Statistics and Application 统计学与应用, 2023, 12(6), 1485-1498
Published Online December 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.126152

Hans汉斯

基于回归方法的鲍鱼年龄预测

孙丽娜

云南财经大学统计与数学学院，云南 昆明

收稿日期：2023年9月25日；录用日期：2023年11月30日；发布日期：2023年12月6日

摘　要

本文基于物理测量确定鲍鱼年龄的方法，根据测量数据，利用R语言，建立线性回归、逻辑回归、岭回归、LASSO回归模型，来预测鲍鱼的年龄。并通过平均绝对误差MAE、均方误差MSE和对称平均绝对百分比误差SMAPE对模型进行评价，结果表明，LASSO回归模型的拟合优度更好。考虑到变量间相关性强，可能存在多重共线性，本文利用偏最小二乘及主成分分析两种方法对变量降维，降维后再进行回归分析，以期消除多重共线性对模型带来的影响。利用MSE评价模型，结果表明，这两种降维方法都没能减小MSE，反而得到模型的MSE更大。

关键词

线性回归，逻辑回归，岭回归，LASSO回归，偏最小二乘，主成分分析

Prediction of Abalone Age Based on Regression Methods

Lina Sun

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Received: Sep. 25th, 2023; accepted: Nov. 30th, 2023; published: Dec. 6th, 2023

Abstract

In this paper, based on physical measurements to determine the age of abalone, linear regression, logistic regression, ridge regression, and LASSO regression models are established to predict the age of abalone based on the measurement data, using R language. The models are evaluated by mean absolute error MAE, mean square error MSE, symmetric mean absolute percentage error SMAPE, and the results show that the LASSO regression model has a better goodness of fit. Considering the strong correlation between the variables and the possible existence of multicollinearity, this paper uses two methods of partial least squares and principal component analysis to reduce the dimensionality of the variables, and then regression is performed after the reduction of dimensionality, in order to

文章引用: 孙丽娜. 基于回归方法的鲍鱼年龄预测[J]. 统计学与应用, 2023, 12(6): 1485-1498.
DOI: 10.12677/sa.2023.126152

eliminate the impact of multicollinearity on the model. Using MSE to evaluate the model, the results show that both methods of dimensionality reduction fail to reduce the MSE, but instead, the MSE of the model is obtained to be larger.

Keywords

Linear Regression, Logistic Regression, Ridge Regression, LASSO Regression, Partial Least Squares, Principal Component Analysis

1. 引言

确定鲍鱼的年龄是一项繁杂又耗时的任务，要先通过锥体切掉其石灰质的外壳，然后进行染色，再利用显微镜观察，记录后计算环数，最后确定年龄。为了避免这样繁杂又耗时的任务，发展了一种更为简便且快速的确定年龄的方法，即通过物理测量来确定鲍鱼的年龄。本文正是基于物理测量的数据结果，建立合适的模型，从而预测鲍鱼的年龄。

2. 理论知识

2.1. 多元线性回归

2.1.1. 总体模型

随机变量 y 与变量 $x_1, x_2, \cdots, x_p$ 的线性回归模型[1]为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{2.1}$$

其中，$\beta_0, \beta_1, \cdots, \beta_p$ 是 $p + 1$ 个未知参数，β_0 是回归常数，$\beta_1, \beta_2, \cdots, \beta_p$ 是回归系数，均为待估参数，ε 是随机误差，一般假定 $\varepsilon \sim N(0, \sigma^2)$。

2.1.2. 样本模型

若有 n 组观测数据 $\left(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i\right)$，则(2.1)式可以表示为：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases} \tag{2.2}$$

写成矩阵的形式为：

$$y = X\beta + \varepsilon \tag{2.3}$$

其中，

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \tag{2.4}$$

$$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{bmatrix}\quad \varepsilon=\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}$$

其中，X 是一个 $n\times(p+1)$阶矩阵，称为回归设计矩阵或资料矩阵。

2.1.3. 参数估计

多元线性回归方程的待估参数 $\beta_0,\beta_1,\cdots,\beta_p$ 根据最小二乘法求得，即寻找参数 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$，使离差平方和：

$$Q\left(\beta_0,\beta_1,\cdots,\beta_p\right)=\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_px_{ip}\right)^2 \tag{2.5}$$

达到极小，亦即寻找 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$，使得：

$$Q\left(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p\right)=\min_{\beta_0,\ \beta_1,\cdots,\beta_p}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_{1i}-\cdots-\beta_px_{ip}\right)^2 \tag{2.6}$$

根据(2.6)求出的 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p$ 就是回归参数 $\beta_0,\beta_1,\cdots,\beta_p$ 的最小二乘估计。

2.2. 逻辑回归

2.2.1. 单变量模型

单变量 X 的逻辑回归模型[2]为：

$$p(X)=\frac{e^{\beta_0+\beta_1X}}{1+e^{\beta_0+\beta_1X}} \tag{2.7}$$

也可以写成：

$$\log\frac{p(X)}{1-p(X)}=\beta_0+\beta_1X \tag{2.8}$$

其中，β_0、β_1 是两个未知参数，β_0 是逻辑回归常数，β_1 是逻辑回归系数，均为待估参数。

2.2.2. 多变量模型

多变量 $X_1,X_2,\cdots X_p$ 的逻辑回归模型为：

$$p(X)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}} \tag{2.9}$$

也可以写成：

$$\log\frac{p(X)}{1-p(X)}=\beta_0+\beta_1X_1+\cdots+\beta_pX_p \tag{2.10}$$

其中，$\beta_0,\beta_1,\cdots,\beta_p$ 是 $p+1$ 个未知参数，β_0 是逻辑回归常数，$\beta_1,\beta_2,\cdots,\beta_p$ 是逻辑回归系数，均为待估参数。

2.2.3. 参数估计

以单变量模型为例，若有 n 组观测数据 $\left(x_1,x_2,...,x_n;y_i\right)\left(i=1,2,...,n\right)$，待估参数 β_0、β_1 的对数似然函数为：

$$l(\beta_0,\beta_1)=\log\left(\prod_{i:y_i=1}p(x_i)\prod_{i':y_{i'}=0}\left(1-p(x_{i'})\right)\right) \tag{2.11}$$

对 β_0、β_1 求偏导：

$$\frac{\partial l(\beta)}{\partial(\beta)}=X'\left(\hat{Y}-Y\right) \tag{2.12}$$

其中，$\beta=(\beta_0,\beta_1)$，$X=\begin{pmatrix}1x_1\\1x_2\\\vdots\\1x_n\end{pmatrix}$，$\hat{Y}=\frac{1}{1+e^{-\beta X}}$，$Y=(y_1,y_2,\cdots,y_n)'$。

再令(2.12)式等于 0，则解得 β 的最大似然估计。

2.3. 岭回归

当变量间出现多重共线性问题时，普通最小二乘法效果明显变差，针对这种情形，霍尔在 1962 年首先提出了一种改进最小二乘估计的方法，即岭估计。

当自变量间存在多重共线性，即$|X'X|\approx 0$ 时，设想给 $X'X$ 加上一个正定矩阵 ki ($k>0$)，那么 $X'X+kI$ 接近奇异的程度就会比 $X'X$ 接近奇异的程度小得多[3]。考虑到变量量纲不一的问题，将数据标准化，则 β 的岭回归估计为：

$$\hat{\beta}(k)=\left(X'X+kI\right)^{-1}X'y \tag{2.13}$$

其中，k 为岭系数。当 $k=0$ 时的岭回归估计 $\beta(0)$就是普通最小二乘估计。$\beta(k)$作为 β 的估计比最小二乘估计 β 稳定。

2.4. LASSO 回归

在原理上，LASSO 回归与岭回归的思想相类似，但惩罚项不是系数的平方而是其绝对值[4]，即在约束条件 $\sum_{j=1}^{p}\left|\beta_j\right|\le s$ 下，需要满足以下条件：

$$\left(\hat{\alpha},\hat{\beta}\right)=\arg\min_{(\alpha,\beta)}\sum_{i=1}^{n}\left(y_i-\alpha-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 \tag{2.14}$$

由于惩罚项取绝对值，LASSO 回归不像岭回归那样压缩系数，而是将系数归为 0，达到变量选择的功效。

2.5. 主成分分析

2.5.1. 总体模型

变量 $X_1,X_2,\cdots,X_p$ 的主成分为：

$$Z=\phi_1X_1+\phi_2X_2+\cdots+\phi_pX_p \tag{2.15}$$

其中，$(\phi_1,\phi_2,\cdots,\phi_p)$为载荷向量，$Z$ 为主成分的得分矩阵。

2.5.2. 样本模型

若有 n 组观测数据$(x_{i1},x_{i2},...,x_{ip})(i=1,2,...,n)$，则(2.15)式可以表示为：

$$\begin{cases} z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \ldots + \phi_{p1}x_{ip} \\ z_{i2} = \phi_{11}x_{i1} + \phi_{22}x_{i2} + \ldots + \phi_{p2}x_{ip} \\ \cdots\cdots \\ z_{im} = \phi_{1m}x_{i1} + \phi_{2m}x_{i2} + \ldots + \phi_{pm}x_{ip} \end{cases} \tag{2.16}$$

优化[5]：

$$\max_{\phi_{11},\phi_{21},\cdots,\phi_{p1}} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \right\}, \quad s.t.\sum_{j=1}^{p}\phi_j^2 = 1 \tag{2.17}$$

2.5.3. 方差贡献率

第 k 个主成分的方差贡献率为：

$$\frac{1}{n}\sum_{i=1}^{n} z_{ik}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jk}x_{ij}\right)^2 \tag{2.18}$$

第 m 个主成分的累积方差贡献率为：

$$\frac{1}{n}\sum_{k=1}^{m}\sum_{i=1}^{n} z_{ik}^2 \tag{2.19}$$

3. 实证分析

3.1. 数据说明与处理

3.1.1. 数据说明

本文选取的数据为 UCI 库的 Abalone 数据集[6]，共有 4177 个样本，9 个变量(表 1)。其中，通过锥体切壳、染色并利用显微镜观察，计算环数，从而确定鲍鱼的年龄。

Table 1. Variable description
表 1. 变量描述

变量名称	变量含义	中文释义		单位
Sex	M, F, and I (infant)	性别	雄性、雌性、婴儿	
Length	longest shell measurement	长度	最长外壳长度	mm
Diameter	perpendicular to length	直径	垂直于长度的直径	mm
Height	with meat in shell	高度	带壳肉的高度	mm
Whole weight	whole abalone	总重量	一只鲍鱼的重量	g
Shucked weight	weight of meat	去壳重量	肉的重量	g
Viscera weight	gut weight (after bleeding)	内脏重量	出血后的肠道重量	g
Shell weight	after being dried	壳重	晒干后的壳重	g
Rings	+1.5 gives the age in years	环数	+1.5 即为鲍鱼年龄	

3.1.2. 数据预处理

1) 变量转化

数据集中的 Sex 变量为定性变量，为便于后续建立模型，将其转化为虚拟变量(表 2)。

Table 2. Virtual variable
表 2. 虚拟变量

变量名称	值	条件
SexI	1	Sex = I
	0	其它
SexM	1	Sex = M
	0	其它

2) 数据划分

将数据划分为训练集和测试集，前 3133 个样本为训练集，后 1044 个样本为测试集。

3.2. 描述分析

3.2.1. 数据概况

表 3 展示了数据中 9 个变量的基本情况，其中 Sex 变量为定性变量，表明鲍鱼的性别，由图 1 可知，鲍鱼中雄性占 36.6%，雌性占 31.3%，婴儿占 32.1%；Rings 变量为鲍鱼的环数，指代鲍鱼的年龄，由图 2 可知，鲍鱼环数从 1 到 9 的频数逐渐增加，至环数为 9 时最多，之后又逐渐减少；其余 7 个变量的最小值、最大值、均值与四分位数均列于表中。

Table 3. Data overview
表 3. 数据概况

Sex	Length	Diameter	Height	Whole
Length: 4177	Min.: 0.075	Min.: 0.0550	Min.: 0.0000	Min.: 0.0020
Class: character	1st Qu.: 0.450	1st Qu.: 0.3500	1st Qu.: 0.1150	1st Qu.: 0.4415
Mode: character	Median: 0.545	Median: 0.4250	Median: 0.1400	Median: 0.7995
	Mean: 0.524	Mean: 0.4079	Mean: 0.1395	Mean: 0.8287
	3rd Qu.: 0.615	3rd Qu.: 0.4800	3rd Qu.: 0.1650	3rd Qu.: 1.1530
	Max.: 0.815	Max.: 0.6500	Max.: 1.1300	Max.: 2.8255

Shucked	Viscera	Shell	Rings
Min.: 0.0010	Min.: 0.0005	Min.: 0.0015	Min.: 1.000
1st Qu.: 0.1860	1st Qu.: 0.0935	1st Qu.: 0.1300	1st Qu.: 8.000
Median: 0.3360	Median: 0.1710	Median: 0.2340	Median: 9.000
Mean: 0.3594	Mean: 0.1806	Mean: 0.2388	Mean: 9.934
3rd Qu.: 0.5020	3rd Qu.: 0.2530	3rd Qu.: 0.3290	3rd Qu.: 11.000
Max.: 1.4880	Max.: 0.7600	Max.: 1.0050	Max.: 29.000

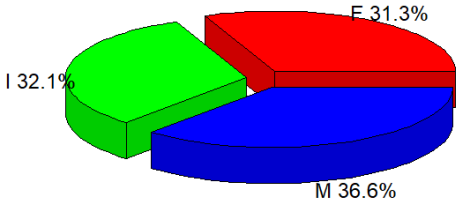

Figure 1. Sex distribution of abalone
图 1. 鲍鱼性别分布

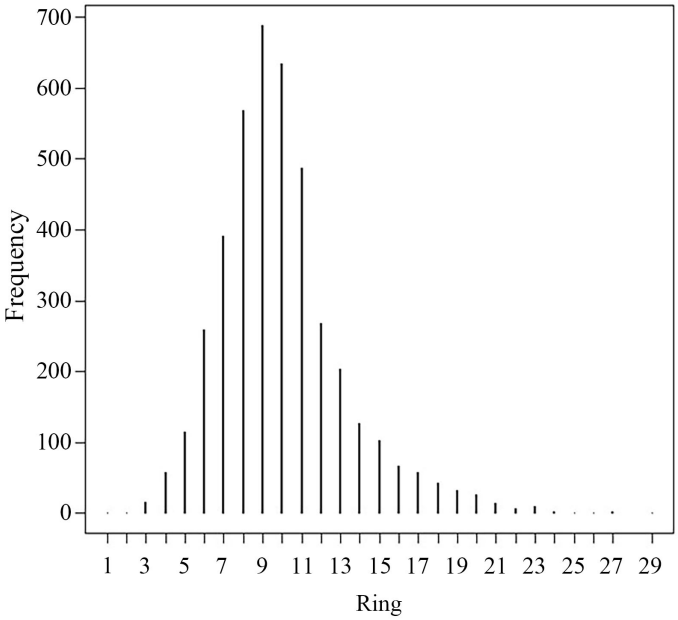

Figure 2. Distribution of abalone rings
图 2. 鲍鱼环数分布

3.2.2. 变量相关阵

由于 Sex 变量为定性变量，所以除去该变量后，再计算剩余 8 个变量的相关矩阵(表 4)，再绘制变量的相关矩阵图(图 3)。

Table 4. Variable correlation matrix
表 4. 变量相关阵

	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Rings
Length	1.000	0.987	0.828	0.925	0.898	0.903	0.898	0.557
Diameter	0.987	1.000	0.834	0.925	0.893	0.900	0.905	0.575
Height	0.828	0.834	1.000	0.819	0.775	0.798	0.817	0.557
Whole	0.925	0.925	0.819	1.000	0.969	0.966	0.955	0.540
Shucked	0.898	0.893	0.775	0.969	1.000	0.932	0.883	0.421
Viscera	0.903	0.900	0.798	0.966	0.932	1.000	0.908	0.504
Shell	0.898	0.905	0.817	0.955	0.883	0.908	1.000	0.628
Rings	0.557	0.575	0.557	0.540	0.421	0.504	0.628	1.000

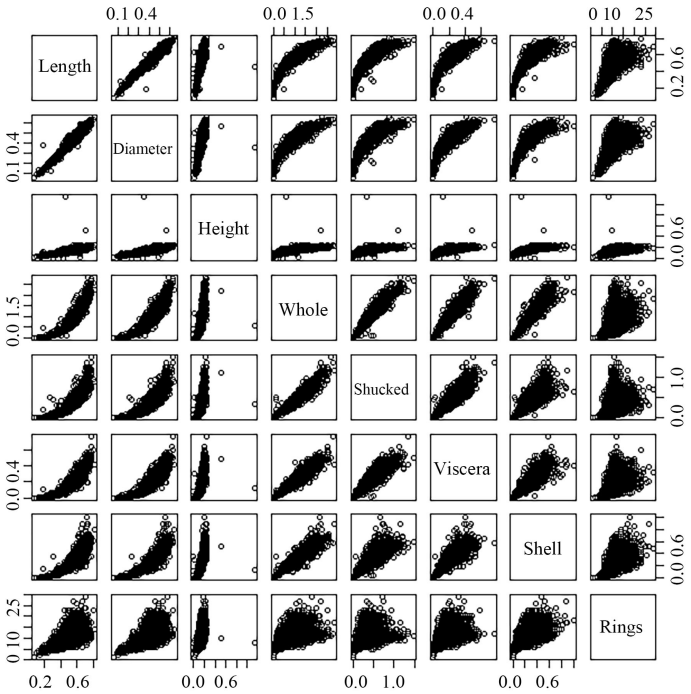

Figure 3. Plot of two-by-two variable correlation matrix
图 3. 两两变量相关矩阵图

3.2.3. 箱线图

单独绘制 Sex 变量与其它 8 个变量的箱线图(图 4)

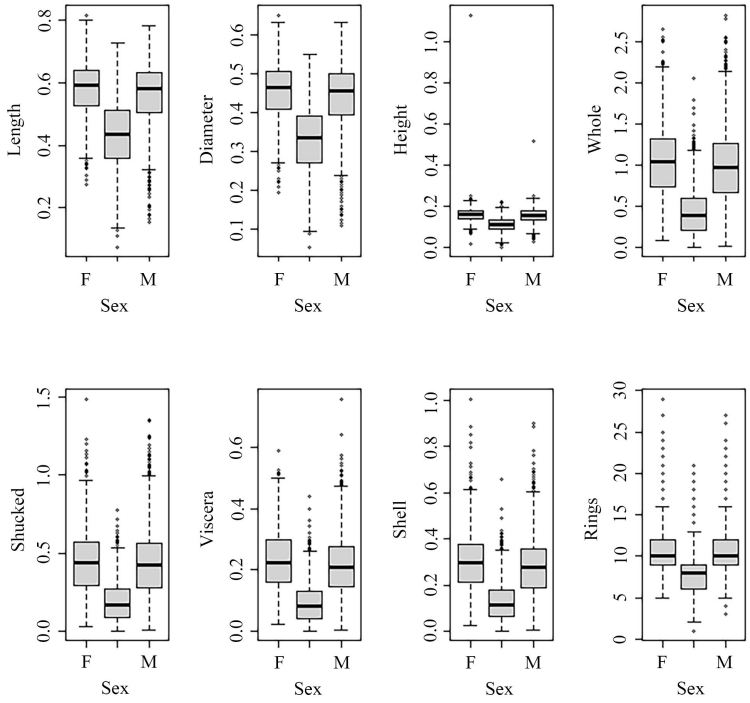

Figure 4. Box plot
图 4. 箱线图

3.3. 回归模型分析

将 Sex 变量转化为虚拟变量后，以 Rings 变量为响应变量，基于训练集数据分别进行线性回归、泊松回归、岭回归及 LASSO 回归，然后利用所建模型，基于测试集数据分别对 Rings 的值进行预测，最后计算各模型在预测方面的均方误差。

3.3.1. 线性回归

由表 5 可知，SexM 变量与 Length 变量不显著，其余变量均在 99.9%的置信水平下显著，表明鲍鱼的年龄受性别及其外壳长度的影响较低。模型的平均绝对误差 MAE 为 1.5936，均方误差 MSE 为 4.5215，对称平均绝对百分比误差 SMAPE 为 0.1551，但可决系数为 0.5429，较小，说明模型拟合程度并不理想。

Table 5. Regression coefficients for the four regression models
表 5. 四个回归模型的回归系数

	线性回归		泊松回归		岭回归	LASSO 回归
(Intercept)	4.0428	***	1.4947	***	4.5489	5.0048
SexI	−0.8475	***	−0.1030	***	−0.9334	−0.8212
SexM	0.0307		0.0074		−0.0174	
Length	−1.4324		0.1946		2.3323	
Diameter	12.2695	***	1.5614	***	5.2324	6.7158
Height	8.7956	***	0.8773	***	11.0120	7.7580
Whole	9.1100	***	0.7751	***	0.5204	
Shucked	−19.6321	***	−1.8160	***	−4.8186	−7.9255
Viscera	−10.8043	***	−0.9134	***	−1.2442	
Shell	8.9358	***	0.4922	**	9.7135	17.6678

注：星号代表显著性水平，*、**、***分别代表在 10%、5%、1%水平下显著。

3.3.2. 泊松回归

由表 5 可知，变量的显著性检验结果与线性回归的结果相差不大，除了 Shell 变量是在 99%的置信水平下显著的。模型的平均绝对误差 MAE 为 1.6242，均方误差 MSE 为 4.7472，对称平均绝对百分比误差 SMAPE 为 0.1582，但 AIC 值为 14255，较大，说明模型拟合程度并不理想。

3.3.3. 岭回归

为了有效避免数据的过拟合问题，在建立岭回归模型时采用了十折交叉验证法。首先，通过十组子样本的交叉验证，绘制回归系数及模型均方误差随岭系数 λ 变化的系数变化图(图 5 左上)和均方误差变化图(图 5 右上)。其次，依据均方误差 MSE 最小原则，选择最优的岭系数 λ，为 0.2076。最后，采用最优岭系数进行预测，并计算出模型的平均绝对误差、均方误差、对称平均绝对百分比误差，分别为 1.6324、4.8270、0.1586。

3.3.4. LASSO 回归

与岭回归相似，也采用十折交叉验证法进行 LASSO 回归。首先，通过十组子样本的交叉验证，绘制回归系数及模型均方误差随惩罚因子 λ 变化的系数变化图(图 5 左下)和均方误差变化图(图 5 右下)。其次，依据均方误差 MSE 最小原则，选择最优的惩罚因子 λ，为 0.0018。最后，采用最优惩罚因子进

行预测，并计算出模型的平均绝对误差、均方误差、对称平均绝对百分比误差，分别为 1.5928、4.5194、0.1551。

3.3.5. 模型比较

由表 6 可知，LASSO 回归模型的 MAE、MSE 和 SMAPE 最小，岭回归模型的 MAE、MSE 和 SMAPE 最大。再对比表 5 中岭回归与 LASSO 回归的系数，可以发现 LASSO 回归模型中有四个变量的系数为 0，这是因为 LASSO 具有执行变量选择的功效，从而让模型变得更容易解释。所以，即使与岭回归一样，当最小二乘估计方差过高，可以减少以偏差小幅增加为代价的方差，LASSO 回归的表现要比岭回归的表现更好。

3.4. 降维

由变量相关阵(表 4)与相关矩阵图(图 3)可知，变量间相关系数较高，可能存在多重共线性问题，所以考虑先对数据进行降维，再建立模型。本文采用偏最小二乘及主成分分析两种降维方法对模型进行优化。在建模前，先对数据进行标准化处理，避免模型结果受方差影响。

Table 6. MSE of the regression model
表 6. 回归模型的 MSE

模型	MAE	MSE	SMAPE
线性回归	1.593591	4.521460	0.1551455
泊松回归	1.624197	4.747157	0.1582436
岭回归	1.632424	4.827036	0.1585987
LASSO 回归	1.592775	4.519386	0.1550553

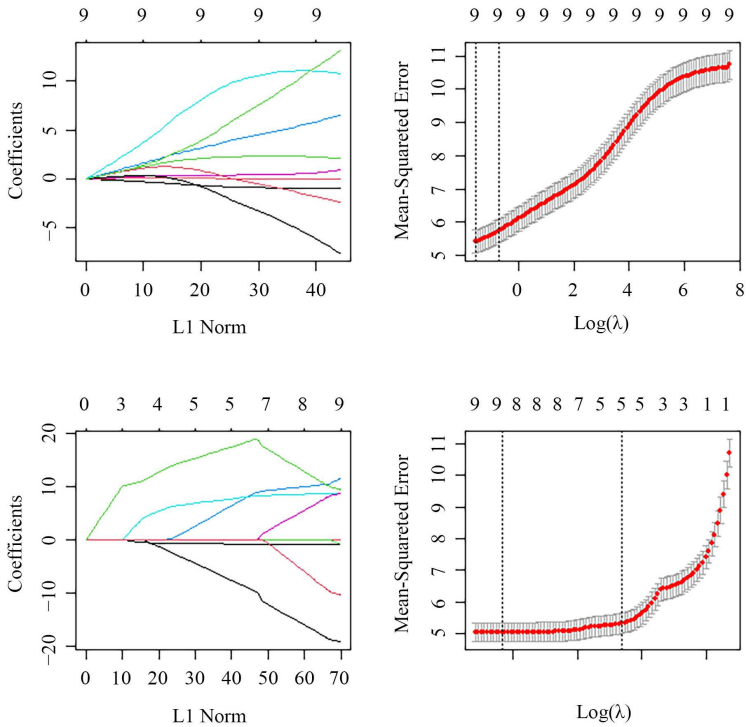

Figure 5. Plot of coefficient changes and MSE changes of ridge regression and LASSO regression
图 5. 岭回归、LASSO 回归的系数变化与 MSE 变化图

3.4.1. 偏最小二乘回归

由表 7 可知，前三个主成分的累积方差贡献率有 89.11%，但前两个主成分的累积方差贡献率已达 79.36%，接近 80%，所以可以考虑主成分个数为 2 或 3 两种情况[7] [8]。当采用 2 个主成分时，MSE 为 5.0441，而当采用 3 个主成分时，MSE 为 4.8048。所以，应选前三个主成分进行建模，系数如表 8 所示。

3.4.2. 主成分回归

由表 9 可知，前两个主成分的特征值都大于 1，且其累积方差贡献率有 88.4%，再根据碎石图(图 6)可知，当主成分个数大于等于 3 时，线的走势变平缓，可以确定选取前两个主成分进行建模[9] [10]，系数如表 8 所示，MSE 为 6.4085。

Table 7. Explanation of variance
表 7. 方差解释情况

	1 comp	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps
X	75.31	79.36	89.11	93.7	96.99	98.44	99.39	99.87	100.00
Rings	34.45	47.29	50.22	52.63	52.87	53.24	53.83	54.26	54.29

Table 8. Regression coefficients of the model after dimensionality reduction
表 8. 降维后模型的回归系数

	偏最小二乘回归	主成分回归
SexI	−0.6346	−0.2239
SexM	−0.3642	0.1590
Length	0.3790	0.2600
Diameter	0.7340	0.2615
Height	0.8929	0.2334
Whole	0.0870	0.2659
Shucked	−1.9577	0.2565
Viscera	−0.4791	0.2602
Shell	2.1021	0.2576

Table 9. Principal component variance contribution ratio
表 9. 主成分方差贡献率

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
Standard deviation	2.610	1.070	0.635	0.524	0.409	0.337	0.252	0.112	0.082
Proportion of variance	0.757	0.127	0.045	0.031	0.019	0.013	0.007	0.001	0.001
Cumulative proportion	0.757	0.884	0.929	0.960	0.978	0.991	0.998	0.999	1.000

Table 10. Load matrix
表 10. 载荷矩阵

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
SexI	−0.256	−0.827	0.737	0.100	−0.619	0.360			
SexM	0.131	0.145	−0.936	0.916	−0.518	0.247			
Length	0.370		0.150		−0.158	−0.849	0.618	−0.328	0.687
Diameter	0.372		0.157		−0.136	−0.723	0.722	−0.140	−0.682
Height	0.333	0.415	0.105	−0.363	−0.689	0.960	−0.426	0.106	
Whole	0.377	−0.320	0.110		0.177	0.275	−0.118	0.390	0.184
Shucked	0.363	−0.633		−0.117		−0.221	−0.397	0.412	−0.163
Viscera	0.368	−0.376			0.244	0.719		−0.702	
Shell	0.367	0.101	0.286	0.241	0.194		−0.545	0.256	

Table 11. Principal component score
表 11. 主成分得分

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9
SexI	0.250	0.522	0.797	0.159	0.029	−0.031	0.047	−0.014	0.004
SexM	−0.127	−0.817	0.550	0.104	0.023	−0.025	0.027	−0.003	0.001
Length	−0.368	0.099	0.033	0.042	0.596	0.083	0.049	0.698	0.021
Diameter	−0.369	0.090	0.009	0.063	0.586	0.004	0.010	−0.713	−0.012
Height	−0.335	0.080	−0.101	0.867	−0.305	0.160	−0.022	0.008	0.000
Whole	−0.375	0.095	0.106	−0.215	−0.231	−0.052	−0.110	−0.017	0.851
Shucked	−0.362	0.095	0.175	−0.312	−0.225	0.489	−0.550	−0.010	−0.373
Viscera	−0.366	0.090	0.074	−0.245	−0.274	0.154	0.807	−0.030	−0.204
Shell	−0.364	0.096	0.047	−0.052	−0.166	−0.837	−0.168	0.058	−0.307

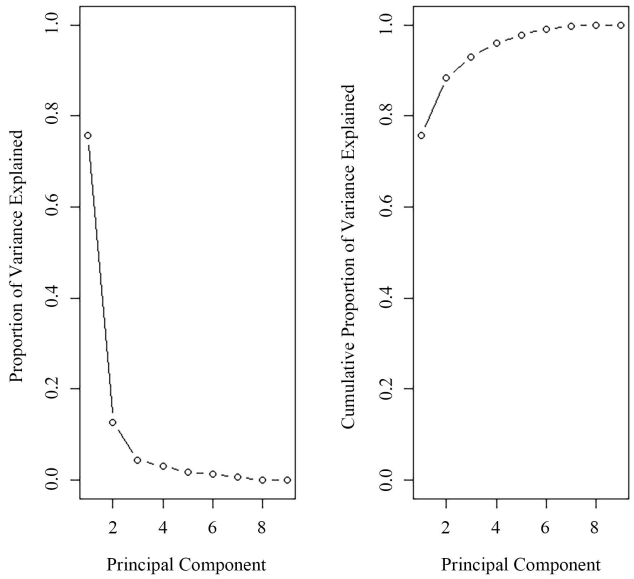

Figure 6. Gravel chart
图 6. 碎石图

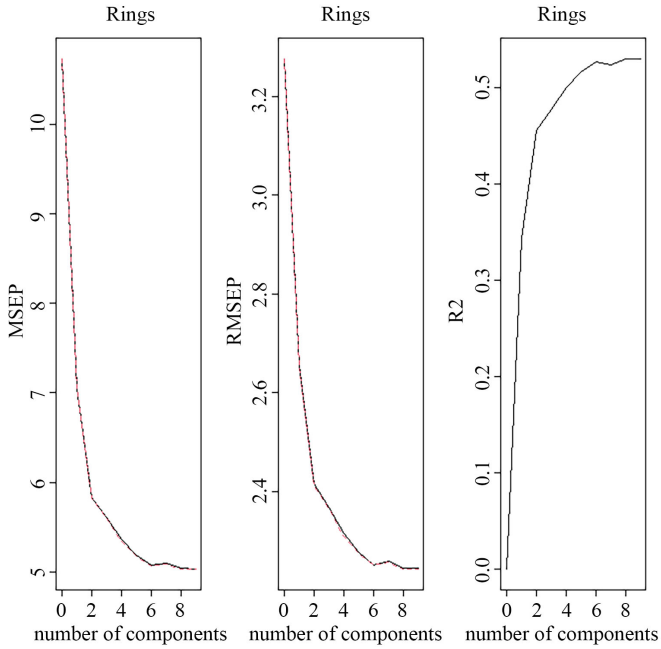

Figure 7. Plot of MSEP, RMSEP and R2 changes for partial least squares regression
图 7. 偏最小二乘回归的 MSEP、RMSEP 与 R2 变化图

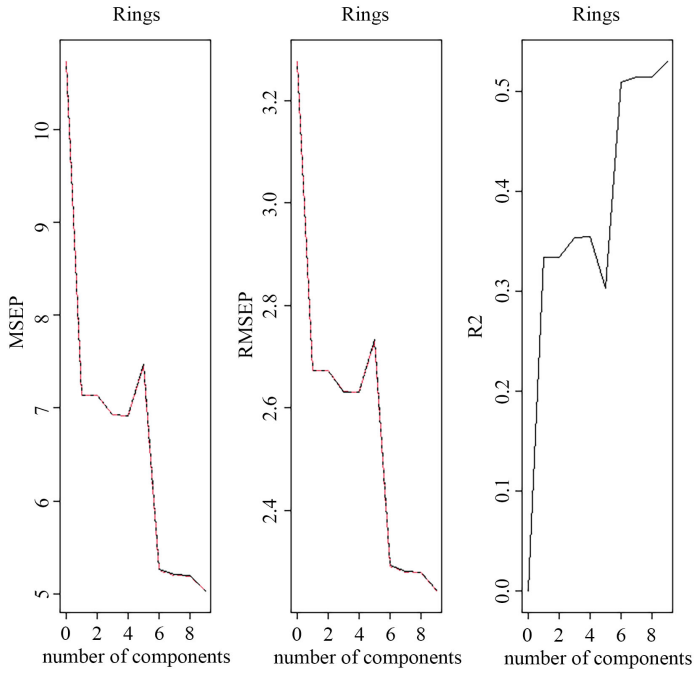

Figure 8. Plot of MSEP, RMSEP and R2 changes for principle component regression
图 8. 主成分回归的 MSEP、RMSEP 与 R2 变化图

3.4.3. 模型比较

对比偏最小二乘回归模型与主成分回归模型的 MSE，偏最小二乘回归模型的 MSE 更小，为 4.8048，所以偏最小二乘法比主成分分析法表现更优。但是通过对比降维后与降维前模型的 MSE，发现降维后模

型的 MSE 要比降维前的更大，说明降维并没有显著优化模型，反而表现欠佳。有两点原因：一为降维后模型的可决系数(图 7 与图 8)甚至不如线性回归模型的好；二为不论是偏最小二乘回归模型的载荷矩阵(表 10)，还是主成分回归的得分矩阵(表 11)，都难以解释[11]。

4. 结论

本文基于 UCI 库的 Abalone 数据集，共 4177 个样本，将其划分为 3133 个样本的训练集和 1044 个样本的测试集，利用训练集样本建立线性回归、逻辑回归、岭回归、LASSO 回归模型，再利用测试集样本分别预测鲍鱼的年龄，最后通过模型评价指标平均绝对误差 MAE、均方误差 MSE 和对称平均绝对百分比误差 SMAPE 来判断模型优劣，对应的值越小，模型越好。实证分析结果表明，LASSO 回归模型的 MAE、MSE 和 SMAPE 最小，分别为 1.5928、4.5194 和 0.1551，岭回归模型的 MAE、MSE 和 SMAPE 最大，分别为 1.6324、4.8270 和 0.1586。原因是 LASSO 具有执行变量选择的功效，即使与岭回归一样，当最小二乘估计方差过高，可以减少以偏差小幅增加为代价的方差，LASSO 回归的表现要比岭回归的表现更好。

结合变量相关阵及相关矩阵图，发现变量间相关性强，有多重共线性存在的可能，为了避免多重共线性对模型评价产生影响，本文采取了两种降维方法，即偏最小二乘法和主成分分析法，以期通过降维再进行回归来消除多重共线性对模型产生的影响。结果表明，即使比主成分回归模型表现更好的偏最小二乘回归模型的均方误差 MSE 都较大，为 4.8048，而主成分回归模型的 MSE 甚至高达 6.4085，两种方法都未达到预期效果。

参考文献

[1] 王学民. 应用多元统计分析[M]. 上海: 上海财经大学出版社, 2017: 328.

[2] 胡雪梅, 谢英, 蒋慧凤. 基于惩罚逻辑回归的乳腺癌预测[J]. 数据采集与处理, 2021, 36(6): 1237-1249. https://doi.org/10.16337/j.1004-9037.2021.06.017

[3] 张瑶瑶, 朱小栋. 基于岭回归极限学习机的微博垃圾用户分类[J]. 计算机与数字工程, 2021, 49(11): 2326-2330.

[4] 方彤, 苏治. 一种基于 LASSO 的多变量混频 GARCH 模型设计与优化算法研究[J]. 数量经济技术经济研究, 2021, 38(12): 146-163. https://doi.org/10.13653/j.cnki.jqte.2021.12.007

[5] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R. Springer, Berlin, 426. https://doi.org/10.1007/978-1-4614-7138-7

[6] Dua, D. and Graff, C. (2019) UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. https://doi.org/10.24432/C55C7W

[7] 王刚, 张福印, 李明辉, 王金龙, 王艺博, 武传伟. 基于偏最小二乘回归算法的空气质量监测系统研究[J]. 传感器与微系统, 2022, 41(1): 37-40+49. https://doi.org/10.13873/J.1000-9787(2022)01-0037-04

[8] Cui, N.N., Wang, G.X., Ma, Q.H., Zhao, T.T., Han, Z.T., Yang, Z. and Liang, L.S. (2021) Evolution of Lipid Characteristics and Minor Compounds in Hazelnut Oil Based on Partial Least Squares Regression during Accelerated Oxidation Process. *LWT*, **150**, Article ID: 112025. https://doi.org/10.1016/j.lwt.2021.112025

[9] 刘鹏飞, 黄仕元, 张鸿钦, 丁志鹏, 李赢杰. 基于主成分分析与灰色预测的新型城镇化综合水平测度——以湖南省为例[J]. 华中建筑, 2021, 39(12): 57-63. https://doi.org/10.13942/j.cnki.hzjz.2021.12.012

[10] 黄佳文, 孙瑞, 阮宇飞. 基于 PCA 与 K-Means 的注射成形制品质量在线检测[J]. 电子技术与软件工程, 2021(21): 117-120.

[11] 赵志挺, 朱亮宇, 高珣洋, 王力. 基于主成分分析协同深度神经网络的带钢板凸度预测[J/OL]. 冶金自动化, 2021: 1-12. http://kns.cnki.net/kcms/detail/11.2067.TF.20211129.1554.004.html