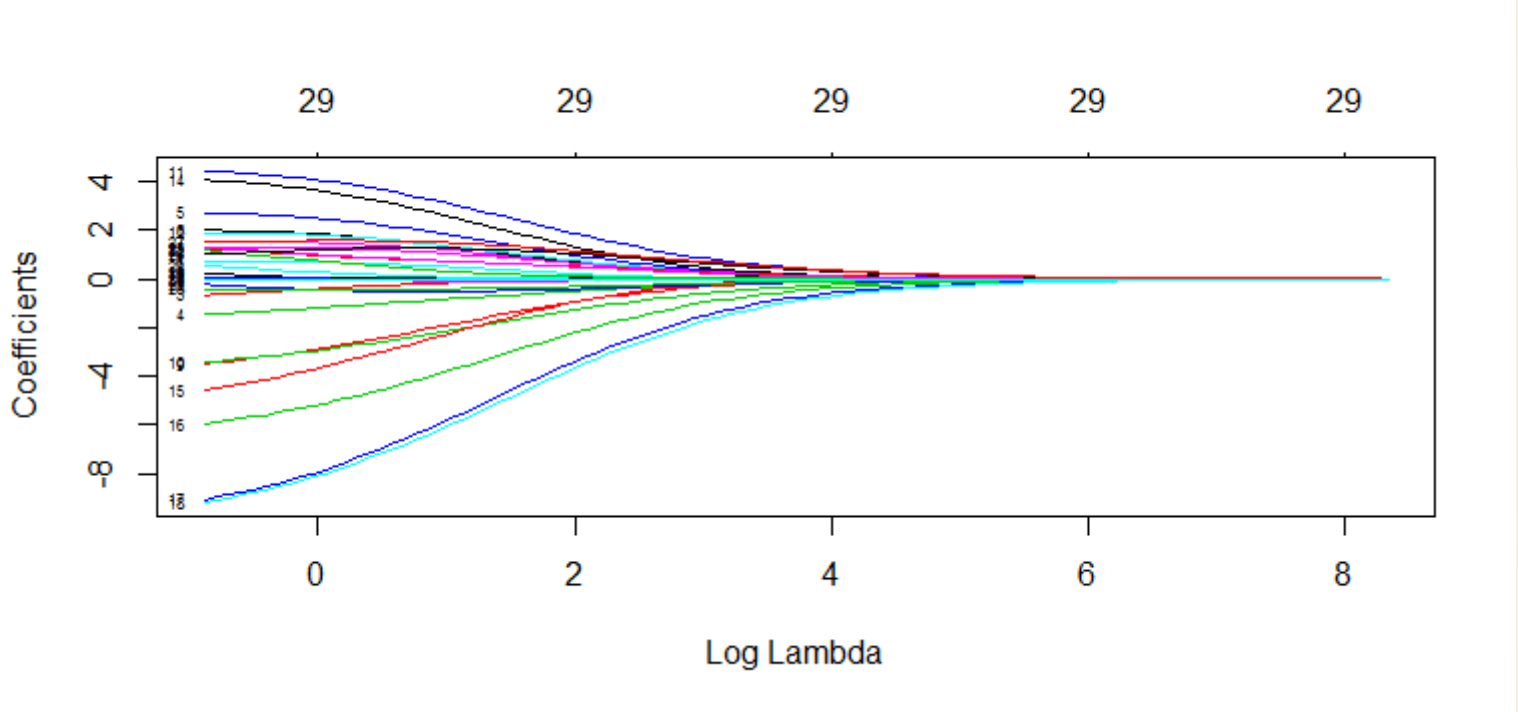
$$\hat{\beta}^{ridge} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$



模型预测—岭回归


系数非0的个数
29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29 29
Mean-Squared Error
5 10 15 20 25
-0.70
0.59(最小一倍标准差的lambda)
-0.79
0.45(均方误差最小的lambda)
0 2 4 6 8
log(Lambda)

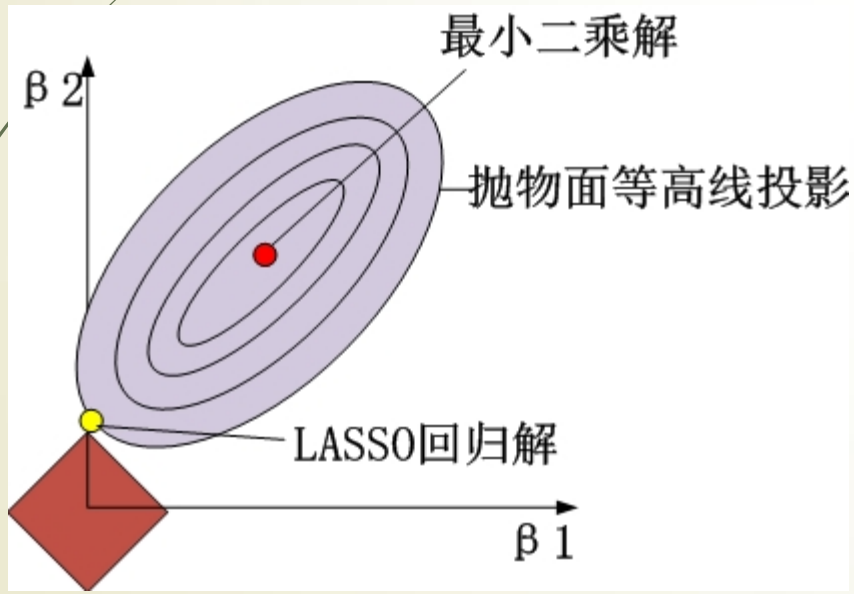
模型预测—岭回归



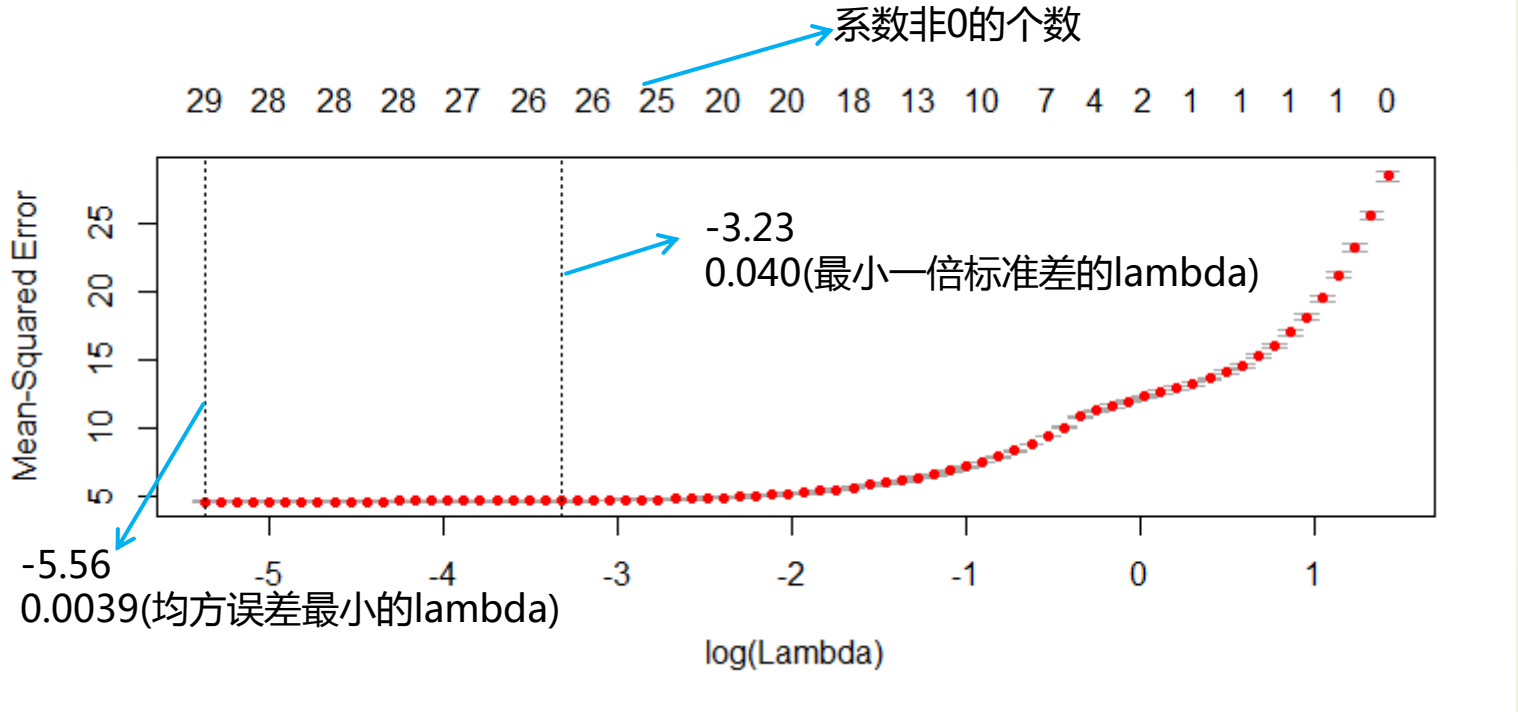
RMSE=2.1190

模型预测—LASSO回归

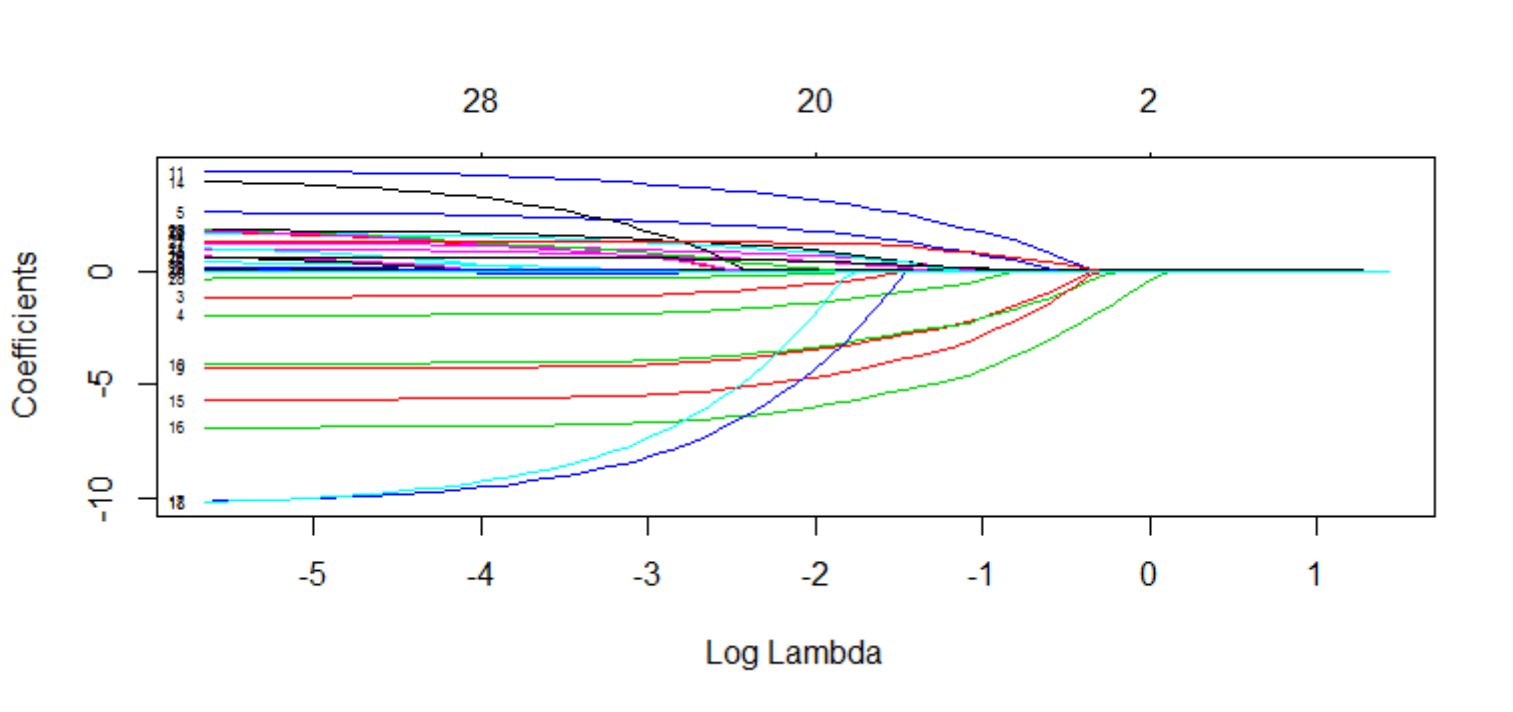
$$\hat{\beta}^{ridge} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\}$$



模型预测—LASSO回归



模型预测—LASSO回归



RMSE=2.0616

各模型系数对照

	lm_coef	ridge_coef	lasso_coef
(Intercept)	10.50033917	11.09842246	10.71454685
size	0.08808992	0.07180904	0.08802539
region闵行	-1.14659260	-0.63383841	-1.14141601
region宝山	-1.95117063	-1.43330250	-1.94598934
region徐汇	2.61390292	2.69480493	2.58196789
region普陀	0.40266320	0.68915528	0.37327038
region杨浦	1.25467750	1.52659208	1.22849049
region长宁	1.83636630	1.98652324	1.80181519
region松江	-4.30674108	-3.43980300	-4.29794352
region嘉定	-4.08743082	-3.37760461	-4.07801087
region黄浦	4.44698042	4.40225153	4.40196941
region虹口	1.68694568	1.89398292	1.65240116
region闸北	0.98810571	1.28459595	0.96217675
region静安	4.13489542	4.02838836	3.93670882
region青浦	-5.68599623	-4.48811462	-5.67308171
region奉贤	-6.90703505	-5.90125676	-6.89208515
region金山	-10.30584394	-8.96884380	-10.14439974
region崇明	-10.40478831	-9.08583747	-10.17071735
direction东	0.86711305	0.21648206	0.63028115
direction东北	0.85701239	0.21273513	0.59895149
direction东南	1.96734321	1.23290395	1.73135762
direction南	1.98399508	1.11560762	1.76844607
direction西	0.39613132	-0.24880620	0.14981291
direction西北	1.24834998	0.51531255	0.98445570
direction西南	1.99387285	1.26368039	1.76157567
rooms	0.63047330	1.01906586	0.62662658
halls	1.29917205	1.51261291	1.30014016
height_type高区	-0.34385797	-0.41924656	-0.33818853
height_type中区	0.12357643	0.04566570	0.11712209
diff	-0.06096338	-0.05454021	-0.05977545

谢谢