**波士顿房价及其影响因素案例分析**

**一、背景描述及数据集介绍**

随着房地产市场的不断发展，房地产业已成为各国国民经济发展的重要支柱，其调整变化将对整个金融市场和社会民生产生较为显著的影响。因此，如何保证房价的稳定性，促进房地产业平稳有序发展成为重要课题。

本案例使用的数据集（Boston House Price Dataset）源自20世纪70年代中期美国人口普查局收集的美国马萨诸塞州波士顿住房价格有关信息。该数据集包含506条观测值，统计了当地城镇人均犯罪率、占地面积超过25,000平方英尺的住宅用地比例等13个指标，第14个指标（即目标变量）给出了住房的中位数报价。具体的变量解释如下：

表1 数据集变量名及解释对应表

|  |  |
| --- | --- |
| 变量名 | 解释 |
| CRIM | 城镇人均犯罪率  per capital crime rate by town |
| ZN | 占地面积超过25,000平方英尺的住宅用地比例  proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | 每个城镇非零售业务的比例  proportion of non-retail business acres per town |
| CHAS | 查尔斯河虚拟变量（如果是河道，则为1;否则为0）  Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| NOX | 一氧化氮浓度（每千万份）  nitric oxides concentration (parts per 10 million) |
| RM | 每间住宅的平均房间数  average number of rooms per dwelling |
| AGE | 1940年以前建造的自住房的年龄比例  proportion of owner-occupied units built prior to 1940 |
| DIS | 加权到波士顿的五个就业中心的距离  weighted distances to five Boston employment centers |
| RAD | 径向高速公路的可达性指数  index of accessibility to radial highways |
| TAX | 每10,000美元的全额物业税率  full-value property-tax rate per 10,000 USD |
| PTRATIO | 城镇的学生与教师比例  pupil-teacher ratio by town |
| BK | 1000(Bk - 0.63) ²，其中Bk是城镇黑人的比例  1000(Bk — 0.63)² where Bk is the proportion of blacks by town |
| LSTAT | 人口状况下降的百分比  lower status of the population |
| MEDV | 自有住房的中位数报价, 单位1000美元  median quotation of owner-occupied units |

**二、描述性统计分析**

首先对数据作简要描述，具体分析结果如表2所示，其中的主要结论如下：

表2 各变量的描述分析

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 观测数 | 均值 | 众数 | 标准差 | 最大值 | 中位数 | 最小值 |
| CRIM | 506 | 3.61 | 0.02 | 8.60 | 88.98 | 0.26 | 0.01 |
| ZN | 506 | 11.36 | 0.00 | 23.32 | 100.00 | 0.00 | 0.00 |
| INDUS | 506 | 11.14 | 18.10 | 6.86 | 27.74 | 9.69 | 0.46 |
| CHAS | 506 | 0.07 | 0.00 | 0.25 | 1.00 | 0.00 | 0.00 |
| NOX | 506 | 0.55 | 0.54 | 0.12 | 0.87 | 0.54 | 0.39 |
| RM | 506 | 6.28 | 5.71 | 0.70 | 8.78 | 6.21 | 3.56 |
| AGE | 506 | 60.08 | 0.00 | 32.19 | 99.30 | 67.70 | 0.00 |
| DIS | 506 | 3.80 | 3.50 | 2.11 | 12.13 | 3.21 | 1.13 |
| RAD | 506 | 9.55 | 24.00 | 8.71 | 24.00 | 5.00 | 1.00 |
| TAX | 506 | 408.24 | 666.00 | 168.54 | 711.00 | 330.00 | 187.00 |
| PTRATIO | 506 | 18.46 | 20.20 | 2.16 | 22.00 | 19.05 | 12.60 |
| BK | 506 | 356.67 | 396.90 | 91.29 | 396.90 | 391.44 | 0.32 |
| LSTAT | 506 | 12.65 | 8.05 | 7.14 | 37.97 | 11.36 | 1.73 |
| MEDV | 506 | 22.53 | 50.00 | 9.20 | 50.00 | 21.20 | 5.00 |

**（一）城镇人均犯罪率（CRIM）**

图1 城镇人均犯罪率与房价散点图

表2显示，城镇人均犯罪率的中位数为0.26，结合图1可以得出，多数样本的人均犯罪率在0附近，即多数地区的治安水平较好，且这类样本的房价分布较为均匀。少数样本的犯罪率在1%-10%，该类样本的房价主要集中于5至30千美元，房价整体水平低于治安较好的地区。同时，其中也存在少部分对应房价极高的样本，可视为异常值。另外，存在极少数样本的犯罪率在10%以上，且大致存在房价随犯罪率升高而降低的趋势。

**（二）占地面积大的住宅用地比例（ZN）**

图2 占地面积大的住宅用地比例与房价散点图

结合表2与图2可以得出，ZN变量的众数及中位数均为0，由此可知，大部分地区没有占地面积较大的住宅。而对于存在该类住宅的地区，由于样本量较少，除ZN在80%以上的地区外，房价在各水平下未呈现出明显的差异性和趋势。

**（三）每个城镇非零售业务的比例（INDUS）**

图3 各城镇非零售业务的比例与房价散点图

由表2可知，INDUS的众数为18，即多数样本的城镇非零售业务比例为18%。此外，图3显示，INDUS介于0到4以及8到30之间时，房价的变化不存在明显规律，即较为平稳；INDUS介于4到8之间时，房价水平显示出小幅度的降低。

**（四）查尔斯河虚拟变量（CHAS）**

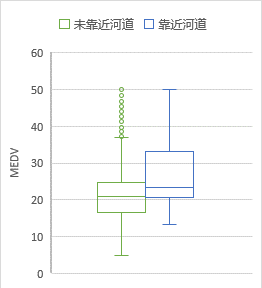


图4 是否靠近查尔斯河的两类样本直方图及对应房价箱线图

图4中，直方图表明大部分地区的住宅都未有查尔斯河流经，仅有35个样本靠近查尔斯河河道。而根据箱线图可以明显得出，靠近河道的样本对应的房价的整体水平及均值高于未靠近河道样本的房价，但同时，未有查尔斯河流经的样本中，存在小部分样本的房价为异常值，即房价水平较高，可能为记录异常或该类样本有其他因素表现较好，弥补了未靠近河道这一缺点。

**（五）一氧化氮浓度（NOX）**

图5 一氧化氮浓度与房价散点图

由表2可知，NOX变量的众数、中位数、均值较为接近，说明数据分布较为集中均匀。图5表明，整体而言，一氧化氮浓度与房价之间表现出，后者随前者的增加而降低的线性趋势。其中，0.4-0.65区间内，存在部分房价水平较高的异常值。此外，可能由于数据量较小的原因，一氧化氮浓度为0.7左右的水平下，房价的整体水平稍有上升，但对整体的线性降低趋势不存在较大影响。

**（六）每间住宅的平均房间数（RM）**

图6 每间住宅的平均房间数与房价散点图

表2及图6表明，RM的众数为5.71，多数样本主要集中于5到7这一区间内。整体来看，除极个别异常值外，房价呈现出随平均房间数增加而升高的趋势。

**（七）1940年以前建造的自住房的比例（AGE）**

图7 1940年以前建造的自住房的比例与房价散点图

由表2可知，AGE变量的众数为0，中位数为67.70，说明存在部分地区的自住房均为1940年后建造，相对而言房屋较新。但同时，也有50%左右的样本的老式住房比例达到七成以上，且根据图7可以发现，该类样本房价随老式自住房比例的增加而缓慢下降。

**（八）加权到波士顿的五个就业中心的距离（DIS）**

图8 加权到波士顿的五个就业中心的距离与房价散点图

由表2可知，DIS变量的大部分数据集中于3左右，而图8表明，对于DIS小于3的样本，整体呈现出房价随距离的增大而增大的趋势，而对于DIS大于3的样本，整体房价水平表现出平稳的态势。这一现象与常识存在一定偏差，推测可能原因在于DIS小于3的样本地区为老城区，老式自住房的比例较高，根据（七）中的分析，老式自住房的房价水平较低，此外，该地区交通条件可能较为良好，距离对房价的反向影响不明显。

**（九）径向高速公路的可达性指数（RAD）**

图9 径向高速公路的可达性指数与房价散点图

由表2可知，RAD变量的众数为24，中位数为5，说明大部分地区的高速公路可达性较好，且结合图9可以得出这类样本的房价水平较为接近且整体来看房价低于可达性指数低的地区。

**（十）每10,000美元的全额物业税率（TAX）**

图10 全额物业税率与房价散点图

由表2可得，TAX的中位数为330，众数为666。结合图10可知，全额物业税率主要集中于200到400之间，该区间内，房价的取值范围较大，但大部分介于20到30之间。该区间之外，全额物业税率高的样本基本取值为666，其对应房价的均值低于税率较低的样本，即高税率对房价产生了抑制作用。

**（十一）城镇的学生与教师比例（PTRATIO）**

图11 城镇的学生与教师比例与房价散点图

由表2可得，大部分地区的生师比为20左右，即大约平均1名老师负责20名学生，样本的最大值为22，最小值为12.6，结合图11可以得出，小部分地区的师资力量显著好于整体的师资力量，其余地区的师资水平差异较小。此外，整体而言，随着城镇学生与教师比例的增大，对应的住房房价水平降低。

**（十二）城镇黑人的比例（BK）**

图12 城镇黑人的比例与房价散点图

由表2及图12可得，大部分样本的BK接近于400，且其房价的取值范围较广，近似呈正态分布。而对于BK小于350的地区，房价整体呈随黑人比例的提升而小幅升高的趋势。

**（十三）人口状况下降的百分比（LSTAT）**

图13 人口状况下降的百分比与房价散点图

由图13可以得出，除0到10区间内极个别异常值外，房价水平整体呈现出随人口状况下降百分比的升高而下跌的线性趋势。

**（十四）自有住房中位数报价（MEDV）**

图14 自有住房中位数报价的直方图

由图14可以得出，MEDV不服从正态分布，且在47-50区间内存在异常值，结合图1至图13的内容分析，考虑删除MEDV为50的样本，得到样本量为490的数据集，对其建立多元线性回归模型。

**三、多元线性回归模型**

**（一）相关系数分析**

表3 各变量相关系数表

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | BK | LSTAT | MEDV |
| CRIM | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ZN | -0.20 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |  |
| INDUS | 0.41 | -0.53 | 1.00 |  |  |  |  |  |  |  |  |  |  |  |
| CHAS | -0.06 | -0.05 | 0.04 | 1.00 |  |  |  |  |  |  |  |  |  |  |
| NOX | 0.42 | -0.51 | 0.77 | 0.09 | 1.00 |  |  |  |  |  |  |  |  |  |
| RM | -0.22 | 0.31 | -0.41 | 0.04 | -0.32 | 1.00 |  |  |  |  |  |  |  |  |
| AGE | -0.04 | -0.38 | 0.34 | 0.07 | 0.35 | -0.01 | 1.00 |  |  |  |  |  |  |  |
| DIS | -0.38 | 0.67 | -0.71 | -0.08 | -0.77 | 0.25 | -0.41 | 1.00 |  |  |  |  |  |  |
| RAD | 0.63 | -0.31 | 0.60 | -0.03 | 0.61 | -0.20 | 0.16 | -0.49 | 1.00 |  |  |  |  |  |
| TAX | 0.58 | -0.30 | 0.72 | -0.07 | 0.67 | -0.28 | 0.20 | -0.53 | 0.91 | 1.00 |  |  |  |  |
| PTRATIO | 0.29 | -0.38 | 0.39 | -0.12 | 0.19 | -0.29 | 0.20 | -0.25 | 0.46 | 0.45 | 1.00 |  |  |  |
| BK | -0.38 | 0.18 | -0.36 | 0.04 | -0.38 | 0.12 | -0.08 | 0.30 | -0.45 | -0.45 | -0.17 | 1.00 |  |  |
| LSTAT | 0.46 | -0.42 | 0.64 | -0.01 | 0.61 | -0.61 | 0.23 | -0.54 | 0.51 | 0.57 | 0.36 | -0.36 | 1.00 |  |
| MEDV | -0.45 | 0.40 | -0.60 | 0.07 | -0.52 | 0.69 | -0.22 | 0.37 | -0.48 | -0.57 | -0.52 | 0.36 | -0.76 | 1.00 |

由表3可知，自有住房的中位数报价与每间住宅的平均房间数的相关系数为0.69，呈较强的正相关性；而其与人口状况下降的百分比的相关系数为-0.76，呈较强的负相关性。

**（二）模型分析**

以MEDV为因变量，其余变量为自变量建立多元线性回归模型：

表4 回归分析结果一

|  |  |
| --- | --- |
| 回归统计 | |
| Multiple R | 0.8847 |
| R Square | 0.7827 |
| Adjusted R Square | 0.7767 |
| 标准误差 | 3.7164 |
| 观测值 | 490 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | SS | MS | F | Significance F |
| 回归分析 | 13 | 23676.7549 | 1821.2888 | 131.8684 | 0.0000 |
| 残差 | 476 | 6574.2329 | 13.8114 |  |  |
| 总计 | 489 | 30250.9878 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 系数 | 标准  误差 | t Stat | P-value | 下限 95.0% | 上限 95.0% |
| Intercept | 31.1059 | 4.0882 | 7.6088 | 0.0000 | 23.0728 | 39.1390 |
| CRIM | -0.1314 | 0.0265 | -4.9545 | 0.0000 | -0.1836 | -0.0793 |
| ZN | 0.0327 | 0.0112 | 2.9345 | 0.0035 | 0.0108 | 0.0547 |
| INDUS | -0.0335 | 0.0493 | -0.6799 | 0.4969 | -0.1303 | 0.0633 |
| CHAS | 0.5050 | 0.7334 | 0.6885 | 0.4914 | -0.9362 | 1.9461 |
| NOX | -13.2037 | 2.9154 | -4.5289 | 0.0000 | -18.9324 | -7.4750 |
| RM | 3.8845 | 0.3530 | 11.0029 | 0.0000 | 3.1908 | 4.5782 |
| AGE | -0.0249 | 0.0062 | -4.0058 | 0.0001 | -0.0371 | -0.0127 |
| DIS | -1.1836 | 0.1534 | -7.7151 | 0.0000 | -1.4851 | -0.8822 |
| RAD | 0.2490 | 0.0524 | 4.7513 | 0.0000 | 0.1460 | 0.3519 |
| TAX | -0.0134 | 0.0030 | -4.5321 | 0.0000 | -0.0193 | -0.0076 |
| PTRATIO | -0.8011 | 0.1048 | -7.6466 | 0.0000 | -1.0069 | -0.5952 |
| BK | 0.0078 | 0.0021 | 3.7180 | 0.0002 | 0.0037 | 0.0120 |
| LSTAT | -0.3680 | 0.0392 | -9.3769 | 0.0000 | -0.4452 | -0.2909 |

由表4可知，模型的拟合优度良好，调整的判定系数为77.67%，整体的检验高度显著，对应值<0.05。相关参数的检验中，INDUS的系数对应的值最大，且INDUS的值为0.4969 > 0.05，未通过检验。因此考虑剔除变量INDUS，即每个城镇非零售业务的比例对房价的中位数报价不具有显著性影响。

对剔除变量INDUS后的其余变量再次建模，得到：

表5 回归分析结果二

|  |  |
| --- | --- |
| 回归统计 | |
| Multiple R | 0.8846 |
| R Square | 0.7825 |
| Adjusted R Square | 0.7770 |
| 标准误差 | 3.7143 |
| 观测值 | 490 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | SS | MS | F | Significance F |
| 回归分析 | 12 | 23670.3704 | 1972.5309 | 142.9801 | 0.0000 |
| 残差 | 477 | 6580.6174 | 13.7958 |  |  |
| 总计 | 489 | 30250.9878 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 系数 | 标准误差 | t Stat | P-value | 下限 95.0% | 上限 95.0% |
| Intercept | 31.2397 | 4.0811 | 7.6547 | 0.0000 | 23.2205 | 39.2590 |
| CRIM | -0.1308 | 0.0265 | -4.9361 | 0.0000 | -0.1828 | -0.0787 |
| ZN | 0.0334 | 0.0111 | 3.0082 | 0.0028 | 0.0116 | 0.0552 |
| CHAS | 0.4712 | 0.7313 | 0.6443 | 0.5197 | -0.9658 | 1.9082 |
| NOX | -13.7537 | 2.7994 | -4.9131 | 0.0000 | -19.2543 | -8.2531 |
| RM | 3.9116 | 0.3506 | 11.1574 | 0.0000 | 3.2227 | 4.6005 |
| AGE | -0.0250 | 0.0062 | -4.0393 | 0.0001 | -0.0372 | -0.0129 |
| DIS | -1.1618 | 0.1499 | -7.7487 | 0.0000 | -1.4564 | -0.8672 |
| RAD | 0.2580 | 0.0507 | 5.0933 | 0.0000 | 0.1585 | 0.3575 |
| TAX | -0.0143 | 0.0027 | -5.2844 | 0.0000 | -0.0196 | -0.0090 |
| PTRATIO | -0.8111 | 0.1037 | -7.8250 | 0.0000 | -1.0148 | -0.6074 |
| BK | 0.0079 | 0.0021 | 3.7477 | 0.0002 | 0.0037 | 0.0120 |
| LSTAT | -0.3705 | 0.0391 | -9.4832 | 0.0000 | -0.4472 | -0.2937 |

由表5可知，模型的拟合优度较前者有所提升，调整的判定系数变为77.70%，整体的检验高度显著，对应值<0.05。相关参数的检验中，CHAS的系数对应的值最大，且CHAS的值为0.5197 > 0.05，未通过检验。因此考虑剔除变量CHAS，即查尔斯河是否流经对房价的中位数报价不具有显著性影响。

对剔除变量CHAS后的其余变量再次建模，得到：

表6 回归分析结果三

|  |  |
| --- | --- |
| 回归统计 | |
| Multiple R | 0.8845 |
| R Square | 0.7823 |
| Adjusted R Square | 0.7773 |
| 标准误差 | 3.7120 |
| 观测值 | 490 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | SS | MS | F | Significance F |
| 回归分析 | 11 | 23664.6436 | 2151.3312 | 156.1316 | 0.0000 |
| 残差 | 478 | 6586.3443 | 13.7790 |  |  |
| 总计 | 489 | 30250.9878 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 系数 | 标准误差 | t Stat | P-value | 下限 95.0% | 上限 95.0% |
| Intercept | 31.2835 | 4.0781 | 7.6712 | 0.0000 | 23.2703 | 39.2966 |
| CRIM | -0.1316 | 0.0264 | -4.9753 | 0.0000 | -0.1836 | -0.0796 |
| ZN | 0.0333 | 0.0111 | 3.0049 | 0.0028 | 0.0115 | 0.0552 |
| NOX | -13.5926 | 2.7865 | -4.8780 | 0.0000 | -19.0679 | -8.1174 |
| RM | 3.9170 | 0.3503 | 11.1830 | 0.0000 | 3.2288 | 4.6053 |
| AGE | -0.0249 | 0.0062 | -4.0225 | 0.0001 | -0.0371 | -0.0127 |
| DIS | -1.1641 | 0.1498 | -7.7705 | 0.0000 | -1.4584 | -0.8697 |
| RAD | 0.2609 | 0.0504 | 5.1731 | 0.0000 | 0.1618 | 0.3600 |
| TAX | -0.0145 | 0.0027 | -5.4052 | 0.0000 | -0.0198 | -0.0092 |
| PTRATIO | -0.8165 | 0.1033 | -7.9074 | 0.0000 | -1.0194 | -0.6136 |
| BK | 0.0079 | 0.0021 | 3.7785 | 0.0002 | 0.0038 | 0.0121 |
| LSTAT | -0.3701 | 0.0390 | -9.4815 | 0.0000 | -0.4468 | -0.2934 |

由表6可知，新模型的拟合优度较前者有所提升，调整的判定系数变为77.73%，整体的检验高度显著，对应值<0.05。在5%的显著性水平下，所有变量的系数均通过检验，因此最终建立的回归方程为

对各变量进行合理性分析，具体解释如下：

**1.对房价有正向影响的变量**

**（1）ZN：占地面积超过25,000平方英尺的住宅用地比例**

ZN对房价的正向影响表明：占地面积较大的住宅比例越高，自有住房的中位数报价越高。主要原因在于占地面积较大的住宅比例高，一般而言该城镇的生活水平也相对更佳，对应的生活服务及环境也更为优质，从而导致住房的中位数报价更高。

**（2）RM：每间住宅的平均房间数**

RM对房价的正向影响表明：每间住宅的平均房间数越多，自有住房的中位数报价越高。每间住宅的平均房间数越多，说明该片区住宅的功能性越强，生活舒适度越高，从而使自有住房的中位数报价更高。

**（3）RAD：径向高速公路的可达性指数**

RM对房价的正向影响表明：径向高速公路的可达性指数越高，自有住房的中位数报价越高。径向高速公路的可达性指数高说明其距离高速公路相对较远，该类区域的城市化水平可能更高，受高速公路的影响较小，从而提高了自有住房的竞争力。此外，该结论与前文描述性分析的结论存在一定的偏差，初步推测可能的原因在于，RAD极高的样本（取值为24）对房价的影响较小，而RAD在0到10区间内的样本，即交通便利度较佳的样本中，距离高速公路越远，对应的房价更高。

**（4）BK：1000(Bk - 0.63) ²，其中BK是城镇黑人的比例**

BK对房价的正向影响表明：城镇中黑人的比例越高，自有住房的中位数报价越高。黑人比例越高的地区，人口密集程度也相对更高，对房产的需求也更高，对应的自有住房的中位数报价也更高。

**2.对房价有负向影响的变量**

**（1）CRIM：城镇人均犯罪率**

CRIM对房价的负向影响表明：城镇人均犯罪率越高，自有住房中位数报价越低。城镇人均犯罪率高表示该地区的社区安全存在较大的隐患，这一特征将极大降低居民的购房意愿，从而反向影响自有住房的中位数报价。

**（2）NOX：一氧化氮浓度（每千万份）**

NOX对房价的负向影响表明：每千万份一氧化氮浓度越高，自有住房中位数报价越低。一氧化氮浓度作为衡量空气质量的重要标准，其数值越高，表明该地区空气质量越差，生活在该地区的居民的健康隐患越大，从而对房价中位数造成负面影响。

**（3）AGE：1940年以前建造的自住房的比例**

AGE对房价的负向影响表明：1940年以前建造的自住房的比例越高，自有住房中位数报价越低。老旧住房的比例高说明该地区住房条件相对较差，且由于需要进行翻新或拆迁工作，房地产商对该地区新住房的建造意愿不强。两种因素共同作用，导致具有老旧住房比例高这一特征的地区的房价更低。

**（4）DIS：加权到波士顿的五个就业中心的距离**

DIS对房价的负向影响表明：加权到波士顿的五个就业中心的距离越远，自有住房中位数报价越低。距离就业中心越远表示居民平均通勤时间更长，对应的该地区居民购房意愿也更低，从而导致自有住房中位数报价降低。该结果与描述性分析的结论存在偏差，初步推断可能是由于DIS数据存在一部分异常值，且DIS大于3的样本间是非平稳的，可能存在负向的影响关系。

**（5）TAX：每10,000美元的全额物业税率**

TAX对房价的负向影响表明：每10,000美元的全额物业税率越高，自有住房中位数报价越低。全额物业税率越高，表明购买该地住房带来的额外生活负担越大，相对应地，居民的购房意愿将降低，从而对房价产生负面影响。

**（6）PTRATIO：城镇的学生与教师比例**

PTRATIO对房价的负向影响表明：城镇的学生与教师比例越高，自有住房中位数报价越低。城镇生师比高表明平均每位教师要负责的学生更多，即师资力量更为不足，学生受教育水平相对更低。对于有孩子的家庭来说，具有该类特征的住房对其的吸引力更低，从而造成房价中位数较低的情况。

**（7）LSTAT：人口状况下降的百分比**

LSTAT对房价的负向影响表明：人口状况下降的百分比越高，自有住房的中位数报价越低。人口状况下降的百分比高说明该地区人口流失的情况更为严重，而人口流失会导致对自有住房需求降低，供需比例增大，从而造成房价中位数降低的结果。

**（三）残差分析**

根据表6对应的模型绘制的残差图如图15所示：

图15 回归分析结果三的残差图

图15表明，残差值与拟合值呈现曲线形态，推断模型可能缺少某个自变量或缺少一个变量的高阶项来解释曲率，即现有变量与因变量MEDV之间可能存在非线性关系，此时考虑引入曲线回归模型。

**四、模型改进**

**（一）模型分析**

通过分析描述性统计中的散点图，可以得出RM与MEDV可能存在二次项相关，LSTAT与MEDV可能存在倒数相关，其他变量与MEDV的相关关系不明显。

重新建立回归模型，在原有模型基础上添加与两项：

表7 回归分析结果四

|  |  |
| --- | --- |
| 回归统计 | |
| Multiple R | 0.9215 |
| R Square | 0.8491 |
| Adjusted R Square | 0.8450 |
| 标准误差 | 3.0969 |
| 观测值 | 490 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | SS | MS | F | Significance F |
| 回归分析 | 13 | 25685.7829 | 1975.8295 | 206.0137 | 0.0000 |
| 残差 | 476 | 4565.2050 | 9.5908 |  |  |
| 总计 | 489 | 30250.9878 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 系数 | 标准  误差 | t Stat | P-value | 下限 95.0% | 上限 95.0% |
| Intercept | 106.7230 | 8.0169 | 13.3123 | 0.0000 | 90.9701 | 122.4758 |
| CRIM | -0.1442 | 0.0221 | -6.5302 | 0.0000 | -0.1876 | -0.1008 |
| ZN | 0.0132 | 0.0095 | 1.3779 | 0.1689 | -0.0056 | 0.0319 |
| NOX | -13.1794 | 2.3256 | -5.6671 | 0.0000 | -17.7490 | -8.6097 |
| RM | -21.9803 | 2.3411 | -9.3889 | 0.0000 | -26.5805 | -17.3801 |
| RM^2 | 2.0088 | 0.1874 | 10.7185 | 0.0000 | 1.6405 | 2.3771 |
| AGE | -0.0116 | 0.0053 | -2.1695 | 0.0305 | -0.0221 | -0.0011 |
| DIS | -0.8673 | 0.1267 | -6.8460 | 0.0000 | -1.1162 | -0.6183 |
| RAD | 0.2045 | 0.0423 | 4.8352 | 0.0000 | 0.1214 | 0.2876 |
| TAX | -0.0119 | 0.0022 | -5.3202 | 0.0000 | -0.0163 | -0.0075 |
| PTRATIO | -0.7160 | 0.0866 | -8.2631 | 0.0000 | -0.8862 | -0.5457 |
| BK | 0.0063 | 0.0018 | 3.6000 | 0.0004 | 0.0029 | 0.0098 |
| LSTAT | -0.2740 | 0.0402 | -6.8222 | 0.0000 | -0.3529 | -0.1951 |
| 1/LSTAT | 20.1864 | 4.3744 | 4.6146 | 0.0000 | 11.5908 | 28.7819 |

由表7可知，新模型的拟合优度有显著提高，调整的判定系数为84.50%，整体的检验高度显著，对应值<0.05。相关参数的检验中，ZN的系数对应的值最大，且ZN的值为0.1689 > 0.05，未通过检验。因此考虑剔除变量ZN，即占地面积超过25,000平方英尺的住宅用地比例对房价的中位数报价不具有显著性影响。

对剔除变量ZN后的其余变量再次建模，得到：

表8 回归分析结果五

|  |  |
| --- | --- |
| 回归统计 | |
| Multiple R | 0.9211 |
| R Square | 0.8485 |
| Adjusted R Square | 0.8447 |
| 标准误差 | 3.0998 |
| 观测值 | 490 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | SS | MS | F | Significance F |
| 回归分析 | 12 | 25667.5731 | 2138.9644 | 222.6039 | 0.0000 |
| 残差 | 477 | 4583.4148 | 9.6088 |  |  |
| 总计 | 489 | 30250.9878 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 系数 | 标准  误差 | t Stat | P-value | 下限 95.0% | 上限 95.0% |
| Intercept | 106.4299 | 8.0216 | 13.2679 | 0.0000 | 90.6678 | 122.1919 |
| CRIM | -0.1425 | 0.0221 | -6.4574 | 0.0000 | -0.1859 | -0.0992 |
| NOX | -13.3964 | 2.3224 | -5.7683 | 0.0000 | -17.9599 | -8.8330 |
| RM | -21.8711 | 2.3420 | -9.3388 | 0.0000 | -26.4730 | -17.2693 |
| RM^2 | 2.0023 | 0.1875 | 10.6771 | 0.0000 | 1.6338 | 2.3708 |
| AGE | -0.0120 | 0.0053 | -2.2419 | 0.0254 | -0.0225 | -0.0015 |
| DIS | -0.7814 | 0.1104 | -7.0774 | 0.0000 | -0.9984 | -0.5645 |
| RAD | 0.1983 | 0.0421 | 4.7113 | 0.0000 | 0.1156 | 0.2810 |
| TAX | -0.0112 | 0.0022 | -5.1365 | 0.0000 | -0.0155 | -0.0069 |
| PTRATIO | -0.7521 | 0.0827 | -9.0967 | 0.0000 | -0.9145 | -0.5896 |
| BK | 0.0063 | 0.0018 | 3.6059 | 0.0003 | 0.0029 | 0.0098 |
| LSTAT | -0.2659 | 0.0398 | -6.6863 | 0.0000 | -0.3440 | -0.1878 |
| 1/LSTAT | 21.5901 | 4.2581 | 5.0703 | 0.0000 | 13.2231 | 29.9571 |

由表8可知，模型的调整的判定系数变为84.47%，整体的检验高度显著，对应值<0.05。在5%的显著性水平下，所有变量的系数均通过检验，因此最终建立的回归方程为

因此，结合新加入的两个变量的分析结果与表6对应的模型分析可以得出新的结论：影响房价的主要因素为每间住宅的平均房间数(RM)和人口状况下降的百分比(LSTAT)的比例。平均房间数越多，房价越高，且呈指数式增长。人口状况下降的百分比越高，房价越低。

**（二）残差分析**

根据表8对应的模型绘制的残差图如图16所示：

图16 回归分析结果五的残差图

图16表明，各残差基本上位于一条水平带中间，这表明表8对应回归模型的线性假定以及对误差项的假定成立。

**五、总结与讨论**

本研究通过自有住房的中位数报价刻画了波士顿住房的房价水平，并通过与之相关的13个影响因素量化地描述了住房的各项特征，以前者为因变量，后者为自变量，构建了多元线性回归模型，且拟合优度良好，但在进行残差分析后，发现残差图呈曲线形态，即自变量与因变量直接拟合的回归模型存在不合理之处。因此，在结合了描述性分析的结果后，将平均房间数的平方与人口状况下降百分比的倒数作为新的自变量加入原模型，新模型的拟合优度显著提升，且残差分析表明模型的线性假定及对误差项的假定均成立。

结合模型结果与描述性分析发现，每间住宅的平均房间数（RM）和城镇黑人比例（BK）与自有住房的中位数报价正相关，城镇人均犯罪率（CRIM）、一氧化氮浓度（NOX）、老式自住房比例（AGE）、全额物业税率（TAX）、城镇的学生与教师比例（PTRATIO）、人口状况下降的百分比（LSTAT）六项因素与自有住房的中位数报价负相关，而占地面积大的住宅用地比例（ZN）、查尔斯河流经该地区（CHAS）、各城镇非零售业务（INDUS）对房价的影响不显著。

此外，径向高速公路的可达性指数（RAD）及加权到波士顿的五个就业中心的距离（DIS）两项因素对房价的影响在描述性分析及回归分析中得出的结论存在部分矛盾的情况。因此，在后续研究中，可以考虑通过分箱处理和引入高阶项或交叉项等方法，对现有模型进行改进，以实现更为良好的预测效果。