基于波士顿房价数据集搭建多元线性回归模型

SY2201230 王昌锐

摘要

本文以多元线性回归统计理论为基础，用 Python 语言和 SPSS 软件对波士顿地区房价数据进行建模分析。首先对数据集进行探索性分析和预处理，随后建立多元线性回归模型并进行显著性检验。考虑可能存在的多重共线性，使用逐步回归方法对模型进行修正，从而提高了模型的解释与预测能力。

关键词：多元线性回归；逐步回归；多重共线性

1 多元线性回归及其数学描述[1]

1.1 多元线性回归

回归分析主要研究变量与变量之间的非确定性关系，又称为相关关系。在对客观事物进行大量试验和观察的基础上，回归分析方法通过建立统计模型来研究变量间的统计规律。根据研究变量的多少，随机变量与解释变量的回归模型可分为一元回归模型、多元回归模型和多元多重回归模型。一元回归模型涉及一个因变量和一个自变量；多元回归模型涉及一个因变量和两个或两个以上的自变量；多元多重回归模型则涉及两个或两个以上的因变量和自变量。如果所建立的统计模型是线性的，称为线性回归模型，否则称为非线性回归模型。

本文主要研究多元线性回归模型。在许多实际问题中，影响事物的因素常常不止一个，因此多元回归具有广泛的应用范围。又由于许多非线性的情形都可以通过某种变量变换化为线性回归问题，或者用多项式来逼近这种非线性关系，因此多元线性回归模型更具普遍性。

在根据实际问题建立多元线性回归分析模型的过程中有几个重要的阶段：

1) 根据研究的目的设置指标变量，包括因变量和自变量；
2) 收集、整理样本数据，并对数据进行预处理；
3) 确定理论回归模型的数学形式；
4) 利用样本数据估计模型的参数；
5) 回归模型和回归系数的检验；

6) 回归模型的运用。

1.2 多元线性回归问题的数学描述

设随机变量y与$p(p \geq 2)$个普通变量$x_1, x_2, ..., x_p$有关，且满足关系式

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

$$E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2 < +\infty$$

其中$\beta_0, \beta_1, ..., \beta_p$和$\sigma^2$是与$x_1, x_2, ..., x_p$无关的未知参数，$\varepsilon$是不可观测的随机变量。称上式为$p$元理论线性回归模型，$\beta_0, \beta_1, ..., \beta_p$为回归系数，$x_1, x_2, ..., x_p$为回归因子。回归系数$\beta$实际反映了对应因子$x$对观察值$y$的贡献大小，因此也称$\beta$为因子$x$的效应。

设有 n 组不全相同的样本观察值（$x_{i1}, x_{i2}, ..., x_{ip}; y_i$），其中$i \in [1, n]$，由上式有

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i$$

$$E(\varepsilon_i) = 0,\ Var(\varepsilon_i) = \sigma^2 < +\infty$$

其中ε_i相互独立。为了方便，常采用矩阵表达式，记

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

则有

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$E(\boldsymbol{\varepsilon}) = 0,\ Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I_n}$$

其中$\boldsymbol{I_n}$为n阶单位矩阵。称$\boldsymbol{Y}$为随机变量的观测向量，$\boldsymbol{\beta}$为未知参数向量，$\boldsymbol{X}$为设计矩阵，$\boldsymbol{\varepsilon}$为n维随机误差向量，$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$。

有了样本数据，就可以使用最小二乘法来求回归系数$\beta_0, \beta_1, ..., \beta_p$的估计。但多元线性回归模型是否真正解释了自变量和因变量的相关关系，还要通过回归效果的显著性检验。

2 回归效果的评估方法[1]

1.1 回归方程的拟合优度检验

由于回归平方和$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$主要反映了自变量$x_1,x_2,...,x_p$对观察值$y$的线性影响，因此$y$与$x_1,x_2,...,x_p$线性相关的密切程度可以用$U$在总平方和$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$中所占的比值来衡量。称$R^2=\frac{U}{L_{yy}}$为决定系数或拟合优度。显然，$0\leq R^2\leq 1$，且$R^2$越接近 1，回归方程拟合数据的程度越好。

1.2 回归系数的显著性检验

对多元线性回归，回归效果的显著性检验主要是检验响应变量对所有的解释变量整体的线性依赖性是否显著，而回归系数的显著性检验主要是检验响应变量对某个解释变量或某几个解释变量的线性依赖性是否显著。下面分别说明 F 检验和 t 检验的检验方法。

对给定的显著性水平α，根据样本观察值计算出F值：

$$F=\frac{n-p-1}{p}\frac{U}{Q}\sim F(p,n-p-1)$$

若$F\geq F_{1-\alpha}(p,n-p-1)$，就认为$y$与$x_1,x_2,...,x_p$之间有显著的线性关系，此时称线性回归效果显著。

当回归效果显著时，仅说明$\beta_1,\beta_2,...,\beta_p$不可能全为 0，但是这并不排斥某一个或几个$\beta_i$为 0。若某个$\beta_i$为 0，这意味着$y$与$x_i$无关，因而可将这个$x_i$从回归方程中剔除掉。对给定的显著性水平$\alpha$，根据样本观察值计算出$t_i$值：

$$t_i=\frac{\widehat{\beta_i}}{\widehat{\sigma}\sqrt{c_{ii}}}\sim t(n-p-1)$$

其中c_{ii}是矩阵$\left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}$主对角线元素。若$|t_i|\geq t_{1-\frac{\alpha}{2}}(n-p-1)$，就认为$y$与$x_i$之间有显著的线性关系。

值得注意的是：t_i的值已经消除了单位的影响，是无量纲量，所以t_i本身除了可以作为检验统计量外，t_i之间也可以直接做比较。粗略地说，哪个t_i的绝对值大，说明相应的自变量x_i对y的影响更显著。

3 多重共线性与逐步回归法

在多元线性回归分析中，由于有多个自变量，首先要考虑的就是如何确定回

归自变量。一般情况下，根据所研究问题的目的，结合相关理论，可以罗列出对因变量可能有影响的一些因素作为自变量。如果遗漏了某些重要的变量，回归方程的效果肯定不会好；但如果考虑过多的自变量，在这些自变量中，某些自变量对问题的研究可能并不重要，有些自变量数据的质量可能很差，有些变量可能和其它变量有很大程度的重叠等。因此，我们需要考虑如何建立“最优”的线性回归模型。

3.1 多重共线性[2]

多元线性回归模型的主要假设之一是自变量彼此之间不存在强相关，否则会出现多重共线性问题。理论上可以证明，多重共线性使回归模型计算复杂且往往会扩大估计方差，降低模型精度，导致多元线性回归系数的显著性偏离真实方向。要判断回归模型是否有多重共线性，最常见的办法是考察方差膨胀因子。方差膨胀因子是解释变量之间存在多重共线性时的方差与不存在多重共线性时的方差之比。自变量x_i的方差膨胀因子记为VIF_i（Variance Inflation Factor），它的计算方法为：

$$VIF_i = \frac{1}{1 - R_i^2}$$

经验判断方法表明：当 $0 < VIF < 10$，不存在多重共线性；当 $10 \leq VIF < 100$，存在较强的多重共线性；当$VIF \geq 100$ 时，存在严重多重共线性。一般认为，如果最大的VIF_i超过 10，多重相关性将严重影响最小二乘法的估计值。

3.2 逐步回归[3]

按自变量x_i对y的影响显著程度（或者说贡献）由大到小地逐个引入回归方程，这种回归分析方法称为逐步回归（Stepwise Regression）。自变量引入的条件是其偏回归平方和U_i经偏 F 检验后是显著的。

回归平方和U是所有自变量对y的总变差的贡献。把剔除了自变量x_i后回归平方和U所减少的数值$\frac{\hat{\beta}_i^2}{c_{ii}}$称为对变量$x_i$的偏回归平方和$U_i$。检验某个自变量$x_i$对$y$的影响是否显著的正规方法是偏 F 检验，检验统计量为

$$F_i = \frac{n - p - 1}{Q} \frac{\hat{\beta}_i^2}{c_{ii}} \sim F(1, n - p - 1)$$

若$F_i \geq F_{1-\alpha}(1,\ n - p - 1)$，说明$x_i$对$y$有显著影响。

先用被解释变量对每一个引入的解释变量做简单回归，然后以对被解释变量

贡献最大的解释变量所对应的回归方程为基础，再逐步引入其余解释变量。每引入一个新的自变量，就要对旧的自变量逐个检验，并剔除偏回归平方和U_i不显著的自变量。边引入，边剔除，直到既无新变量引入也无旧变量删除为止。经过逐步回归，使得最后保留在模型中的解释变量既是重要的，又没有严重的多重共线性。逐步回归的实质是建立“最优”的多元线性回归方程。

删除变量是一种针对数据多重共线性的处理方法，虽然能提高模型的解释和预测能力，但样本中多重共线性的混杂并没有改变。减少混杂最直接有效的办法仍是提高样本量或样本质量，又或者采取主动实验设计。

4 案例分析

4.1 背景介绍

本文选用经典数据集——波士顿（boston）房价数据集[4]做多元线性回归分析。该数据集取自卡内基梅隆大学的 StaLib library，描述了 1978 年波士顿大区不同调查行政区的房价及其可能的影响因素。数据集共 14 列 506 行（表 1），每一列的含义分别如下：

1. CRIM，per capita crime rate by town
2. ZN，proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS，proportion of non-retail business acres per town
4. CHAS，Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX，nitric oxides concentration (parts per 10 million)
6. RM，average number of rooms per dwelling
7. AGE，proportion of owner-occupied units built prior to 1940
8. DIS，weighted distances to five Boston employment centres
9. RAD，index of accessibility to radial highways
10. TAX，full-value property-tax rate per $10,000
11. PTRATIO，pupil-teacher ratio by town
12. B，1000(Bk - 0.63)^2 where Bk is the proportion of black people by town
13. LSTAT，% lower status of the population
14. MEDV，Median value of owner-occupied homes in $1000's

其中 CHAS 和 RAD 是虚拟变量（Dummy Variables），又称虚设变量、名义变量或哑变量，是反映质属性的一个人工变量，通常取值为 0、1 或其它整数。引入哑变量可使线性回归模型变得更复杂，但对问题描述更简明。

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88	11.9

506 rows × 14 columns

表 1 数据示意图

4.2 数据分析及预处理

正式建立模型之前，我们希望对数据集本身有一个快速直观的了解。另外，为确保进行多元线性回归时自变量与因变量之间具有较为显著的相关性，还需要对各变量进行相关性检验。本文使用 Python 语言的第三方库 pandas、matplotlib 和 seaborn 绘制了 Pearson 相关性分析热力图（图 1）和反映因变量（MEDV）随不同自变量变化趋势的散点图矩阵（图 2）。

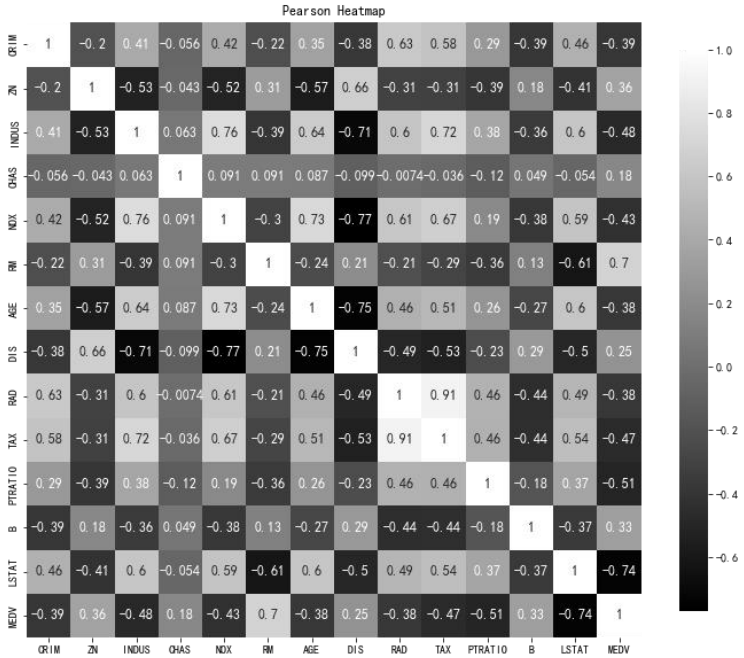

图 1 Pearson 相关性分析热力图

从图 1 的最后一行可以看出，绝大多数的变量与 MEDV 的相关系数都超过了 0.3，认为它们之间有着较为明显的相关关系。特别地，对相关系数绝对值大于 0.5 的三个变量（RM、LSTAT、PTRATIO），我们进行了放大展示（图 3）。

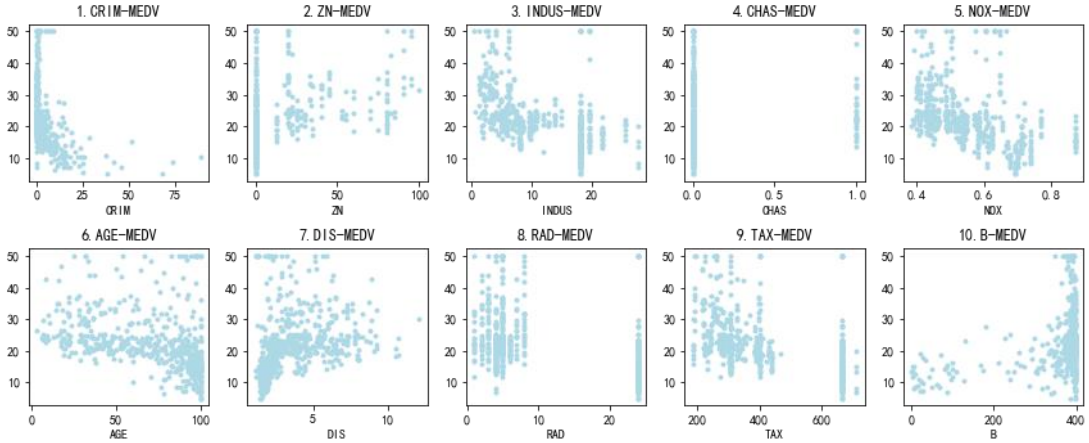

图 2 反映因变量（MEDV）随不同自变量变化趋势的散点图

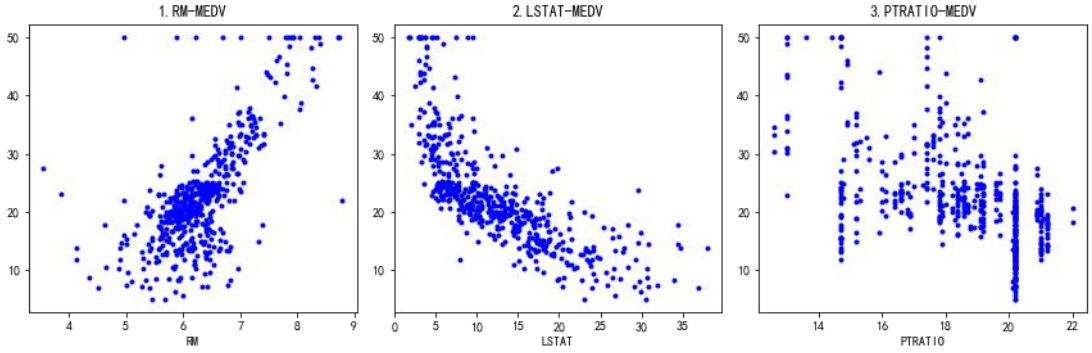

图 3 相关系数绝对值大于 0.5 的三个变量对应的散点图

4.3 降低偏度

由于假设样本所在总体分布要求是正态分布的，所以在做分析的时候需要通过偏度和峰度两个指标来检验样本数据是否属于正态分布。偏度是统计数据分布偏斜方向和程度的度量，是统计数据分布非对称程度的数字特征。峰度则是表征统计数据分布在平均值处峰值高低的特征数。

我们观察 CRIM 和 ZN 两个数值变量对应的散点分布（图 2），发现它们的特征偏度很大，所以需要减少偏度。对高度偏态的数据，不妨对其取对数。特别地，ZN 数据包含 0 值，因此需要先加一个极小量如10^{-5}再取对数。处理后发现 CRIM 从偏度分布变成了峰度分布，而 ZN 则变成了双峰分布。处理后的数据如

图 4 所示。

必须指出，对原始数据做变换后，所得结果不能用第 2 节介绍的相应方法来直接进行假设检验。但利用原始数据和变换后数据求出的R^2有类似的性质，即R^2越接近 1，回归效果越好。

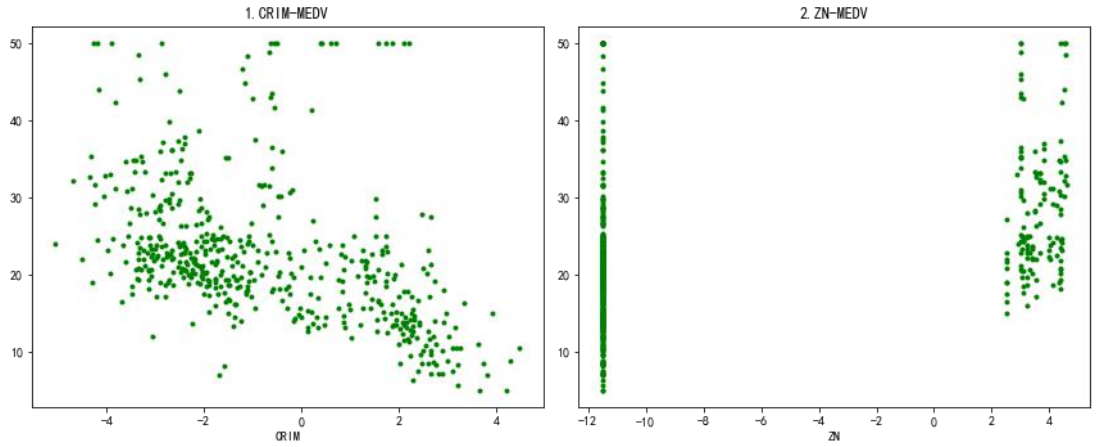

图 4 CRIM 和 ZN 属性降低偏度后的散点图

注意到 CHAS 和 RAD 的分布同样极度不均，但考虑到它们是虚拟变量，因此不需要降低偏度。CHAS 取值只有 0（471 个值）和 1（35 个值），而 RAD 的统计分布如图 5 所示。

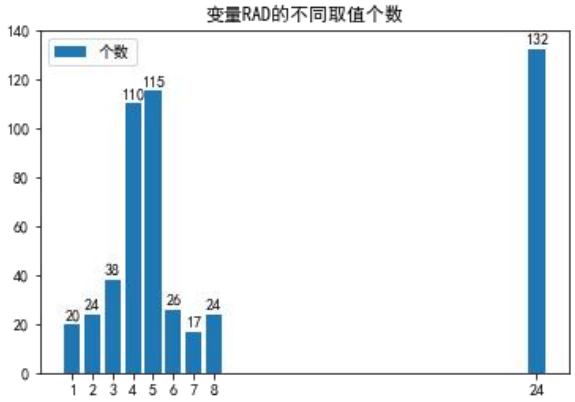

图 5 RAD 不同取值个数的统计直方图

4.4 建立多元线性回归模型并检验

利用 SPSS 软件进行多元线性回归分析，得到如表 2 所示的回归模型。回归模型的数学表达式为：

$$Y = 36.845 + 0.269x_1 + 0.087x_2 + 0.011x_3 + 2.899x_4 - 18.234x_5 + 3.968x_6 - 0.005x_7 - 1.281x_8 + 0.181x_9 - 0.01x_{10} - 0.972x_{11} + 0.011x_{12} - 0.553x_{13}$$

其中$x_i(i = 1, \ldots, 13)$分别对应 13 个自变量 CRIM、ZN、INDUS、CHAS、NOX、RM、AGE、DIS、RAD、TAX、PTRATIO、B 和 LSTAT。表 2 还给出了不同系数的 95%置信区间的上下限。

模型		未标准化系数		标准化系数			B 的 95.0% 置信区间		共线性统计	
		B	标准误差	Beta	t	显著性	下限	上限	容差	VIF
1	(常量)	36.845	5.268		6.995	.000	26.495	47.195		
	CRIM	.269	.277	.063	.970	.333	-.276	.813	.129	7.779
	ZN	.087	.051	.063	1.717	.087	-.013	.187	.401	2.495
	INDUS	.011	.063	.008	.180	.857	-.112	.135	.249	4.018
	CHAS	2.899	.875	.080	3.312	.001	1.179	4.619	.933	1.072
	NOX	-18.234	4.012	-.230	-4.545	.000	-26.116	-10.352	.213	4.686
	RM	3.968	.423	.303	9.369	.000	3.136	4.800	.521	1.919
	AGE	-.005	.013	-.016	-.399	.690	-.032	.021	.320	3.126
	DIS	-1.281	.204	-.293	-6.276	.000	-1.682	-.880	.250	4.005
	RAD	.181	.078	.172	2.336	.020	.029	.334	.101	9.896
	TAX	-.010	.004	-.176	-2.567	.011	-.017	-.002	.116	8.609
	PTRATIO	-.972	.138	-.229	-7.030	.000	-1.243	-.700	.515	1.942
	B	.011	.003	.105	3.846	.000	.005	.016	.737	1.357
	LSTAT	-.553	.052	-.429	-10.738	.000	-.654	-.452	.341	2.934

a. 因变量：MEDV

表 2 模型回归系数表

在 SPSS 中，复相关系数 R 和拟合优度R^2用于检验回归方程对样本观测值的拟合程度，越接近 1 说明随机误差所占比重越小，回归的效果越显著。表 3 为模型相关系数表。$R = 0.855$，表明因变量与自变量之间的线性相关程度十分密切。$R^2 = 0.732$，说明因变量变化的 73.2%可由这 13 个自变量给出解释。对多元线性回归模型，R^2很难贴近 1，因此 0.75 左右的拟合程度可以接受。

模型	R	R 方	调整后 R 方	标准估算的误差	德宾-沃森
1	.855[a]	.732	.725	4.8260	1.059

a. 预测变量：(常量), LSTAT, CHAS, B, PTRATIO, DIS, RM, RAD, ZN, AGE, INDUS, NOX, CRIM, TAX

b. 因变量：MEDV

表 3 模型摘要

通过 F 检验对回归方程的显著性进行检验。选定显著性水平$\alpha = 0.05$，利用 SPSS 软件得到方差分析表，结果如表 4 所示。表中数据“显著性”一栏即为 P 值，P 值小于显著性水平α，因此认为回归方程是显著的。

模型		平方和	自由度	均方	F	显著性
1	回归	31257.576	13	2404.429	103.238	.000[b]
	残差	11458.720	492	23.290		
	总计	42716.295	505			

a. 因变量：MEDV

b. 预测变量：(常量), LSTAT, CHAS, B, PTRATIO, DIS, RM, RAD, ZN, AGE, INDUS, NOX, CRIM, TAX

表 4 方差分析 ANOVA

回归分析中假定残差服从正态分布，残差检验就是检验残差是否服从正态分布。表 3 中德宾·沃森统计量取值 1.059<2，说明残差项间存在一定正相关。表 5 所示为残差数据，显示了残差及标准残差的部分统计量。图 6（左）是以标准化残差为横坐标建立的标准化残差-频率直方图，可以看到残差分布大致呈现正态分布（平均值-3.19E-15，标准差 0.987）。图 6（中）是预期累积概率和实测累积概率的 P-P 图，图中斜线对应均值为零的正态分布，散点基本分布在该斜线附近，也证实了残差分布的正态性。图 6（右）所示散点图处于无序状态，表明回归模型的标准化残差与标准化预测值之间没有系统的相关关系，即同方差性成立。

残差统计[a]

	最小值	最大值	平均值	标准偏差	个案数
预测值	-1.493	43.995	22.533	7.8674	506
残差	-15.3727	27.0793	.0000	4.7635	506
标准预测值	-3.054	2.728	.000	1.000	506
标准残差	-3.185	5.611	.000	.987	506

a. 因变量：MEDV

表 5 残差数据

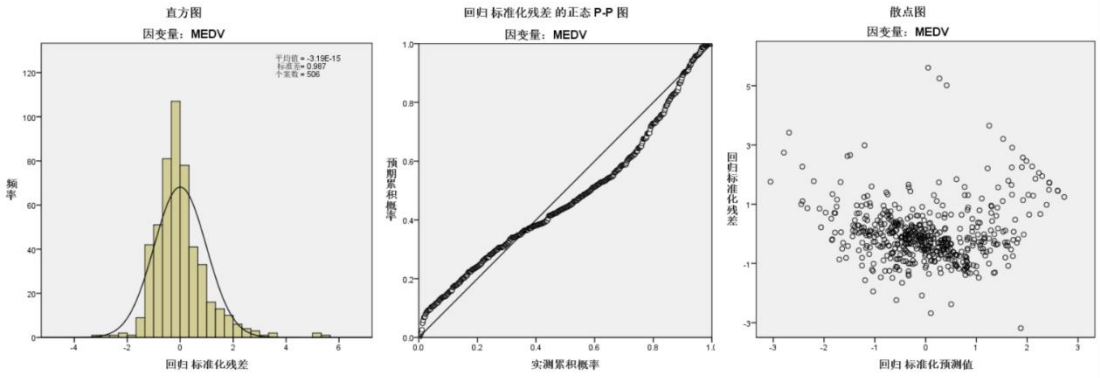

图6 (左)标准化残差-频率直方图；(中)正态P-P图；(右)标准化的残差-预测值散点图

4.5 逐步回归建立模型并检验

根据表 2 回归系数显著性检验的结果，自变量 CRIM、ZN、INDUS 和 AGE 的 P 值均大于显著性水平 0.05，因此有理由怀疑它们是多余变量。再重复 4.3 节

多元回归分析模型建模步骤，只是 SPSS 程序中自变量由“输入”改为“步进”，将自变量逐个引入，即 3.2 节介绍的逐步回归分析方法。引入顺序依次为常量、LSTAT、RM、PTRATIO、DIS、NOX、CHAS、B、RAD、TAX，剔除的变量恰是 CRIM、ZN、INDUS 和 AGE。结果如表 6、7、8 所示。

对比表 3，表 6 最后一步（第 9 步）的$R = 0.854$，$R^2 = 0.730$，与旧模型相比基本没有变化。同时，表 7 中 F 检验的显著性小于 0.05，说明这些自变量与因变量具有显著的线性关系。对比表 2 和表 8，发现每个自变量对应的方差膨胀因子 VIF 均有所下降，t 检验显著性均小于 0.05，模型的解释和预测能力变强。

模型	R	R 方	调整后 R 方	标准估算的误差	德宾-沃森
1	.738[a]	.544	.543	6.2158	
2	.799[b]	.639	.637	5.5403	
3	.824[c]	.679	.677	5.2294	
4	.831[d]	.690	.688	5.1386	
5	.841[e]	.708	.705	4.9939	
6	.846[f]	.716	.712	4.9326	
7	.850[g]	.722	.718	4.8818	
8	.852[h]	.726	.721	4.8569	
9	.854[i]	.730	.725	4.8240	1.053

表 6 新模型摘要

模型		平方和	自由度	均方	F	显著性
9	回归	31174.099	9	3463.789	148.849	.000[j]
	残差	11542.197	496	23.271		
	总计	42716.295	505			

表 7 新模型最后一步的方差分析 ANOVA

模型		未标准化系数		标准化系数			B 的 95.0% 置信区间		共线性统计	
		B	标准误差	Beta	t	显著性	下限	上限	容差	VIF
9	(常量)	35.957	5.144		6.990	.000	25.850	46.065		
	LSTAT	-.545	.048	-.423	-11.436	.000	-.639	-.451	.398	2.514
	RM	4.015	.409	.307	9.807	.000	3.210	4.819	.557	1.795
	PTRATIO	-1.074	.123	-.253	-8.715	.000	-1.316	-.832	.647	1.545
	DIS	-1.138	.163	-.260	-6.968	.000	-1.458	-.817	.390	2.565
	NOX	-17.739	3.568	-.224	-4.972	.000	-24.750	-10.729	.270	3.710
	CHAS	2.852	.869	.079	3.283	.001	1.145	4.559	.946	1.057
	B	.010	.003	.102	3.805	.000	.005	.016	.756	1.323
	RAD	.230	.062	.218	3.728	.000	.109	.351	.159	6.272
	TAX	-.009	.003	-.172	-2.794	.005	-.016	-.003	.143	6.994

a. 因变量：MEDV

表 8 新模型回归系数表

5 结论

综上所述，可得知以下信息。

1）优化后的多元线性回归模型结果为：

$$Y = 35.957 - 0.545x_1 + 4.015x_2 - 1.074x_3 - 1.138x_4 - 17.739x_5 + 2.852x_6 + 0.01x_7 + 0.23x_8 - 0.009x_9$$

其中$x_i(i = 1, ..., 9)$分别对应 9 个自变量 LSTAT、RM、PTRATIO、DIS、NOX、CHAS、B、RAD 和 TAX。

2）$R^2 = 0.730$，说明房价变动的 73%可由这 9 个自变量来解释。

3）9 个自变量的 P 值均小于 1%，呈现出较强的显著性。

4）城镇低社会地位人口比例 LSTAT 对应的系数为-0.545，说明在其它因素相同的情况下，LSTAT 每增加一个单位，预测房价会下降 545 美元。实际上 LSTAT 越高，说明该地区生活质量越低，房价自然也低。

5）城镇每户平均房间数 RM 对应的系数为 4.015。RM 指数越高，该地区房屋面积平均值越大，甚至可能为别墅区。由于房价最直接的影响因素就是房屋面积，RM 每增加一个单位，预测房价提高 4015 美元。

6）城镇学生/教师比例 PTRATIO 反映了该地区的教育程度和学校的密集程度，对应的系数为-1.074。PTRATIO 越低，说明该地区的教育资源更好，甚至可能是学区房，故与房价呈现负相关。

7）与五个波士顿劳动力聚集区的加权距离指数 DIS 对应的系数为-1.138，说明城镇离中心区越远，该地区的房价越低。

8）氮氧化物浓度指数 NOX 对应的系数为-17.739。实际上，NOX 高的地区主要在大型工厂附近，自然环境差且距离市中心较远，房价受其影响较大。NOX 指数每上升 1%，预测房价会下降 177.39 美元

9）靠近河流指数 CHAS 对应的系数为 2.852。实际上，有河流经过的城镇自然环境会更好，因此 CHAS 指数与房价呈现正相关。

10）$B = 1000 * (Bk - 0.63)^2$，对应的系数为 0.01。Bk 指代城镇中黑人的比例，Bk 下降，B 就会增加，该地区的平均房价也更高。事实上，美国黑人聚集的地区治安条件相对更差，因此房价普遍偏低。

11）与辐射式公路的接近指数 RAD 对应的系数为 0.23，说明在其它因素相同的情况下，交通越便利，房价越高。

12）全值财产税指数 TAX 对应的系数为-0.009。实际上，更少人愿意去财产税高的城镇居住，故而该地区的房价与 TAX 呈现负相关。每一万美元的财产税增加 1 美元，预测该地区的房价会随之下降 9 美元。

13）剔除的四个变量 CRIM（城镇人均犯罪率）、ZN（大面积住宅用地比例）、INDUS（非零售商业用地比例）和 AGE（1940 年建成的自用房屋比例）或多或少与其它的自变量存在相关关系，因此剔除这些变量是可以接受的。如果认为这些变量是重要的，那么可以考虑使用主成分分析对变量进行压缩，或者给模型添加正则项的方法如岭回归、Lasso 回归等，本文不涉及这些回归分析方法。

参考文献

[1]孙海燕,周梦,李卫国,冯伟. 数理统计[M]. 北京: 北京航空航天大学数学系, 2011

[2]王惠文. PLS 回归在消除多重共线性中的作用[J]. 数理统计与管理, 1996, 15(6):5.

[3]余继峰,张涛,宋召军主编. 高等教育“十三五”规划教材 数学地质方法与应用[M]. 徐州：中国矿业大学出版社, 2019.02.53 页

[4]Belsley, Kuh & Welsch, 'Regression diagnostics: Identifying Influential Data and Sources of Collinearity', Wiley, 1980. 244-261.