DOI:10.19699/j.cnki.issn2096-0298.2017.04.068

大数据时代房地产价格预测模型的研究

盐城工学院管理学院　徐颖
盐城工学院数理学院　黄素珍

摘　要：大数据时代推动了房地产市场思维方式、管理方式和商业模式的变革。本文利用百度搜索数据，以江苏省南京市为例建立了新建住宅销售价格指数的多元线性回归、完全二次多项式回归和逐步回归模型，仿真结果表明，逐步回归模型预测精度高，稳定性好。

关键词：大数据　百度指数　新房价格　逐步回归模型
中图分类号：F726　　文献标识码：A　　文章编号：2096-0298(2017)02(a)-134-04

1 引言

2016年12月中旬，中央经济工作会议提出，要坚持“房子是用来住的，不是用来炒的”的定位，要求回归住房居住属性。住房价格高居民承受能力低，居住水平和居住质量会由此下降；反之住房价格水平低，能增强居民的购房能力，相应提高居民的居住水平和居住质量。因此，住房价格的高低成为关系到居民切身利益的重大经济问题和社会问题。

影响房地产价格的因素有很多，如果把所有可能影响的因素全部考虑进去，所建立起来的回归方程却不一定是最好的。首先由于自变量过多，使用不便，而且在回归方程中引入无意义的量，会使误差方差的估计值增大，降低预测的精确性及回归方程的稳定性。另一方面，通常希望回归方程中包含的变量尽可能多一些，特别是对房价有显著影响的自变量，这样会减小误差方差的估计值，从而提高预测的精度。

本文尝试利用百度引擎提供的影响房价的因素搜索指数，建立新建住宅销售价格指数和二手住宅销售价格指数的多元线性回归、完全二次多项式回归和逐步回归模型，仿真结果表明，逐步回归模型预测精度高，稳定性好。

2 房价预测模型构建

一般来说，房价由三部分构成：土地成本、建筑成本(建安成本，税费成本，营销成本等)和开发商的预期利润。根据这些成本费用，我们基本上就可估算出一个项目的成本价格是多少，最终售价大约会在什么范围。房价构成模型可以表示为下式：

$$P = P_0 + C + D \tag{1}$$

式(1)中P表示最终售价，P_0表示楼面地价，C表示建筑成本价，D表示开发商利润。设地价为P_X元/平方米，土地容积率为$\tau = \frac{P_X}{P_0}$，预期利润率为$d = \frac{D}{P_0 + C}$，因此房价计算公式可以表示为

$$P = (1+d)(\frac{P_X}{\tau} + C) \tag{2}$$

则

$$\ln P = \ln(1+d) + \ln(\frac{P_X}{\tau} + C) \tag{3}$$

设$Z_1 = 1 + d$，$Z_2 = \frac{P_X}{\tau} + C$，得

$$\ln P = \ln Z_1 + \ln Z_2 \tag{4}$$

设某月房价为P，下月房价为P'，相应地Z_1变化为Z_1'，Z_2变化为Z_2'，则$\ln P' = \ln Z_1' + \ln Z_2'$，因此有

$$\ln\left(1+\frac{P'-P}{P}\right) = \ln\left(1+\frac{Z_1'-Z_1}{Z_1}\right) + \ln\left(1+\frac{Z_2'-Z_2}{Z_2}\right)$$

设$Y = \ln\left(1+\frac{P'-P}{P}\right)$，$X_1 = \ln\left(1+\frac{Z_1'-Z_1}{Z_1}\right)$，$X_2 = \ln\left(1+\frac{Z_2'-Z_2}{Z_2}\right)$，则

$$Y = X_1 + X_2 \tag{5}$$

式(4)和(5)表明，房价P与各项因子Z_1、Z_2之间并非线性关系，但是经过处理后的Y与各项因子X_1、X_2之间是线性关系，其中X_1代表与政策有关的因素，X_2代表与成本有关的因素，则基于房价构成模型的多元回归预测模型如下式：

$$Y = a_0 + a_1X_1 + a_2X_2 + \varepsilon(\text{其中}\varepsilon\text{为误差项}) \tag{6}$$

然而，影响房价的因素除了政策因素、成本因素以外，还有其他更加复杂的因素，包括经济因素、社会因素、行政因素、周边环境因素以及房屋内在因素等。因此将上述模型做如下改进：

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m + \varepsilon \tag{7}$$

式(7)中$X_1, X_2, \ldots, X_m$，表示影响Y的m个因素，a_0表示常数项，$a_1, a_2, \ldots, a_m$分别表示各种因素的权重。对于第i个观测值：

$$\begin{cases} Y_1 = a_0 + a_1X_{11} + a_2X_{12} + \cdots + a_mX_{1m} + \varepsilon_1 \\ Y_2 = a_0 + a_1X_{21} + a_2X_{22} + \cdots + a_mX_{2m} + \varepsilon_2 \\ \cdots\cdots \\ Y_n = a_0 + a_1X_{n1} + a_2X_{n2} + \cdots + a_mX_{nm} + \varepsilon_n \end{cases}$$

(其中$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$　$i = 1,2,\ldots,n$)

3 数据的获取与处理

因变量分别是新建商品住宅销售价格指数。采用2014年1月～2016年8月共32个月的月度同比数据，来源于国家统计局网站。

解释变量是与新房价格相关的某些关键词的网络搜索指数。根据董倩等在文中的方法[2]，对于新房价格预测，最终选取了8个关键词，分别是房价走势、房源、装修、房产网、公积金、房贷利率、新楼盘、保障房。

为了与因变量保持一致，我们对所有关键词的搜索指数做如下处理：首先将根据日搜索指数计算月度平均搜索指数，然后将月度平均搜索指数转换为同比数据，最终得到所有关键词从2014年1月到2016年8月的月度同比数据。

最后将因变量和解释变量的月度同比数据先加1再取对数。

4 仿真分析

本文采用matlab技术，以南京市为例建立和分析新房价格预测模型。

4.1 可视化相关性分析

对表1中的数据按照上述方法进行数据处理，并在matlab中读入处理后的数据，建立因变量矩阵和解释变量矩阵，计算变量间的相关系数矩阵R和线性相关性检验的P值矩阵P，绘制相关系数矩阵图如图1。

$$R=\begin{bmatrix}
1.0000 & 0.7779 & 0.5515 & 0.4755 & 0.3086 & 0.2345 & 0.1859 & -0.2481 & 0.5858\\
0.7779 & 1.0000 & 0.6359 & 0.6419 & 0.5605 & 0.6692 & 0.6005 & -0.0027 & 0.8110\\
0.5515 & 0.6359 & 1.0000 & 0.6730 & 0.7585 & 0.4984 & 0.4211 & 0.3052 & 0.5900\\
0.4755 & 0.6419 & 0.6730 & 1.0000 & 0.8948 & 0.8283 & 0.8261 & 0.0963 & 0.8521\\
0.3086 & 0.5605 & 0.7585 & 0.8948 & 1.0000 & 0.8228 & 0.8032 & 0.3326 & 0.7593\\
0.2345 & 0.6692 & 0.4984 & 0.8283 & 0.8228 & 1.0000 & 0.9677 & 0.2088 & 0.8267\\
0.1859 & 0.6005 & 0.4211 & 0.8261 & 0.8032 & 0.9677 & 1.0000 & 0.1439 & 0.8212\\
-0.2481 & -0.0027 & 0.3052 & 0.0963 & 0.3326 & 0.2088 & 0.1439 & 1.0000 & 0.1312\\
0.5858 & 0.8110 & 0.5900 & 0.8521 & 0.7593 & 0.8267 & 0.8212 & 0.1312 & 1.0000
\end{bmatrix}$$

$$P=\begin{bmatrix}
1.0000 & 0.0000 & 0.0011 & 0.0060 & 0.0857 & 0.1965 & 0.3084 & 0.1710 & 0.0004\\
0.0000 & 1.0000 & 0.0001 & 0.0001 & 0.0008 & 0.0000 & 0.0003 & 0.9882 & 0.0000\\
0.0011 & 0.0001 & 1.0000 & 0.0000 & 0.0000 & 0.0037 & 0.0164 & 0.0894 & 0.0004\\
0.0060 & 0.0001 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.6001 & 0.0000\\
0.0857 & 0.0008 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.0000 & 0.0629 & 0.0000\\
0.1965 & 0.0000 & 0.0037 & 0.0000 & 0.0000 & 1.0000 & 0.0000 & 0.2515 & 0.0000\\
0.3084 & 0.0003 & 0.0164 & 0.0000 & 0.0000 & 0.0000 & 1.0000 & 0.4319 & 0.0000\\
0.1710 & 0.9882 & 0.0894 & 0.6001 & 0.0629 & 0.2515 & 0.4319 & 1.0000 & 0.4742\\
0.0004 & 0.0000 & 0.0004 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4742 & 1.0000
\end{bmatrix}$$

图1 相关系数矩阵图

表1 江苏省南京市新房销售价格指数和关键词月度搜索指数同比涨跌数据

时间	同比涨跌	房价走势	房源	装修	房产网	公积金	房贷利率	新楼盘	保障房
2014.1	16.94%	148.72%	32.26%	67.69%	26.50%	50.00%	95.59%	20.83%	62.50%
2014.2	17.71%	323.33%	100.00%	131.45%	51.97%	96.32%	175.41%	221.43%	141.03%
2014.3	15.95%	129.17%	13.51%	77.18%	12.82%	59.90%	130.88%	-58.14%	69.84%
2014.4	14.75%	134.78%	100.00%	93.75%	55.17%	84.28%	127.42%	42.86%	83.64%
2014.5	12.04%	17.48%	-12.50%	-11.11%	-11.86%	-5.96%	-4.32%	7.14%	-19.63%
2014.6	9.77%	10.68%	22.73%	-13.79%	-13.98%	-6.21%	-4.38%	69.23%	-13.68%
2014.7	7.24%	-0.91%	-23.64%	-12.17%	-19.70%	-13.61%	-9.66%	42.86%	-1.00%
2014.8	6.34%	4.04%	-25.00%	-17.05%	-21.11%	-16.29%	-10.27%	54.17%	-1.83%
2014.9	5.80%	-28.21%	-34.43%	-11.97%	-20.00%	-15.14%	-2.10%	87.50%	5.10%
2014.10	4.78%	-15.32%	-55.17%	-8.02%	-19.17%	3.85%	17.53%	17.86%	9.18%
2014.11	3.14%	-24.11%	-52.63%	-1.75%	-13.25%	7.26%	20.86%	60.00%	-13.51%
2014.12	-0.20%	-23.96%	-12.00%	5.36%	-7.64%	-3.50%	9.93%	14.81%	-11.32%
2015.1	-0.50%	-24.74%	-36.59%	16.06%	-6.76%	9.68%	15.79%	58.62%	3.30%
2015.2	-2.08%	-63.78%	-68.97%	-33.45%	-38.86%	-27.19%	-19.64%	-55.56%	-36.17%
2015.3	-3.17%	-31.82%	4.76%	-6.06%	-19.89%	2.93%	4.46%	200.00%	-13.08%
2015.4	-3.28%	-29.63%	-30.43%	-12.90%	-20.00%	-5.12%	0.71%	-36.67%	-28.71%
2015.5	-1.29%	-51.24%	-40.48%	-6.67%	-12.28%	-5.28%	9.77%	33.33%	-9.30%
2015.6	-0.17%	-49.12%	-18.52%	-1.78%	0.63%	1.39%	14.50%	65.91%	2.44%
2015.7	0.75%	-34.86%	40.48%	9.96%	8.18%	2.73%	18.32%	80.00%	-11.11%
2015.8	3.06%	-31.07%	0.00%	18.26%	1.27%	7.72%	28.24%	124.32%	-14.02%
2015.9	3.68%	-20.24%	0.00%	10.09%	1.92%	9.09%	10.00%	28.33%	-16.50%
2015.10	4.50%	-22.34%	42.31%	-6.64%	-1.28%	-14.81%	-7.18%	136.36%	-23.36%
2015.11	5.38%	-37.65%	40.74%	9.78%	4.17%	-5.85%	-4.76%	47.92%	-2.08%
2015.12	6.84%	-35.62%	-6.82%	37.71%	3.45%	3.96%	9.68%	51.61%	4.26%
2016.1	8.90%	-54.79%	0.00%	56.13%	0.72%	-9.48%	17.53%	-36.96%	1.06%
2016.2	11.29%	41.30%	116.67%	64.92%	24.58%	16.31%	27.41%	105.00%	33.33%
2016.3	15.63%	22.67%	13.64%	47.58%	19.15%	0.63%	17.07%	-9.26%	9.68%
2016.4	20.38%	15.79%	31.25%	37.50%	6.94%	2.16%	10.56%	121.05%	26.39%
2016.5	24.29%	67.80%	92.00%	16.52%	8.67%	1.49%	16.44%	40.00%	32.05%
2016.6	28.06%	113.79%	59.09%	9.95%	-11.80%	-5.50%	-5.33%	-34.25%	14.29%
2016.7	30.43%	61.97%	-18.64%	7.09%	-22.09%	-12.68%	-9.03%	-44.44%	9.09%
2016.8	32.67%	83.10%	46.15%	23.94%	-13.21%	-6.23%	-11.31%	-19.28%	10.87%

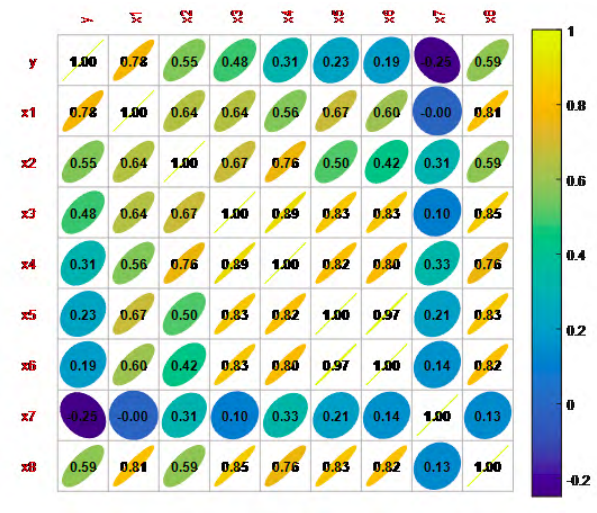

图2 相关系数矩阵图

从上面计算的矩阵R、矩阵P和图2知，Y与X_1, X_2, X_3, X_8的线性相关性是显著的，X_7与$X_1, X_2, X_3, X_4, X_5, X_6, X_8$的线性相关性均是不显著的。

4.2 多元线性回归

第一步，模型的建立。

这里先尝试作8元线性回归，建立Y关于$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$的回归模型如下：

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + a_5X_5 + a_6X_6 + a_7X_7 + a_8X_8 + \varepsilon$$

第二步，调用linearmodel类的fit方法求解模型。

根据计算结果可以写出经验回归方程如下：

$$\hat{Y} = 0.10068 + 0.12313X_1 + 0.00081406X_2 + 0.1388X_3 + 0.022894X_4 - 0.18963X_5 - 0.15343X_6 - 0.030922X_7 + 0.10932X_8 \quad (8)$$

对回归方程进行显著性检验的p值为$1.71\times10^{-9} < 0.05$，说明该方程是显著的。

第三步，残差分析与异常值诊断。

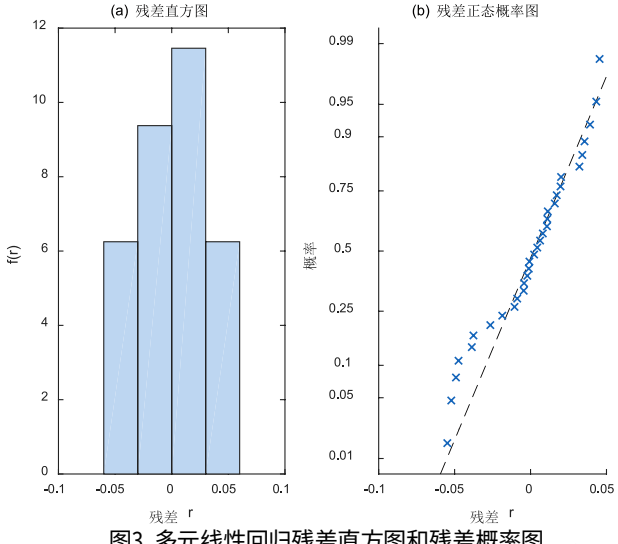

图3 多元线性回归残差直方图和残差概率图

从图3可以看出，残差基本服从正态分布。

第五步，模型改进。

下面去除异常值，将最不显著的线性项X_2和X_4去掉，重新建立回归模型

$$Y = a_0 + a_1X_1 + a_3X_3 + a_5X_5 + a_6X_6 + a_7X_7 + a_8X_8 + \varepsilon$$

再次调用fit函数作6元线性回归。得剔除异常值和线性项后的经验回归方程为

$$Y = 0.097303 + 0.124X_1 + 0.1534X_3 - 0.18758X_5 - 0.15205X_6 - 0.028553X_7 + 0.1054X_8 \quad (9)$$

对整个回归方程进行显著性检验的p值为$5\times10^{-11} < 0.05$，说明该方程是显著的。

4.3 完全二次多项式回归

假设Y关于$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$的理论回归方程为：

$$Y = a_0 + \sum_{i=1}^{8} a_iX_i + \sum_{i=1}^{7}\sum_{j=i+1}^{8} a_{ij}X_iX_j + \sum_{i=1}^{8} a_{ii}X_i^2$$

这是一个完全二次多项式方程(包括常数项、线性项、交叉乘积项和平方项)。利用fit函数求方程式中的未知参数的估计值。

$$\begin{aligned}\hat{Y} = {} & 0.069233 + 0.366X_1 - 0.077362X_2 + 0.10436X_3 + 0.43958X_4 + 0.056272X_5 \\ & - 0.59195X_6 + 0.18389X_7 - 0.12117X_8 + 0.03299X_1^2 + 0.14035X_1X_2 \\ & + 0.05315X_2^2 - 0.58236X_1X_3 - 0.38959X_2X_3 + 0.24775X_3^2 + 0.32531X_1X_4 \\ & + 2.2024X_2X_5 - 2.813X_3X_5 - 0.073084X_1X_6 - 1.2051X_2X_6 \\ & + 1.7655X_6^2 - 0.26649X_1X_7 + 0.24924X_2X_7 - 0.4686X_3X_7 \\ & + 0.39276X_4X_7 - 0.096452X_6X_7 - 0.054238X_7^2 + 0.15298X_1X_8 \\ & - 0.058246X_2X_8 - 0.83127X_5X_8 + 0.56732X_7X_8 + 0.36931X_8^2\end{aligned} \quad (10)$$

4.4 拟合效果图

上面调用fit函数作了8元线性回归拟合、6元线性回归拟合和完全二次多项式拟合，得出了3个经验回归方程(8)(9)(10)。拟合效果图如图4所示。

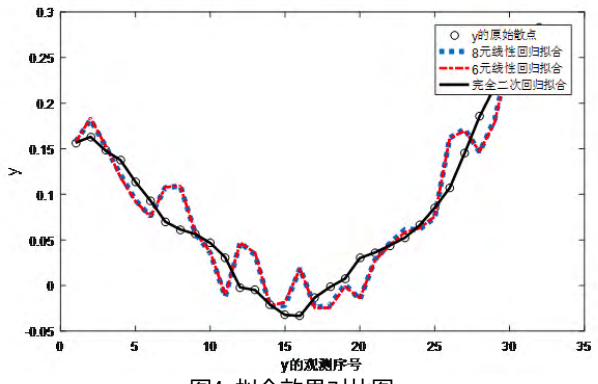

图4 拟合效果对比图

单纯从拟合的准确性来看，完全二次多项式回归拟合的拟合效果较好，8元和6元线性回归拟合的拟合效果差不多，相对都比较差。

4.5 逐步回归

在完全二次多项式回归模型的基础上，利用linearmodel类对

文化旅游产业的融资模式选择①

西安欧亚学院　葛联迎　刘甲瑢

摘　要：旅游业作为陕西省的支柱产业，新兴的文化旅游也成为一个新的经济增长点，企业对资金的需求越来越大。然而，相比其他行业，文化旅游业因为无形资产占比较大，投资回收期长，盈利能力弱、融资方式单一、融资模式陈旧、融资效率低下等诸多问题为文化旅游企业的长期发展埋下了隐患。融资方式的合理化，不仅有利于优化资源配置，而且有利于企业资本结构和治理结构的完善，从而推动文化旅游企业经营机制的转变和经济效益的提高。本文通过对我国文化旅游业融资方式现状的研究，通过层次分析法利用专家打分的方式对不同发展阶段的文化旅游企业融资的影响因素作出分析，为不同发展阶段的旅游文化企业融资提供一定依据。

关键词：文化旅游　阶段特点　融资模式　层次分析法

中图分类号：F592　　文献标识码：A　　文章编号：2096-0298(2017)02(a)-137-03

文化旅游景区作为我国旅游产业发展的主体，其特点为一般企业规模较大，在旅游行业中处于利润率较高的行业，盈利能力较强。资产规模较大，现金流量也较多，由于旅游景区投资时间长，旅游景区的融资优势是固定资产规模较大，比旅行社集团、酒店集团的盈利能力要强，现金流量也较多，然而目前这种大型文化景区经营主要仍是以门票经济为主，融资劣势在于盈利方式较为单一，经营收入主要来自门票，同时，由于文化旅游行业受外界波动性影响大，导致系统性风险较高，因此文化旅游产业对于融资模式的选择至关重要。

多数产业的发展都有其独具特点的生命周期，由于文化旅游产业发展的时间还很短，因此我们依据其不同的发展阶段对其分为初创阶段、成长阶段、成熟阶段三个阶段，从经营特性、风险特性、财务指标这三个方面分别分析其每个阶段的不同之处。

1　初创阶段文化旅游产业的阶段性特点

1.1　经营特性

初创阶段的文化旅游产业由于刚刚起步，缺乏独立的自主开发能力，文化旅游产业的“文化创意”缺乏，同时产业基础薄弱，创新技术水平低下，这就制约了文化旅游产业的发展。所以该阶段的

①基金项目：陕西省教育厅课题“陕西文化旅游企业创新融资模式研究”(14JK2064)。

象的stepwise方法，经过19次回归，得到二次多项式回归方程如下：

$$\hat{Y}=0.26671-0.032165X_1-0.00010849X_2-0.10223X_3+0.58273X_4+0.51205X_5-0.519X_6-0.48516X_7+0.10566X_8-0.26434X_1X_3+0.13904X_1X_5+0.80225X_1X_7+1.3048X_2X_6-0.58617X_2X_7-0.58001X_2X_8-6.2347X_3X_6+1.2803X_3X_7+2.8893X_3X_8+0.25259X_4X_7+3.1135X_4X_8-3.5079X_5X_6-2.5473X_5X_7-7.4893X_5X_8+1.7618X_6X_7+6.2328X_6X_8-1.9517X_7X_8+3.2397X_5^2+3.35X_6^2+0.42057X_7^2-3.1594X_8^2$$

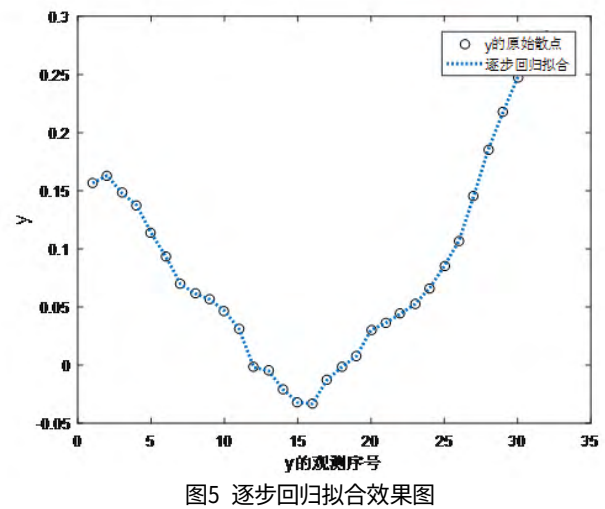

图5　逐步回归拟合效果图

对整个回归方程进行显著性检验的p值为$5.84\times10^{-6}<0.05$，说明该方程是显著的。模型拟合效果图如图5所示。

5　结语

房地产行业多年畸形发展催生了畸高的房价，从严调控、保持房地产市场的平稳健康发展被写入多个省份的政府工作报告。本文以百度搜索数据为基础，建立了房价预测的“最优”模型——逐步回归模型，该模型不但可以即时预测房价的走势，而且为地方政府制定调控房价措施提供有效的参考依据。另外，本文建立的房价预测模型还可以拓展到其他城市新房价格的预测和二手房价格的预测，可以预见，在人们利用网络搜索房产信息越来越多的将来，由于搜索数据量越来越大，从而预测的精度会越来越高.

参考文献

[1] 谢中华.MATLAB统计分析与应用[M].北京：北京航空航天大学出版社，2015.
[2] 董倩，等.基于网络搜索数据的房地产价格预测[J].统计研究，2014(10).
[3] 成鸿飞，等.基于MATLAB的房价预测与调控模型研究[J].科技论坛，2010(6).
[4] 杨志辉，等.基于MATLAB的房地产销售预测的科学计算[J].统计与决策，2005(1).
[5] 刘悦婷，等.基于MATLAB的兰州市商品住宅价格变动分析及预测[J].甘肃科学学报，2011(9).