Operations Research and Fuzziology 运筹与模糊学, 2023, 13(6), 7130-7138
Published Online December 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.136699

基于弹性网降维的两步估计回归模型的上证50股票指数追踪研究

杨丽鑫，戴家佳

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2023年10月23日；录用日期：2023年12月16日；发布日期：2023年12月26日

摘　要

随着我国金融产业的飞速发展，股票投资成为大众最为青睐的一种理财方式。如何较为有效地对股票指数进行追踪对各投资机构以及众多散户来说至关重要。文章对上证50综合指数日收盘价数据建立基于弹性网降维的两步估计回归模型，第一步先采用绝对约束估计和弹性约束估计对原始变量进行降维，再根据误差分析结果，选择使得指数追踪误差更小的解释变量作为指数追踪的研究对象，第二步用最小二乘估计建立经验回归方程，再使用逐步回归剔除不显著变量，寻找部分股票构成的最优的追踪组合。实证结果表明：弹性网降维的两步估计回归模型能更有效地对股票价格进行预测，指数追踪效果最好。

关键词

两步估计，弹性网回归，指数预测模型，Lasso回归

Shanghai 50 Stock Index Tracking Research Based on Elastic Net Dimensionality Reduction Two-Step Estimation Regression Model

Lixin Yang, Jiajia Dai

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Oct. 23rd, 2023; accepted: Dec. 16th, 2023; published: Dec. 26th, 2023

Abstract

With the rapid development of China's financial industry, stock investment has become the most

文章引用: 杨丽鑫, 戴家佳. 基于弹性网降维的两步估计回归模型的上证 50 股票指数追踪研究[J]. 运筹与模糊学, 2023, 13(6): 7130-7138. DOI: 10.12677/orf.2023.136699

popular way of financing. How to track the stock index more effectively is very important for each investment institution and many retail investors. This paper establishes a two-step estimation regression model based on elastic net dimensionality reduction for the daily closing price data of Shanghai Composite Index 50. In the first step, absolute constraint estimation and elastic constraint estimation are used to reduce the dimensionality of the original variables. Then, according to the error analysis results, explanatory variables with smaller index tracking errors are selected as the research objects of index tracking. The second step is to establish the empirical regression equation with the least square estimation, and then use stepwise regression to eliminate the insignificant variables to find the optimal tracking combination of some stocks. The empirical results show that the two-step regression model of elastic net dimensionality reduction can predict the stock price more effectively, and the index tracking is the best.

Keywords

Two-Step Estimation, Elastic Net Regression, Exponential Prediction Model, Lasso Regression

1. 引言

计算机科学的快速发展使得当代科学研究能够很容易地收集到海量数据集，特别是在金融领域。而金融领域指数型衍生品创新在金融危机中表现突出，被大多数投资者和机构所采用，根据指数来选择股票成为投资者重点关注之一，因此，如何用部分股票来追踪目标指数就成为是否采用指数法进行投资的关键所在，指数跟踪技术的目的是使投资者获得较高利益。股票指数追踪研究的内容是通过权重的优化再配置来寻找使得该组合指数的追踪误差最小部分股票，实现组合收益与指数涨跌基本一致，此研究具有高精度、低交易成本、且能保证追踪组合的高流动性，具有重要的意义[1]。

本文在解决股票指数追踪问题时，第一步先采用绝对约束估计和弹性约束估计对原始变量进行降维，再根据误差分析结果，选择使得指数追踪误差更小的解释变量作为指数追踪的研究对象，第二步用最小二乘估计建立经验回归方程，保证系数的估计是无偏的，有更小的残差平方和，寻找部分股票构成的最优的追踪组合。

2. 材料与方法

2.1. 数据采集

本文数据来源于金融 choice，选取上证 50 成分股收盘价数据。表 1 展示上证 50 包含的 50 家上市公司，如伊利股份、中国电信、山西汾酒、贵州茅台、招商银行、中信证券等。

Table 1. List of stocks in the Shanghai 50 Index
表 1. 上证 50 指数成分股列表

1	2	3	4	5
包钢股份	中国石化	中信证券	三一重工	招商银行
上汽集团	北方稀土	复星医药	恒瑞医药	万华化学

Continued

片仔癀	通威股份	贵州茅台	海螺水泥	用友网络
保利发展	恒力石化	恒生电子	海尔智家	闻泰科技
山西汾酒	海通证券	伊利股份	航发动力	长江电力
三峡能源	隆基绿能	中信建投	中国神华	兴业银行
国泰君安	农业银行	中国平安	工商银行	中国太保
中国人寿	长城汽车	中国建筑	华泰证券	中国电信
中国石油	中国中免	紫金矿业	中远海控	中金公司
药明康德	海天味业	韦尔股份	华友钴业	兆易创新

该数据集一共有 243 个样本，包括 50 个解释变量和 1 个响应变量(上证 50 指数)，部分数据展示如下表，可以看到宝钢股份和中国石化的收盘价几乎在 5 元以下，而兆易创新的收盘价则超过 100 元，这里说明了不同企业的收盘价会存在较大差异，需要在建模前要对数据进行标准化。

2.2. 统计分析方法

在金融大数据统计分析中，首要的问题就是变量的选择问题，由于变量的影响、数据收集的成本和分析的时效不同，并不总是需要尽可能多的收集全部的变量，且证券市场中很多变量是相互依存的，这时也没有必要将高度关联的变量都考虑进来。变量选择较为常用的方法有以下 3 种：逐步回归法、绝对约束估计、弹性约束估计。

逐步回归法的基本思想是将变量逐个引入模型，每引入一个解释变量后都要进行 F 检验，并对已经选入的预测变量逐个进行 t 检验，当原来引入的预测变量由于后面预测变量的引入变得不再显著时，则将其删除。以确保每次引入新的变量之前回归方程中只包含显著性变量。这是一个反复的过程，直到既没有显著的预测变量选入回归方程，也没有不显著的预测变量从回归方程中剔除为止。以保证最后所得到的预测变量集是最优的。

绝对约束估计(Lasso)通过构造一个惩罚函数得到一个较为精炼的模型，使得它压缩一些回归系数，即强制系数绝对值之和小于某个固定值[2]；同时设定一些回归系数为零。因此保留了子集收缩的优点，是一种处理具有复共线性数据的有偏估计。Lasso 就是在普通的线性回归模型的残差平方和后加入 1 范数惩罚项，具体数学表达式如下，

$$\begin{aligned} &Q(\beta)=\|y-X\beta\|^2+\lambda\|\beta\|_1 \\ &\Leftrightarrow \min\|y-X\beta\|^2 \ \text{s.t.}\ \sum|\beta_j|\le s \end{aligned} \tag{1}$$

s 值控制了收缩情况。s 值越大，变量收缩较小；s 值越小，变量收缩越大，部分变量系数变成 0，所以 Lasso 具有重要的稀疏性质。Lasso 的复杂程度由 λ 来控制，λ 越大对变量较多的线性模型的惩罚力度就越大，从而最终获得一个变量较少的模型，故一个合适 λ 值对建立指数追踪模型尤为重要，在后续的建模中将使用交叉验证选取最优的 λ 值。

弹性约束估计(Elastic Net)结合了岭回归和 Lasso 的正则化方法通过两个参数 λ_1 和 λ_2 来控制惩罚项的大小[3]，具体数学表达式定义如下：

$$\tilde{\beta}=\arg\min_{\beta}\left(\sum_{i}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2+\lambda_1\sum_{j=1}^{p}\left|\beta_j\right|+\lambda_2\sum_{j=1}^{p}\beta_j^2 \tag{2}$$

当 $\lambda_1=0$ 时，弹性约束估计就是岭回归，当 $\lambda_2=0$ 时，弹性约束估计就是绝对约束估计，因此，弹性约束估计同时具有绝对约束估计和岭估计的特点[4]。

普通的线性回归模型要求满足高斯马尔可夫条件[5]，即误差项同方差、0 均值、且不相关，而对线性回归模型中回归系数或误差方法进行估计，通常采用最小二乘估计，因为其估计出来的参数具有优良的性质，设多元线性回归方程如下：

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_p x_p \tag{3}$$

普通最小二乘法拟合线性模型，本质上是要寻找使得残差平方和达到最小的系数估计，这时所估计的系数满足无偏性、有效性，即最小二乘估计的方差最小[6]。

3. 描述分析及多重共线性检验

3.1. 描述分析性统计分析

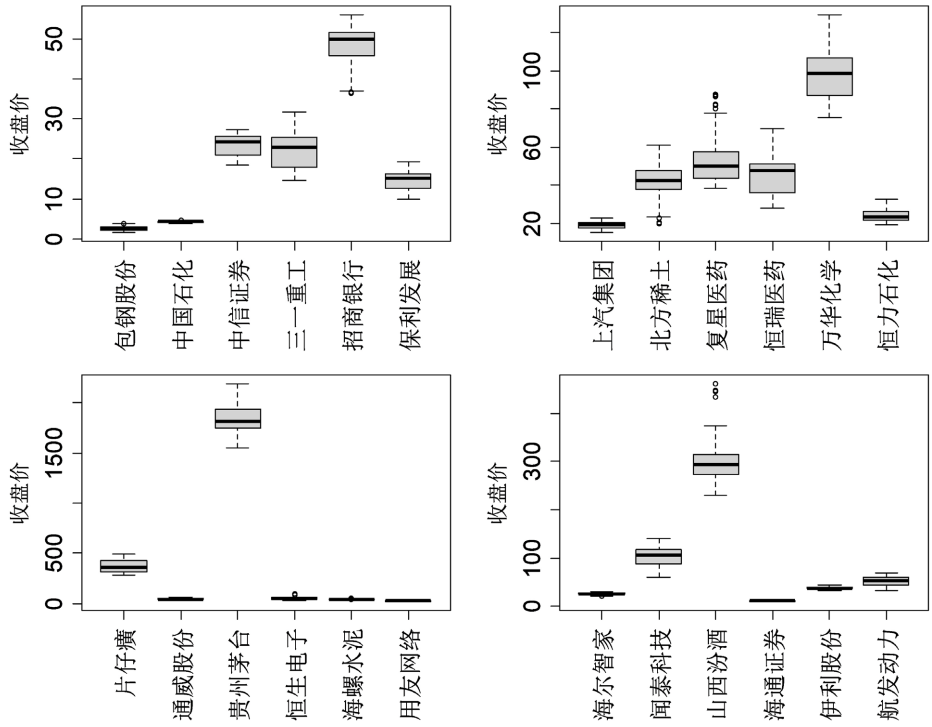

Figure 1. Analysis of box chart of closing price of some enterprises
图 1. 部分企业收盘价箱线图分析

从图 1 可以发现，招商银行、万华化学、贵州茅台、山西汾酒 4 个企业相比于同组的其它公司，其收盘价均值、最大值、最小值、第 1 四分位数、第 3 四分位数的表现都遥遥领先，其中贵州茅台的收盘价最高，均值超过 1500 元。从上图的第 1 幅箱线图可看到中国石化收盘价整体上略高于宝钢股份，但总体收盘价水平差不多，中信证券相比于三一重工收盘价均值略高，但三一重工收盘价浮动较大，中信证券更加稳定；第 2 幅箱线图中，北方稀土、复星医药、恒瑞医药 3 个企业差别不大，其中北方稀土和复

星医药存在较多异常值，同组内上汽集团收盘价最低，均值在 20 左右；第 3 幅箱线图可以发现通威股份、恒生电子、海螺水泥、用友网络这四个企业收盘价表现大致相同；第 4 幅箱线图中，除闻泰科技和山西汾酒外，其它企业收盘价这段时间内收盘价一直低于 100，最差的是海通证券，且发现山西汾酒数据存在较多异常值。对于上述分析中存在异常值的变量都借助 R 软件中的 cooks.distance()函数计算其库克距离，再结合现实情况采取剔除、修正或保留等措施。

3.2. 多重共线性检验

为了说明哪几个自变量之间有一定的多重共线性的关系存在，接下来使用方差扩大因子法来诊断多重共线性[7]，表 2 给出了 49 个企业所对应的方差膨胀因子。没有列在表格内的韦尔股份的方差膨胀因子为 122.59，一般认为回归方程的多重共线性的存在就是由方差膨胀因子超过 10 的这几个变量引起的[8]，但是针对本例，没有低于 10 的方差膨胀因子，最低的是隆基绿能的 14.71，通过查看方差膨胀因子较大的企业与其它公司的相关系数，发现这些企业会与多数的企业具有较强的相关性，而方差膨胀因子较小的隆基绿能等企业就出现与以上情况相反的情况，故重点关注方差膨胀因子较大的特征变量，如宝钢股份(37.01)、三一重工(89.90)、长城汽车(73.3)、中国电信(88.124)、等，可以结合 lasso 和 Elastic Net 降维的结果综合考虑采用剔除这些变量等方法来解决多重共线性问题。

Table 2. Variance inflation factor of Shanghai 50 enterprises
表 2. 上证 50 企业的方差膨胀因子情况

包钢股份	中国石化	中信证券	三一重工	招商银行	保利发展	上汽集团
37.01	17.72	115.7	89.90	57.39	35.08	35.77
北方稀土	复星医药	恒瑞医药	万华化学	恒力石化	华友钴业	片仔癀
35.48	32.14	163.08	48.36	27.14	15.08	38.62
通威股份	贵州茅台	恒生电子	海螺水泥	用友网络	海尔智家	闻泰科技
16.72	25.43	35.33	16.21	66.74	19.24	58.08
山西汾酒	海通证券	伊利股份	航发动力	兆易创新	长江电力	三峡能源
18.46	113.24	20.80	45.3	17.65	28.014	15.312
隆基绿能	中信建投	中国神华	兴业银行	国泰君安	农业银行	中国平安
14.71	58.73	97.91	24.79	106.50	31.31	102.98
工商银行	中国太保	中国人寿	长城汽车	中国建筑	华泰证券	中国电信
22.98	153.54	55.32	73.30	36.82	122.20	88.124
中国石油	中国中免	紫金矿业	中远海控	中金公司	药明康德	海天味业
17.8403	43.6949	7.5307	22.6997	83.2568	43.4336	47.870

4. 上证 50 股票指数追踪实证分析

4.1. 绝对约束估计

选取上证 50 成分股收盘价数据对 Lasso 和 Elastic Net 的变量选择方法在指数跟踪方面的应用效果进

行详细分析。本文将样本数据集划分为训练集(2/3)和测试集(1/3)：共有 161 个拟合样本，80 个预测样本。利用绝对约束估计得到了回归方程的系数估计，其中宝钢股份、北方稀土、恒力石化、恒生电子、三峡能源、中国神华、工商银行、长城汽车、中国建筑、华泰证券、中国电信、中国石油、中远海控、华友钴业等 14 个变量系数为 0，即 Lasso 方法剔除了 14 个解释变量，还剩余 36 个解释变量。

从图 2 可以发现，在训练集上证 50 指数跟踪图中，星号代表拟合值，实线代表真实值，可看出实际走势和 36 只成分股的跟踪走势基本重合，偏离程度较小，除拐点外，两条曲线在一定程度上几乎重合，说明该时间段内目标指数的跟踪效果良好。而对于指数跟踪效果，不能只看模型在训练集中的表现，还需要用测试集数据进行验证，从预测残差图看到 1~65 号样本预测残差在 0 附近波动，最高在 20 附近，而从 65 号样本之后预测残差呈直线上升趋势，最高超过 60，结合测试集指数跟踪情况，发现 40 号样本之后，预测指数和实际指数趋势虽然相同，但是基本没有重合，而是在上下波动，故 Lasso 回归(绝对约束估计)在预测集取得良好的指数跟踪效果，而在测试集的表现不尽如人意。

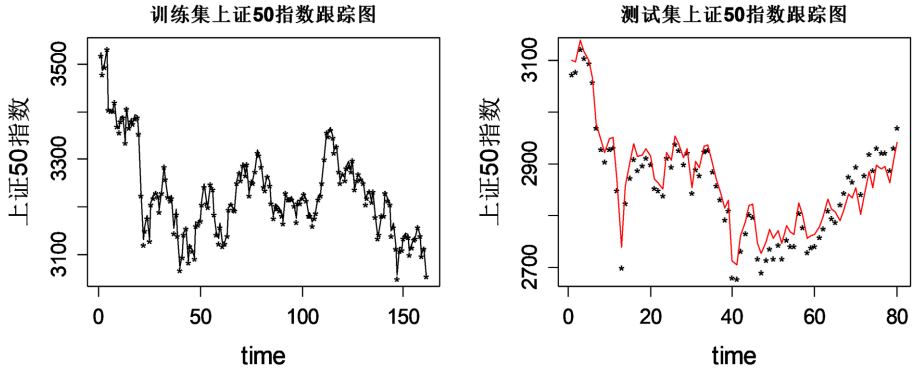

Figure 2. Absolute constraint estimation residuals and visualization of tracking results
图 2. 绝对约束估计残差及跟踪结果可视化

4.2. 弹性约束估计

利用弹性约束估计得到了回归方程的系数估计，其中宝钢股份、北方稀土、恒力石化、恒生电子、三峡能源、中国神华、工商银行、中国人寿、长城汽车、中国建筑、中国电信、中国石油、中远海控、华友钴业等 14 个变量系数为 0，Elastic Net 方法(弹性约束估计)剔除的变量多数与 Lasso 方法相同，唯一不同的是 Elastic Net 方法剔除了中国人寿，而 Lasso 方法删除了华泰证券。

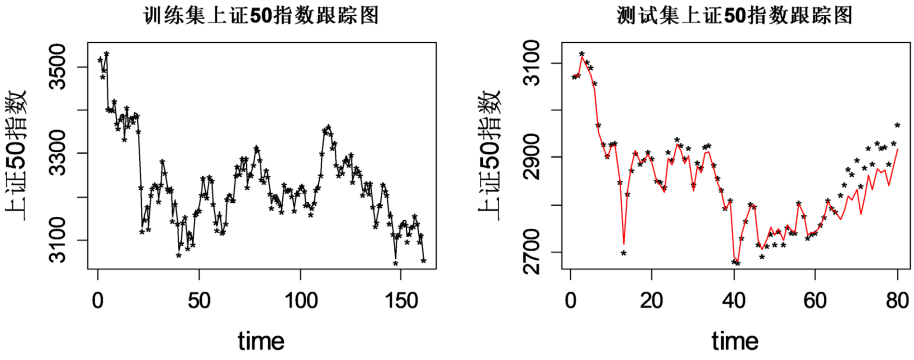

Figure 3. Residual estimation with elastic constraints and visualization of tracking results
图 3. 弹性约束估计残差及跟踪结果可视化

图 3 中对于训练集指数跟踪效果，Elastic Net 方法预测残差图和 Lasso 方法看上去大致相同，都是从 65 号样本之后预测残差呈直线上升趋势，最高超过 60，而从 Elastic Net 测试集上证 50 指数跟踪图可以看到，65 号样本之前实际走势和预测走势基本重合，比 Lasso 方法追踪效果好，而从 65 号样本之后，预测误差很大，基本脱离真实值。

4.3. 基于弹性网降维的两步估计回归模型

前面部分的分析中，Elastic Net 方法为我们选择了包括中国石化、中信证券、三一重工等 36 个解释变量，使用这 36 个解释变量和响应变量上证 50 指数建立一般的线性回归模型，通过最小二乘估计获得拟合模型，使得系数估计具有无偏性。

其估计残差及跟踪结果展示如图 4 所示，训练集的表现效果与单独使用 Elastic Net 和 Lasso 方法看起来相差不大，但是可以发现基于弹性网降维的两步估计模型(后简称两步估计)明显的改善了预测集的上证 50 指数追踪效果，预测残差总的在−10 到 10 之间波动，65 号样本之后的残差不再呈直线上升趋势，最大值也从之前的 60 几降低到 10 几。从测试集上证 50 指数跟踪图可以看到预测曲线和真实曲线基本重合，取得良好的指数跟踪效果。

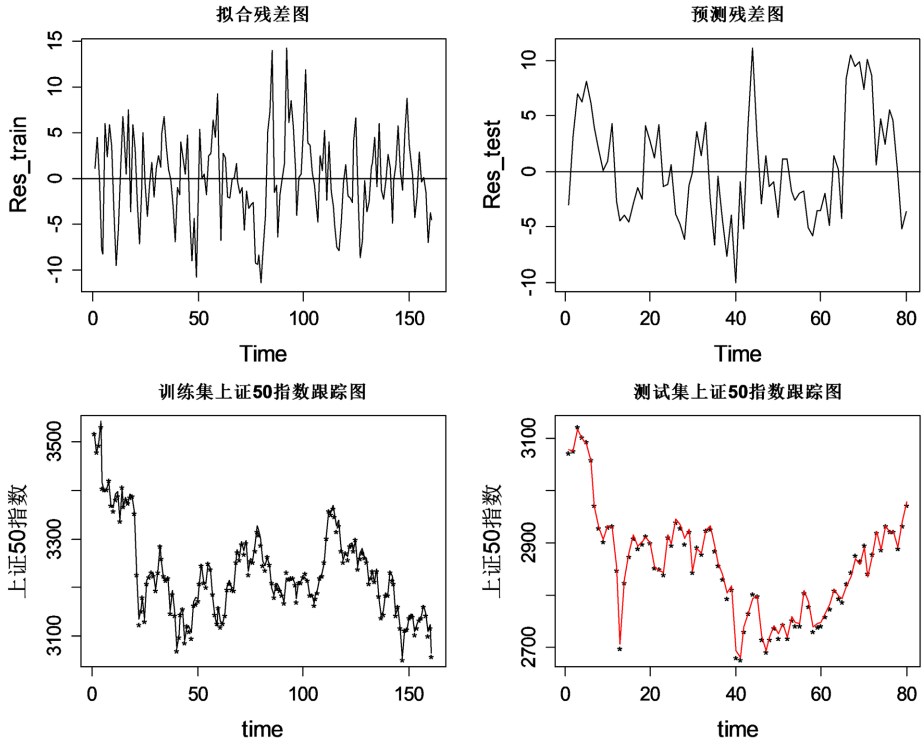

Figure 4. Residual estimation with elastic constraints and visualization of tracking results
图 4. 基于弹性网降维的两步估计残差及跟踪结果可视化

表 3 给出了弹性网降维的两步估计回归模型的误差分析，对比分析在逐步回归删除不显著变量前后模型于四个评价指标的上的表现[9]。训练集上两者变现差不多，而在测试集上，两步估计 + 逐步回归模型(30 个自变量)略优于两步估计模型(36 个自变量)，且考虑作为基金公司，需要用最少的变量达到对指数的跟踪，从而实现股票与股指期货的对冲，故两步估计 + 逐步回归模型(TSERS)将作为我们上证 50 指数追踪的最终模型。

Table 3. Variance inflation factor of Shanghai 50 enterprises
表 3. 上证 50 企业的方差膨胀因子情况

评价指标	训练集 (两步估计)	训练集 (两步估计 + 逐步回归)	测试集 (两步估计)	测试集 (两步估计 + 逐步回归)
SSE	3608.705	3683.268	1800.851	1781.589
MSE	22.368	22.83	22.510	22.270
RMSE	4.729	4.778	4.744	4.719
MAE	3.673	3.728	3.88	3.887

最终 TSERS 模型解释变量及其系数估计如下，此时回归系数是无偏的，且所有的变量均显著，模型的 R-squared 为 0.9994，Adjusted R-squared 为 0.9994，p-value 为 2.2e−16，远远小于 0.05，说明模型是显著的，且具有较好的拟合效果。下表 4 给出了最终的参数估计结果。

Table 4. Two-step estimation of dimensionality reduction of elastic network + stepwise regression model coefficient
表 4. 弹性网降维的两步估计 + 逐步回归模型系数

变量	系数	变量	系数
中国石化	20.247547***	海通证券	8.182230***
中信证券	2.533406*	伊利股份	2.321856***
三一重工	1.067891*	航发动力	1.340668***
招商银行	4.613060***	长江电力	5.147644***
保利发展	2.667730***	隆基绿能	0.395359***
上汽集团	4.920164***	中信建投	3.088349***
复星医药	0.778536***	兴业银行	9.450583***
恒瑞医药	2.902463***	中国平安	4.464143***
万华化学	1.279498***	中国太保	3.062012**
通威股份	2.902954***	中国中免	0.369060***
贵州茅台	0.301109***	紫金矿业	8.952732***
用友网络	1.019569***	中金公司	−0.789409*
海尔智家	1.985032***	药明康德	0.724065***
闻泰科技	−0.139162.	韦尔股份	0.388946***
山西汾酒	0.133853***	兆易创新	0.307495***

为了更好地对比各个模型的效果，表 5 给出了四个模型的误差对比分析。

可以发现两步估计 + 逐步回归在 4 个指数追踪评价指标上的表现都一致的好，进一步说明该方法的优势。

Table 5. Comparative analysis of errors on four regression model test sets
表 5. 四个回归模型测试集上的误差对比分析

评价指标	lasso 回归	弹性网	两步估计	两步估计 + 逐步回归
SSE	61009.86	59975.63	1800.851	**1781.589**
MSE	762.6233	749.70	22.510	**22.270**
RMSE	27.616	27.38	4.744	**4.719**
MAE	20.63395	18.362	3.88	**3.887**

5. 总结

股票指数追踪作为投资组合策略，能帮助投资者从证券市场获得超出预期的收益，在一定程度上降低投资风险，受到很多投资者青睐。本文以上证 50 指数及其成分股为研究对象，构建了基于弹性网降维的两步估计回归指数追踪模型。第一步采用弹性约束估计对原始成分股进行降维，第二步建立最小二乘估计经验回归方程，并且采用逐步回归方法来剔除不显著变量。其中还包括了对原始数据的异常值检验和变量间的多重共线性检验。通过指数追踪实证研究验证明，本文提出的模型在拟合性、预测性、误差最小性上都取得较好的结果，实现了较优的指数追踪效果，并使得投资组合的稀疏性达到最优，一定程度上降低了投资者的投资成本，具有一定的实际指导意义。

参考文献

[1] 秦晔玲, 朱建平. 基于自适应 Lasso 变量选择方法的指数跟踪[J]. 统计与决策, 2018, 34(16): 141-145.

[2] 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.

[3] Sloboda, W.B., Pearson, D. and Etherton, M. (2023) An Application of the LASSO and Elastic Net Regression to Assess Poverty and Economic Freedom on ECOWAS Countries. *Mathematical Biosciences and Engineering*: *MBE*, **20**, 12154-12168. https://doi.org/10.3934/mbe.2023541

[4] Lakmini, W. and Liwan, L. (2023) Mind the Large Gap: Novel Algorithm Using Seasonal Decomposition and Elastic Net Regression to Impute Large Intervals of Missing Data in Air Quality Data. *Atmosphere*, **14**, Article 355. https://doi.org/10.3390/atmos14020355

[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. https://doi.org/10.27352/d.cnki.gylgu.2021.001055

[6] Lukman, A.F., Farghali, R.A., Kibria, B.M.G., *et al*. (2023) Robust-Stein Estimator for Overcoming Outliers and Multicollinearity. *Scientific Reports*, **13**, Article No. 9066. https://doi.org/10.1038/s41598-023-36053-z

[7] 王瑶. 基于多重共线性修正下的多元线性回归[D]: [硕士学位论文]. 伊宁: 伊犁师范大学, 2023.

[8] 安雨晴, 杨宇, 王莉. 一种基于多元逐步回归的裂纹定量监测模型[J]. 工程与试验, 2023, 63(3): 42-47.

[9] Hongbing, L. and Mi, H. (2023) Forecast of China's Natural Gas Demand Based on the Double-Logarithmic Model with Stepwise Regression Method. *Energy Sources, Part A*: *Recovery, Utilization, and Environmental Effects*, **45**, 8491-8506. https://doi.org/10.1080/15567036.2023.2227584