Statistics and Application 统计学与应用, 2023, 12(1), 40-52
Published Online February 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.121006

Hans汉斯

基于线性模型的帆船性能影响因素分析

姜仁学

云南财经大学，云南 昆明

收稿日期：2023年1月15日；录用日期：2023年2月5日；发布日期：2023年2月17日

摘　要

本研究利用代尔夫特船舶流体力学实验室的帆船数据，利用线性模型量化了各变量与帆船单位重量排水的剩余阻力之间的关系，并基于该数据提出了帆船设计的优化方案。首先，本文探究了帆船数据集中各变量间的相关关系，发现帆船的单位重量排水剩余阻力与各变量间存在较强的对数线性关系。当对帆船的单位重量剩余阻力以2为底取对数后发现，取对数后的单位重量排水剩余阻力与各变量间存在一定的线性关系，但与各可控因素之间的线性关系均不显著。其次，本文探究了各可控因素与弗劳德数之间的交互作用对帆船单位重量排水剩余阻力的影响。将交互作用引入后，各自变量间产生了较强的共线性，于是分别用岭回归和Lasso拟合数据，发现Lasso在此数据上表现出了较好的效果。其拟合的取对数数据的回归均方为0.220，原始数据的回归均方为4.64。且其结果中，取对数后数据的回归残差可以均匀地分布在某个范围内。最后，根据得到的对数线性模型可知，在弗劳德数一定的情况下，各指标中梁宽吃水量比对帆船的单位重量排水剩余阻力影响最大，其次是船长排水量比，最后是浮心纵坐标。故在实际生产中，设计者更应该关注帆船的船长排水量比。

关键词

线性关系，对数线性，多重共线性，岭回归，Lasso

Analysis of Influencing Factors of Sailing Ship Performance Based on Linear Model

Renxue Jiang

Yunnan University of Finance and Economics, Kunming Yunnan

Received: Jan. 15th, 2023; accepted: Feb. 5th, 2023; published: Feb. 17th, 2023

Abstract

In this study, the relationship between the variables and the residual resistance of the vessel per

文章引用: 姜仁学. 基于线性模型的帆船性能影响因素分析[J]. 统计学与应用, 2023, 12(1): 40-52.
DOI: 10.12677/sa.2023.121006

unit weight is quantified by linear model using the sailing data from the Laboratory of Ship Fluid Mechanics in Delft, and based on the data, an optimization scheme of sailing design is proposed. First of all, this paper explores the correlation between variables in the sailing data set, and finds that there is a strong log-linear relationship between the residual resistance per unit weight of sailing boat and each variable. When the logarithm of the residual resistance per unit weight of a sailing boat is taken as base 2, it is found that there is a certain linear relationship between the residual resistance per unit weight after logarithm and various variables, but the linear relationship between the residual resistance per unit weight and various controllable factors is not significant. Secondly, this paper explores the interaction between the controllable factors and Froude number on the residual resistance of the vessel per unit weight. After the interaction was introduced, a strong collinearity was generated between the variables. Therefore, ridge regression and Lasso were used to fit the data respectively, and it was found that Lasso showed a better effect on the data. The fitting regression mean square of logarithmic data is 0.220, and that of original data is 4.64. In the results, the regression residuals of logarithmic data can be evenly distributed in a certain range. Finally, according to the obtained log-linear model, under the condition of constant Froude number, the ratio of beam width to draft of each index has the greatest influence on the residual resistance of the vessel per unit weight, followed by the ratio of captain displacement, and finally the ordinate of center of buoyancy. Therefore, in the actual production, designers should pay more attention to the skipper displacement ratio.

Keywords

Linear Relation, Logarithmic Linearity, Multicollinearity, Ridge Regression, Lasso

1. 背景介绍

帆船是利用风力前进的船，是继舟、筏之后的一种古老的水上交通工具，起源于居住在海河区域的古代人的水上交通工具，已有 5000 多年的历史。现代帆船始于荷兰，经过船舶技术的不断发展，帆船的运输功能被现代的机器动力船所取代而逐渐发展成为了一种娱乐项目或体育竞技项目。现代帆船按桅杆数可分为单桅帆船、双桅帆船和多桅帆船，按船底型可分为平底和尖底帆船，按船首型可分为宽头、窄头和尖头帆船。帆船作为一种娱乐项目或体育竞技项目在大众之间广受欢迎，因而“高性能帆船的设计”逐渐引起人们的重视。

在文献[1]中，张益采用有限元分析的方法研究了竞赛帆船龙骨盒结构的参数优化设计问题，在该文章中作者采用有限元分析的方法计算了龙骨盒结构的变形和应力，根据帆船的实际损坏区域提出了基于响应面法的龙骨盒形状、尺寸优化方法。最后，作者采用遗传算法对玻璃钢层合板的铺层顺序进行优化，发现在新的铺层顺序下龙骨盒强度和承载能力有了较大提高。

在文献[2]中，胡健康等人采用可变质量变化对大倾角稳性影响的计算方法研究了载荷位置的变化对龙骨帆船大倾角稳性的影响程度，研究表明，在设计、检验和驾驶帆船时要注意人员布局对帆船稳性的影响。

在文献[3]中，许小颖等研究了三体船阻力性能预报实验及仿真模拟的问题，在该文章中作者采用多元线性回归模型对相关数据进行研究，得到了单体船在不同弗劳德数(Fr)下的阻力系数的经验计算估算公

式。之后作者采用计算机模拟的方式验证了回归公式应用于三体船阻力性能预报的可能性。

在人们的研究中发现，帆船性能的重要衡量指标“单位重量排水的剩余阻力”和帆船的一些结构数据指标之间存在一定的数量关系。然而对其数量关系的研究却存在明显空白。本文意在根据已有数据探索帆船的“单位重量排水的剩余阻力”和帆船的一些结构数据指标之间数量关系，据此得到一种可以有效降低帆船“单位重量排水的剩余阻力”的设计方案。

2. 研究目的和意义

帆船在当今的娱乐项目及体育竞技中具有举足轻重的地位，且具有极高的经济效益。一款高性能的帆船能够直接带来可观的经济收益，对于经济发展和人们业余生活的丰富具有极大地影响。

本文意在使用统计模型探索帆船的“单位重量排水剩余阻力”和帆船设计时的一些结构数据指标之间数量关系，并据此提出对帆船的性能改进具有一定意义的方案，以图对帆船的性能进行一定程度的优化。

3. 数据介绍

在本文中所使用的数据集为 Yacht Hydrodynamics Data Set，它来源于著名的 UCI 数据库，数据链接为 http://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics。该数据集在代尔夫特船舶流体力学实验室通过实验得到，共包括 308 个观测，其中每个观测具有七个特征，均为数值型数据，其具体信息如下表 1。这些实验中共涉及 22 种不同的船体形式，均源自于弗朗斯·马斯设计的“标准 43”密切相关的母体形式。

Table 1. Variable names in Chinese and English
表 1. 变量名称中英文对照表

变量名称(英文)	变量名称(中文)
Longitudinal position of the center of buoyancy	浮心纵坐标(1)
Prismatic coefficient	菱形系数(2)
Length-displacement ratio	船长排水量(3)比
Beam-draught ratio	梁宽吃水量(4)比
Length-beam ratio	船长梁宽比
Froude number	弗劳德数(5)
Residuary resistance per unit weight of displacement	单位重量排水的剩余阻力(6)

(1) 浮心纵坐标亦称“浮心纵向位置”，是船的浮力中心在整个帆船上的纵向位置，可通过船的设计进行调整。(2) 菱形系数是指与基平面相平行的任一水线面下的船的型排水体积与对应的船长及中横剖面面积的乘积所表示的棱柱体体积之比。也等于船的方形系数与中横剖面系数之比。表示了排水量沿船长的分布，也表示了船的首尾部相对于中部的尖瘦程度。对船的快速性影响较大，对船的耐波性及建造工艺等也有影响。(3) 排水量是指船装满货物后排开水的重力，也就是船满载后受到水的浮力。根据物体漂浮的条件可得出下列公式：排水量(浮力) = 船自身的重量 + 满载时货物的重量(所受的重力 = 浮力)。(4) 吃水量是指水线面与船底基平面之间的垂直距离，即水线面至船底龙骨板下缘的垂直距离。(5) 弗劳德数，符号为 Fr，是水的惯性力与重力之比，是用来确定水流动态如急流、缓流的一个量纲为一的数。(6) 剩余阻力是指船舶收到的总阻力减去摩擦阻力。

在本文中，设定 Residuary resistance per unit weight of displacement 为因变量，设定 Longitudinal position of the center of buoyancy，Prismatic coefficient，Length-displacement ratio，Beam-draught ratio，Length-beam ratio，Froude number 为自变量，通过建立统计模型来研究这些自变量与因变量之间的关系。

4. 多元线性回归模型介绍

4.1. 多元线性回归模型

多元线性回归模型常用来表述变量 y 和 x 之间的随机线性关系，它的一般形式为[4]：

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

其中 $\varepsilon\sim N\left(0,\sigma^2I\right)$ 是随机误差项，x_i 为非随机变量且相互之间不相关，y 为随机变量。如果对 y 和 x 进行了 n 次观测，得到 n 组观测值 $y_i,x_{1i},x_{2i},\cdots,x_{pi},i=1,\cdots,n$，则他们满足以下关系式：

$$y_i=\beta_0+\beta_{1i}x_{1i}+\beta_{2i}x_{2i}+\cdots+\beta_{pi}+\varepsilon_i,i=1,\cdots,n$$

其矩阵表示为：

$$Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix},X=\begin{bmatrix}1&x_{11}&\cdots&x_{p1}\\1&x_{12}&\cdots&x_{p2}\\\vdots&\vdots&\ddots&\vdots\\1&x_{1n}&\cdots&x_{pn}\end{bmatrix},\beta=\begin{bmatrix}\beta_0\\\beta_1\\\vdots\\\beta_p\end{bmatrix},\varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix},$$

此时模型可写作：

$$Y=X\beta+\varepsilon$$

4.2. 模型的参数估计方法

在正态假定下，如果 X 是列满秩的，则普通线性回归模型参数的最小二乘估计为：

$$\hat{\beta}=\left(X^{\mathrm{T}}X\right)^{-1}X^{\mathrm{T}}Y,$$

则 Y 的估计值为：

$$\hat{Y}=X\hat{\beta},$$

记残差向量为 $e=Y-\hat{Y}=Y-X\hat{\beta}$，则随机误差方差 σ^2 的最小二乘估计为：

$$\hat{\sigma}^2=\frac{e^{\mathrm{T}}e}{n-p-1}.$$

4.3. 可能存在的问题及检测方法

4.3.1. 自变量与因变量间为非线性关系

线性回归模型假定自变量与因变量之间为线性关系。如果自变量与因变量的真实关系是非线性的，那么我们得出的几乎所有结论都是不可信的，而且模型的预测精度也可能显著降低。

残差图是一种很有用的图形工具，可用于识别自变量与因变量之间的非线性关系。在一元线性回归中，我们可以绘制残差 $e_i=y_i-\hat{y}_i$ 与自变量 x_i 之间的散点图。在多元线性回归中，因为有多个预测变量，我们转而绘制残差与预测值 $\hat{y}_i$ 的散点图。理想情况下残差图显示不出明显的规律。若存在明显规律，则表明线性模型的某些方面可能存在问题。

4.3.2. 误差项自相关

回归系数和拟合值的标准误的计算都基于误差项不相关的假设。如果误差项之间有相关性，那么估计标准误往往低估了真实标准误。因此，置信区间和预测区间比真实区间窄。

误差项之间的相关关系往往存在于时间序列数据，即在离散时间点测量得到的观测构成的数据中。

确定一个给定数据集是否存在这类问题，可以根据模型绘制作为时间函数的残差。如果误差项不相关，那么图中应该没有明显规律。应用中，我们常使用残差序列的样本自相关系数构造统计量通过假设检验的方式判断误差项的自相关性。

4.3.3. 误差项方差非恒定

误差项方差非恒定是指误差项的方差随着自变量的变化而变化。线性模型中的假设检验和标准误差、置信区间的计算都依赖于误差项方差恒定的假设。对于这种情况我们可以通过绘制残差项与因变量的残差图来检测误差项方差的性质，当残差项随因变量的变化而表现出某种趋势时可认定残差项可能存在异方差性。

4.3.4. 自变量间存在多重共线性或近似多重共线性

共线性是指两个或更多的自变量之间高度相关。在回归中，共线性的存在可能会引发问题，因为可能难以分离出单个变量对响应值的影响。共线性降低了回归系数的准确性，使得 $\hat{\beta}_i$ 的标准误变大。当线性模型的显著性检验显著而回归系数的显著性检验几乎均不显著时，自变量间往往存在共线性。检测共线性的一个简单方法是看自变量的相关系数矩阵。该矩阵中出现绝对值大的元素表示有一对儿变量高度相关，因此数据存在共线性问题。但并非所有的相关性问题都可以通过检查相关系数矩阵检测到：即使没有某对儿变量间存在特别高的相关性，有可能三个或更多变量之间存在共线性。这种情况被称为多重共线性。

变量间的多重共线性可以由变量间的方差膨胀因子反映，方差扩大因子是拟合全模型时的系数 $\hat{\beta}_j$ 的方差除以单变量回归中 $\hat{\beta}_j$ 的方差所得的比例。方差扩大因子的最小取值为 1，表示完全不存在共线性。通常情况下，在实践中总有少数自变量间存在共线性。一个经验法则是，方差扩大因子超过 5 或 10 就表示有共线性问题。方差扩大因子的计算公式为：

$$VIF\left(\hat{\beta}_j\right)=\frac{1}{1-R^2_{X_j|X_{-j}}}$$

其中 $R^2_{X_j|X_{-j}}$ 是 X_j 对所有自变量回归的 R^2。如果 $R^2_{X_j|X_{-j}}$ 接近于 1，那么存在共线性，且方差扩大因子很大。

4.4. 问题的解决与模型选择

4.4.1. 变量间非线性关系的处理

当变量间存在非线性关系时，一种简单的方法是使用自变量的非线性变换来拟合模型，另外我们也可以使用多项式回归、阶梯函数、回归样条、广义可加模型等来拟合变量间的非线性关系。

4.4.2. 误差项自相关的处理

当误差项存在自相关性时我们可以更换模型，重新对数据进行拟合，或者使用差分法、迭代法来去除数据中误差项的相关性。

4.4.3. 误差项方差非恒定的处理

当误差项方差非恒定时，一方面我们可以使用凹函数对因变量 y 做变换，比如 $\log y$ 和 $\sqrt{y}$。这种变换使得较大的响应值有更大的收缩，降低了异方差性。有时，我们也可以对每个因变量的方差进行估计，利用加权最小二乘法拟合模型。

4.4.4. 自变量间存在共线性的处理

当自变量间存在共线性时，一方面，我们可以将部分问题变量从模型中剔除，这通常并未对回归拟

合做出太多的妥协，因为共线性的存在意味着在其他变量存在的前提下，此变量提供的有关因变量的信息是多余的；另一方面，我们可以使用降维技术将共线性的变量组合成一个变量或者使用正则化的技术对模型进行改进。

4.5. 岭回归

多元线性回归通过最小化如下函数对 $\beta_0,\beta_1,\cdots,\beta_p$ 进行估计来拟合最小二乘回归：

$$RSS=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2,$$

岭回归与最小二乘回归十分相似，其系数估计值 $\hat{\beta}^R$ 通过最小化下式得到：

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2,$$

其中 $\lambda\geq 0$ 是一个调节参数，将单独确定[5]。

岭回归的回归系数的最小平方估计量为：

$$\hat{\beta}^R=\left(X'X+\lambda I\right)^{-1}X'Y.$$

4.6. Least Absolute Shrinkage and Selection Operator (LASSO)

Lasso 与岭回归类似[6]，其系数估计值 $\hat{\beta}_\lambda^L$ 通过求解下式得到：

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|=RSS+\lambda\sum_{j=1}^{p}\left|\beta_j\right|.$$

因为该函数为非光滑凸函数，故 Lasso 回归的系数估计值多数情况下不存在解析形式[7]，对于多变量情形一般通过循环坐标下降法进行求解。

5. 数据分析与建模

5.1. 变量间的相关性分析

该数据为实验获得，由于在实验过程中实验帆船类型有限及实验时的经费限制，这些数据并不能在其可能取值的范围中连续取值，故样本中各变量的取值往往表现出离散分布的特征，但在本文的研究中均将这些数据视为连续型数据进行建模。

由于本文主要探究数据集中自变量与因变量之间的关系，以图利用合适的模型拟合自变量与因变量之间的关系，从而实现用自变量合理地预测因变量。首先，计算原数据中各变量的相关系数矩阵如下表 2 (其中 $V1$-$V7$ 分别与变量名称中英文对照表(表 1)中的各变量相对应)：

经观察该相关系数矩阵可以发现，除弗劳德数与单位重量排水的剩余阻力之间存在明显的线性关系外，其他变量与单位重量排水的剩余阻力之间均可认为不存在线性相关关系。

经观察原始数据发现，实验人员在试验时共设置了弗劳德数的 14 个水平，每个水平上分别进行了 22 次实验，由于上述结果显示，弗劳德数与单位重量排水的剩余阻力之间存在较强的线性相关关系。故绘制原数据集中弗劳德数与单位重量排水的剩余阻力的折线图如下图 1，其中横轴代表弗劳德数，纵轴代表单位重量排水的剩余阻力。

Table 2. Correlation coefficient matrix of variables in the original data set
表 2. 原数据集中各变量相关系数矩阵

	*V*1	*V*2	*V*3	*V*4	*V*5	*V*6	*V*7
*V*1	1.00	−0.01	0.00	0.00	0.00	0.00	0.02
*V*2	−0.01	1.00	−0.03	0.35	−0.08	0.00	−0.03
*V*3	0.00	−0.03	1.00	0.38	**0.68**	0.00	0.00
*V*4	0.00	0.35	0.38	1.00	−0.38	0.00	−0.01
*V*5	0.00	−0.08	**0.68**	−0.38	1.00	0.00	0.00
*V*6	0.00	0.00	0.00	0.00	0.00	1.00	**0.81**
*V*7	0.02	−0.03	0.00	−0.01	−0.00	**0.81**	1.00

根据该图像可以得出，随着弗劳德数的不断增加，单位重量排水的剩余阻力呈现指数上升的趋势，其中图像中的红色曲线为对该序列以其弗劳德数为自变量，以其单位重量排水的剩余阻力为因变量拟合得到的指数拟合曲线，拟合结果为，拟合优度故。故接下来的主要工作是探索自变量与因变量之间的对数线性关系。

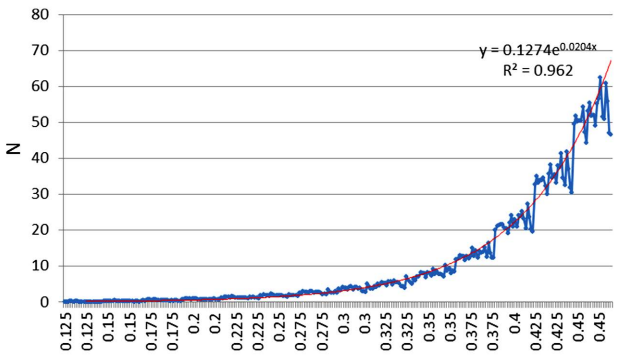

Figure 1. Froude number and residual resistance per unit weight drainage line chart
图 1. 弗劳德数与单位重量排水剩余阻力折线图

5.2. 变量间的对数线性关系分析

对数据集中单位重量排水的剩余阻力取对数，记为 *V*8，绘制弗劳德数与取对数后的单位重量排水的折线图如下图 2，其中横轴代表弗劳德数，纵轴代表取对数后的单位重量排水的剩余阻力。

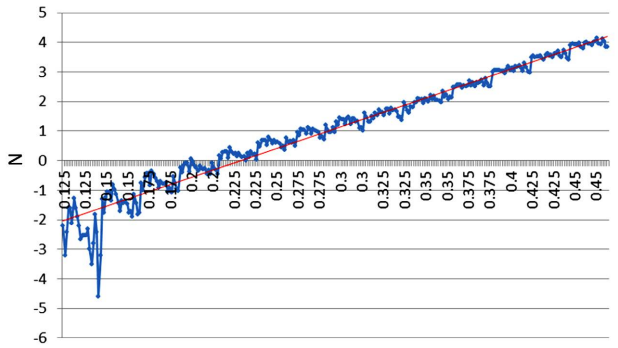

Figure 2. Line diagram of Froude number and residual resistance per unit weight of drainage after logarithm
图 2. 弗劳德数与取对数后单位重量排水剩余阻力折线图

根据上图可以发现，取对数后的单位重量排水的剩余阻力与弗劳德数之间存在较明显的线性关系。计算取对数后的单位重量排水的剩余阻力与各自变量间的相关系数如下表 3，其中 $V8$ 表示取对数后的单位重量排水的剩余阻力。

Table 3. Take the correlation coefficient between logarithmic unit weight residual resistance and each variable
表 3. 取对数单位重量剩余阻力与各变量的相关系数

	$V1$	$V2$	$V3$	$V4$	$V5$	$V6$	$V8$
$V8$	0.02	−0.01	0.01	0.03	−0.01	0.98	1.00

由上表可得，在经过对数处理后，所有的变量中仍旧只有弗劳德数与取对数后的单位重量排水的剩余阻力之间存在高度的线性关系，且线性关系比之前有所提高，但其他自变量与取对数后的单位重量排水的剩余阻力之间的线性关系仍旧不显著。

5.3. 变量间交互作用的引入

在接下来的研究中，本文考虑到弗劳德数在实际应用中为不可控因素，而除弗劳德数之外的变量与取对数后的单位重量排水的剩余阻力之间可能不是简单的线性关系。各自变量与弗劳德数之间可能存在交互作用，二者交互作用的结果与取对数后的单位重量排水的剩余阻力之间存在明显的线性关系。记 $V9$-$V13$ 分别为浮心纵坐标、菱形系数、船长排水量比、梁宽吃水量比、船长梁宽比分别与弗劳德数交互作用的结果，计算当前已有变量间的相关系数矩阵如下表 4。

Table 4. Correlation coefficient matrix of each variable
表 4. 各变量相关系数矩阵

	$V1$	$V2$	$V3$	$V4$	$V5$	$V6$	$V7$	$V8$	$V9$
$V1$	1.00	−0.01	0.00	0.00	0.00	0.00	0.02	0.02	**0.84**
$V2$	−0.01	1.00	−0.03	0.35	−0.08	0.00	−0.03	−0.01	−0.01
$V3$	0.00	−0.03	1.00	0.38	**0.68**	0.00	0.00	0.01	0.00
$V4$	0.00	0.35	0.38	1.00	−0.38	0.00	−0.01	0.03	0.00
$V5$	0.00	−0.08	**0.68**	−0.38	1.00	0.00	0.00	−0.01	0.00
$V6$	0.00	0.00	0.00	0.00	0.00	1.00	**0.81**	**0.98**	−0.46
$V7$	0.02	−0.03	0.00	−0.01	0.00	**0.81**	1.00	**0.79**	−0.35
$V8$	0.02	−0.01	0.01	0.03	−0.01	**0.98**	**0.79**	1.00	−0.44
$V9$	**0.84**	−0.01	0.00	0.00	0.00	−0.46	−0.35	−0.44	1.00
$V10$	0.00	0.12	0.00	0.04	−0.01	**0.99**	**0.80**	**0.98**	−0.46
$V11$	0.00	0.00	0.15	0.06	0.10	**0.99**	**0.80**	**0.97**	−0.45
$V12$	0.00	0.13	0.14	0.37	−0.14	**0.92**	**0.74**	**0.91**	−0.42
$V13$	0.00	−0.02	0.15	−0.08	0.22	**0.97**	**0.79**	**0.96**	−0.45

	$V10$	$V11$	$V12$	$V13$
$V1$	0.00	0.00	0.00	0.00
$V2$	0.12	0.00	0.13	−0.02
$V3$	0.00	0.15	0.14	0.15
$V4$	0.04	0.06	0.37	−0.08

Continued

*V*5	−0.01	0.10	−0.14	0.22
*V*6	**0.99**	**0.99**	**0.92**	**0.97**
*V*7	**0.80**	**0.80**	**0.74**	**0.79**
*V*8	**0.98**	**0.97**	**0.91**	**0.96**
*V*9	−0.46	−0.45	−0.42	−0.45
*V*10	1.00	**0.98**	**0.93**	**0.96**
*V*11	**0.98**	1.00	**0.93**	**0.99**
*V*12	**0.93**	**0.93**	1.00	**0.86**
*V*13	**0.96**	**0.99**	**0.86**	1.00

由上述相关系数矩阵可得，在引入交互效应后，除弗劳德数及单位重量排水的剩余阻力与取对数后的单位重量排水的剩余阻力之间有较高的线性关系之外，菱形系数、船长排水量比、梁宽吃水量比、船长梁宽比与弗劳德数的交互作用与取对数后的单位重量排水的剩余阻力之间也存在明显的线性关系。但伴随的问题是，随着交互项的引入，变量间出现了明显的多重共线性，此时传统的多元线性回归模型已经无法被用来拟合这些数据，于是本文接下来的研究从线性回归模型的修正版本岭回归及 Lasso 回归展开。

5.4. 岭回归

利用 R 软件拟合岭回归模型[6] [8]，通过交叉验证的方式选择最佳的 λ 值，结果如下：

首先，我们得到了回归均方与 $\log(\lambda)$ 的关系图如下图 3，由该图可以发现，随着 $\log(\lambda)$ 的值不断减小，在交叉验证过程中回归均方不断减小，且其离散程度不断降低。基于此我们得到了最佳的 $\log(\lambda)$ 值为 0.26。

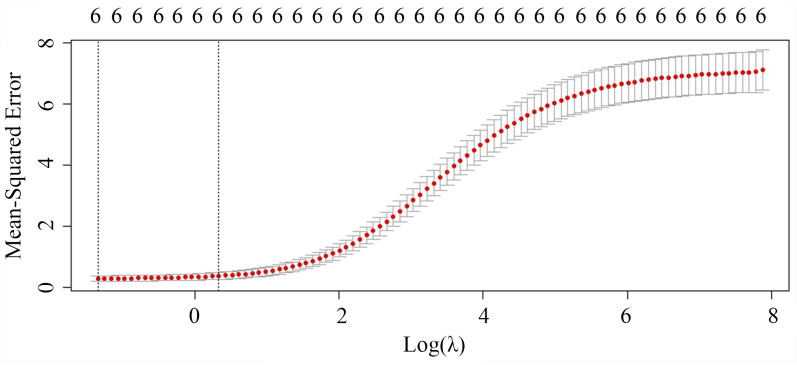

Figure 3. Ridge regression regression mean square and logarithmic penalty parameter relationship
图 3. 岭回归回归均方与对数惩罚参数关系图

其次，我们得到岭回归参数的估计结果如下表 5。

Table 5. Ridge regression parameter estimation results
表 5. 岭回归参数估计结果

(Intercept)	*V*6	*V*9	*V*10	*V*11	*V*12	*V*13
−5.778	7.519	0.009	10.229	0.973	0.745	1.386

于此我们可得，取对数后的单位重量排水的剩余阻力与各变量间的线性关系为：

$$\hat{V}8 = -5.778 + 7.519V6 + 0.009V9 + 10.229V10 + 0.973V11 + 0.745V12 + 1.386V13.$$

由上述岭回归结果可以发现，在可控因素中对取对数后的单位重量排水的剩余阻力影响最大的是棱型系数。

同时有：

$$\hat{V}7 = 2^{\hat{V}8}$$

其预测的均方误差为：

$$MSE1 = \frac{\sum_{i=1}^{n}\left(V8 - \hat{V}8\right)^2}{n} = 0.271$$

该预测下原变量的预测均方误差为：

$$MSE2 = \frac{\sum_{i=1}^{n}\left(V7 - \hat{V}7\right)^2}{n} = 23.641$$

即利用该模型对单位重量排水的剩余阻力预测的平均误差为 4.862。

5.5. LASSO

利用 R 软件拟合 Lasso 回归模型[6] [8]，通过交叉验证的方式选择最佳的 λ 值，结果如下：

首先，我们得到了回归均方与 $\log(\lambda)$ 的关系图如下图 4，由该图可以发现，随着 $\log(\lambda)$ 的值不断减小，在交叉验证过程中回归均方不断减小，且其离散程度不断降低。基于此我们得到了最佳的 $\log(\lambda)$ 值为 0.009。

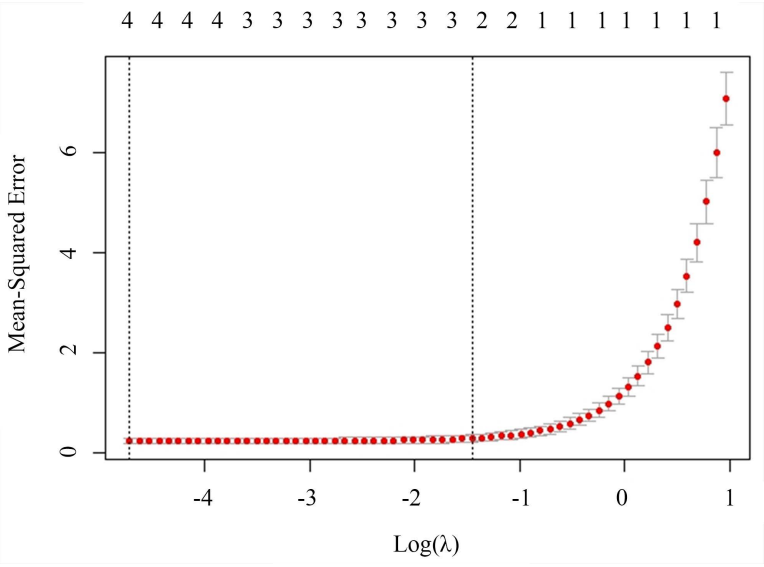

Figure 4. Lasso regression mean square and logarithmic penalty parameter relationship

图 4. Lasso 回归均方与对数惩罚参数关系图

其次，我们得到岭回归参数的估计结果如下表 6。

Table 6. Lasso parameter estimation results
表 6. Lasso 模型参数估计结果

(Intercept)	$V6$	$V9$	$V10$	$V11$	$V12$	$V13$
−5.969	24.454	0.058	.	0.161	0.227	.

根据上述结果可知，菱形系数、船长梁宽比与弗劳德数的交互作用的回归系数在 Lasso 中被压缩至零，并且我们可以得到取对数后的单位重量排水的剩余阻力与各变量间的线性关系为：

$$\hat{V}8 = -5.969 + 24.454V6 + 0.058V9 + 0.161V11 + 0.227V12 .$$

由该结果可以发现，在可控因素中对取对数后的单位重量排水的剩余阻力影响最大的因素为梁宽吃水量比。

其预测的均方误差为：

$$MSE1 = \frac{\sum_{i=1}^{n}\left(V8 - \hat{V}8\right)^2}{n} = 0.220$$

该预测下原变量的预测均方误差为：

$$MSE2 = \frac{\sum_{i=1}^{n}\left(V7 - \hat{V}7\right)^2}{n} = 4.64$$

即利用该模型预测单位重量排水的剩余阻力的平均误差为 2.154。

5.6. 模型比较与选择

由上述结果可知，Lasso 在原数据集上的预测均方误差为 4.64，远低于岭回归在原数据集上的预测方差 23.641。故 Lasso 回归有比岭回归更好的预测效果。

分别绘制各模型在原数据集及取对数后数据集中的预测残差图如图 5~8。

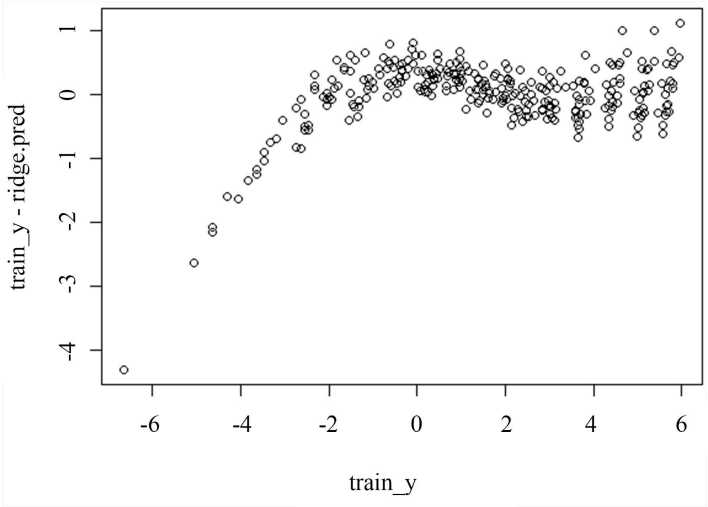

Figure 5. Ridge regression linear model fits residual graph
图 5. 岭回归线性模型拟合残差图

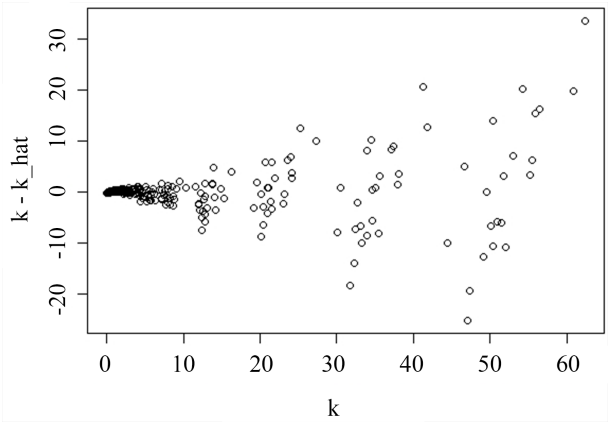

Figure 6. Ridge regression log-linear model fits residual graph
图 6. 岭回归对数线性模型拟合残差图

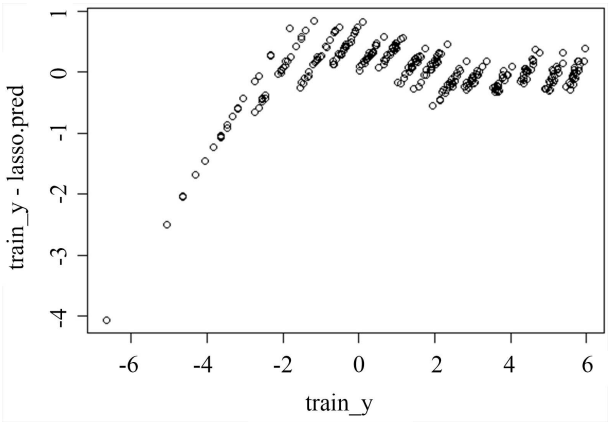

Figure 7. Lasso linear model fits residual graph
图 7. Lasso 线性模型拟合残差图

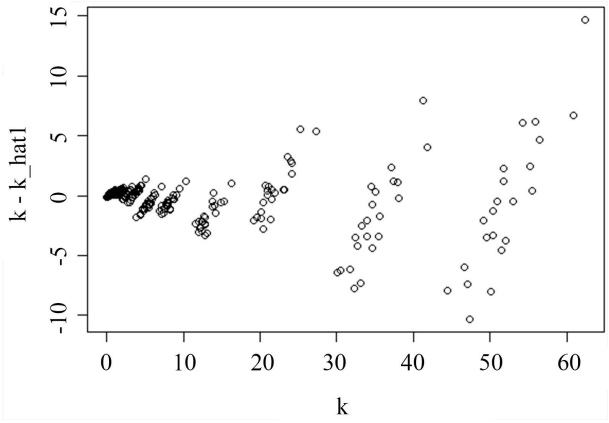

Figure 8. Lasso log-linear model fits residual graph
图 8. Lasso 对数线性模型拟合残差图

根据上述结果可以发现，在对取对数后的单位重量排水的剩余阻力的拟合中，Lasso 结果的残差比岭回归结果的残差更加稳定，即 Lasso 的结果中残差上下波动的范围更加稳定，Lasso 可以更好的利用自变量中的信息来预测因变量中的非随机信息，但两种回归结果的残差均表现出了自相关性。由于 Lasso 中

参数的估计不存在解析形式而是迭代求解，故该问题对 Lasso 中的参数的估计结果不会产生太大影响，Lasso 的结果完全可以用来对该问题中涉及的变量间的关系进行分析。

5.7. 结论与建议

根据上述结果，Lasso 在此数据集中表现出了较好的预测效果，故接下来从 Lasso 的结果出发讨论单位重量排水的剩余阻力的优化问题，即帆船的结构优化问题。由于取对数前后数值的大小关系不变，故我们从 $V8$ 与各变量的关系出发，讨论此优化问题。由于，

$$\hat{V}8 = -5.969 + 24.454V6 + 0.058V9 + 0.161V11 + 0.227V12,$$

且弗劳德数在驾驶帆船过程中为不可控因素。故由上述结果可知，浮心纵坐标每减少一个单位，对数化的单位重量排水的剩余阻力平均减少 0.058*$V6$ 个单位，船长排水量比每下降一个单位，对数化的单位重量排水的剩余阻力平均减少 0.161*$V6$ 个单位，梁宽吃水量比每下降一个单位，对数化的单位重量排水的剩余阻力平均减少 0.227*$V6$ 个单位。故实际应用中对单位重量排水的剩余阻力影响最大的是梁宽吃水量比，其次是船长排水量比，最后是浮心纵坐标。

在实际生产中，当船只的排水量一定时，船长与梁宽往往存在反比例关系，故在船只设计中我们应当合理调整船长与梁宽的大小关系，使得船只单位重量排水的剩余阻力可以达到最小。

致　谢

本文写作过程中得到了多位老师、同学的帮助，在此对其表示感谢！

基金项目

本文受到云南省教育厅科学研究基金项目(NO.2022Y546)的支持。

参考文献

[1] 张益. 竞赛帆船龙骨盒结构优化设计[D]: [硕士学位论文]. 厦门: 集美大学, 2016.

[2] 胡健康, 林少芬, 陈清林, 朱兆一. 荷载位置对龙骨帆船大倾角稳性影响[J]. 集美大学学报(自然科学版), 2016, 21(2): 130-135.

[3] 许小颖, 高丽莎. 基于多元回归分析的三体船阻力性能预报[J]. 应用科技, 2020, 47(5): 1-5.

[4] 王松桂, 等. 线性模型引论[M]. 北京: 科学出版社, 2004: 147-183.

[5] [美] Freund, R.J. 回归分析-因变量统计模型[M]. 沈崇麟, 译. 重庆: 重庆大学出版社, 2012: 151-189.

[6] [美] James, G. 统计学习导论-基于 R 应用[M]. 王星, 等, 译. 北京: 机械工业出版社, 2015: 140-180.

[7] Belloni, A. and Chernozhukov, V. (2013) Least Squares after Model Selection in High-Dimensional Sparse Models. *Bernoulli*, **19**, 521-547. https://doi.org/10.3150/11-BEJ410

[8] [美] Kabacoff, R.I. R 语言实战[M]. 王小宁, 等, 译. 北京: 人民邮电出版社, 2016: 135-155.