Statistics and Application 统计学与应用, 2020, 9(3), 335-344
Published Online June 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.93036

Hans汉斯

Analysis of the Correlation between Housing Price Data in Boston Based on the Regression Method

Ran Zhao

Qufu Normal University, Qufu Shandong
Email: 486814141@qq.com

Received: May 8th, 2020; accepted: May 23rd, 2020; published: Jun. 1st, 2020

Abstract

According to the variables in the Boston housing price data set, a linear regression model was established for the Boston housing price by using R software. The significance test of the regression equation and regression coefficient was carried out. The model was established after the Box-Cox transformation was used for the case that the basic assumptions were violated. Lasso regression was used to simplify the equation appropriately, but the regression coefficient of the model established by lasso regression was small, because the variables in this data were not multicollinearity, which was consistent with the judgment results of R software. Finally, the response variable in the data and the independent variable whose absolute value of its correlation coefficient is greater than 0.5 establish a linear regression equation and predict the housing price. Because the distribution range of housing price in Boston will change with the change of influencing factors, and the median has certain robustness, we establish a regression model for the median of housing price, namely quantile regression model.

Keywords

Linear Regression Model, Box-Cox Transformation, Lasso Regression, Prediction

基于回归方法分析波士顿房价数据间的相关关系

赵　冉

曲阜师范大学，山东　曲阜
Email: 486814141@qq.com

收稿日期：2020年5月8日；录用日期：2020年5月23日；发布日期：2020年6月1日

摘　要

根据波士顿房价数据集中的变量使用R软件对波士顿房价建立线性回归模型，对回归方程和回归系数进

文章引用: 赵冉. 基于回归方法分析波士顿房价数据间的相关关系[J]. 统计学与应用, 2020, 9(3): 335-344.
DOI: 10.12677/sa.2020.93036

行显著性检验，针对违背基本假设的情况使用Box-Cox变换后再建立模型。为适当精简方程使用Lasso回归，但其建立的模型回归系数很小，原因是此数据中的变量并没有多重共线性，与使用R软件判断结果一致。最后，数据中的响应变量与其相关系数的绝对值大于0.5的自变量建立线性回归方程，并对房价进行预测。由于波士顿房价的分布范围会随着影响因素的变化而发生变化，且中位数具有一定的稳健性，因而我们对房价的中位数建立回归模型，即分位数回归模型。

关键词

线性回归模型，Box-Cox变换，Lasso回归，预测

1. 引言

波士顿房价数据集是统计的 20 世纪 70 年代中期波士顿郊区房价的中位数，统计了当时教区部分的犯罪率、房产税等共计 13 个指标，统计出房价，试图能找到指标与房价的关系并进行预测。

2. 材料与方法

2.1. 变量名称与建模目的

2.1.1. 变量名称简介

分析波士顿房价数据集(Boston House Price Dataset)可知影响响应变量 MEDV 的因素可能有 13 个，以下为各个属性的介绍，见表 1。

Table 1. Introduction of related variables
表 1. 相关变量的介绍

变量缩写	变量含义
RIM	城镇人均犯罪率
ZN	占地面积超过 25,000 平方英尺的住宅用地比例
INDUS	每个城镇非零售业务的比例
CHAS	Charles River 虚拟变量(如果是河道，则为 1；否则为 0)
NOX	一氧化氮浓度(每千万份)
RM	每间住宅的平均房间数
AGE	1940 年以前建造的自住单位比例
DIS	加权距离波士顿的五个就业中心
RAD	径向高速公路的可达性指数
TAX	每 10,000 美元的全额物业税率
PTRATIO	城镇的学生与教师比例
B	1000 (Bk-0.63)^2 其中 Bk 是城镇黑人的比例
LSTAT	人口状况下降%
MEDV	自有住房的中位数报价，单位 1000 美元

本例是属于回归模型的案例，在数据集中包含 506 组数据。通过对波士顿房地产数据进行初步的观察并分析找出影响房价中位数的因素，希望建立一个能够预测房屋价值的多元线性回归模型。

2.1.2. 多元线性回归模型的一般形式

设随机变量 y 与一般变量 $x_1,x_2,\cdots,x_p$ 的线性回归模型为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

式中，$\beta_0,\beta_1,\cdots,\beta_p$ 是 $p+1$ 个未知参数，β_0 称为回归常数，$\beta_1,\cdots,\beta_p$ 称为回归系数。y 称为解释变量(因变量)，$x_1,x_2,\cdots,x_p$ 是 p 个可以精确测量并控制的一般变量，称为解释变量(自变量)。

ε 是随机误差，并且假定

$$\begin{cases}E(\varepsilon)=0\\ \mathrm{var}(\varepsilon)=\sigma^2\end{cases}$$

2.2. 问题解决方法与知识依托

在本例中我们使用 R 软件解决相应问题。部分代码见参考文献[1]。

2.2.1. 预处理

首先将数据导入 R 软件中，为了消除量纲不同和数量级差异带来的影响，就需要对数据进行标准化处理，然后用最小二乘法估计未知参数，求得标准化回归系数。

样本数据的标准化公式为：

$$x_{ij}^*=\frac{x_{ij}-\overline{x}_j}{\sqrt{L_{jj}/n}},\ i=1,2,\cdots,n;j=1,2,\cdots,p$$

$$y_i^*=\frac{y_i-\overline{y}}{\sqrt{L_{yy}/n}},\ i=1,2,\cdots,n$$

式中

$$L_{jj}=\sum_{i=1}^{n}\left(x_{ij}-\overline{x}_j\right)^2$$

是自变量 $x_j\left(j=1,2,\cdots,p\right)$ 的离差平方和。用最小二乘法求出标准化的样本数据的经验回归方程，记为：

$$\hat{y}^*=\hat{\beta}_1^*x_1^*+\hat{\beta}_2^*x_2^*+\cdots+\hat{\beta}_p^*x_p^*$$

式中，$\hat{\beta}_1^*,\hat{\beta}_2^*,\cdots,\hat{\beta}_p^*$ 是 y 对自变量 $x_1,x_2,\cdots,x_p$ 的标准化回归系数。

2.2.2. 回归参数的普通最小二乘估计

即寻找参数 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p$，使离差平方和 $Q\left(\beta_0,\beta_1,\cdots,\beta_p\right)=\sum^{n}\left(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip}\right)^2$ 达到极小。

当 $\left(X'X\right)^{-1}$ 存在时，即得回归参数的最小二乘估计为：

$$\hat{\beta}=\left(X'X\right)^{-1}X'y$$

2.2.3. 回归方程、回归系数的检验

1) F 检验

对多元线性回归方程的显著性检验就是要看自变量 $x_1,x_2,\cdots,x_p$ 从整体上对随机变量 y 是否有明显的影响。

原假设 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$

构造 F 检验统计量如下：

$$F=\frac{SSR/p}{SSE/(n-p-1)}$$

当原假设成立时，F 服从自由度为 $(p,n-p-1)$ 的 F 分布。

当 $F>F_{\alpha}(p,n-p-1)$ 时，拒绝原假设 H_0，否则认为在显著性水平 α 下，y 与 $x_1,x_2,\cdots,x_p$ 有显著的线性关系，即回归方程是显著的。

2) t 检验

检验 x_j 是否显著等价于检验

$$H_{0j}:\beta_j=0,\quad j=1,2,\cdots,p$$

如果接受原假设，则 x_i 不显著；如果拒绝原假设，则 x_j 是显著的。

据此可以构造 t 统计量

$$t_j=\frac{\hat{\beta}}{\sqrt{c_{jj}}\hat{\sigma}}$$

式中

$$\hat{\sigma}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2}$$

2.2.4. 违背基本假设情况的检验

1) 异方差性

违背了回归模型的基本假定，即

$$\mathrm{var}(\varepsilon_i)\neq\mathrm{var}(\varepsilon_j)，\ 当 i\neq j 时$$

诊断方法：绘制残差图 等级相关系数法

解决方法：多元加权最小二乘估计 BOX-COX 变换

2) 自相关性

违背基本假设，即

$$\mathrm{cov}(\varepsilon_i,\varepsilon_j)\neq 0，\ 当 i\neq j 时$$

诊断方法：图示检验法 自相关系数法 DW 检验

解决方法：迭代法 差分法 BOX-COX 变换

2.2.5. 多重共线性

1) 共线性诊断

① 方差扩大因子法

$c_{jj}=\frac{1}{1-R_j^2}$ 作为方差扩大因子的定义，证明见参考文献[2]，当 $VIF_j\geq 10$ 时，说明自变量 x_j 与其余自变量之间有严重的多重共线性。(注意：有些教材认为 $vif>4$ 即存在多重共线性。详见参考文献[3]。)

② 条件数

记 $X'X$ 的最大特征根为 λ_m，称

$$k_i=\sqrt{\frac{\lambda_m}{\lambda_i}},\quad i=0,1,\cdots,p$$

为特征根 λ_i 的条件数。

通常认为 $0 < k < 10$ 时，设计矩阵 X 没有多重共线性；$10 \le k < 100$ 时，存在较强的多重共线性；$k \ge 100$ 时，存在严重的多重共线性。

2) 解决方法

剔除不重要的解释变量，在此例中我们将看到不显著的回归系数，当回归系数不显著时，剔除变量。

由于样本量足够大，因而增大样本量已经无法解决问题。

岭回归，详见参考文献[4]。

主成分回归与偏最小二乘估计。

2.2.6. Lasso 回归

Lasso 回归又称为套索回归，并提供了从零开始到最小二乘拟合的系数和拟合的整个序列。Lasso 是一种收缩估计方法，其基本思想是在回归系数的绝对值之和小于一个常数的约束条件下，使残差平方和最小化，从而能够产生某些严格等于 0 的回归系数，进一步得到可以解释的模型。R 语言中有多个包可以实现 Lasso 回归，这里使用 lars 包实现。

3. 结果与分析

3.1. 回归方程的建立

3.1.1. 回归方程的初步建立

由于数据为多元的，因而无法用一元回归分析的方法绘制散点图。为探究各个属性与响应变量的关系，我们先对其建立线性回归模型，讨论模型的合理性。回归系数及 p 值见表 2。

Table 2. Coefficients of regression equation and their p values
表 2. 回归方程的系数及其 p 值

	Estimate	Std. Error	t value	Pr(>|t|)
CRIM	−0.101	3.07E−02	−3.287	0.001087
ZN	0.118	3.48E−02	3.382	0.000778
INDUS	0.0153	4.59E−02	0.334	0.738288
CHAS	0.0742	2.38E−02	3.118	0.001925
NOX	−0.224	4.81E−02	−4.651	4.25E−06
RM	0.291	3.19E−02	9.116	<2e−16
AGE	0.00212	4.04E−02	0.052	0.958229
DIS	−0.338	4.57E−02	−7.398	6.01E−13
RAD	0.290	6.28E−02	4.613	5.07E−06
TAX	−0.226	6.89E−02	−3.28	0.001112
PTRATIO	−0.224	3.08E−02	−7.283	1.31E−12
B	0.0924	2.67E−02	3.467	0.000573
LSTAT	−0.407	3.94E−02	−10.347	<2e−16

统计量 $F=108.1$，$p<2.2\text{e}{-16}$，给定显著性水平 $\alpha=0.05$，则 $p<\alpha$，因而拒绝原假设，认为回归方程是显著的。但是，根据上表可知部分回归系数不显著。此时残差的标准差为 0.516，$R^2=0.7406$，调整的 $R^2=0.7338$，拟合效果一般。

3.1.2. 回归方程的进一步分析

由于回归方程中部分回归系数不显著，因而剔除不显著的变量。首先剔除变量中 p 值最大的，进行回归分析，然后在剩下的变量中剔除最大的，进行分析，依次进行，直至回归方程中所有的回归系数都显著为止。回归方程与逐步回归选择变量结果相同。

建立的回归方程为：

$$\begin{aligned}\hat{M}=&-0.101\hat{CR}+0.116\hat{Z}+0.075\hat{CH}-0.219\hat{N}+0.290RM\\&-0.342\hat{D}+0.284RA-0.216\hat{T}-0.223\hat{P}+0.092\hat{B}-0.406\hat{L}\end{aligned}$$

说明：由于变量较多，为适当精简方程，将变量的首字母作为变量，同时回归系数保留三位小数。

此时，统计量 $F=128.5$，$p<2.2\text{e}{-16}$，回归方程显著，调整的 $R^2=0.7348$，并且回归方程的各个回归系数都已显著。

3.2. 违背基本假设情况的检验与解决

回归方程的残差为

$$\hat{\varepsilon}=Y-\hat{Y}=(I-H)Y$$

其中

$$H=X(X'X)^{-1}X'$$

称 H 为帽子矩阵。

在得到回归方程后，计算残差，可以对残差进行正态性检验。

检验结果 $p<\alpha$，因而认为残差不满足正态性假设。另外由残差的 QQ 图(见图 1)也可以看出残差不满足正态性假设。

左上图为残差与拟合图，用来检验线性，若散点集中分布在一条直线附近，则表示线性关系良好；

右上图为 QQ 图，用来检验正态性，若散点集中分布在 Q-Q 图中的直线上，则表示残差正态性良好；

左下图为位置尺度图，用来检验同方差性，若点在曲线周围随机分布，则表示同方差性成立；

右下图为残差与杠杆图，可以观测出离群点、高杠杆点和强影响点。独立性是无法从图中分辨出来的。

从图上可以看出该模型的残差并未随机分布，而是呈现异方差的问题。

同样，未标准化的模型也具有异方差性，下面对未标准化的数据进行变换。

对模型进行 Box-Cox 正态变换，求得 $Est\ Power=0.1158$，这里取 0.116。

记变换后的 MEDV 为 y，回归系数见表 3，则回归方程为

$$\begin{aligned}y=&5.0748-0.01354\text{CRIM}+1.7157\text{e}{-03}\text{ZN}+0.1516\text{CHAS}-1.0424\text{NOX}\\&+0.1422\text{RM}-0.0761\text{DIS}+0.0191\text{RAD}-7.896\text{e}{-04}\text{TAX}\\&-0.0542\text{PTRATIO}+5.917\text{e}{-04}\text{B}-0.0397\text{LSTAT}\end{aligned}$$

回归方程、回归系数皆通过显著性检验，且 $\hat{\sigma}=0.2688$，比之前模型有所降低，$R^2=0.7884$。比之前有所提高，绘制残差图发现残差也有所改善，因而模型较之前有所改善。变化前后的残差比较见图 2。

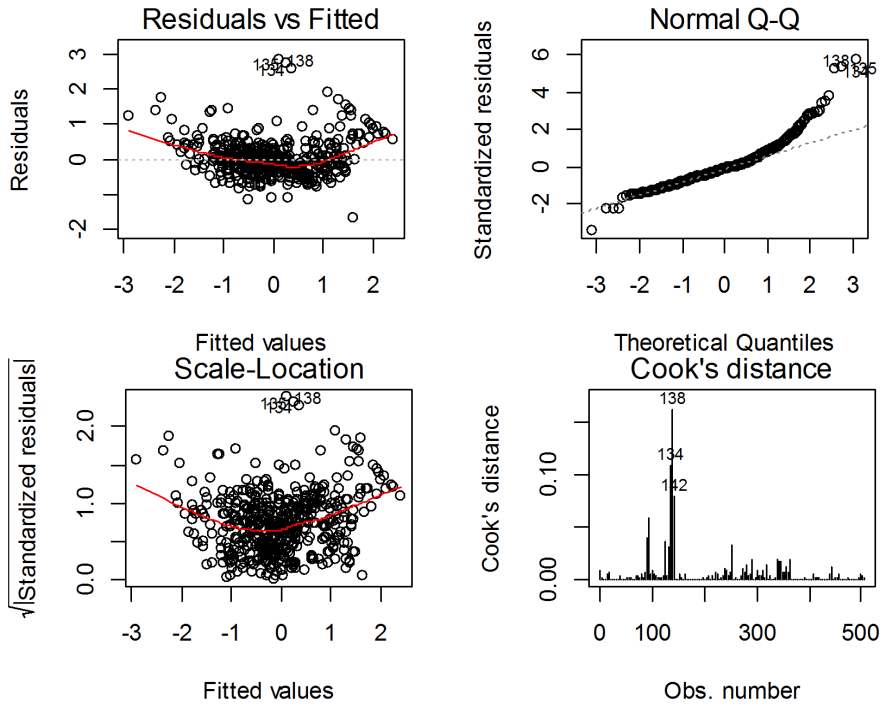

Figure 1. Residual and fitting diagram, QQ diagram, position scale diagram, residual and lever diagram are drawn
图 1. 绘制残差与拟合图、QQ 图、位置尺度图、残差与杠杆图

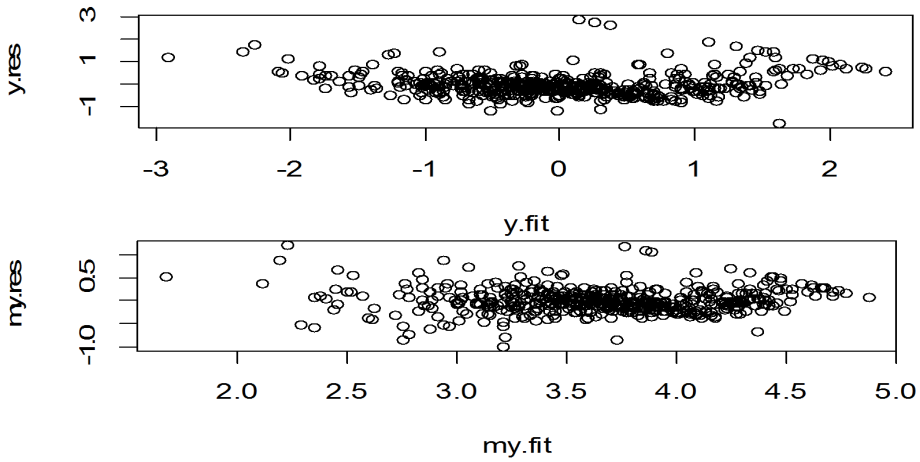

Figure 2. Comparison of two residuals before and after transformation
图 2. 变换前后两残差的比较

Table 3. Regression coefficient after Box-Cox transformation
表 3. Box-Cox 变换后的回归系数

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.0748	0.2876	17.646	<2e−16
CRIM	−0.0135	1.8603e−03	−7.278	1.34e−12
ZN	1.7157e−03	7.674e−04	2.236	0.02582
CHAS	0.1516	0.0485	3.128	0.00186

Continued

NOX	−1.0424	0.2006	−5.196	2.99e−07
RM	0.1422	0.0231	6.167	1.44e−09
DIS	−0.0761	0.0105	−7.216	2.03e−12
RAD	0.0191	3.5981e−03	5.309	1.67e−07
TAX	−7.896e−04	1.914e−04	−4.126	4.34e−05
PTRATIO	−0.0542	7.3246e−03	−7.399	5.94e−13
B	5.917e−02	1.517e−04	3.899	0.00011
LSTAT	−3.9745e−02	2.6914e−03	−14.767	<2e−16

由表 3 可以看出，部分回归系数较小，且自变量较多。

3.3. 多重共线性的诊断

使用函数 vif(myfit)，可以求出各个自变量的方差扩大因子。方差扩大因子均小于 10，不存在多重共线性。

使用条件数求得 $k = 62.47931$，$\sqrt{k} = 7.904386$，设计矩阵 X 没有多重共线性，同样由岭迹图也可以看出。

3.4. 降维

虽然以上求得的模型通过了检验，但是自变量数量较多，尝试使用达到降维的目的。

通过 Lasso 回归得到的 $R^2 = 0.788$，截距项为 4.332224。

Lasso 回归后不为零的回归系数见表 4。

Table 4. Non-zero Lasso regression coefficient
表 4. 不为零的 Lasso 回归系数

CRIM	CHAS	NOX	RM	DIS
−9.672307e−03	1.374909e−01	−6.228825e−01	1.671858e−01	−3.813086e−02
RAD	TAX	PTRATIO	B	LSTAT
1.103018e−03	−3.993305e−05	−4.779697e−02	4.879095e−04	−3.971597e−02
ZN				
0.0000000000				

从结果可以看到，ZN 项的系数为 0，TAX 系数的绝对值是剩下的所有项中值最大的，这里也可以看出来，其他项虽然系数都非常小但不为 0，这是因为这些项之间的关系是非线性的，无法用线性组合互相表示。

由图 3 可以看到图中的竖线对应于 Lasso 中迭代的次数，对应的系数值不为 0 的自变量即为选入的，竖线的标号与 step 相对应。

在进行 Lasso 回归后，自变量的数量只减少一个，且各回归系数取值较小，不好处理。

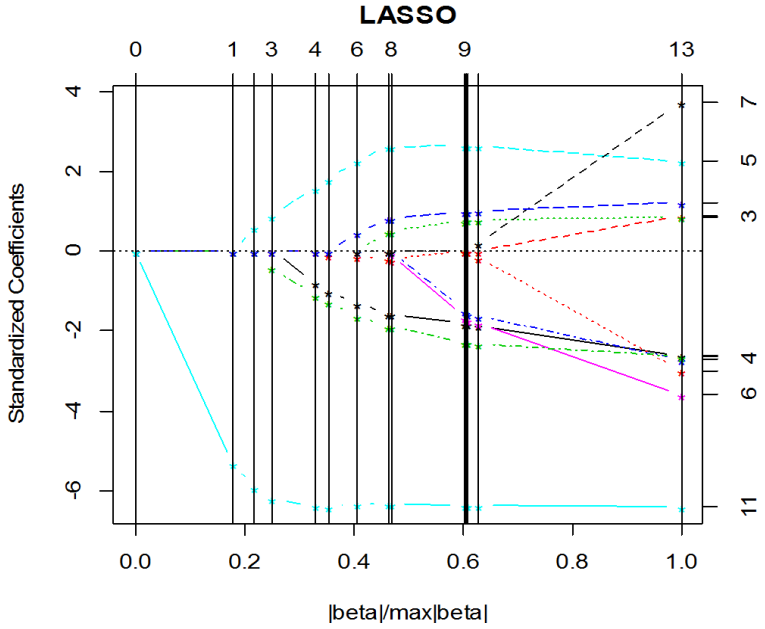

Figure 3. Shows the order in which the independent variables are selected
图 3. 展示自变量被选入的顺序

4. 讨论

4.1. 响应变量与部分自变量的回归模型

1) 通过计算自变量与响应变量的相关系数，可以发现与响应变量有较大相关关系的有 RM、PTRATIO、LSTAT 三个变量，因此对其建立线性回归模型。

方法与上同，经过 Box-Cox 变换后的回归方程 $R^2 = 0.7167$，$\hat{\sigma} = 0.3975$。

回归方程为：

$$yy = 4.889430 + 0.228333\text{RM} - 0.072931\text{PTRATIO} - 0.061199\text{LSTAT}$$

2) 回归系数的解释

RM 增加，MEDV 也会增加。因为随着房屋数量的增加，相对房屋价格应该会减小。

LSTAT 增加，MEDV 会减小。因为低收入者多的地方，他们居住的地区房屋价格会低一些。

PTRATIO 增加，MEDV 会减小。因为师生数量比表明了一个地方教育发展状况，比值越大，说明该地区缺老师，教育状况较差，因此该地区房价也会低。

4.2. 利用回归模型对自有住房的中位数 MEDV 进行预测

假设王某是一个在波士顿地区的房屋经纪人，使用此模型对客户进行评估他们想要出售的房屋的中位数报价，王某会建议每位客户的房屋销售价格大约为多少？客户的信息见表 5，建议见表 6。

Table 5. Information collected by three customers
表 5. 三个客户收集到的信息

自变量	客户 1	客户 2	客户 3
房屋内房间数	5	4	8
人口状况下降%	17%	32%	3%
城镇的学生与教师比例	15:1	22:1	12:1

Table 6. Suggests that the mean of the median house price is
表 6. 建议房价中位数的均值为

客户 1	客户 2	客户 3
17.83637	6.365289	43.99365

4.3. 模型分析

虽然模型的误差标准差较小，但是模型的拟合优度一般，可能是线性回归模型不合适，也有可能采集的数据不能充分解释响应变量的值，应该尝试建立非线性模型或其他模型提高拟合优度；数据中可以看到异常值，既不能盲目删除，也不应该置之不理，应具体分析，具体情况具体讨论。

4.4. 使用性探讨

1978 年采集的数据，在考虑通货膨胀的前提下，由于相关的政策发生了变化，因此在今天不适用；

以上数据所采集的变量，不能够完全描述一个房屋，房屋价值还受房屋外观、新旧程度等因素的影响；

像波士顿这样的大城市，回归模型仅适用于它本身，不能适用于其他乡镇。

参考文献

[1] 薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007.

[2] 周纪芗. 回归分析[M]. 上海: 华东师范大学出版社, 1993.

[3] Kabacoff, R.I. R 语言实战[M]. 王小宁, 刘撷芯, 黄俊文, 等, 译. 北京: 人民邮电出版社, 2016: 181.

[4] 何晓群, 刘文卿. 应用回归分析[M]. 第 5 版. 北京: 中国人民大学出版社, 2019.