**Advances in Social Sciences** 社会科学前沿, 2020, 9(8), 1123-1131
Published Online August 2020 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2020.98156

Hans汉斯

# Application of Nonparametric Regression in House Price Forecast

**Yunying Lin**

Jiangxi University of Finance and Economics, Nanchang Jiangxi
Email: 786359959@qq.com

Received: Jul. 26th, 2020; accepted: Aug. 7th, 2020; published: Aug. 14th, 2020

## Abstract

**Housing is closely related to human life, which is an important part of the total wealth of residents and also affects people's happiness index to a certain extent. Therefore, great importance is attached to the qualitative and quantitative research on housing prices both at home and abroad. Based on the Boston house price data of Harrison and Rubinfeld, this paper discusses the comparative analysis of OLS regression and nonparametric regression in house price prediction by using R software. The results show that the OLS regression model is against the OLS regression statistical hypothesis, and OLS regression is not in line with the theoretical basis. Based on the characteristics of nonparametric regression, it is more suitable to use nonparametric regression (Lasso regression and Ridge regression) to predict house prices, and Bootstrap method and circulation method are used to select the model. When using multiple linear regression to analyze the data, we can't ignore the premise hypothesis when the multiple linear regression is established. However, the data in reality are often not ideal, so the applicability of nonparametric regression is wider.**

## Keywords

**Nonparametric Regression, Bootstrap, Housing Price**

# 非参数回归在房价预测上的应用

**林赟英**

江西财经大学，江西 南昌
Email: 786359959@qq.com

收稿日期：2020年7月26日；录用日期：2020年8月7日；发布日期：2020年8月14日

## 摘 要

**住房与人类生活息息相关，是居民总财富的重要组成部分，在一定程度上也影响着人们的幸福指数，因**

文章引用: 林赟英. 非参数回归在房价预测上的应用[J]. 社会科学前沿, 2020, 9(8): 1123-1131.
DOI: 10.12677/ass.2020.98156

**此国内外都很重视对房价定性和定量的研究。本文利用哈里森和鲁宾菲尔德的波士顿房价数据，使用R软件，探讨OLS回归和非参数回归在房价预测上的比较分析。实验结果表明：使用OLS回归模型预测房价违反了OLS回归统计假设，使用OLS回归是不符合理论依据的。基于非参数回归的特性，更适合采用非参数回归(Lasso回归和Ridge回归)对房价进行预测，并使用Bootstrap法和循环法对模型进行选择。在使用多元线性回归对数据进行分析时，不能忽略其多元线性回归成立时的前提假设，而现实中的数据往往是非理想化的，因此非参数回归的适用性更广。**

## 关键词

**非参数回归，Bootstrap，房价**



## 1. 引言

对于大多数国家来说，住房与人类生活息息相关，住房是居民总财富的重要的组成部分。易成栋、任建宇和高璇对房价和幸福感进行实证分析发现，住房会影响居民的幸福感[1]。不仅如此，房价是支撑国民经济的重要产业，房价的波动对我国的经济的影响巨大。对房价进行合理地预测分析，有利于国家政策的调控，对相关主体产业的发展也有借鉴意义。

## 2. 文献综述

国内外都很重视对房价定性和定量的研究。国内外学者对于房价定性预测的研究成果比较多，如国内学者周佳琪和金百锁基于空间网络自回归变点模型对房价的影响因素进行分析，认为商业区、地铁等会影响房价[2]。薛建谱和王卫华基于均衡分析认为引起房价上涨的主要因素是收入[3]。范允奇和王艺明基于二阶段局部动态调整模型认为土地成本是引起房价上涨的重要因素等[4]。但是对于房价定量预测的文献较少。在房价定量研究的文献中，Malpezzi 对美国重复交易住宅价格指数用时间序列截面回归分析[5]。国内学者邬嘉怡、王思玉、史宏炜、李虎森、楼凯达和崔丽鸿对北京房屋价格采用多小波变换进行分析[6]。唐晓彬、张瑞和刘立新对北京二手房价格指数采用蝙蝠算法的 SVR 模型进行分析[7]。

通过文献回顾可以看出国内外理论界进行房价预测的方法主要有时间序列截面回归分析等。本文将通过对 13 个变量进行 OLS 回归，通过指出 OLS 回归的不足之处，采用非参数回归模型预测房价。本文的创新之处在于使用 Bootstrap 法和循环法选择较佳的非参数回归方法来预测房价。

## 3. 研究方法

总体多元线性回归模型的公式如下：

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_a x_{ta} + \varepsilon_t$$

其中，$t = 1, 2, \cdots, n$，$n$ 为样本容量，$y_t$ 是响应变量，$x_{t1}, \cdots, x_{ta}$ 是预测变量，$\beta_0$ 是截距项，$\beta_1, \cdots, \beta_a$ 是总体多元线性回归模型中 $x_{t1}, \cdots, x_{ta}$ 的回归系数，$\varepsilon_t$ 是除了 $x_{t1}, \cdots, x_{ta}$ 以外可以影响 $y_t$ 的其它不可观测因素，称为扰动项。

样本多元线性回归模型的公式如下：

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \cdots + \hat{\beta}_a x_{ta} + \hat{\varepsilon}_t$$

样本多元线性回归函数的公式如下：

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \cdots + \hat{\beta}_a x_{ta}$$

其中，$t = 1, 2, \cdots, n$，$\hat{\beta}_0$、$\hat{\beta}_1, \cdots, \hat{\beta}_a$是样本多元线性回归模型中$x_{t1}, \cdots, x_{ta}$的回归系数，$\hat{y}$是预测变量的估计值，$\hat{\varepsilon}_t$是残差项，可以通过$\hat{\varepsilon}_t = y - \hat{y}$测量出。

### 3.1. OLS 回归

采用普通最小二乘法(OLS)来估计多元线性回归模型中的$\hat{\beta}_0$、$\hat{\beta}_1, \cdots, \hat{\beta}_a$。OLS 回归对样本数据的要求比较高，需要拟合的线性回归满足四条统计假设：1) 预测变量呈正态分布，则残差值服从均值为 0 的正态分布；2) 预测变量间相互独立；3) 响应变量和预测变量线性相关，即预测变量和残差值相互独立；4.预测变量的方差不随响应变量的变化而变化[8]。

### 3.2. 非参数回归

相比于 OLS 回归，非参数回归的优点是对参数估计的方法不同，不需要知道总体分布情况下进行统计推断回归，灵活性较高，所分析的基础完全依赖于数据。

Lasso 回归的参数估计方法：$\min \sum_{t=1}^{n} \{y_t - \beta_0 + \beta_1 x_{t1} + \cdots + \beta_a x_{ta}\}^2 + \alpha \sum_{j=1}^{d} |\beta_j|$。Lasso 回归会将不显著的变量的系数压缩至 0，惩罚力度$\alpha$越大，减少的变量越多，Lasso 回归可以起到降维的目的。

Ridge 回归的参数估计方法：$\min \sum_{t=1}^{n} \{y_t - \beta_0 + \beta_1 x_{t1} + \cdots + \beta_a x_{ta}\}^2 + \alpha \sum_{j=1}^{d} \beta_j^2$。随着惩罚力度$\alpha$的增大，Ridge 回归使得预测变量的系数收缩至 0 但不会变成 0 [9]。

### 3.3. 随机性处理

非参数回归模型具有两处随机性：一是对训练集和测试集的抽取具有随机性；二是交叉验证对 lambda 的选择具有随机性，导致每次的结果都会存在些许偏差。

为了克服样本抽取的随机性，采用 Bootstrap 法，对总样本重复随机替换抽样，生成一系列 Lasso 回归模型和 Ridge 回归模型，比较所拟合的 Ridge 和 Lasso 回归模型预测能力的差值，记为“VS”。若 VS 置信区间是异于 0 的，则表明这两种模型的预测效果存在显著差异。若预测区间小于 0 的，则表明 Ridge 回归模型的预测效果比 Lasso 回归模型的预测效果要好，若预测区间大于 0，则反之。

为了克服交叉验证的随机性导致每次的最优 lambda 的值不同，采用循环语句，运行 1000 次交叉验证的结果，产生 1000 个不同的 Lasso 回归和 Ridge 回归，比较 Lasso 回归和 Ridge 回归的预测能力，并将结果储存在“value”中。若 value 置信区间是异于 0 的，则表明这两种模型的预测效果存在显著差异。若预测区间小于 0 的，则表明 Ridge 回归模型的预测效果比 Lasso 回归模型的预测效果要好，若预测区间大于 0，则反之。

## 4. 数据来源及预处理

### 4.1. 数据来源

本文的数据来源于哈里森和鲁宾菲尔德的波士顿房价数据，共有 506 个观测值，其中有 14 个变量，分别为城镇人均犯罪率(CRIM)、住宅用地超过 25000 平方英尺的比例(ZN)、城镇非零售商业用地比例(INDUS)、查尔斯河哑变量(CHAS)、一氧化氮浓度(NOX)、每个住宅的平均房间数(RM)、1940 年以前建造的自住房屋的比例(AGE)、到波士顿五个中心区域的加权距离(DIS)、径向公路通达性指数(RAD)、每

10,000 美元的全值财产税率(TAX)、城镇师生比例(PTRATIO)、1000(Bk − 0.63)$^2$、Bk 是按城镇划分的黑人比例(B)、人口中地位低下者的比例(LSTAT)、自有住房的中位数价值(以 1000 美元计)(MEDV)，其中 MEDV 为响应变量，其余为预测变量。

### 4.2. 数据预处理

由于 ZN、AGE、TAX、B 四个变量的方差较大，数据的波动性较大，为了不改变变量的变化趋势，分别对这四个变量取对数，其中，由于 ZN 的范围是从 0~100，对 0 取对数是无意义的，因此对 ZN 处理方式为“ZN = log(ZN + 1)”，同样不改变 ZN 的变化趋势。取对数之后的四个变量，变化趋势相对之前较为平稳。

为了建模的需要，将数据划分为训练集和测试集。为了验证方便，保证所取的训练集和测试集都是一样的，设立种子。随机抽取总观测值的三分之二的数据作为训练集，三分之一的数据作为测试集。

## 5. 多元线性回归模型

### 5.1. OLS 回归

#### 5.1.1. 拟合 OLS 回归模型

使用逐步回归，对回归子集进行选择。本文使用 leaps()函数实现逐步回归，对模型的选择依据的是调整后的 $R^2$，调整后的 $R^2$ 值越大即预测变量解释响应变量的程度越大。根据调整后的 $R^2$ 大小，选择调整后的 $R^2$ 最大的值为 0.7233854。其对应的 OLS 回归(1)模型为如表 1 所示，由于 ZN 和 AGE 的系数不显著，剔除这两个变量，重新拟合 OLS 回归(2)模型，如表 1 所示。虽然调整后的 $R^2$ 值减小了，但是只减少了 0.0011，减少幅度不大。则拟合的 OLS 回归模型为：

**Table 1.** The coefficients and significance of OLS regression (1) and (2) models
**表 1.** OLS 回归(1)、(2)模型系数及其显著性

| | OLS 回归(1)系数 | OLS 回归(2)系数 |
|---|---|---|
| 截距 | 52.57004*** | 50.04133*** |
| CRIM | −0.11131** | −0.10626** |
| ZN | 0.33454 | . |
| CHAS | 2.73477* | 2.77260* |
| NOX | −19.25825*** | −20.87625*** |
| RM | 4.32148*** | 4.34167*** |
| AGE | −0.79081 | . |
| DIS | −1.54062*** | −1.26070*** |
| RAD | 0.27064*** | 0.27666*** |
| TAX | −3.96964** | −3.73767** |
| PTRATIO | −1.00185*** | −1.10388*** |
| B | 1.38320* | 1.36656* |
| LSTAT | −0.47606*** | −0.48774*** |
| 调整后的 $R^2$ | 0.7234 | 0.7223 |

Signif. codes: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1

$$\widehat{\text{MEDV}} = 50.04133 - 0.10626 * \text{CRIM} + 2.77260 * \text{CHAS} - 20.87625 * \text{NOX} + 4.34167 * \text{RM} - 1.26070 * \text{DIS} + 0.27666 * \text{RAD} - 3.73767 * \text{TAX} - 1.10388 * \text{PTRATIO} + 1.36656 * \text{B} - 0.48774 * \text{LSTAT}$$

Adjusted $R$-squared = 0.7223

### 5.1.2. OLS 回归统计假设检验

(1) 从图 1 Normal Q-Q 上可以看出，标准化残差散点大部分都没有落在 45˚角的直线上，且散点双侧严重偏离直线，违反了 OLS 回归残差服从正态性的假设。

(2) 从图 1 Residuals vs Fitted 上可以看出，残差值和拟合值有明显的曲线关系，这说明残差项里面还存在着未被提取出来的与拟合值线性相关的变量，即违反了多元线性回归自变量和因变量线性相关的假设。

(3) 从图 1 Scale-Location 上可以看出，残差方差随着拟合值水平的变化而变化，标准化残差散点并不是随机分布的，即该多元线性回归违反了同方差性。

(4) 使用 Durbin-Watson 检验函数检验残差的序列相关性，检验结果显示 D-W 统计量的值为 1.742645，p 值为 0.024。在 5%的显著性水平下，拒绝残差值之间相互独立的原假设，残差值之间是相关的，违反了多元线性回归残差独立性的统计假设。

使用 OLS 回归来预测房价固然其预测误差较小，但是其实都是经不起推敲的。从上述 OLS 回归统计假设的检验可以看出，其实现实中的数据并不符合 OLS 回归的统计假设，因此相比 OLS 回归更适合使用非参数回归拟合模型来预测房价。

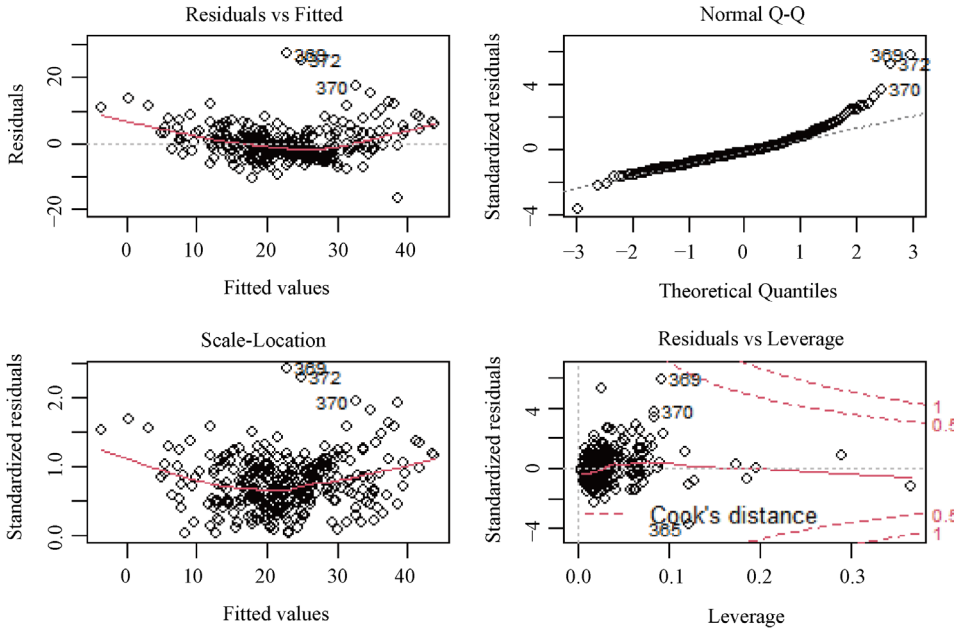


**Figure 1.** OLS regression diagnosis chart
**图 1.** OLS 回归诊断图

## 5.2. 非参数回归

### 5.2.1. Lasso 回归

为了比较模型之间预测能力的优劣，将数据集划分为训练集和测试集，在测试集上拟合模型，并利

用测试集预测变量来检验模型的预测能力。cv.glmnet()函数自带交叉检验的功能，用 cv.glmnet()函数选择拟合 Lasso 回归来确定最优的 lambda (惩罚力度)值，交叉验证结果如图 2 所示，最优的 log(lambda)值在(−1, 0.5)区间内，即最优 lambda 值在(0.368，1.649)区间内。由“lambda_Lasso = Lasso.regression$lambda.1se”选择误差在最小值的 1 个标准误差内的 lambda 值，则最优的 lambda 的值为 0.451622。使用最优的 lambda，用 glmnet()函数确定最终的 Lasso 回归，结果如图 3 所示。Lasso 回归将 ZN、INDUS、AGE、RAD 的系数压缩至 0。对所建立的 Lasso 回归的预测能力进行模型评估，得到的 MSE 为 23.42612。该 Lasso 回归所得房价的预测值和实际值误差较小。

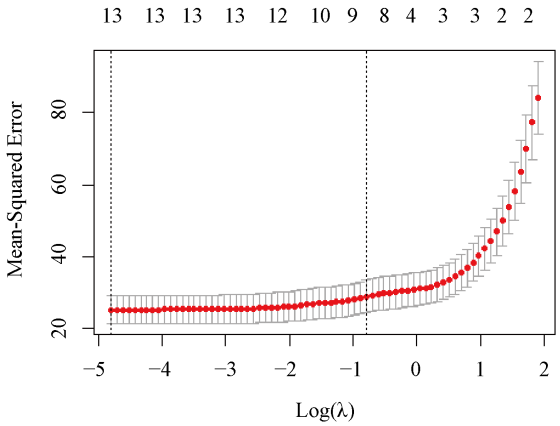


**Figure 2.** Model error of different lambda values based on Lasso regression
**图 2.** 基于 Lasso 回归不同 lambda 值所对应的模型误差

```
> coef(model.lasso)
14 × 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  9.07926404
CRIM        −0.02524566
ZN           .
INDUS        .
CHAS         1.49303144
NOX         −2.06321252
RM           4.77857424
AGE          .
DIS         −0.25290880
RAD          .
TAX         −0.25941895
PTRATIO     −0.69165505
B            1.01881490
LSTAT       −0.47609211
```

**Figure 3.** Coefficients of predictive variables in Lasso regression
**图 3.** Lasso 回归预测变量系数值

### 5.2.2. Ridge 回归

与 Lasso 回归一样，使用十折交叉验证法确定最优的 lambda 值，十折交叉验证结果显示(如图 4)，最优的 lambda 值在(2.718，7.389)区间。最终提取最优的 lambda 值为 5.19255。根据最优的 lambda 值确定最终的 Ridge 回归模型，大部分响应变量的系数都被压缩至接近于 0，结果如图 5 所示。Ridge 回归将变量 CRIM、ZN、INDUS、RAD 等系数压缩至接近于 0，但是并不等于 0，因此 Ridge 回归适用于研究不希望去掉当中的任何一个变量。对所建立的 Ridge 回归进行模型评估，得到的 MSE 为 24.30997。

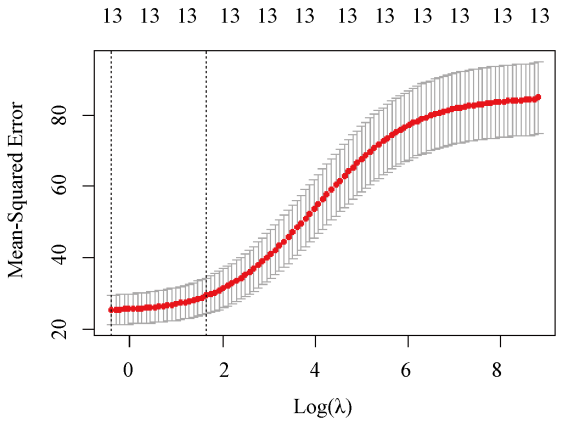


**Figure 4.** Model error of different lambda values based on Ridge regression
**图 4.** 基于 Ridge 回归不同 lambda 值所对应的模型误差

```
> coef (model. ridge)
14 × 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  22.46709791
CRIM         −0.06549373
ZN            0.13129784
INDUS        −0.06659681
CHAS          2.84929501
NOX          −5.18858015
RM            3.70174826
AGE          −0.43436444
DIS          −0.44307234
RAD           0.02859241
TAX          −1.30534525
PTRATIO      −0.64760958
B             1.26121699
LSTAT        −0.31591274
```

**Figure 5.** Coefficients of predictive variables in Ridge regression
**图 5.** Ridge 回归预测变量系数值

## 5.3. 模型选择

在上述分析中，Lasso 回归的预测能力为 23.4255 比 Ridge 回归的小，但是仔细观察发现，其实两个值之间相差不大。为了选择预测能力最强的模型，以下将对模型进行选择。

从图 6 上看，VS 不呈正态分布。它的置信区间通过代码“boot.ci(results, type=c("perc", "bca"))”，结果显示，使用 Percentile 方法生成的置信区间为(0.859，2.732)，用 BCa 方法生成的置信区间为(0.570, 2.119)。两种方法的置信区间都表明，Ridge 回归模型的预测效果比 Lasso 回归模型的预测效果存在差异，且 Lasso 回归模型的预测效果相较之下更好一点。

比较多次交叉验证所得出的不同 lambda 值，实验结果显示，“value”值显著异于 0，即表示 Lasso 回归和 Ridge 回归在预测能力上存在显著性差异，95%的置信区间为(0.1341764, 0.3513815)，表明 Ridge 回归的预测误差大于 Lasso 回归的预测误差，即 Lasso 回归的拟合效果更好。

综上所述，为了避免样本随机性和交叉验证带来的随机性，对 Lasso 回归和 Ridge 回归的预测能力进行比较分析。实验结果均表明，Lasso 回归的拟合效果更好且不受随机性的影响。

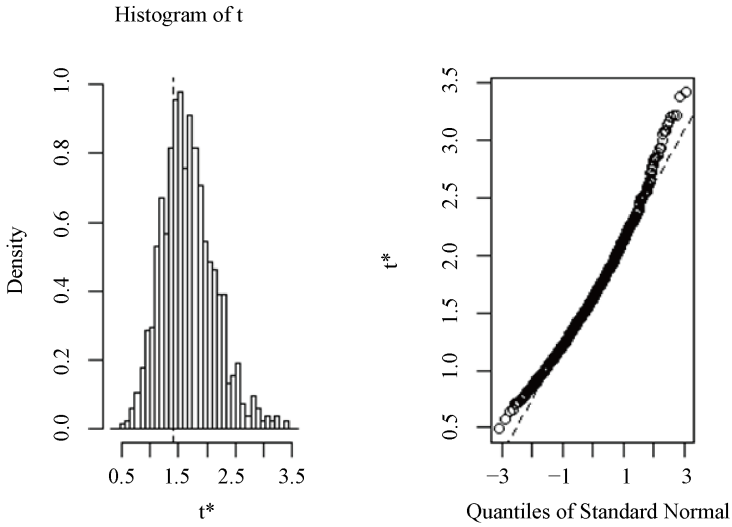


**Figure 6.** The mean value of MSE difference (VS) between Ridge and Lasso regression obtained by self-help method
**图 6.** 自助法所得的 Ridge 与 Lasso 回归 MSE 差(VS)的均值

## 5.4. 模型的拟合效果

通过上述分析，Lasso 回归的拟合效果更好，因此本文将采取 Lasso 回归对模型进行拟合。为了更好的拟合模型，采取全部的数据重新拟合 Lasso 回归，交叉验证选择的最优惩罚力度为 0.4564，则拟合的 Lasso 回归模型为：

$$\begin{aligned}\widehat{\text{MEDV}} = {} & 16.92617 - 0.00816 * \text{CRIM} + 1.66474 * \text{CHAS} + 4.26095 * \text{RM} \\ & - 0.15784 * \text{DIS} - 0.79626 * \text{TAX} - 0.71296 * \text{PTRATIO} \\ & + 0.66186 * \text{B} - 0.52029 * \text{LSTAT}\end{aligned}$$

其拟合效果如图 7。图中的散点为响应变量真实值和拟合值的差，除了个别异常值，图中散点基本在 0 处上下波动，拟合效果是很不错的。

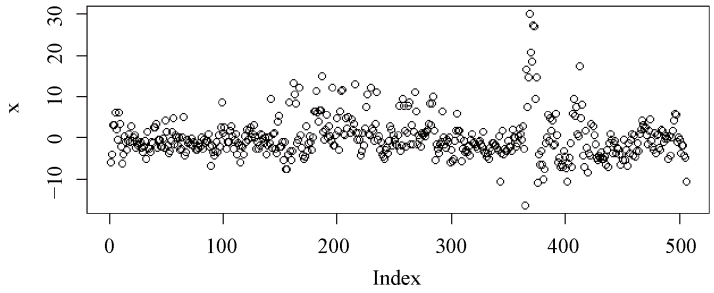


**Figure 7.** The fitting effect of Lasso regression
**图 7.** Lasso 回归的拟合效果

## 6. 结论

本文通过探讨 OLS 回归在房价中的应用，发现其受到诸多的限制，由于现实中的数据并非是理想化的，因此使用 OLS 回归对房价进行预测极易违反满足 OLS 回归成立的统计假设。而非参数回归不会受到总体分布情况的影响，其拟合的模型完全取决于数据。考虑其拟合的结果可能会受到训练集和交叉验

证结果随机性的影响，采取 Bootstrap 法和循环法对其进行验证，并选定最终的模型，其预测效果和真实效果相差不大。因此，选择非参数回归对预测房价是可取的。

## 致　谢

感谢本文撰写期间给出指导的师哥，感谢授课经济计量、非参数统计和 R 软件的老师们，感谢参考文献中的作者们给予的允许转载和引用权的资料，感谢提出本文研究思想的学者们。

## 参考文献

[1] 易成栋, 任建宇, 高璇. 房价、住房不平等与居民幸福感——基于中国综合社会调查 2005、2015 年数据的实证研究[J]. 中央财经大学学报, 2020(6): 105-117.

[2] 周佳琪, 金百锁. 基于空间网络自回归变点模型的合肥市房地产价格影响因素分析[J]. 中国科学院大学学报, 2020, 37(3): 398-404.

[3] 薛建谱, 王卫华. 基于均衡模型的我国商品房价格影响因素分析[J]. 统计与决策, 2013(22): 118-121.

[4] 范允奇, 王艺明. 中国房价影响因素的区域差异与时序变化研究[J]. 贵州财经大学学报, 2014(1): 62-67.

[5] Malpezzi, S. (1999) A Simple Error Correction Model of House Prices. *Journal of Housing Economics*, **8**, 27-62. https://doi.org/10.1006/jhec.1999.0240

[6] 邬嘉怡, 王思玉, 史宏炜, 李虎森, 楼凯达, 崔丽鸿. 基于多小波的北京市房屋市场价格的分析预测[J]. 北京化工大学学报(自然科学版), 2019, 46(5): 101-106.

[7] 唐晓彬, 张瑞, 刘立新. 基于蝙蝠算法 SVR 模型的北京市二手房价预测研究[J]. 统计研究, 2018, 35(11): 71-81.

[8] 黄文, 王正林. 数据挖掘-R 语言实战[M]. 北京: 电子工业出版社, 2014: 160-169.

[9] 张守一, 葛新权, 王斌. 非参数回归及其应用[J]. 数量经济技术经济研究, 1997(10): 60-65+87.