**房价预测模型**

**——基于波士顿数据的多变量回归模型**

**摘 要**

在经济发展和城市化进程加速的背景下，探究各类因素对房价影响的研究成为公众关注的热点。针对波士顿房价特性，本文通过分析20世纪70年代中期各城镇非零售商业土地比例、低收入人口比例等13个变量，进行探索性数据分析。接着，构建多元线性回归模型，并对所建立的模型进行诊断，剔除异常值、高杠杆点和强影响点。通过散点图和Q-Q图等检验模型基本假设，对房价预测影响较小的变量进行剔除以优化模型，最终得出结论。此类方法可用于预测国内房价，为购房者提供价格参考。

**关键词：**波士顿房价 多元线性回归 回归诊断 房价预测

**一、问题描述**

1.问题背景

在“房住不炒”的政策环境下，房价问题备受社会关注，成为广大居民关切的核心议题。研究房价的复杂性需要综合考虑经济和非经济因素。为深入了解波士顿地区房价受多种因素影响的情况，本文选用了由Harrison和Rubinfeld教授收集的波士顿地区房价数据，并运用多元线性回归模型进行详尽分析。

2.研究目标

通过波士顿房价数据集的深入研究，该数据集记录了20世纪70年代中期波士顿郊区房价，并包含了多个可能影响房价的关键指标，如城镇人均犯罪率、地区二氧化碳浓度等。通过建立这些指标与房价之间的多元线性回归模型，本研究的目标在于深入探讨各因素对房价的影响。通过分析各指标的权重和关系，揭示不同因素在决定房价方面的相对重要性。通过建立多元线性回归模型，我们还将探讨城镇人均犯罪率、地区二氧化碳浓度等因素对房价的具体影响，并基于模型对房价变化趋势进行预测。这将为制定房地产政策、了解市场动态以及为潜在投资者提供参考提供有力的依据。

3.研究价值

通过对波士顿房价的深入预测研究，我们将更加熟练地掌握多元线性回归分析的方法和步骤。随着波士顿房价预测的成功，我们有望在国内房价预测分析中借鉴相同的方法。此外，本文还致力于探讨实际社会意义，为缓解社会问题、维护房地产经济稳定做出积极贡献。通过深入理解房价形成机制，我们能够为未来的房地产政策制定和社会发展方向提供有实质性的洞察。

**二、变量选择与数据收集**

|  |  |  |  |
| --- | --- | --- | --- |
| 变量 | 符号 | 含义 | 类型 |
|  | MEDV | 自住房的平均房价（千美元） | 连续型 |
|  | CRIM | 城镇人均犯罪率 | 连续型 |
|  | ZN | 大于25,000平方英尺的地块划分为住宅用地的比例 | 连续型 |
|  | INDUS | 每个城镇非零售商业土地的比例 | 连续型 |
|  | CHAS | 查尔斯河虚拟变量（边界是河流则为1，否则为0） | 离散型 |
|  | NOX | 一氧化氮浓度（每千万） | 连续型 |
|  | RM | 每间住宅的平均房间数 | 连续型 |
|  | AGE | 自住房屋在1940年之前建造的比例 | 连续型 |
|  | DIS | 到加州五个就业中心的加权距离 | 连续型 |
|  | RAD | 对径向高速的可达性指数 | 离散型 |
|  | TAX | 每10,00美元的全价物业税 | 连续型 |
|  | PTRATIO | 城镇中学生与教师比例 | 连续型 |
|  | B | 1000(Bk−0.63)2 ，Bk为城镇中黑人比例 | 连续型 |
|  | LSTAT | 低社会阶层人口比例 | 连续型 |

**三、探索性数据分析**

1.描述性分析

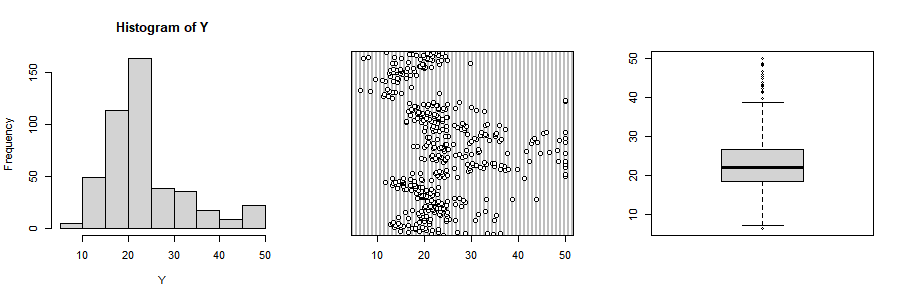
首先通过探索性分析得出各个变量的平均值、标准差、最小值、最大值、下四分位点、中位数、上四分位数，描述性统计见表。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变量 | 最小值 | 下四分位点 | 中位数 | 上四分位数 | 最大值 | 平均值 |
|  | 6.30 | 18.50 | 21.95 | 26.60 | 50.00 | 23.75 |
|  | 0.00000 | 0.04944 | 0.14466 | 0.81962 | 9.96654 | 1.26920 |
|  | 0.00000 | 0.00000 | 0.00000 | 18.10000 | 100.00000 | 13.30000 |
|  | 0.00000 | 3.44000 | 6.96000 | 18.10000 | 27.74000 | 9.20500 |
|  | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 1.00000 | 0.14080 |
|  | 0.385 | 0.449 | 0.538 | 0.647 | 7.313 | 1.101 |
|  | 0.356 | 0.596 | 0.632 | 6.949 | 100.000 | 15.680 |
|  | 1.137 | 3.200 | 6.525 | 8.998 | 100.000 | 5.875 |
|  | 1.130 | 2.431 | 3.926 | 6.332 | 24.000 | 6.173 |
|  | 1.000 | 4.000 | 5.000 | 24.000 | 666.000 | 78.060 |
|  | 20.200 | 254.000 | 307.000 | 403.000 | 711.000 | 339.300 |
|  | 2.600 | 17.000 | 18.900 | 20.200 | 396.900 | 42.620 |
|  | 0.320 | 365.000 | 390.660 | 395.620 | 396.900 | 332.790 |
|  | 1.730 | 6.878 | 10.380 | 15.015 | 34.410 | 11.538 |

2.数据可视化

（1）响应变量

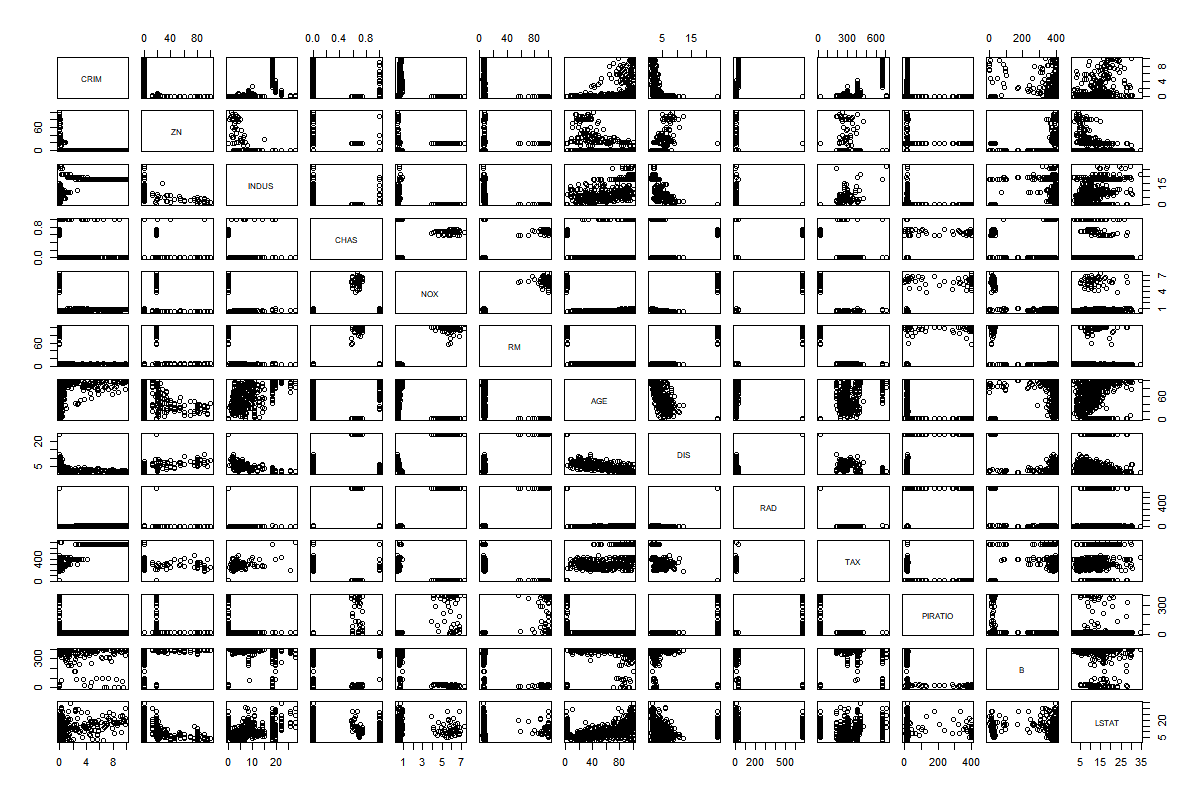
绘制响应变量的直方图、散点图、箱线图如图所示：



从直方图中可以看出，响应变量大致服从正态分布，数据集中分布在10-30之间；从箱线图中可以看出，响应变量的中位数在22附近，数据集中分布在18-26附近。

（2）预测变量

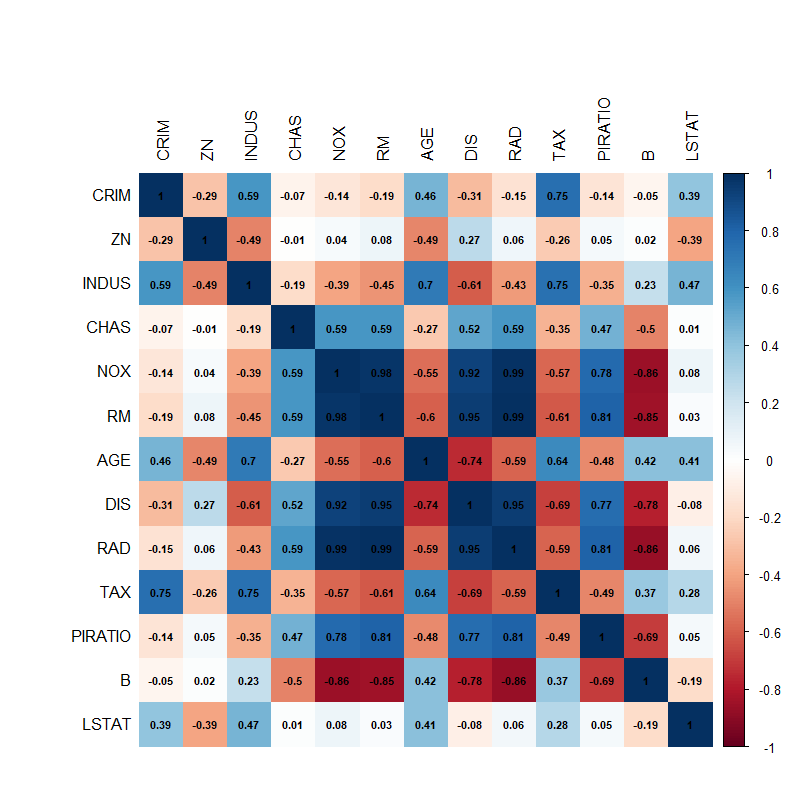
绘制各预测变量与其他预测变量的散点图矩阵如图：



观察此散点图矩阵可知，大部分变量在平面内均呈现随机分布，由于X4为定性变量，取值为0和1，所以X4与其他预测变量之间的散点图分布在0和1，而X6和X7的散点图以及X6和X8的散点图在平面内未成均匀分布，说明变量之间具有一定的相关关系。

（3）相关系数

下面绘制响应变量关于各预测变量的相关系数热力图如图所示：



由图可知，响应变量与预测变量之间均具有一定的相关性，初步判断可以进行回归分析。

**四、模型设定**

多元线性回归是一种广泛应用于统计学的方法，旨在研究两个或更多自变量与因变量之间的关系。通过建立一个线性方程，该方法能够揭示这些变量之间的复杂关联，并提供对因变量的可靠预测。本研究旨在深入探究影响房屋售价的多个因素，包括经济、交通、教育、自然环境、安全和社会等方面。综合文献综述的基础上，我们选择采用多元线性回归模型，以Y表示因变量，X1至X13表示自变量，对波士顿地区的房价进行全面而深入的分析。

这一模型设定具备明显的优势。通过前文描述性统计分析，我们发现部分自变量与因变量之间存在着线性关系。多元线性回归模型结构简明，易于理解和阐释，使我们能够捕捉这些线性关系，更为有效地解释房价的变动。考虑到波士顿地区房价数据集涵盖了多个可能影响房价的指标，采用多元线性回归模型可以对这些指标进行全面综合分析，从而更全面地探讨各种因素对房价的影响。

这一模型设定的科学性和合理性为我们提供了一个有力的工具，用以深入剖析波士顿地区房价的形成机制。通过精心选择自变量，我们可以更准确地理解各个影响因素之间的相互作用，为房地产市场的决策提供科学依据。

**五、模型建立**

1.全模型分析

（1）回归结果

首先我们选取所有预测变量进行多元回归分析，初步了解模型的拟合情况以及各自变量对因变量的解释程度。全模型拟合结果如表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量 | 系数 | 标准误 | t检验 | p值 |
| 常数项 | 20.87 | 5.48 | 3.81 | 0.00 |
|  | -0.24 | 0.22 | -1.06 | 0.29 |
|  | 0.04 | 0.01 | 2.86 | 0.00 |
|  | 0.05 | 0.06 | 0.86 | 0.39 |
|  | 2.44 | 0.83 | 2.94 | 0.00 |
|  | -11.66 | 3.93 | -2.97 | 0.00 |
|  | 5.11 | 0.45 | 11.24 | 0.64 |
|  | -0.01 | 0.01 | -0.47 | 0.00 |
|  | -1.27 | 0.20 | -6.50 | 0.00 |
|  | 0.29 | 0.09 | 3.45 | 0.00 |
|  | -0.01 | 0.00 | -3.15 | 0.00 |
|  | -0.83 | 0.13 | -6.54 | 0.00 |
|  | 0.01 | 0.00 | 3.51 | 0.00 |
|  | -0.52 | 0.06 | -9.07 | 0.00 |

由表可知，除X3、X7、X12外，大部分预测变量的p值均小于0.05，说明对响应变量具有显著影响；且模型Ra2为0.7328，说明模型初步拟合效果较好。

全模型拟合结果为：

1. 结果解读

城镇人均犯罪率（CRIM）对因变量的系数为-0.238025，但p值为0.29>0.10，这意味着在其他变量保持不变的情况下，城镇人均犯罪率对自住房平均房价的影响是不显著的。原因可能是感知和实际安全性的差异。虽然这可能是购房者判断社区安全性的一个指标，但这种感知可能与实际犯罪对社区房价的影响不同。购房者可能在评估社区时考虑的其他因素，如社区服务、教育资源和邻里关系等，都可能在一定程度上抵消了人均犯罪率对房价的潜在负面影响。

地块划分为住宅用地的比例（ZN）对因变量的系数为0.038221，p值<0.01，这说明在其他条件不变的情况下，地块划分为住宅用地的比例对自住房平均房价有显著的正向影响。这可能是因为土地用途的合理配置，较高比例的住宅用地划分可能代表城市规划中对居住区的合理规划，即更多的土地用于住宅建设，可能带来更好的社区环境，如公园、绿地等，提升了居住区的舒适性和吸引力，从而推动了房价的上升。

非零售商业土地的比例（INDUS）对因变量的系数为0.051356，p值为0.39>0.10，这表明在其他条件不变的情况下，非零售商业土地的比例对自住房平均房价的影响可能不是统计上显著的。这可能是由于非零售商业土地反映了社区的整体规划和用地结构，而购房者更可能关注居住环境，例如邻里关系、安全性和配套服务，而非零售商业土地的比例在购房者决策中可能相对次要。

查尔斯河（CHAS）对因变量的系数为2.435048，p值<0.01，这表示查尔斯河是否流经对自住房平均房价有显著正向影响。这可能反映了河流存在对房价产生影响的特殊经济机制。河流通常提供了宜人的自然景观，具有观赏价值，具备水景的房产可能被认为更具吸引力，因为它们提供了独特的居住体验。这种独特性可能诱发了购房者对河边房产的高度需求，从而推动了沿河房价的上升。

一氧化氮浓度（NOX）对因变量的系数为-11.657986，p值<0.01，这说明一氧化氮浓度对自住房平均房价有显著负向影响。这可能反映了环境质量对房地产市场的影响。较高的一氧化氮浓度通常与环境污染相关，购房者往往对居住环境的质量非常敏感，而高浓度的一氧化氮可能影响居住者的健康和生活品质，因此可能减弱了对这些地区房产的需求，进而导致房价下跌。

房每间住宅的平均房间数（RM）对因变量的系数为5.110158，p值<0.01，这表示每间住宅的平均房间数对自住房平均房价有显著正向影响。可能反映了房屋内在特性对房价的影响。较多的房间数通常意味着更大的住房空间，提供更多的功能区域。这可能提高了住房的居住舒适性和生活质量，从而增加了购房者对这类住房的需求，因此购房者更愿意支付更高的价格以获得更多的房间和空间。

房屋建造年份（AGE）对因变量的系数为-0.006094，p值为0.64>0.10，这说明在其他条件不变的情况下，房屋建造年份对自住房平均房价的影响可能不是统计上显著的。这可能是由于屋建造年份可能受到市场对老房屋价值的认知的影响。购房者可能更注重房屋的实际状态、维护情况和现代化程度，而不仅仅是建造年份。如果市场认为老房屋具有历史价值、独特设计或经过有效翻新后可以继续安全居住，那么建造年份的影响就不会显著。

到就业中心的距离（DIS）对因变量的系数为-1.271514，p值<0.01，这表示到就业中心的距离对自住房平均房价有显著负向影响。这可能反映了地理位置与就业中心距离的经济机制。较远的距离意味着更高的通勤成本，包括时间和金钱，购房者通常愿意支付更高的房价以换取更便捷的通勤体验。而较远的距离可能削弱了购房者对这些地区房产的支付意愿，从而对房价产生了负面影响。

对径向高速的可达性指数（RAD）对因变量的系数为0.294444，p值<0.01，这说明对径向高速的可达性对自住房平均房价有显著正向影响。对自住房平均房价的显著正向影响可能反映了交通可达性对地产市场的影响。对径向高速的良好可达性通常代表了便捷的交通网络，使得居民更容易快速、高效地到达城市中心或其他主要区域。这种便捷性可以提高购房者对该地区房产的兴趣，从而支持较高的房价水平。

每10,00美元的全价物业税（TAX）对因变量的系数为-0.011360，p值<0.01，这表示每10,00美元的全价物业税对自住房平均房价有显著负向影响。对自住房平均房价的显著负向影响可能反映了物业税对地产市场的经济影响。高物业税增加了购房者的总体住房成本，因为购房者需要支付更高的税款。另外高额物业税可能降低了房产的投资回报率，因为税收占据了总收益的较大部分，投资者可能会寻求更有利可图的投资机会。这些原因都会导致对该地区房产的需求减少，从而对房价产生负面影响。

城镇中学生与教师比例（PTRATIO）对因变量的系数为-0.831030，p值<0.01，这说明学生与教师比例对自住房平均房价有显著负向影响。这可能是由于教育资源与房地产市场之间的复杂关系所导致的。学生与教师比例被认为是评估学校教育质量的一个重要指标。较高的学生与教师比例可能被视为教育资源分配不足，从而影响了购房者对该地区学区的认可度。购房者通常愿意支付更高的价格以获得更好的教育资源，而较高的学生与教师比例可能减弱了该社区对家庭购房者的吸引力，进而减少了购房者对该地区房产的支付意愿，最终对房价产生了负面影响。

城镇中黑人比例（B）对因变量的系数为0.012314，p值<0.01，这表示城镇中黑人比例对自住房平均房价有显著正向影响。这可能是因为可能高黑人比例社区具有更为包容和文化丰富的特点，并且反映了社区的社会和谐性，即不同族裔群体之间的良好关系。使得购房者更愿意在这样的社区中购房，从而支持了房价水平的上升。

低社会阶层人口比例（LSTAT）对因变量的系数为-0.520753，p值<0.01，这说明低社会阶层人口比例对自住房平均房价有显著负向影响。这可能是因为较高的低社会阶层人口比例可能与社区服务和设施的限制相关，购房者通常希望购房区域提供高质量的学校、医疗服务和公共设施，而低社会阶层人口比例较高可能意味着较差的社区服务，从而降低了购房者对该地区房产的支付意愿。

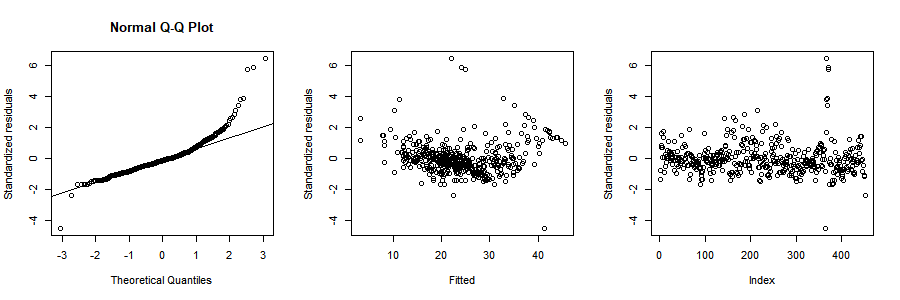
2.回归诊断

（1）假设验证

将全模型拟合得到的标准化残差按从大到小的顺序排序后，绘制标准化残差的正态概率图——Q-Q图，由图可知改散点图近似于一条截距为0、斜率为1的直线，这个结果表明残差服从标准正态分布，此模型满足正态性假定。

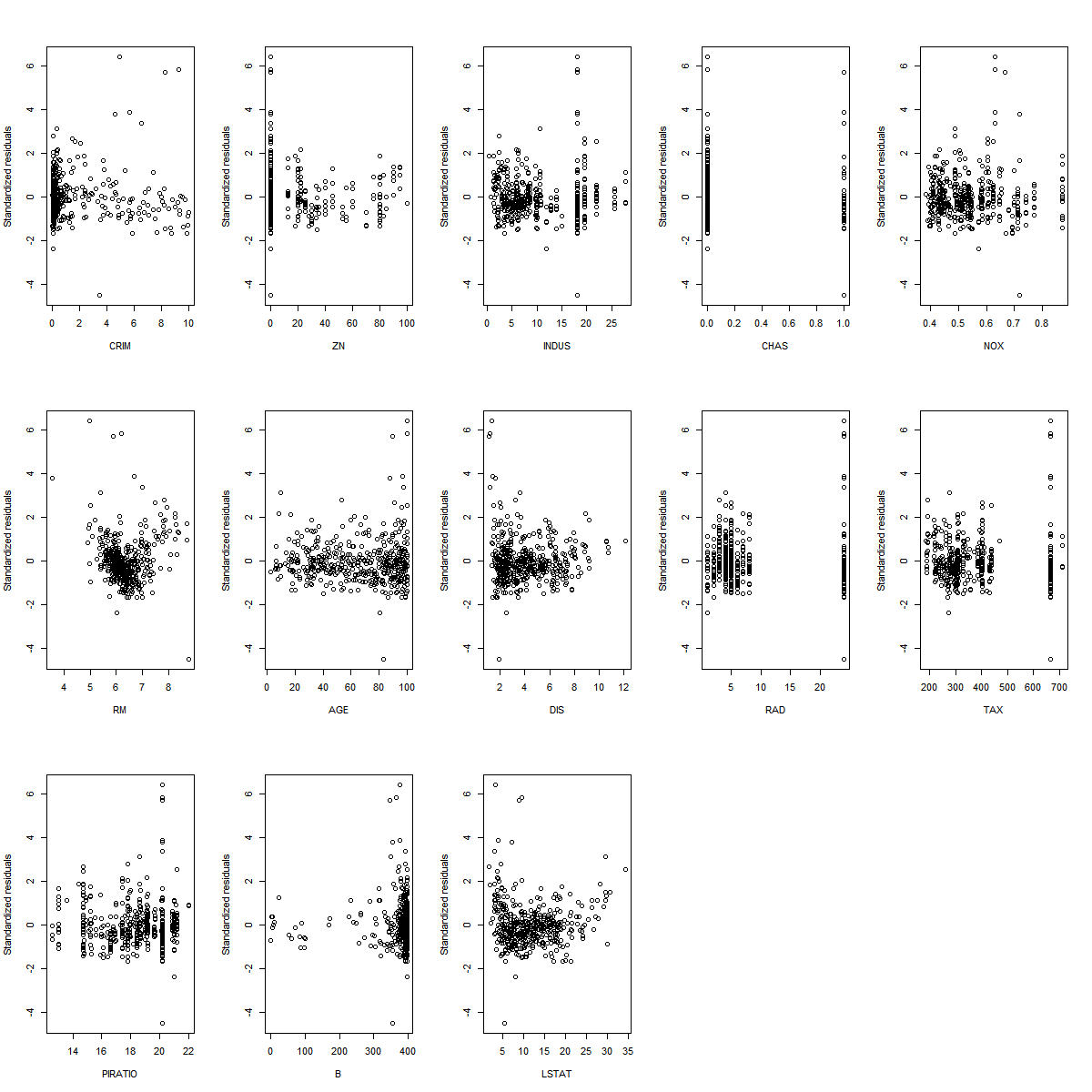
绘制标准化残差关于拟合值的散点图，图中的点呈现随机分布，说明标准化残差与拟合值之间不相关。

绘制标准化残差序列图，由图可知残差在0附近随机分布，说明标准化残差符合独立性假定。



（2）标准化残差关于每个自变量的散点图

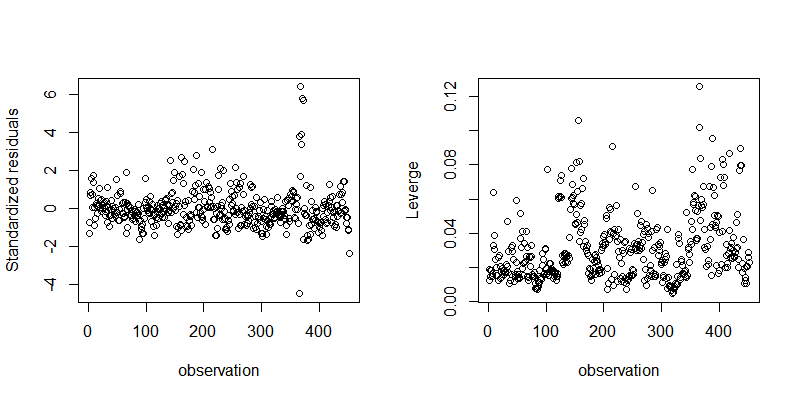
绘制标准化残差关于每个自变量的散点图，检验标准化残差与自变量之间的相关性。



由图可知，各自变量之间的标准化残差都大致服从随机分布，证明线性假定成立。

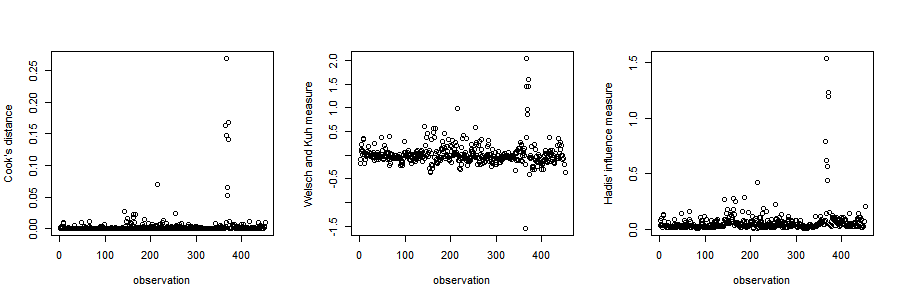
（3）异常点以及高杠杆点识别

绘制标准化残差的序列图以及杠杆值序列图。



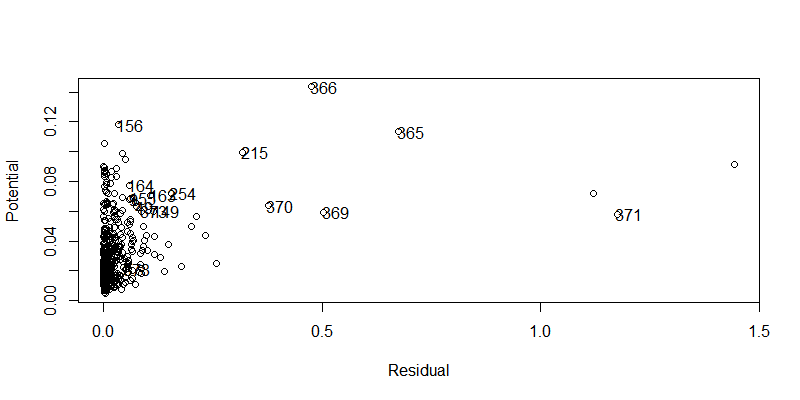
（4）强影响点识别

由于强影响点会对模型的参数估计产生较大影响，因此本文采用三种强影响点判别方法分别是Cook距离、Welsch-Kuh度量和Hadi影响度量来综合判断强影响点。



（5）残差位势图

采用残差位势图来判断观测值中是否含有异常点或者高杠杆点，或者两者都是的点。由图可知点371为异常点，点149、156、163、164、254为高杠杆值点，点215、365、366、369、370既为异常点也是高杠杆值的点。



综合使用上述多种判别指标和方法，我们最终筛选出以下18个强影响点：

9,49,149,155,156,163,164,215,254,365,366,369,370,371,373,378,467,491

（6）自变量的显著性检验

为了检验异常点的存在是否对之前的全模型回归结果的显著性产生影响，我们在剔除所有强影响点后，使用剩余样本继续对全魔性的所有变量进行t检验，判断其系数是否在95%的置信水平下具有显著性。所得结果如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量 | 系数 | 标准误 | t-值 | p-值 |
| 常数项 | 0.32 | 4.70 | 0.07 | 0.95 |
|  | -0.42 | 0.18 | -2.30 | 0.02 |
|  | 0.03 | 0.01 | 2.61 | 0.01 |
|  | 0.04 | 0.05 | 0.79 | 0.43 |
|  | 1.08 | 0.72 | 1.49 | 0.14 |
|  | -4.16 | 3.23 | -1.29 | 0.20 |
|  | 7.27 | 0.42 | 17.43 | 0.00 |
|  | -0.03 | 0.01 | -3.04 | 0.00 |
|  | -1.03 | 0.16 | -6.53 | 0.00 |
|  | 0.24 | 0.07 | 3.47 | 0.00 |
|  | -0.01 | 0.00 | -4.15 | 0.00 |
|  | -0.72 | 0.10 | -7.16 | 0.00 |
|  | 0.01 | 0.00 | 4.13 | 0.00 |
|  | -0.29 | 0.05 | -5.34 | 0.00 |

由表可知，变量X3、X4、X5的p值均大于0.05，在95%置信水平下认为其回归系数不显著，因此删去显著性检验不显著的自变量。

**六、简化模型**

1.简化模型结果

在去除全模型回归中的不显著变量，并剔除所有强影响点后，再次进行回归，所得简化模型的拟合效果如表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 变量 | 系数 | 标准误 | t-值 | p-值 |
| 常数项 | -2.36 | 3.85 | -0.61 | 0.54 |
| X1 | -0.48 | 0.18 | -2.67 | 0.01 |
| X2 | 0.03 | 0.01 | 2.66 | 0.01 |
| X6 | 7.32 | 0.41 | 18.04 | 0.00 |
| X7 | -0.04 | 0.01 | -3.51 | 0.00 |
| X8 | -1.01 | 0.14 | -7.09 | 0.00 |
| X9 | 0.25 | 0.07 | 3.68 | 0.00 |
| X10 | -0.01 | 0.00 | -4.69 | 0.00 |
| X11 | -0.69 | 0.09 | -7.30 | 0.00 |
| X12 | 0.01 | 0.00 | 4.27 | 0.00 |
| X13 | -0.28 | 0.05 | -5.20 | 0.00 |

由回归结果表可知，部分系数大小发生微小变化；此外简化模型的R2a为0.8173，相较于最初的全模型回归的R有显著提升，拟合效果有较大改善，说明在剔除强影响点并删去不显著自变量后，模型预测更为精准。

因此简化模型的表达式为：

2.简化模型分析

计算三次回归模型的AIC、BIC和SBC，如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | AIC | BIC | SBC |
| lm1 | 2668.788 | 2730.494 | 2730.494 |
| lm2 | 2344.680 | 2405.776 | 2406.385 |
| lm3 | 2343.012 | 2391.889 | 2392.376 |

可以发现简化模型（lm3）的AIC、BIC和SBC值均小于全模型（lm2），说明简化模型与全模型相比是一个更优的模型。

3.简化模型的F检验

接下来对模型进行F检验，比较全模型和简化模型对数据的拟合效果。

由检验结果可知不能拒绝原假设，因此我们认为简化模型是恰当的。

**七、结论**

通过对房价影响因素的深入回归模型分析，我们得出结论，房价的波动受到多方面因素的综合影响。这些因素跨足经济、交通、教育、自然环境、安全、社会等多个领域，呈现出复杂而多元的关联。在此基础上，我们提出一系列结论和建议，旨在促进房地产市场的健康发展和社会全面进步。

首先，房价的合理稳定需要综合施策。在经济、交通、教育、环境等多方面共同努力下，我们可以加强政策调控，实施限购、限贷等政策以降低市场投机性需求；增加对保障性住房的投入，以满足低收入群体的基本住房需求；通过优化城市规划，提高城市交通效率，改善生活品质，从而影响房价的稳定。

其次，打击炒房行为是确保市场健康发展的关键。严厉打击违法违规行为，如捂盘惜售、虚假宣传等，通过完善监管体系和法制建设，确保市场秩序的良性发展，防止不良投机行为对市场造成干扰。

此外，我们应该借鉴国内外成功经验，结合我国实际情况，制定更为切实可行的调控政策。例如，可以学习美国的房地产税制度，通过税收手段引导市场行为；借鉴新加坡的组屋制度，通过大力推动保障性住房建设，满足广大市民的基本居住需求。

展望未来，通过这些综合性的措施，我们有望实现房地产市场的健康有序发展，提高人民的生活幸福感和满意度，为全面建设社会主义现代化强国奠定坚实基础。在不断总结经验教训的过程中，我们能够更好地应对市场变化，不断推动住房市场的良性循环，实现经济和社会的双赢。