E-Commerce Letters 电子商务评论, 2024, 13(4), 5359-5374
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341771

Hans汉斯

# 基于弹性网络回归的股票价格预测

**朱　灿，谢学琴**

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2024年9月23日；录用日期：2024年10月16日；发布日期：2024年11月27日

## 摘　要

**本文基于弹性网络回归模型对股票价格进行了预测分析。通过收集国内某酒厂738个交易日的数据，选取开盘价、最高价、最低价、交易量、涨跌幅等作为自变量，以收盘价为因变量，分别应用线性回归、岭回归、Lasso回归及弹性网络回归四种模型进行预测分析。研究结果表明，弹性网络回归模型在处理多重共线性问题和变量筛选方面具有显著优势。通过交叉验证确定了最优的惩罚参数，使得模型的预测误差最小。最终，最高价、最低价、交易量和涨跌幅被筛选为影响股票收盘价的主要因素。本文的研究为股票价格预测提供了有效的方法和工具，并为金融市场的投资决策提供了重要参考。**

## 关键词

**弹性网络回归，股票价格预测，多重共线性，岭回归，Lasso回归**

# Stock Price Prediction Based on Elastic Net Regression

**Can Zhu, Xueqin Xie**

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Sep. 23rd, 2024; accepted: Oct. 16th, 2024; published: Nov. 27th, 2024

## Abstract

**This study conducts a stock price prediction analysis based on the Elastic Net regression model. Using data from 738 trading days of a domestic brewery, the study selects opening price, highest price, lowest price, trading volume, and price change percentage as independent variables, with the closing price as the dependent variable. Four models are applied for prediction analysis: linear regression, ridge regression, Lasso regression, and Elastic Net regression. The results show that the Elastic Net regression model has significant advantages in handling multicollinearity issues and variable selection. The optimal penalty parameters are determined through cross-validation, minimizing the**

文章引用: 朱灿, 谢学琴. 基于弹性网络回归的股票价格预测[J]. 电子商务评论, 2024, 13(4): 5359-5374.
DOI: 10.12677/ecl.2024.1341771

**prediction error. Ultimately, the highest price, lowest price, trading volume, and price change percentage are identified as the key factors influencing stock closing prices. This research provides an effective method and tool for stock price prediction and offers important insights for financial market investment decisions.**

## Keywords

**Elastic Net Regression, Stock Price Prediction, Multicollinearity, Ridge Regression, Lasso Regression**



## 1. 引言

股票作为资本市场的重要组成部分，其历史可追溯至数百年前。随着资本主义工业的发展，股票作为集资手段应运而生，既解决了企业资金短缺问题，也为投资者提供了新的投资方式。在现代经济中，股票作为重要的融资工具，为企业筹集资金，推动生产扩张和研发活动。此外，股票市场价格的波动反映了企业经营状况和市场环境的变化，直接影响投资者的决策和企业的融资能力。融资融券业务是证券市场的重要组成部分，能够提高市场的活跃度和价格发现效率。我国融资融券业务经历了从禁止到逐步开放的发展过程。1998 年《证券法》明确禁止此类交易，2005 年修订的《证券法》[1]则允许在监管批准下开展业务。2010 年，融资融券交易正式进入市场操作阶段。本研究以国内某酒厂 738 个交易日的数据为样本，综合分析了开盘价、收盘价、涨跌幅及融资融券等因素，旨在揭示市场动态和投资决策的相关性，为制定投资策略提供理论支持。旨在为投资者、金融从业者和学术界提供有关机器学习在股票价格预测中应用的深入理解，并为未来的研究和实践提供有益的启示方面具有一定价值和意义。

## 2. 文献综述

国内众多学者对股票价格预测展开了深入研究。赵莹莹(2023) [2]利用自适应弹性网模型，筛选出对上证 50 指数影响较大的成分股，并通过这些成分股的表现，追踪上证 50 指数的实际收盘价走势。魏巍(2023) [3]以 Lasso 分位数回归为对照组，探讨了弹性网分位数回归在筛选沪深 300 指数成分股中的应用，并结合 Mallows 模型平均法和 Jackknife 模型平均法对筛选出的成分股进行预测。研究发现，弹性网分位数回归结合模型平均法能够灵活处理高维股指追踪问题，表现出较好的变量选择与模型预测能力，追踪效果显著。袁子杨(2022) [4]指出，马科维兹投资组合模型在实际应用中存在稳健性不足及高换手率的问题。为解决这一问题，作者将自适应弹性网引入马科维兹模型，并结合 Huber 损失函数，构建了稀疏且稳健的投资组合。研究结果显示，该模型在高频交易中表现出色，能够有效降低换手率和交易成本，并提升外样本的稳健性。邢艳雅(2021) [5]针对时间序列数据，提出了两类新的自适应弹性网模型，通过引入相关系数作为权重，解决了传统时间序列分析无法进行变量选择的问题，得到了简洁且有效的回归模型。韩情和汪子琦(2020) [6]基于上证综合指数日收盘价数据，构建自回归模型，分别将 Lasso 回归、自适应 Lasso 回归和弹性网回归方法应用于自回归模型，并对未来十个交易日的股票价格进行预测。周政瑜(2021) [7]同样基于上证综合指数日收盘价数据，比较了 Lasso、自适应 Lasso 及弹性网回归在自回归模型中的表现，结果表明，弹性网方法在股票价格预测中的表现更为优越。宋家辉(2020) [8]通过 Jensen $\alpha$ 指数和分位数回归分析股票和基金的收益能力，利用逐步回归、Lasso 惩罚及弹性网惩罚探讨股票收益风

格及基金投资风格，最后通过 Sharpe 指数和 Sortino 指数对股票和基金组合的绩效进行评价，验证了弹性网分位数回归在收益评价中的优越性。基于当前国内股票市场的情况及研究现状，本文通过构建弹性网回归模型，对某酒厂的股票价格进行分析，旨在探索经济发展模式和规律。研究结果不仅能够为广大投资者提供新的投资分析方法，还对金融证券公司在股票投资领域具有重要的参考价值。

## 3. 数据来源与指标设计

本文数据来源于国内某酒厂自 2020 年来每日开盘价、最高价、最低价、交易量、涨跌幅、融券余量、融资偿还额、融资融券余额、融资净买入共 738 条数据。本文以收盘价作为因变量，以开盘价、最高价、最低价、交易量、涨跌幅、融券余量、融资偿还额、融资融券余额、融资净买入这些指标作为自变量来建立模行[9]，探究市场的活跃度和股票的价格动态，旨在于揭示数据背后的趋势、模式和潜在问题(表 1)。

**Table 1.** Index explanation
**表 1.** 指标解释

| 变量 | 名称 | 名词解释 |
|---|---|---|
| $X_1$ | 开盘价(元) | 交易日股票开始交易时的价格 |
| $X_2$ | 最高价(元) | 最高价是交易日内股票成交的最高 |
| $X_3$ | 最低价(元) | 最低价是交易日内股票成交的最低价格 |
| $X_4$ | 交易量(百万) | 某时间段内股票买卖的总数量 |
| $X_5$ | 涨跌幅 | 股票与前一日收盘价的涨跌百分比 |
| $X_6$ | 融券余量(万股) | 尚未被买回或卖出的融券数量 |
| $X_7$ | 融资偿还额(亿元) | 投资者偿还的融资借款金额 |
| $X_8$ | 融资融卷余额(亿元) | 投资者融资买入与融券卖出后未偿余额 |
| $X_9$ | 融资净买入(万元) | 融资买入与融资偿还的差额，若差值为正，则为净买入 |
| $Y$ | 收盘价(元) | 交易日股票结束交易时的价格 |

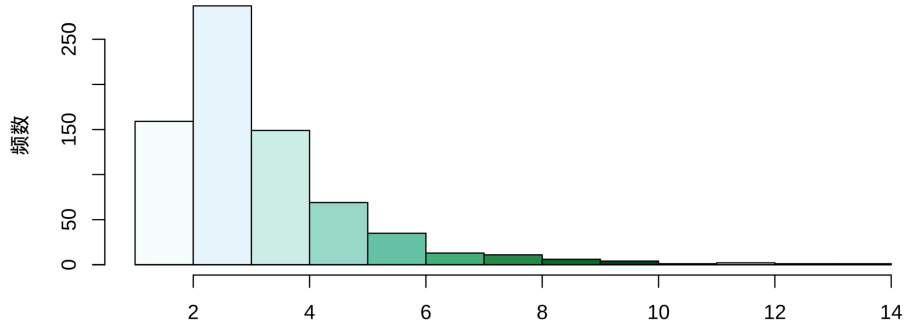
## 4. 描述分析

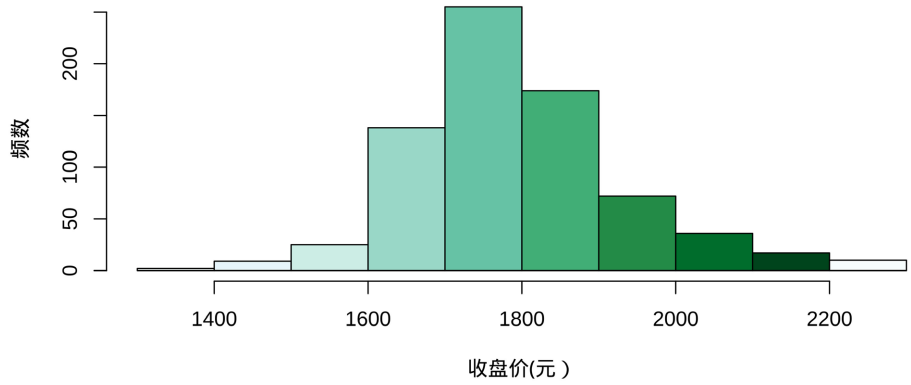
### 4.1. 样本分布

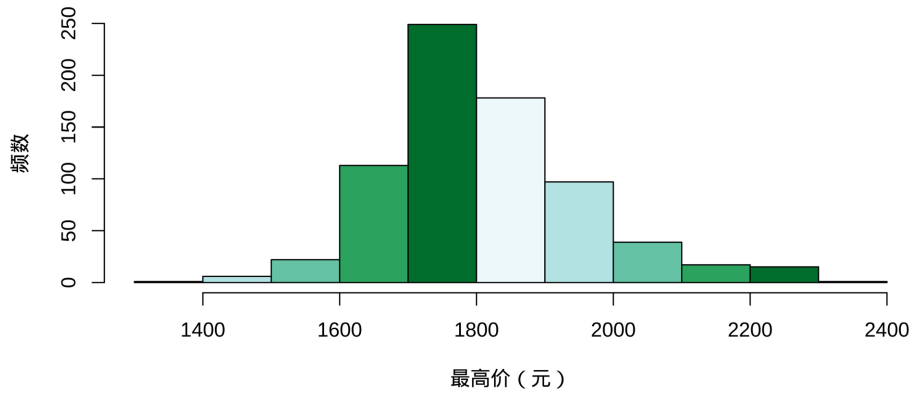
为了深入理解这些数据的分布特性，为每一个解释变量绘制直方图。首先考虑开盘价、收盘价、最高价、最低价、涨跌幅、融资买入额，见图 1。这些变量的直方图均呈现出近似的正态分布形态，表明了数据的集中趋势和市场的稳定性。其中，开盘价、收盘价、最高价、最低价主要集中在[1700, 1900]元的区间内，显示了市场的相对稳定和价格的一致性；涨跌幅的波动范围较小，位于[−0.025, 0.025]之间，反映了市场的平稳运行和投资者情绪的相对稳定；融资买入额在[−25,000, 25,000]的区间内波动，同样体现了市场的稳定性和可预测性；这些都是一个积极的信号，均没有出现极端的高低偏离。

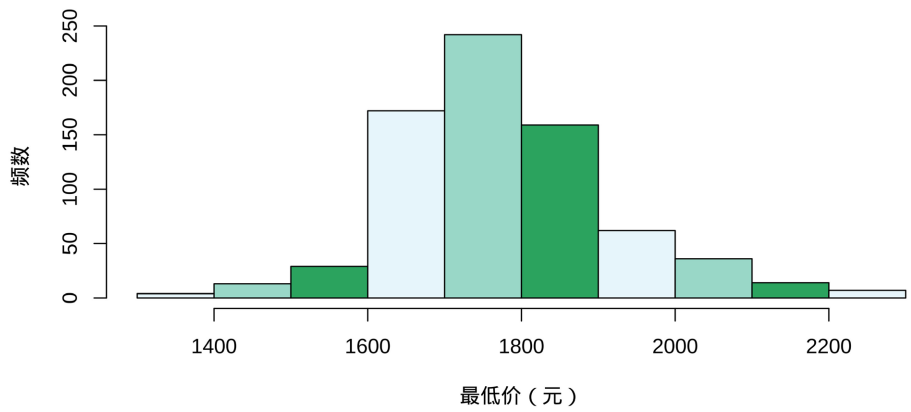
对融券余量、交易量、融券偿还量、融资融券余额 4 个自变量也做类似直方图分析，见图 2。这四个变量的分布均呈现出一定程度的右偏形态，这种右偏分布往往意味着大部分样本的取值集中在较小的一侧，而少数样本则拥有较大的取值，这在一定程度上反映了市场中的“二八定律”。其中，融券余量主要集中在[0, 30]万股的范围内，显示出市场的融券交易量总体保持在一个较为稳定的水平；交易量主要集中在[0, 300]百万的区间内，表明市场的交易活跃度适中；融券偿还量主要集中在[2, 6]的范围内，反映了市场融券偿还的常态水平；融资融券余额主要集中在[170, 210]之间；这些变量的分布范围均处于合理区间

内，为后续的深入研究提供了良好的基础。



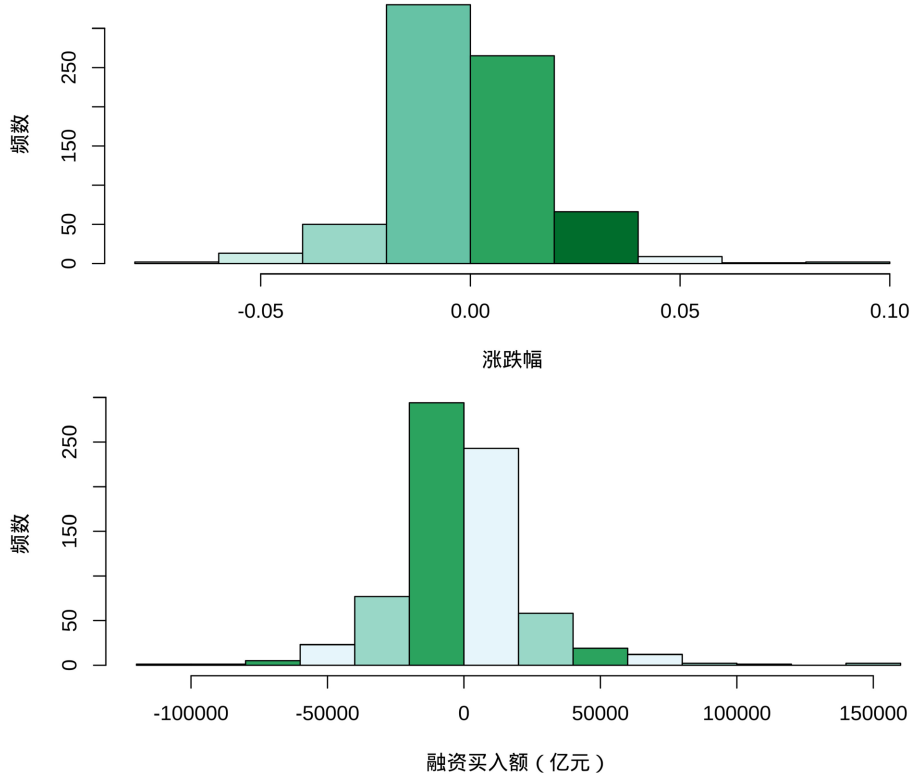
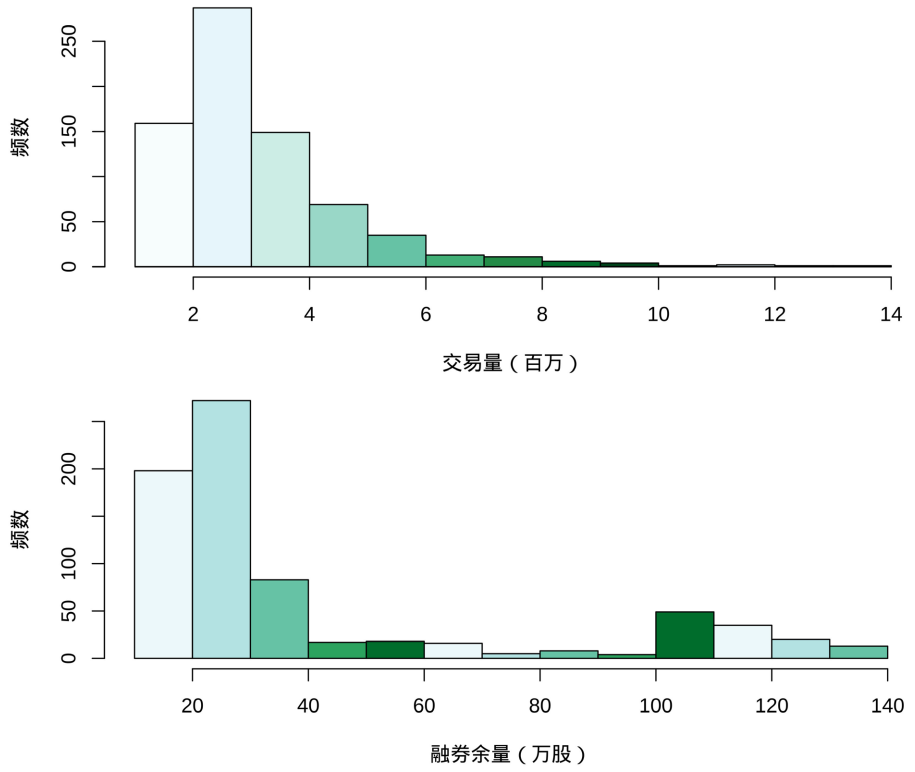



**Figure 1.** Histogram analysis of opening price, closing price, highest price, lowest price, increase or decrease, and margin purchase amount

**图 1.** 开盘价、收盘价、最高价、最低价、涨跌幅、融资买入额直方图分析



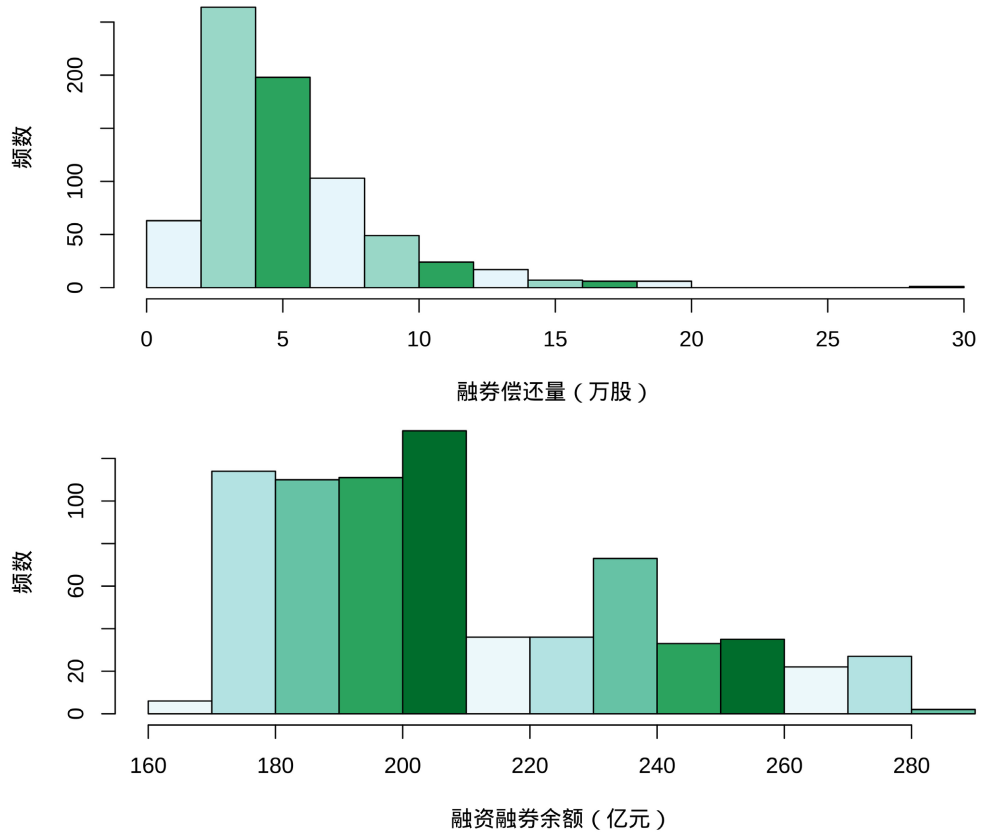
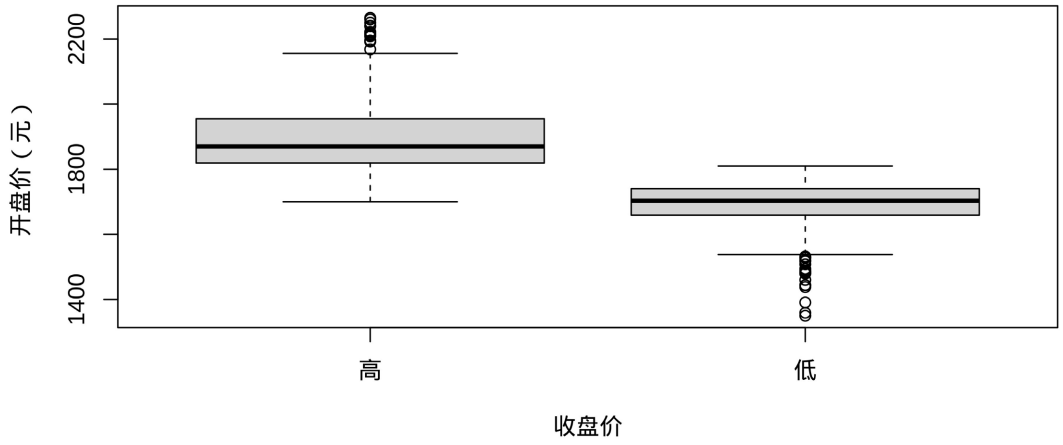



**Figure 2.** Histogram analysis of margin lending balance, trading volume, margin lending repayment volume, and margin lending balance
**图 2.** 融券余量、交易量、融券偿还量、融资融券余额直方图分析

引入了因变量(收盘价)的高低作为分类标准，将数据集划分为两类：收盘价低(设为 0)，收盘价高(设为 1)。为更直观地比较不同因变量取值下各自变量的分布情况，采用箱线图进行对比分析，见图 3。从中可以看到，开盘价、最高价、最低价在两组中的分布情况存在显著差异，显示出这些变量与收盘价的高低具有较强的相关性。而融券余量、融券偿还量在两组中的差异稍小，但也表现出一定的规律性。同时，对于收盘价高的组的中位数要明显高于收盘价低的组。说明高于开盘价、最高价、最低价、融券余量、融券偿还量的股票，往往对应着较高的收盘价。



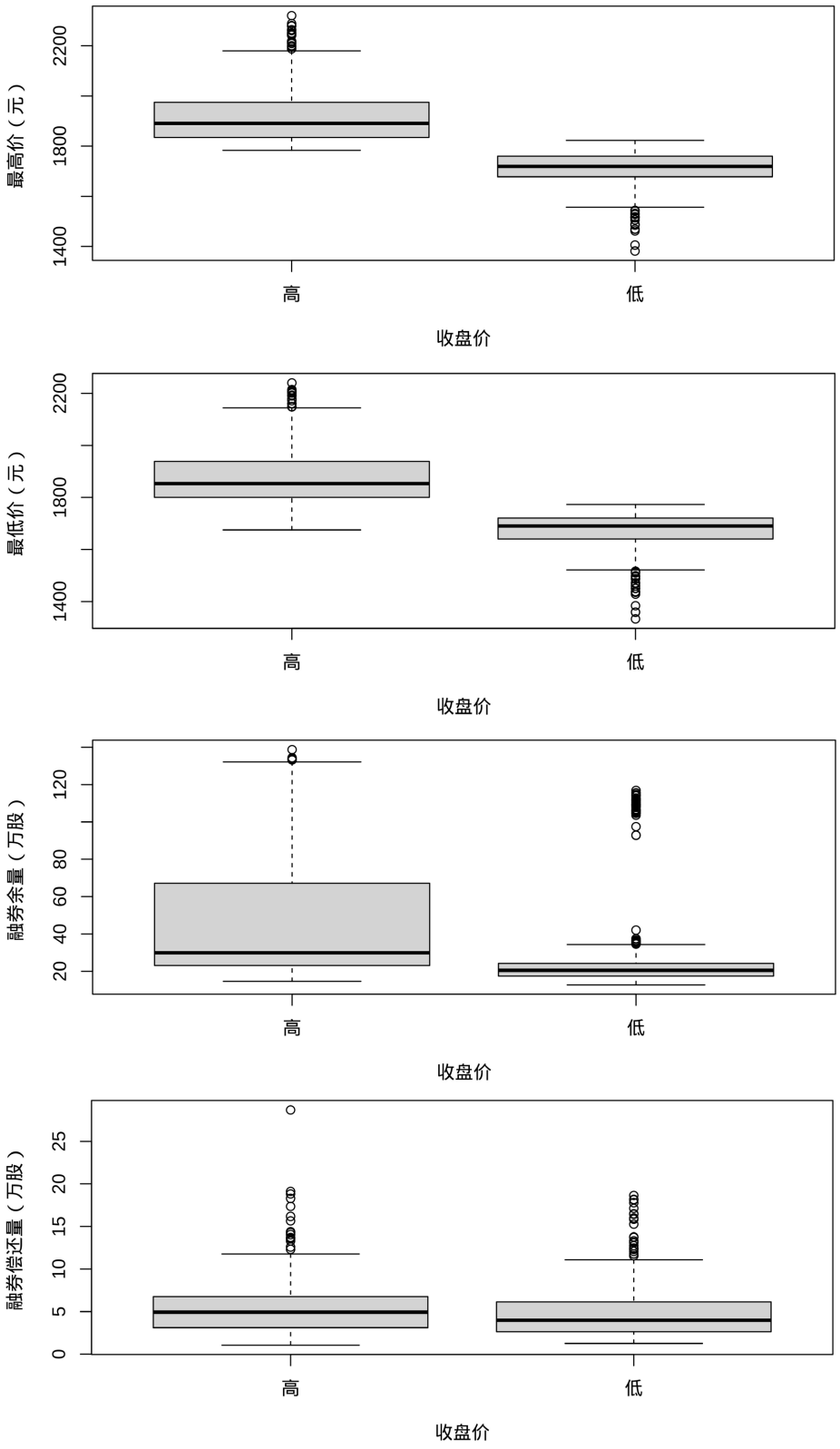



**Figure 3.** Box chart analysis of opening price, highest price, lowest price, margin balance, and margin repayment

**图 3.** 开盘价、最高价、最低价、融券余量、融券偿还量箱线图分析

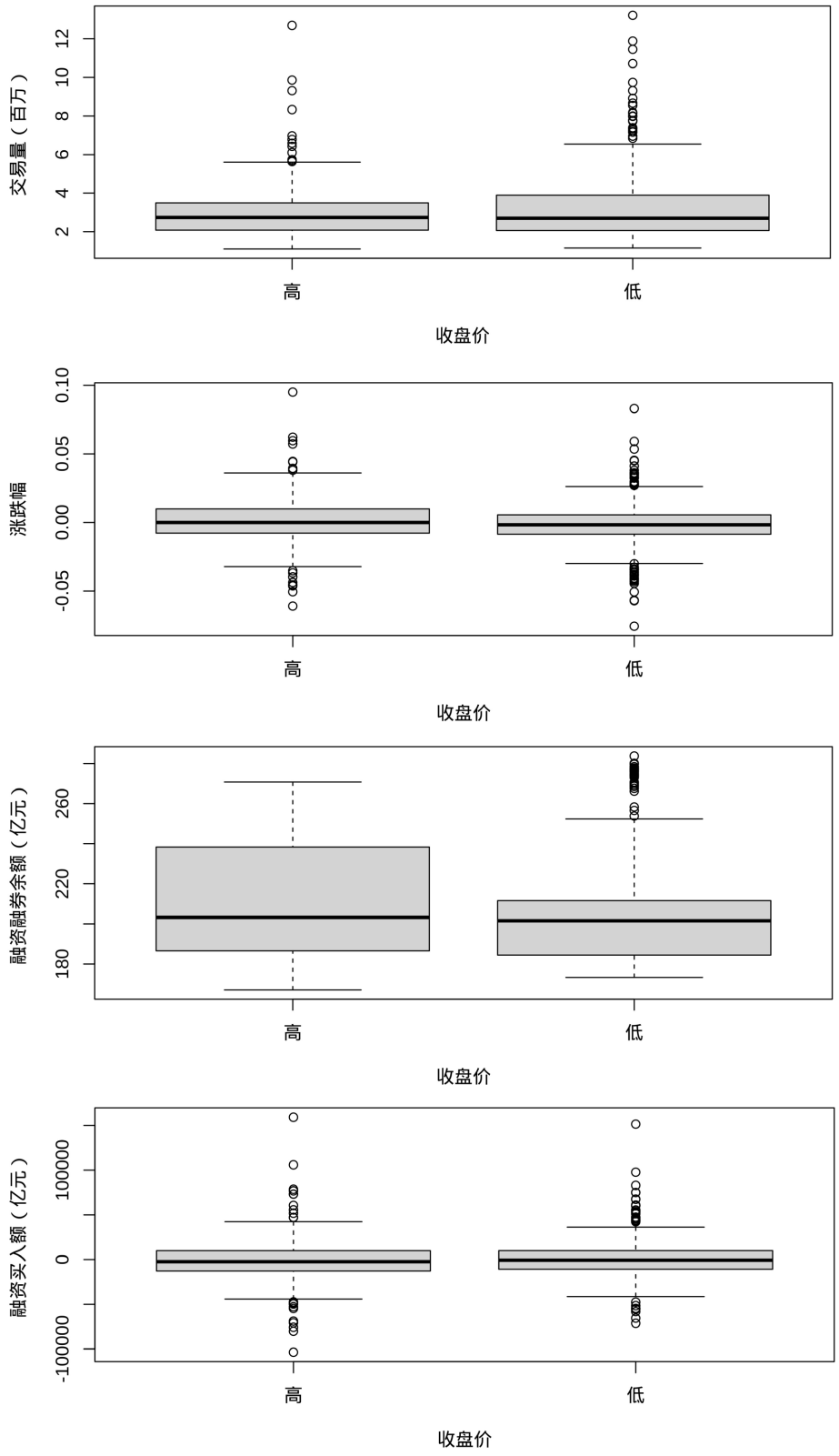



**Figure 4.** Box chart analysis of opening price, highest price, lowest price, margin balance, and margin repayment

**图 4.** 开盘价、最高价、最低价、融券余量、融券偿还量箱线图分析

对交易量、涨跌幅、融资融券余额、融资买入额 4 个连续型自变量做类似分析，见图 4。从图中可知，这四个变量在收盘价高低两组之间的分布差异并不显著，但均表现出与收盘价一定程度的相关性。

为了更全面地了解数据的整体特征，最后对所有变量做一个描述统计分析报表，见表 2。从上往下依次是：开盘价、最高价、最低价、交易量、涨跌幅、融券余量、融资偿还额、融资融券余额、融资净买入、收盘价。整个数据集包含了 738 条样本。从表中可以看出，开盘价、收盘价、最低价、收盘价之间最小值、中位数、最大值之间差异并不显著，显示出这四个价格指标在市场中相对稳定性。特别是开盘价和收盘价最低值均为 1350 元，最高值为 2265 元和 2271 元，相差较小，进一步证实了市场稳定性。

**Table 2.** Descriptive statistical analysis of variables
**表 2.** 变量描述统计分析

| 变量 | 样本量 | 均值 | 标准差 | 最小值 | 中位数 | 最大值 |
|---|---|---|---|---|---|---|
| $X_1$ | 738 | 1.7953 | 1.4335 | 1350.0000 | 1778.1500 | 2265.0000 |
| $X_2$ | 738 | 1.8153 | 1.4403 | 1382.0000 | 1793.6750 | 2320.0000 |
| $X_3$ | 738 | 1.7759 | 1.4067 | 1333.0000 | 1760.0000 | 2240.0100 |
| $X_4$ | 738 | 3.1200 | 1.6229 | 1.1100 | 2.7100 | 13.2100 |
| $X_5$ | 738 | −1.0407 | 1.7326 | −0.0756 | −0.0008 | 0.0950 |
| $X_6$ | 738 | 4.0852 | 3.5016 | 12.7200 | 23.3300 | 138.7000 |
| $X_7$ | 738 | 5.2277 | 3.3457 | 1.0600 | 4.4450 | 28.6900 |
| $X_8$ | 738 | 2.0896 | 2.8425 | 167.1000 | 202.0700 | 282.7100 |
| $X_9$ | 738 | −4.6872 | 2.4066 | −103700.0000 | −1652.2900 | 159100.0000 |
| $Y$ | 738 | 1.7950 | 1.4234 | 1350.0000 | 1776.4200 | 2271.0000 |

### 4.2. 相关性分析

为更直观地分析各变量之间的关系，对数据进行了标准化处理，并绘制散点图矩阵[10]，见图 5。该矩阵包含十个子图，每个子图都展示了两个变量之间的散点图关系。从图中可以清晰看到，$Y$ 与 $X_1$、$X_2$、$X_3$ 之间呈现出较强的相关性，表明这些价格指标在市场中往往同步波动，相互影响。同时，$X_1$、$X_2$、$X_3$ 变量之间也表现出了显著的相关性，进一步强调了在市场走势中的一致性。而 $Y$ 与其他变量之间的相关性较弱，在构建模型时，需要综合考虑这些变量之间的复杂关系。

## 5. 模型建立

### 5.1. 线性回归

线性回归是回归分析中最基本的一类回归问题，若使用最小二乘求解回归系数，线性回归又可称为最小二乘回归(Least Square Regression, LS) [11]。对于一般的线性回归模型来说，假设预测变量的个数为 $p$，样本容量为 $n$，则：

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ \varepsilon_i \sim N\left(0, \sigma^2\right), i = 1, 2, \cdots, n \end{cases} \tag{1}$$

若记：

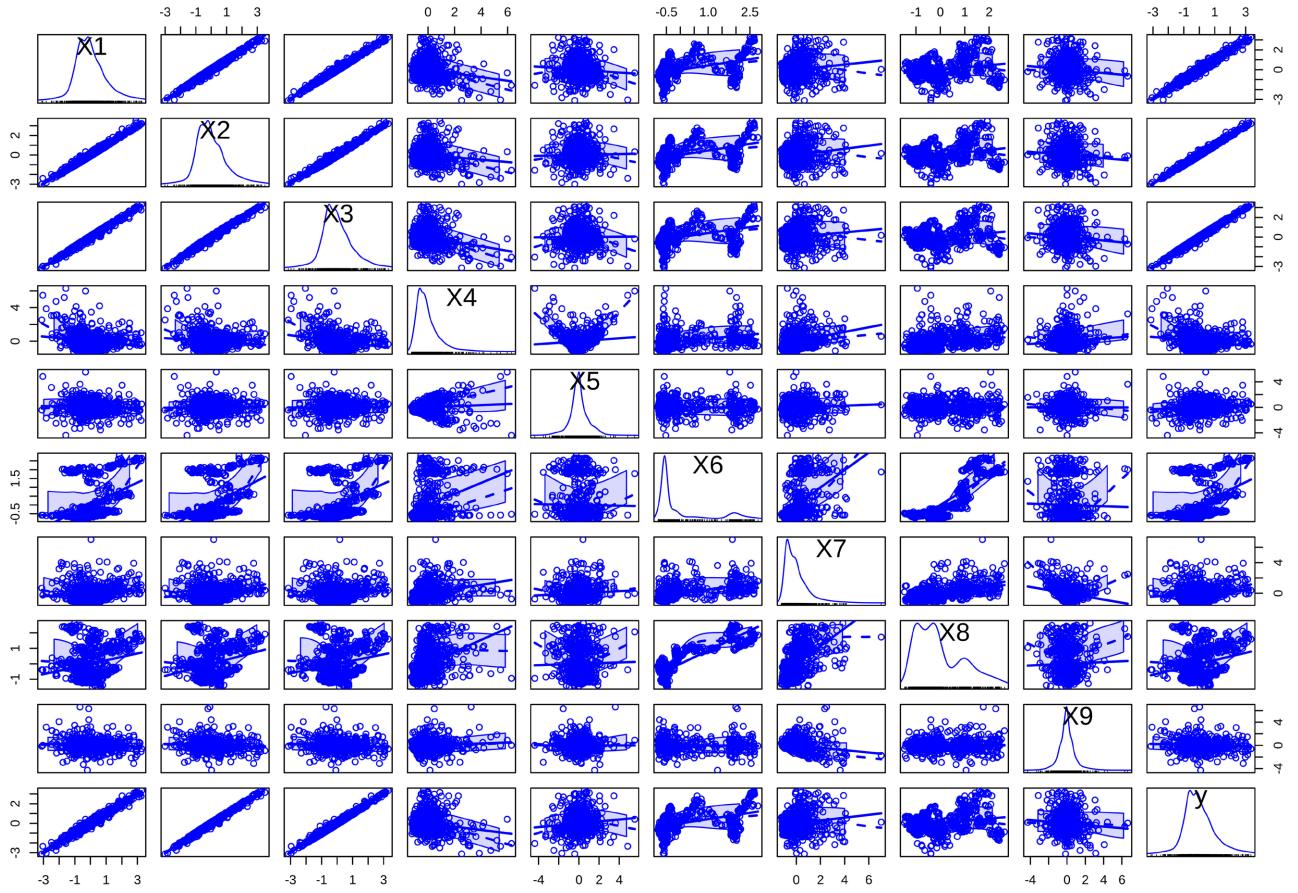



**Figure 5.** Establishment of correlation analysis model after data standardization
**图 5.** 数据标准化后的相关关系分析模型建立

$$\boldsymbol{Y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{2}$$

则式(2)可以用矩阵表示为：

$$\begin{cases} \boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}\sim N\left(\boldsymbol{0},\sigma^2\boldsymbol{I}_n\right) \end{cases} \tag{3}$$

故回归系数的最小二乘估计为：

$$\hat{\boldsymbol{\beta}}^{LS}=\arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\} \tag{4}$$

也即 $\hat{\boldsymbol{\beta}}^{LS}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ 。

在实际应用中， $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的特征值中存在接近于零的值，是导致自变量间产生多重共线性的重要原因，因此可用 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 的条件数

$$k=\frac{\lambda_1}{\lambda_p} \tag{5}$$

来度量原始自变量 $x_1, x_2, \cdots, x_p$ 之间多重共线性的严重程度，其中分别表示矩阵的最大特征值和最小特征值。一般若 $k<100$，则认为多重共线性程度较小；若 $100 \le k \le 1000$，则认为存在中等或偏强的多重共线性；若 $k>1000$，则认为存在严重的多重共线性。

当特征之间不存在多重共线性或特征较少时，LS 回归可以得到准确的系数估计。然而，当特征之间存在多重共线性时，LS 回归估计可能会变得不稳定，导致模型过拟合。

本文借助 R 语言，运用 kappa()函数对原始自变量进行检验，结果显示 $k=1125.524$，原始自变量间存在严多重共线性[12]。为进一步建立更加精准的预测模型，引入岭回归、Lasso 回归和弹性网络回归。

## 5.2. 岭回归

针对多重共线性的问题，Hoerl & Kennard 于 1970 年提出了岭回归[13]。岭回归估计的定义为：

$$\hat{\boldsymbol{\beta}}^{Ridge} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{6}$$

其中，$\lambda \ge 0$ 为惩罚参数，$\lambda$ 取值越大，回归系数收缩越大。特别地，当 $\lambda = 0$ 时，岭回归退化为 LS 回归。在惩罚项中，并没有对常数项 $\beta_0$ 进行惩罚，对每一个响应加上一个常数，不会对回归系数造成影响。

进一步得岭回归的解为

$$\hat{\boldsymbol{\beta}}^{Ridge} = \left( \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{7}$$

由式(7)可以看出，岭回归的解是在 LS 回归解的基础上，加了一个正的惩罚参数 $\lambda$，故当矩阵 $\boldsymbol{X}$ $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}$ 的某些列向量近似线性相关时，矩阵的奇异性要比 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 低，从而降低了估计值的方差，提高了估计精度。然而，岭回归也有一定的局限性，它的回归结果中包含所有的预测变量，岭回归对所有特征都进行缩减，但不会将系数缩减为零，故而没有进行变量选择，因此会影响模型的准确性。

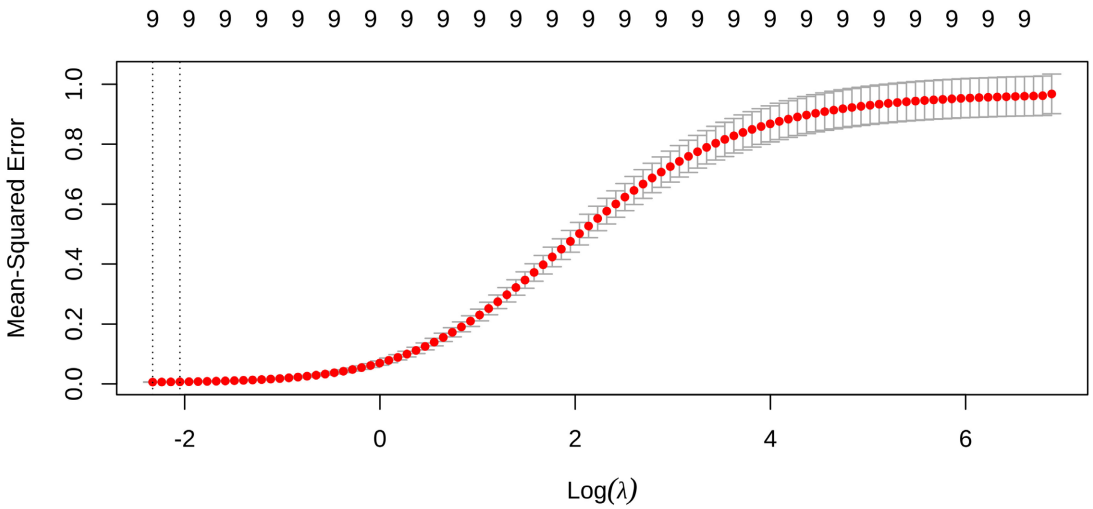



**Figure 6.** Cross validation results of Ridge regression
**图 6.** Ridge 回归的交叉验证结果图

见图 6，下方 $x$ 轴表示 Ridge 回归中惩罚项 $\lambda$ 值取对数 $(\log(\lambda))$，上方 $x$ 轴的数字表示每个 $\lambda$ 值对应的非 0 系数的变量个数，这些数字对应的值是说：不同 $\lambda$ 值计算得到模型中所有变量系数不为 0 的变量的个数，变量间不存在相关关系(系数为 0)被筛选掉了， $y$ 轴表示 MSE 值，每个红色点表示数据在进行

交叉验证过程中，每个 $\lambda$ 对应的 MSE 值，每条竖线(误差线)表示数据在进行交叉验证过程中，每个 $\lambda$ 对应的 MSE 值，左边虚线表示评价指标最佳的 $\lambda$ 值(lambda.min)，右边虚线表示评价指标在最佳值 1 个标准误范围的模型的 $\lambda$ 值(lambda.1se)。

通过 10 折交叉验证选取最优的惩罚项参数，得到图 6，随着 $\lambda$ 的变大，模型误差在逐渐变大，在图中可以找到误差最小时的 $\lambda$，通过计算得到 $\lambda_{\min}=0.1005$。

### 5.3. Lasso 回归

针对岭回归中没有变量选择问题，Tibshirani 在 1996 年提出 Lasso 回归，Lasso 估计的定义为[14]：

$$\hat{\boldsymbol{\beta}}^{Lasso}=\arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\} \tag{8}$$

与岭回归的二次惩罚项 $\lambda\sum_{j=1}^{p}\beta_j^2$ 相比，Lasso 的一次惩罚项 $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ 既能把非 0 的预测变量系数 $\beta_j$ 向 0 收缩，又能选择出那些很有价值的预测变量($\left|\beta_j\right|$ 值大的预测变量)。这是因为相对于 $\lambda\sum_{j=1}^{p}\beta_j^2$ 来说，$\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ 对变量系数 $\beta_j$ 的收缩程度要小，因此 Lasso 能选出更精确的模型。这意味着 Lasso 回归可以自动地选择重要的特征，排除掉不相关或不重要的特征，从而简化模型，提高模型的可解释性。

Lasso 回归与 LS 回归相比虽然大大降低了预测方差，达到了系数收缩和变量选择的目的，但是也有一定的局限性。譬如：

① 在 Lasso 回归求解路径中，对于 $n\times p$ 的设计矩阵来说，最多只能选出 $\min\{n,p\}$ 个变量，当 $p>n$ 的时候，最多只能选出 $n$ 个预测变量。

② 对于通常的 $n>p$ 情形，如果预测变量中存在很强的共线性，Lasso 的预测表现受控于岭回归。

通过 10 折交叉验证选取最优的惩罚项参数，见图 7，随着 $\lambda$ 的变大，模型误差在逐渐变大，在图中可以找到误差最小时的 $\lambda$，通过计算得到 $\lambda_{\min}=0.004151$，并且最终筛选出 4 个重要的特征变量，分别是最高价($X_2$)、最低价($X_3$)、交易量($X_4$)、涨跌幅($X_5$)。

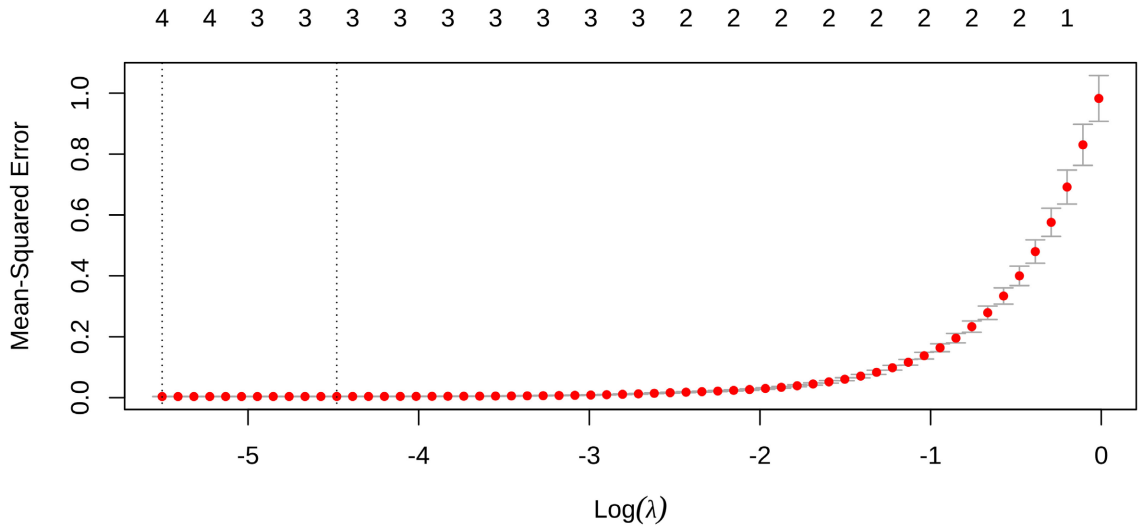



**Figure 7.** Cross-validation results of Lasso regression
**图 7.** Lasso 回归的交叉验证结果图

### 5.4. 弹性网络回归

基于 Lasso 回归的局限性，Zou & Hastie 在 2005 年提出了弹性网络回归方法，回归系数表达式为[15]：

$$\hat{\boldsymbol{\beta}}^{EN}=\arg\min_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\left[\alpha\sum_{j=1}^{p}\left|\beta_j\right|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right]\right\} \tag{9}$$

由此可知，弹性网络回归的惩罚项 $\lambda\left[\alpha\sum_{j=1}^{p}\left|\beta_j\right|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right]$ 恰好为岭回归惩罚项和 Lasso 回归惩罚项的一个凸线性组合。当 $\alpha=0$ 时，弹性网络回归即为岭回归；当 $\alpha=1$ 时，弹性网络回归即为 Lasso 回归，因此，弹性网回归兼有 Lasso 回归和岭回归的优点。

见图 8，通过 10 折交叉验证选取最优的惩罚项参数最佳 $\alpha$ 取值为 0.951，随着 $\lambda$ 的变大，模型误差在逐渐变大，在图中可以找到误差最小时的 $\lambda$，通过计算得到，并且最终筛选出 4 个重要的特征变量，分别是最高价($X_2$)、最低价($X_3$)、交易量($X_4$)、涨跌幅($X_5$)。

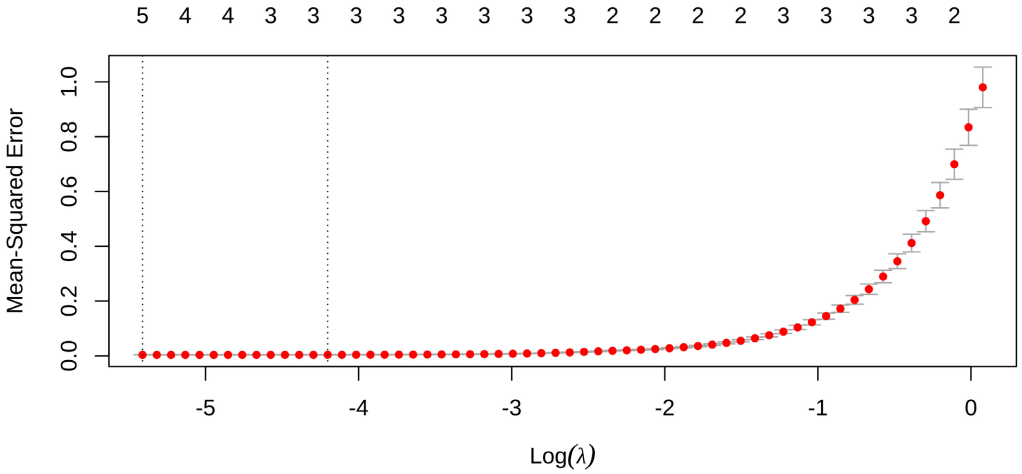



**Figure 8.** Cross-validation results of elastic net regression
**图 8.** 弹性网络回归的交叉验证结果图

## 6. 模型应用

### 6.1. 异常样本检验

在实际建模过程中，如果存在异常样本，可能会对模型准确性和可靠性产生较大的影响，因此在建模之前采用 Cook 距离、残差图、正态性检验对原始数据进行异常样本检验[16] [17]。

Cook 距离是由著名的统计学家库克(D.R.Cook)提出的，实际应用中通过对比每一个样本 Cook 距离的相对大小，从而对各样本的影响力做出大概的判断。如果发现有少数一两个样本的 Cook 距离从量级上明显比其他样本的 Cook 距离大特别多，则定义该样本为异常样本并剔除。见图 9，第 142 号样本的 Cook 距离远超出其他样本，因此将第 142 号样本定义为异常样本。

残差图以残差为纵坐标，以其他适宜的量(如拟合值、自变量的观察值或数据观测序号等)为横坐标。通过观察残差的离散程度，可以发现模型在不同位置的拟合情况。如果某些点的残差明显偏离其他点，那么这些点可能被视为异常样本。在正态性检验中，QQ 图(Quantile-Quantile 图)是一种非常有用的工具，如果样本数据来自正态分布，那么数据点应该大致沿着一条直线分布。在 QQ 图中，如果有一些数据点明显偏离了这条线，那么这些偏离的数据点可能是异常样本。

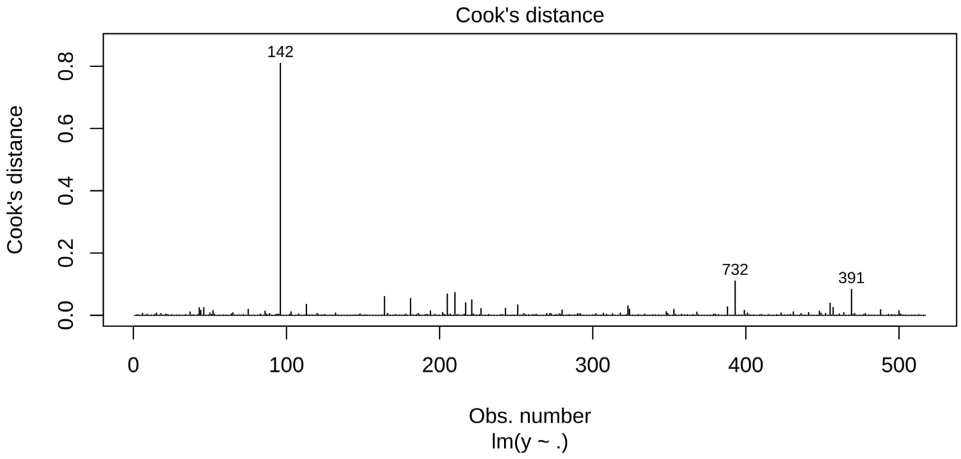
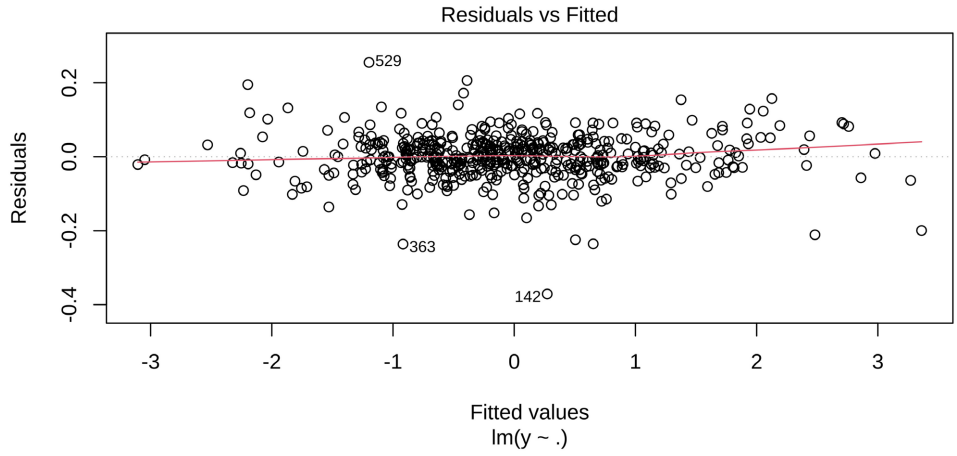
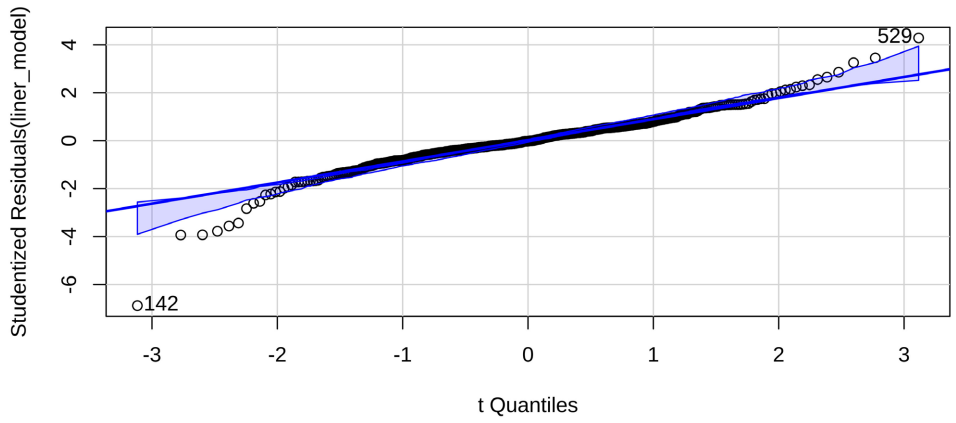



**Figure 9.** Cook distance of each sample
**图 9.** 各个样本的 Cook 距离

见图 10，第 142、529 号样本明显异常。综上，原始数据中的第 142 和 529 号样本为异常样本，故剔除后再进行回归建模。




(a)




(b)

**Figure 10.** Residual plot (a) and QQ plot (b)
**图 10.** 残差图(a)和 QQ 图(b)

### 6.2. 模型选择

基于剔除异常样本后的数据，运用标准线性回归、岭回归、Lasso 回归、弹性网络回归建立回归预测模型，得到最终的变量筛选及各变量的回归系数结果，见表 3。

**Table 3.** Variable screening and regression coefficients of each model
**表 3.** 各模型变量筛选及回归系数

| 模型和变量 | 线性回归 | 岭回归 | Lasso 回归 | 弹性网络回归 |
|---|---|---|---|---|
| 截距项 | −00002076 | 0.0005841 | −0.0002544 | −0.0002420 |
| $X_1$ | 0.0672455 | 0.3016001 | . | . |
| $X_2$ | 0.5631928 *** | 0.3302317 | 0.5894744 | 0.5899235 |
| $X_3$ | 0.3607701 *** | 0.3190983 | 0.4004924 | 0.4002317 |
| $X_4$ | −0.0129309 ** | −0.0135599 | −0.0073044 | −0.0076799 |
| $X_5$ | 0.0876680 *** | 0.1004186 | 0.0769397 | 0.0773040 |
| $X_6$ | 0.0089858 | 0.0290736 | . | . |
| $X_7$ | 0.0007566 | −0.0000631 | . | . |
| $X_8$ | −0.0070447 | −0.0068311 | . | . |
| $X_9$ | 0.0044297 | 0.0012680 | . | . |

为进一步验证各模型的性能，利用均方误差(MSE)作为评价指标，结果显示弹性网络回归模型的 MSE 和 MAE 值均最小，见表 4，选原则弹性网络回归模型作为最终的预测模型[12]，即

$$Y = -0.000242 + 0.5899235X_2 + 0.4002317X_3 - 0.0076799X_4 + 0.0773044X_5 \quad (10)$$

进一步表明最高价($X_2$)、最低价($X_3$)、交易量($X_4$)、涨跌幅($X_5$)都对茅台集团股票每天收盘价格产生显著影响，且除 $X_4$ 外，其他变量对收盘价影响都是正向的，其中 $X_2$ 和 $X_3$ 对收盘价的影响最大。

**Table 4.** Comparison of prediction accuracy of each model
**表 4.** 各模型预测精度比较

| 评价指标 | 线性回归 | 岭回归 | Lasso 回归 | 弹性网络回归 |
|---|---|---|---|---|
| MSE | 0.0028181 | 0.0045067 | 0.0028018 | **0.0027965** |
| MAE | 0.0407640 | 0.0518940 | 0.0407532 | **0.0407170** |

## 7. 结论与建议

### 7.1. 结论

本文比较多种回归模型(线性回归、岭回归、Lasso 回归和弹性网络回归)预测效果，结果表明弹性网络回归模型在处理多重共线性、筛选重要特征变量方面具有显著优势，能够更准确地预测股票收盘价。基于均方误差(MSE)和平均绝对误差(MAE)指标，弹性网络回归模型在所有模型中表现最佳。因此，本文认为弹性网络回归是预测股票价格有效方法之一，并具有较高应用前景。研究发现，最高价、最低价、交易量和涨跌幅这四个变量对收盘价影响最为显著，尤其是最高价和最低价。表明股票市场中这些变量变化对投资者决策有重要参考价值，特别是在短期股价预测中，它们能够提供较为可靠的指引。因此，未来研究和实务中，应重点考虑这些变量作用，并在制定投资策略时加以充分利用。本文通过模型的构

建和验证，证明了在特定条件下，股票的短期价格是具有一定可预测性的。尽管股票市场存在波动性，但通过科学的数据分析和模型选择，可以提高对市场动态的理解和预测能力，为投资者提供有力支持。未来的研究可以进一步扩展样本数据，验证模型的适用性，探索更加复杂的市场环境中的预测方法。

### 7.2. 建议

本文结果表明，在应用弹性网络回归模型预测股票价格时，最高价、最低价、交易量和涨跌幅是影响收盘价最显著变量，最高价和最低价波动对股价波动具有显著直接作用。因此，建议在未来研究与实际应用中，应特别关注这些关键指标变化，尤其是最高价和最低价的相对波动性。同时，结合交易量和涨跌幅的动态变化进行深入分析，有利于提升股价预测模型精度，进而为投资决策提供更可靠依据。

本文通过交叉验证发现，弹性网络回归模型能够有效结合岭回归与 Lasso 回归的优势，克服了传统线性模型中的多重共线性问题，并对特征变量进行了筛选和系数有效收缩。与单一使用岭回归或 Lasso 回归相比，弹性网络回归在处理复杂、多维变量环境时，能够在偏差和方差之间取得平衡，降低模型过拟合风险。因此，建议在股票价格预测研究中优先选用弹性网络回归模型。相比于线性回归、岭回归和 Lasso 回归等传统方法，弹性网络回归在多变量复杂背景下表现出更高的预测准确性和稳健性，具有较强的推广性和应用价值。

## 参考文献

[1] 周亚萍, 刁训娣. 融资融券与上证 50 ETF 期权收益[J]. 中央财经大学学报, 2024(6): 65-75.

[2] 赵莹莹. 基于自适应弹性网的上证 50 指数追踪研究[D]: [硕士学位论文]. 重庆: 西南大学, 2023.

[3] 魏巍. 基于弹性网分位数回归的模型平均方法在股指追踪中的应用[D]: [硕士学位论文]. 合肥: 安徽大学, 2023.

[4] 袁子杨. 基于自适应弹性网 Aenet 的稀疏稳健投资组合研究[D]: [硕士学位论文]. 深圳: 深圳大学, 2022.

[5] 邢艳雅. 一种新的弹性网模型及在时间序列预测中的应用研究[D]: [硕士学位论文]. 太原: 太原理工大学, 2021.

[6] 韩情, 汪子琦, 耿文静. 基于弹性网-自回归模型的股票价格研究[J]. 广西质量监督导报, 2020(10): 194-195.

[7] 周政瑜. 基于自适应弹性网 Expectile 回归的投资组合决策[D]: [硕士学位论文]. 长沙: 湖南大学, 2021.

[8] 宋家辉. 基于弹性网分位数的金融投资组合绩效评价[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2020.

[9] 严凌. 融资融券对我国证券市场股票交易影响研究[D]: [硕士学位论文]. 昆明: 昆明理工大学, 2011.

[10] 应建联, 潘斌, 陈文. 基于回归分析的新产品上市初期量价关系研究与实践[J]. 现代商业, 2024(8): 63-66.

[11] 马京晶. 股票价格趋势预测中的回归与分类研究[J]. 电脑知识与技术, 2024, 20(12): 12-14, 23.

[12] 何秀丽. 多元线性模型与岭回归分析[D]: [硕士学位论文]. 武汉: 华中科技大学, 2005.

[13] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. https://doi.org/10.1080/00401706.1970.10488634

[14] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[15] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[16] 卢铁男. 中国股票发行定价研究[D]: [博士学位论文]. 上海: 华东师范大学, 2002.

[17] 袁忠智, 陈柠. 分析测试中的数据处理和结果表述——正态样本异常值检验和正态性检验[J]. 兵工标准化, 1996(5): 24-26.