Statistics and Application 统计学与应用, 2022, 11(3), 583-594
Published Online June 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.113062

Hans汉斯

基于变量选择的房地产价格影响因素分析与预测

赖新林

云南财经大学，云南 昆明

收稿日期：2022年5月24日；录用日期：2022年6月15日；发布日期：2022年6月27日

摘　要

房地产作为我国国民经济的支柱性产业，其衍生的房价问题一直以来都是一大民生问题。然而近几年由于国内大中小城市房价普遍居高不下，普通人想要买房变得越来越艰难，这就使得研究房价的影响因素十分有必要。本文基于北京市2017年1月至2018年1月份期间的二手房历史交易数据，从15个维度上对北京市的房价进行建模分析。为了挖掘出影响房价的主要因素，进而对房地产行情进行有效估计和预测，本文首先运用多元线性回归方法对房价数据进行建模，并基于逐步选择和Lasso回归方法对初始的十五个预测变量进行变量选择。最后通过提取各个模型的有效信息并比较不同模型的解释性效果和预测效果，得出对房价有较强影响的因素是：社区均价、房屋面积、装修状态、关注人数以及交易时间。同时，在预测性能方面，逐步回归和Lasso回归方法的表现比多元线性回归有较明显的优势。

关键词

影响因素，房价预测，逐步选择，Lasso回归

Analysis and Prediction of Real Estate Price Influencing Factors Based on Variable Selection

Xinlin Lai

Yunnan University of Finance and Economics, Kunming Yunnan

Received: May 24th, 2022; accepted: Jun. 15th, 2022; published: Jun. 27th, 2022

Abstract

As one of the pillar industries of China's national economy, real estate has always been a major li-

文章引用: 赖新林. 基于变量选择的房地产价格影响因素分析与预测[J]. 统计学与应用, 2022, 11(3): 583-594.
DOI: 10.12677/sa.2022.113062

velihood issue. In recent years, due to the increasing house prices in large, medium and small cities in China, it has become more and more difficult for most of people to buy a house, which makes it necessary to study the contributing factors of house price. Based on the historical transaction data of second-hand houses in Beijing from January 2017 to January 2018, this paper models and analyzes the house prices in Beijing from fifteen dimensions. To dig out the main factors affecting house prices and then effectively estimate and predict the house price, we first analyze house prices by using multiple linear regression model, and then select the important variable among the fifteen predictors based respectively on stepwise selection and Lasso regression. Finally, by extracting the effective information of each model and comparing the explanatory effect and prediction effect of different models, the factors that can have an important impact on house prices are: community average price, the square of the house, the renovation condition, followers of the house and the trade time. At the same time, in terms of prediction performance, stepwise regression and Lasso regression have significant advantages over multiple linear regression.

Keywords

Influencing Factors, House Price Forecast, Stepwise Selection, Lasso Regression

1. 引言

房地产价格也称房价，一般是指在一定区域内的房屋成交的平均价格，是根据当前市场情况专门制定的，常用每平米多少元表示，代表着某一区域的房地产市场整体价位水平，一定程度上能反映该区域的物价水平和居民收入情况。基于区域大小不同，其实际意义也不一样。以楼盘为单位，是给购房者作为参考的；以城市或地区为单位，则是给城市或国家决策者参考的。近些年来，我国在经济发展方面所取得的成就日益受到世界的瞩目，同时我国城市化进程的推进也相较前几年有了长足进步，并由此带来了房地产业的不断发展壮大，我国房地产行业的支柱地位在国民经济中也表现的越来越明显，逐渐成为新的经济增长点。然而，由于众多大城市房价过热，随之引发了大城市买房难和租房难的问题，居民的生活压力急剧增长，这就使得人们越来越关注房地产行业的各种相关数据和信息。值此严峻形势下，为解决居民住房问题，针对房地产的数据分析就显得尤为重要。

目前，针对房价影响的相关研究受到各方的关注，比如：李晨(2010)分析了影响我国各地区房价的重要因素并提出相关建议和措施，通过计算因子得分筛选出了经济、预期、房地产以及交通环境等 4 个代表性指标，但其研究缺乏一个可以量化的统计模型[1]。张侠等(2018)运用 OLS 和逐步回归方法对土地价格、人口数量、收入和其他 6 个房屋外部因素建立了房地产价格因素分析模型[2]。陈将浩(2014)以全国以及 31 个大陆省份或直辖市的相关数据为样本，利用 k 均值聚类将 31 个地区分为四类，建立起了房价与影响因素的回归方程，然而其残差未通过等方差、独立性检验，意味着可能遗漏部分较为重要的指标，致使未能得出较为理想的回归模型[3]。杨沐晞(2012)，李晓童等(2017)和李函谕等(2021)引入随机森林方法对特征价格模型进行回归预测并建立二手房价格评估模型，尽管这些文献里随机森林得到的预测效果较好，但是该模型无法给出明确的函数表达式，从而不能直观地得出各个变量对房价影响的大小[4] [5] [6]。因此找到一种具备高预测能力又能直观地对模型做出解释的模型对与房价的评估和影响因素分析，是非常有实际意义的。另外，现有研究大多关注于影响房价的外在因素，而少有探究房屋自身因素对房价的

影响。本文综合考虑影响房地产价格的外在和内在因素进行实证研究，并对这些因素的影响程度进行分析，这将对于给居民提供指导性意见和提升居民幸福指数具有重要意义。我们基于北京市 2017/01/01 至 2018/01/31 期间的二手房交易数据，通过两种不同的变量选择方法，从初始变量中筛选得到了对房价具有显著性影响的因素，并建立了房价预测模型，该模型通过了异方差性、正态性、异常值和共线性等检验。

2. 理论准备

2.1. 多元线性回归

2.1.1. 多元线性回归模型

多元线性回归是线性回归的一种，是研究一个因变量依赖两个或两个以上自变量而变化的数学回归模型[7]。事实上，一种现象的发生常常是多个因素共同作用的结果，相比只用一个自变量来对因变量进行预测或估计，无疑，由多个自变量的最优组合共同来预测或估计会更有效，也更符合实际。在一般情况下，假设有 p 个不同的预测变量，则多元线性回归模型的形式为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon \tag{1}$$

其中，Y 表示响应变量，$X_1,\cdots,X_p$ 是 p 个可以精确测量并可控制的一般变量，称为解释变量，β_0 称为截距项，$\beta_1,\cdots,\beta_p$ 称为回归系数，ε 为随机误差，常被假定为 $E(\varepsilon)=0, var(\varepsilon)=\sigma^2$。当 $p=1$ 时，式(1)即为一元线性回归模型。

2.1.2. 回归系数估计

在实践中，式(1)中的回归系数 $\beta_0,\beta_1,\cdots,\beta_p$ 都是未知的，需要进行估计。然后对给定的任意一组 $(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)$，可由如下公式对 Y 进行预测，预测值标记为 $\hat{y}$：

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1X_1+\cdots+\hat{\beta}_PX_P \tag{2}$$

根据最小二乘理论，模型(1)的系数估计 $(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)$ 为使得如下的残差平方和达到最小：

$$RSS=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\cdots-\hat{\beta}_Px_{ip}\right)^2 \tag{3}$$

的一组系数值，即为多元回归系数的最小二乘估计。更一般的，对于响应变量是多维型的，其多元回归的系数估计形式可用矩阵代数来表示。

2.2. 逐步回归

逐步回归法是一种常用的变量选择方法，其基本思想是将变量逐一引入模型中，引入的条件是使得其偏回归平方和经验是显著的，同时对已经选入的解释变量逐个进行 t 检验，将经检验后认为不显著的变量删除，以保证回归方程中只包含显著性变量。如此重复经过若干步，直到回归方程中既没有新的显著的解释变量引入，也没有不显著的解释变量剔除为止。

常用的逐步回归法有向前、向后逐步选择法以及这两种的混合方法。向前逐步法的思想是变量由少变多，根据不同的准则(如 AIC、BIC、调整 R^2 等)依次在模型中添加变量，直到没有可引入的变量为止。向后逐步法则正好相反，先将全部自变量选入回归模型，逐次迭代，每次移除一个对模型拟合结果最不利的变量[7]。向前和向后逐步选择的混合方法，是与向前逐步方法类似，逐次将变量加入模型中，然而在加入新变量的同时也会移除不能提升模型拟合效果的变量。如此混合方法在试图达到最优子集选择效果的同时也保留了前两种方法在计算效率上的优势。故本文在后续的房地产数据分析中也将采用向前和

向后逐步回归的变量选择方法，来获取对房价产生重要影响的因素。

2.3. Lasso 回归

Lasso 回归(Lasso Regression)和岭回归是统计学中比较常用且经典的压缩估计方法，它们都是通过生成一个惩罚函数对模型中的变量系数进行压缩，以达到防止模型过度拟合的目的。相较于最优子集选择、逐步回归方法通常选择涉及变量减少的模型，岭回归的最终模型将包含所有预测变量[7]。岭回归因其采用 l_2 惩罚项使得系数将系数往 0 的方向进行缩减的同时，并不会将其中任何一个都精确压缩至 0，因此当变量个数非常大时该方法不利于模型解释。而 Lasso 回归采用了 l_1 惩罚原则，能够有效克服岭回归上述缺点，使得其在模型特征选择及预测预测中都具备较大优势。

Lasso 回归的系数可通过求解下式的最小值得到：

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|. \tag{4}$$

从式(4)可知：Lasso 回归系数的估计值受调节参数λ的影响，当 λ 足够大时，式中的惩罚项会迫使一些系数估计(对于模型的贡献很小)强制设定为 0，故而需对其进行恰当选择。同时这将使得 Lasso 回归最终得到一个只包含部分变量的简单易解释模型。另外，由于 λ 一般是可以通过交叉验证得到的，其在一定程度上能够保证最终模型的模型精度且使得解释性较强，因此该方法在实际生活中的应用非常广泛。

3. 数据整理与分析

3.1. 数据来源

为了研究影响房价的主要因素并更好地对房价做出预测，以便帮助更多的购房者尽可能地提供科学合理的购房指导意见，我们以链家网站上挂牌出售的北京市从 2010 年 1 月至 2018 年 1 月的二手房交易数据为基础进行深入的统计分析。本案例所涉及的数据取材于 Kaggle 平台提供的“Housing price in Beijing”数据集，参考网址如下：https://www.kaggle.com/ruiqurm/lianjia。该数据是开源数据，为链家平台自身数据，是其进行市场分析的重要支撑，透明性和真实性较高。

3.2. 数据描述和预处理

该北京房地产数据一共记录 318,851 单房屋出售交易，其中包含 26 个不同的变量：如房子所在的经纬度、挂牌出售时长、房产面积、房产总价、房子每平米价格、装修状态、梯户比等。在 26 个变量中，我们首先剔除了 Url (房子对应的链家网址)、id (房子在链家网上的 id)、Cid (社区代码)、DOM (挂牌出售时长)等对房价预测无关的变量。此外，考虑到 kitchen (厨房数量)、livingRoom (客厅数量)、drawingRoom (书房数量)和 bathRoom (卫生间数量)等 4 个指标作为住宅的必备配置，其数值通常固定，对房价产生实质性影响的可能性较低，因而也将不予考虑；由于房产总价与房价(即房屋每平米价格)之间存在较高的共线性问题，而该问题的存在将大大影响模型的准确性，因此房产总价也被排除在以下的模型之外。同时，我们认为房子所在地的经纬度对房价的影响主要来自各行政区的房价差异，查阅相关资料可知，北京市的房价变化趋势大体呈现辐射状分布，即由中心呈环状向外递减。因此，本文考虑以天安门广场所在坐标为中心，将数据中的经度(Lng)、纬度(Lat)变量重新组合成一个新的距离变量(distance)，以便更好地分析影响房价变化的决定因素，该距离变量衡量了从各个房源到天安门广场的直线距离。

在上述处理后，本文还考察了数据的缺失情况，由于数据缺失值占比很小，我们考虑直接删除缺失数据所在的交易单。另外，考虑到本文选取的各变量指标多是与房屋自身相关的硬性指标，并不会随交易时间而发生变化，为探究交易时间与房价变化背后的联系，我们选取该数据集中 2017/01/01 至

2018/01/31 期间的房地产交易数据进行分析，最终得到了 42,929 × 16 维的数据矩阵，这 16 个变量的相关解释见表 1。

Table 1. Variable descriptions
表 1. 变量说明

编号	变量名	类型	说明
1	price	连续型	房价(每平米价格)
2	distance	连续型	房子距故宫的距离
3	tradeTime	日期型	交易时间
4	followers	连续型	关注人数
5	square	连续型	面积
6	floor	离散型	所在楼层类型：0-未知，1-底，2-低，3-中，4-高，5-顶
7	buildingType	离散型	建筑形式：1-塔式，2-平房， 3-蝶式，4-板式
8	constructionTime	连续型	建筑时间
9	renovationCondition	离散型	装修状态：1-其他，2-毛坯，3-简装，4-精装
10	building Structure	离散型	建筑结构：1-未知，2-混合，3-木砖结构，4-砖混凝土结构，5-钢构，6-钢筋混凝土结构
11	ladderRatio	连续型	梯户比(电梯数与每层楼住户数的比例)
12	elevator	离散型	是否有电梯：1-有，0-没有
13	fiveYearsProperty	离散型	是否已过五年产权期限：1-是，0-否
14	subway	离散型	是否有地铁：1-是，0-否
15	district	离散型	所属区：1-东城，2-丰台，3-大兴，4-亦庄开发区，5-房山，6-昌平，7-朝阳，8-海淀，9-石景山，10-西城，11-通州，12-顺义，13-门头沟
16	communityAverage	连续型	社区均价

接下来，本文将以房价(price)作为响应变量，通过对该变量关于其他变量建立不同的统计模型来分析、探讨房价的影响因素，并最终确立一个合理有效的房价预测模型。

4. 房价的影响因素的建模分析与预测

为了合理地挖掘房价的影响因素并有效预测房价，本文将前文所得数据集的 42929 个数据样本随机分成两部分：一部分作为训练集——令其占原有样本量的 99%，用于每个模型的估计和拟合；剩余部分作为测试集——采用训练集得到的估计模型对测试集中的房价进行预测，进一步得到房价真实值与预测值的偏差平方和平均的开根(即均方根误差，简写为：RMSE)，以便评价各个模型的预测效果。本文所有数值结果皆是基于 R 软件计算所得。

4.1. 普通最小二乘法(OLS)估计模型

针对前文所得到的房地产数据，在其中的训练集里建立房价影响因素分析的多元线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{5}$$

其中，Y 为房价，$X_1, X_2, \cdots, X_p$ 分别表示表 1 中除房价以外的 15 个变量，这里 $p=15$。表 2 为多元回归模型(5)的最小二乘估计结果。

由表 2 中各自变量系数的 t 检验中对应的 p 值可知，在其它自变量不变的情况下，对房价有强显著性影响的变量是社区均价(communityAverage)、交易时间(tradeTime)、面积(Square)、关注人数(followers)和装修状态(renovationCondition)这 5 个变量；房屋是否已满五年产权(fiveYearsProperty)变量也对房价有着一定的影响；其他变量如距离(distance)、楼层(floor)、建筑类型(buildingType)等，对房价的影响不显著。另外，回归方程 F 检验的 p-value: < 2.2e−16，这意味着，在给定显著性水平 α 为 0.05 下，这个模型整体是显著的，其中拟合优度为 0.9155，调整的 R^2 达到了 0.9124，也进一步表明了该回归方程拟合效果较优良。

Table 2. Linear regression fitting results
表 2. 线性回归拟合结果

	Coefficients	Std. Error	t value	Pr (>|t|)	
(Intercept)	4.12E+05	5.60E+04	7.356	1.03E−12	***
distance	−5.10E−02	3.62E−02	−1.409	0.15967	
tradeTime	−2.34E+01	3.24E+00	−7.212	2.64E−12	***
followers	−1.38E+01	4.81E+00	−2.872	0.00429	**
square	−8.59E+01	1.23E+01	−6.971	1.25E−11	***
floor	−5.75E+02	3.55E+02	−1.621	0.10568	
buildingType	1.13E+01	3.87E+02	0.029	0.97684	
constructionTime	1.06E+01	5.45E+01	0.195	0.84548	
renovationCondition	1.22E+03	3.87E+02	3.152	0.00174	**
buildingStructure	−3.72E+01	3.75E+02	−0.099	0.92085	
ladderRatio	3.18E+03	2.91E+03	1.091	0.27569	
elevator	6.60E+02	1.52E+03	0.435	0.6635	
fiveYearsProperty	1.34E+03	7.51E+02	1.789	0.07429	.
subway	−5.78E+01	7.89E+02	−0.073	0.94162	
district	−4.91E+01	1.43E+02	−0.344	0.73135	
communityAverage	1.06E+00	1.92E−02	55.102	<2E−16	***

Signif. Codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1; Residual standard error: 7402 on 414 degrees of freedom; Multiple R-squared: 0.9155, Adjusted R-squared: 0.9124; F-statistic: 299.1 on 15 and 414 DF, p-value: < 2.2e−16.

然而，尽管从调整 R^2 来看该模型具备很高的解释性，但该模型是基于全部 15 个预测变量得到的，其可能包含了多个不重要的变量。无关变量的加入将增大模型各系数估计的方差，并可能导致估计结果存在较大变异，这将进一步导致模型的预测性能变差。因此，为了更精准地获得能够对房价产生重要影响的变量，接下来本文将运用两种变量选择方法来做进一步分析，细节见 4.2 节和 4.3 节。

4.2. 逐步选择(Stepwise Selection)——筛选重要变量

本节将应用 2.2 节中的逐步回归理论对房地产数据的训练集重新分析。根据逐步回归的思想，本节

实现了与房价无关的变量的筛选，并得到了与房价真正相关的变量。我们采用基于 AIC 准则的向前和向后逐步选择混合方法最终得到了一个只包含 8 个预测变量的模型，该模型选择以距离(distance)、交易时间(tradeTime)、受关注人数(followers)、面积(square)、楼层(floor)、装修状态(renovationCondition)、是否已过五年产权年限(fiveYearsProperty)、社区均价(communityAverage)作为房价的解释变量。另外结合全模型分析，根据两个模型的整体 AIC 值以及各模型增减变量后的 AIC 值和 RSS 的表现，逐步回归得到的最终模型在 AIC 和 RSS 方面优于全变量模型。同时，为了进一步观察逐步回归的模型解释力，表 3 给出了由逐步回归方法得到的最终模型的各变量系数结果。

Table 3. Stepwise regression results
表 3. 逐步回归结果

	Coefficients	Std. Error	t value	Pr (>|t|)	
(Intercept)	4.04E+05	5.52E+04	7.314	1.32E−12	***
distance	−5.39E−02	3.19E−02	−1.687	0.092283	.
tradeTime	−2.28E+01	3.18E+00	−7.166	3.49E−12	***
followers	−1.43E+01	4.75E+00	−3.015	0.002724	**
square	−7.85E+01	1.08E+01	−7.276	1.70E−12	***
floor	−6.06E+02	3.49E+02	−1.737	0.083133	.
renovationCondition	1.28E+03	3.78E+02	3.385	0.000778	***
fiveYearsProperty	1.29E+03	7.28E+02	1.774	0.07673	.
communityAverage	1.05E+00	1.64E−02	64.048	<2E−16	***

Signif. Codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1; Residual standard error: 7358 on 421 degrees of freedom; Multiple R-squared: 0.9151, Adjusted R-squared: 0.9135; F-statistic: 567.1 on 8 and 421 DF, p-value: < 2.2e−16.

根据表 3 的系数结果来看，逐步选择得到的回归模型(含 8 个预测变量)整体是非常显著的，且其调整 R^2 达到了 0.9135 (多元回归模型的调整 R^2 为 0.9124)。再观察各变量的显著性程度，我们可以发现逐步回归得到的与房价显著相关的几个变量和 4.1 节中的结果几乎一致，但本节的方法还将距离(distance)和楼层(floor)视为对房价有较强影响的两个因素。

4.3. Lasso 回归模型——变量降维

本节采用了 Lasso 回归方法来分析房地产数据。在这里，值得一提的是 Lasso 回归的两个作用：一是识别重要变量，二是得到预测模型。本节以房价的预测作为前提条件，挖掘了对房价产生重要影响的变量并对它们构建新的回归模型。为实现此目的，本文先选择在 $\lambda = 10^4$ 至 $\lambda = 10^{-2}$ 的范围内对房地产数据中的训练集据进行 Lasso 回归，并得到了相应的 Lasso 系数变化图，见图 1；再利用十折交叉验证方法获得使交叉验证误差达到最小的 λ 值，同时得到所对应模型的变量系数；最后对非零系数变量关于房价做进一步的建模分析，其模型估计结果见表 4。

图 1 展示了随着λ的对数的增大，Lasso 回归非零系数的变量在不断减少，甚至部分系数直接为 0。通过模拟计算得：当 $\lambda = 270.8776$ 时，所对应模型的交叉验证效果最优，且模型筛选出了 8 个重要变量。表 4 为经过交叉验证拟合选出的最优λ值所对应的 8 个非零系数变量及其 Lasso 回归系数。通过和表 3 中逐步选择方法得到的拟合模型进行比较，我们发现 Lasso 得到的非零系数变量和逐步选择得到的结果相一致。根据表 4 中的变量，进一步关于房价做普通最小二乘回归，所得结果将与表 3 一致。

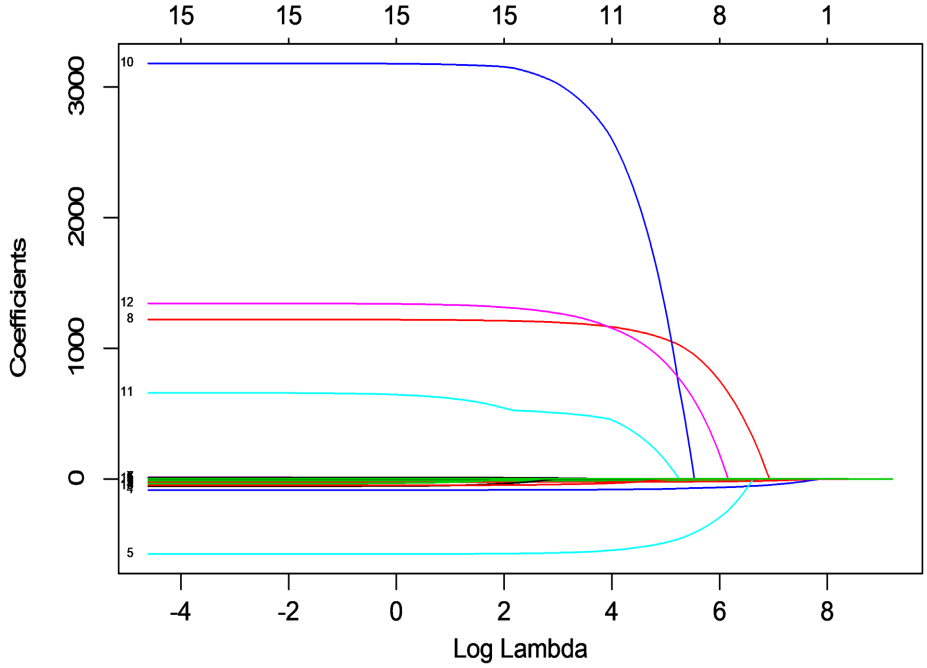

Figure 1. Plot of the Lasso regression coefficients as λ changes
图 1. Lasso 回归系数随 λ 的变化图

Table 4. Non-zero coefficient variables and coefficients for Lasso regression
表 4. Lasso 回归的非零系数变量及系数

Variable	Coefficients
(Intercept)	3.638886e+05
distance	−3.070959e−02
tradeTime	−2.055111e+01
followers	−9.898148e+00
square	−6.980960e+01
floor	−3.970193e+02
renovationCondition	9.245955e+02
fiveYearsProperty	5.499683e+02
Community Average	1.043899e+00

4.4. 模型的性能比较

本节将对 4.1、4.2、4.3 节所得模型的残差标准误、调整 R^2 以及各个模型的预测效果进行综合比较分析，以便选出最优的房价预测模型。本文中，模型的预测效果由拟合模型在测试集中得到的房价估计值与其真实值的偏差平方和平均的开根(记为 RMSE)来表示，数值结果见表 5。

纵观表 5 的调整 R^2 数据可知，OLS 和逐步回归模型的预测变量对房价的解释能力几乎是一样好的，尽管各模型的回归变量个数不一致。从残差标准误来看，逐步回归模型稍优于 OLS 模型，这说明逐步回归在模型的系数估计的准确性方面较有优势。此外，从 RMSE 的角度可以看出，逐步回归拟合和 Lasso

拟合的方法在该数据集中对房价的预测性能上大大优于 OLS 拟合模型。综合表 5 中的各项指标来看，逐步回归无论是从模型的解释能力、模型估计的精度还是房价的预测方面都是较有优势的。

Table 5. Performance comparison of models
表 5. 模型的性能比较

模型	预测变量个数	残差标准误	调整 R^2	RMSE	显著性变量 a，b 及其个数
OLS 模型	15	7402	0.9124	154376748	3, 4, 5, 9, 13, 16 (6 个)
逐步选择	8	7358	0.9135	8044.984	2, 3, 4, 5, 6, 9, 13, 16 (8 个)
Lasso 回归	8	-	-	8002.162	-

[a] 备注：这里的显著性按 p 值小于 0.1 算；[b] 备注：由于表格篇幅所限，相关变量按表 1 中的编号指代。

综合前文知识，本文拟建立的房价预测模型为：

$$\begin{aligned}\text{Price} = 404000 - 0.054\text{distance} - 22.8\text{tradeTime} - 14.3\text{followers} - 78.5\text{square} - 606\text{floor} \\ + 1280\text{renovationCondition} + 1290\text{fiveYearsProper} + 1.05\text{communityAverage}\end{aligned} \tag{6}$$

其中，各变量释义可参考表 1 内容，各变量系数值参考表 3。根据表 3 可知，给定显著性水平 $\alpha = 10\%$ 下，模型(6)的所有解释变量系数的 t 统计量的 p 值均小于 0.1，故系数均显著。为了检验该模型的有效性，本文将在下一章检验所提模型的性质。

5. 模型检验

本章将对逐步回归(或者 Lasso)选出来的模型(见(6)式)进行多方面的检验，如共线性检验、异方差检验、正态性检验和异常点检验。

通过对各自变量做多重共线性检验发现，该模型所有变量的 VIF 值(方差膨胀因子，简称 VIF)均小于 5，因此认为该模型的变量不存在严重的多重共线性。图 2 左刻画了训练集下逐步回归方法下的房价拟合值与对应的残差估计值的散点图，易看出该数据集并未有明显的异方差现象。图 2 中为经过标准化变换后的残差的分位数图，直观上看较为明显的呈一条直线，说明残差数据是符合正态性的。图 2 右显示该

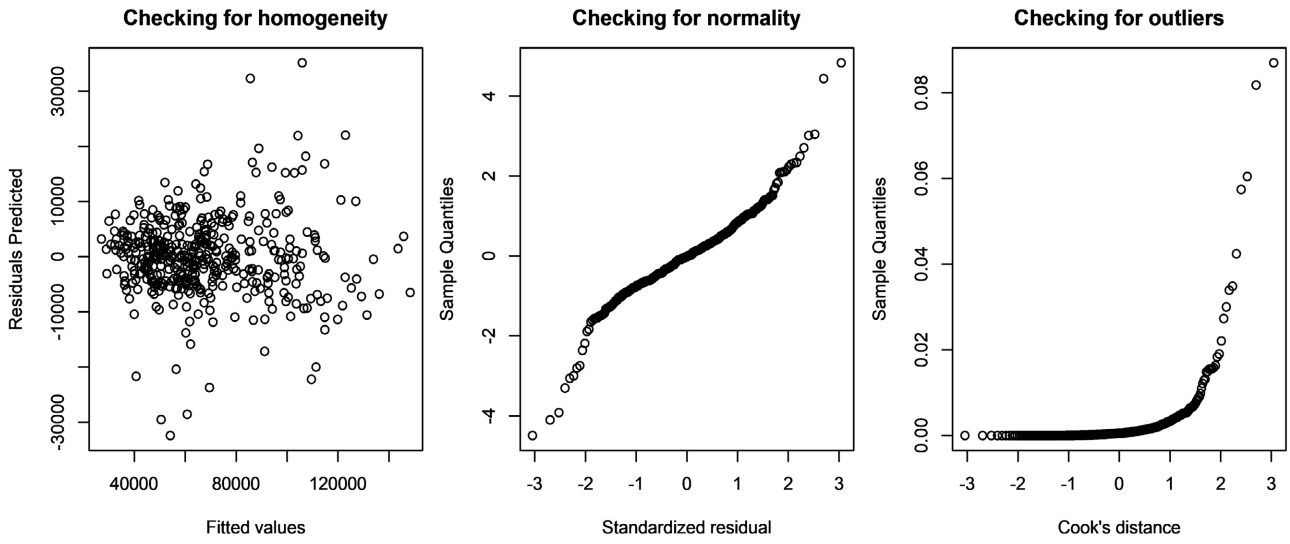

Figure 2. Homogeneity test, normality test, outlier test
图 2. 异方差检验、正态性检验、异常点检验图

数据集或存在五个较为异常的数据点，在对这五个异常点进行检验分析后，发现它们的异常性是极其显著的，其对应的 p-值均小于 0.001，故可判定其为异常点，直接删除这 5 个异常数据。

在删除异常数据后，本文再次对模型(6)关于房价数据做回归拟合，结果见表 6，最终得到的预测模型如下：

$$\begin{aligned}\text{Price} = 420000 - 0.006\text{distance} - 23.7\text{tradeTime} - 11.2\text{followers} - 81.7\text{square} - 732\text{floor} \\ + 999\text{renovationCondition} + 601\text{fiveYearsProper} + 1.04\text{communityAverage}\end{aligned}$$

可见，删除异常点后，该拟合模型的残差标准误降低了近 11%，其调整 R^2 达到了 0.929，该值高于此前的逐步回归模型，而且该模型的 RMSE 也低于异常点删除之前，说明模型得到了进一步的改善。

Table 6. The result of the fit after deleting the anomalous data
表 6. 删除异常数据后的拟合结果

	Coefficients	Std. Error	t value	Pr (>|t|)	
(Intercept)	4.20E+05	4.93E+04	8.523	2.89E−16	***
distance	−6.36E−03	2.90E−02	−0.219	0.82667	
tradeTime	−2.37E+01	2.84E+00	−8.352	1.01E−15	***
followers	−1.12E+01	4.26E+00	−2.616	0.00922	**
square	−8.17E+01	9.91E+00	−8.242	2.23E−15	***
floor	−7.32E+02	3.11E+02	−2.355	0.01898	*
renovationCondition	9.99E+02	3.39E+02	2.953	0.00333	**
fiveYearsProperty	6.01E+02	6.51E+02	0.923	0.35665	
communityAverage	1.04E+00	1.47E−02	71.058	<2E−16	***

Signif. Codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1; Residual standard error: 6543 on 416 degrees of freedom; Multiple R-squared: 0.9304, Adjusted R-squared: 0.929l F-statistic: 694.9 on 8 and 416 DF, p-value: < 2.2e-16; RMSE: 8015.496.

另外，观察表 6 的结果可知，异常值对于一个模型的影响是很大的。通过与表 3 中的结果对比发现：floor(楼层)的显著性增强了，其 p-值从原来的 0.083 降到了 0.019，而且其系数估计值约为−732，这说明楼层对房价的估计很重要；是否满五年产权期限和距离对于房价的变化而言反而不是那么重要；社区均价、房屋面积、受关注人数和房子的装修状态始终都是被视为房价参考的重要因素；在我们的模型里，交易时间对房价的影响也很显著，其也能在房价走势图(见图 3)体现。

图 3 展示了从 2017 年到 2018 年北京每月平均房价的变动情况，自 2017 年 3 月份以来，北京的平均房价呈现明显下降趋势，该下降情况在一定程度上可能是受到了房价调控的影响[8]。同时，在本文的模型里，房价调控所在时期则被可以解释为交易时间，于是在调控政策未变的情况下，交易时间的影响就会变得异常显著。虽然近年来房地产调控政策不时变化，但是理论上只要不是频繁的出台变化极大的调控政策，我们的模型就能通过交易时间来解释当前调控对房价的影响力。显然，短期频繁的政策变化在生活中并不常见，因此，对于每一个时间段，我们的方法可以得到一个对应的最佳房价预测模型。

6. 结束语

6.1. 结论

本文运用了三种统计模型对北京房价做了影响因素分析及预测分析。综合比较得出来的结论有：普

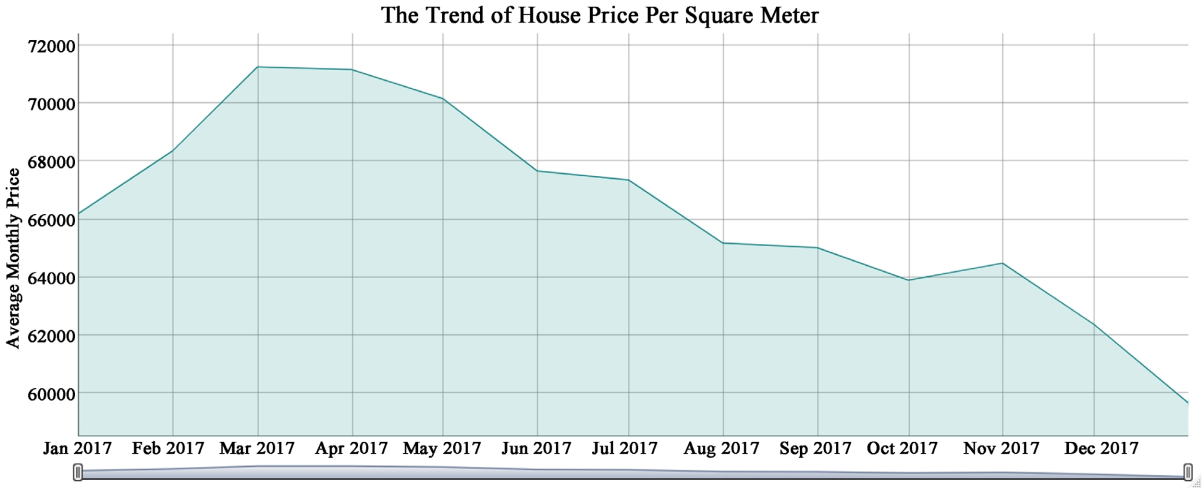

Figure 3. Trend of monthly average house prices in Beijing
图 3. 北京月平均房价变化趋势图

通最小二乘回归的预测性能较差、且考虑的变量个数较多，不适合高维数据的直接分析；逐步选择方法和 Lasso 回归可以高效地获取与房价有重要关系的变量，同时也提升了模型的预测能力，从而达到变量选择的目的。综合逐步选择和 Lasso 回归方法，本文得到的能够影响房价的主要因素有五个：交易时间、社区均价、房屋面积、装修状态和关注人数，而次要因素则有一个：楼层。最后，本文提出了一个预测模型，从模型来看，距离、交易时间、关注人数、面积、楼层均与房价呈负相关，而装修状态、是否已过五年产权年限、社区均价与房价呈正相关。相关性大致都与实际情况相吻合，值得注意的是，面积与房价的负相关性一方面可能是受市场供求关系的影响，小户型总价低，空间利用率高，是很多人置业的优先选择；另一方面是由于相同条件下小户型房子在建设过程中单位面积的建筑投入要高于大户型，单位成本的提高自然使得其单价增高。

6.2. 建议

房价是一个复杂的经济范畴，影响地区房屋均价的因素除了本文所研究的因素外，更多的还有诸如经济形势和国家政策影响等，这些也是造成地区房价居高不下的重要原因。有鉴于此，我们为居民购房提供几个建议：

第一，针对自身需求，合理定位产品。房子所附带的区位优势，如：交通、位置、周围环境、外部配套设施等往往备受购房者关注，但实际中并非所有人都能利用到这些优势，因此需要我们审视自身的需求，做到合理定位。

第二，摆正自身心态，明确购房目的。房价的持续上涨很大一部分原因在于众多开发商疯狂拿地、炒房者恶意抬高房价。一方面作为刚需购房者，应做好资金预算，坚定本心；另一方面，需明确购房目的，不作无谓幻想，理性购买。

第三，关注国家政策，了解市场形势。房地产市场的稳定有赖于政府的宏观调控，借助行政手段对市场进行干预。购房者需紧跟国家政策，努力提高对市场变化的敏感度，以便及时了解行情和未来市场走向。

基金项目

云南财经大学研究生创新基金项目(2022YUFEYC071)。

参考文献

[1] 李晨. 基于因子分析法的中国房价影响因素分析[J]. 经济研究导刊, 2010(16): 158-159.

[2] 张侠, 吴晶晶, 孙道助. 基于线性回归模型的安徽省房价影响因素分析[J]. 阜阳师范学院学报(自然科学版), 2018, 35(4): 73-77.

[3] 陈将浩. 房价影响因素及 R 语言实现[D]: [硕士学位论文]. 合肥: 中国科学技术大学, 2014.

[4] 杨沐晞. 基于随机森林模型的二手房价格评估研究[D]: [硕士学位论文]. 长沙: 中南大学, 2012.

[5] 李晓童, 郭萱, 王成杰. 基于随机森林方法的北京市二手房价格研究[J]. 数据挖掘, 2017, 7(2): 37-45. https://doi.org/10.12677/HJDM.2017.72004

[6] 李函谕, 魏嘉银, 卢友军. 基于随机森林的深圳二手房价格预测与分析[J]. 现代信息科技, 2021, 5(15): 100-104.

[7] [美]加雷斯.詹姆斯, 等, 著. 统计学习导论——基于 R 运用[M]. 王星, 等, 译. 北京: 机械工业出版社, 2015.

[8] 潘慧峰, 刘曦彤. 限购政策对房地产价格及供求的调控效果研究——以北京市为例[J]. 价格理论与实践, 2017(8): 48-51.