Operations Research and Fuzziology 运筹与模糊学, 2022, 12(4), 1420-1438
Published Online November 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.124151

Hans汉斯

基于回归模型的股票指数追踪问题实证研究

周　洁

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2022年10月9日；录用日期：2022年11月5日；发布日期：2022年11月11日

摘　要

随着股票价格指数的发展与演变，股指对于投资的作用显得尤为重要。本文采用最小二乘估计、岭估计、绝对约束回归(Lasso)、弹性约束估计及两步估计对深证区块链50指数进行指数追踪，得到相应的投资组合，并将Cp准则和CV准则下的Lasso和岭估计作对比，得出结论：Cp准则下岭估计更好，CV准则下Lasso更好，两步估计下Lasso进行变量选择后用刘估计进行回归效果较好。

关键词

指数追踪，岭估计，Lasso估计，刘估计，弹性约束估计

An Empirical Study of the Stock Index Tracking Problem Based on a Regression Model

Jie Zhou

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Oct. 9th, 2022; accepted: Nov. 5th, 2022; published: Nov. 11th, 2022

Abstract

With the development and evolution of stock price index, the role of stock index in investment is particularly important. This paper uses least squares estimation, ridge estimation, absolute constraint regression (Lasso), elastic constraint estimation and two-step estimation to track the SZSE Blockchain 50 Index to obtain the corresponding investment portfolio. The Lasso and ridge estimation are compared, and it is concluded that the ridge estimation is better under the Cp criterion, the Lasso under the CV criterion is better, and the Lasso performs variable selection under the two-step estimation.

文章引用: 周洁. 基于回归模型的股票指数追踪问题实证研究[J]. 运筹与模糊学, 2022, 12(4): 1420-1438.
DOI: 10.12677/orf.2022.124151

Keywords

Index Tracking, Ridge Estimation, Lasso Estimation, Liu Estimation, Elastic Constraint Estimation

1. 引言

指数追踪通过追踪股票市场基准指数收益构建投资组合，是一种被动型投资策略，其目的是追踪一个股票指数的持仓或盈利表现，试图最小化跟踪误差。投资者以指数成分股为投资对象，通过购买一部分或全部的某指数中的股票来构建投资组合，以此来使投资组合的变动趋势与该指数相一致，取得与指数大致相同的收益率。

杨楠(2004)用岭回归解决多重共线性问题，对分析各变量间的关系具有独特帮助[1]。薛宏刚，张锐敏等人(2012)将岭回归的方法应用到套期保值技术中，发现该方法能有效提高样本外的套期保值效率[2]。张家茂，杨思思(2017)在对房地产股价线性模型的变量选择进行研究时，运用弹性约束估计实现成分股变量选择问题[3]。张慧伟(2018)基于弹性估计筛选出部分成分股来进行股指追踪，结果表明可以用较少的成分股来吻合指数的走势[4]。杨思思(2018)建立股指与其成分股的线性回归模型，利用岭估计和弹性约束估计探讨模型中的多重共线性问题，不断修正得到有效的回归模型[5]。王琪，冷林峰等人(2018)在研究股指跟踪时采用两步估计，先用弹性约束估计筛选一部分变量再做回归[6]。J Ranstam, J A Cook (2018)提出 Lasso 回归旨在识别变量和相应的回归系数，从而形成最小化预测误差的模型[7]。韩笑，滕兴虎等人(2020)采用正回归、绝对约束估计和弹性约束估计选择变量，得出银行类指数及其成分股的线性回归方程[8]。

2. 数据与描述

2.1. 数据说明

深证区块链 50 指数(代码 399286.SZ)由深圳证券交易所和深圳证券信息有限公司于 2019 年 12 月 24 日正式对外发布，是以深交所上市公司中，业务领域涉及区块链产业上中下游的公司为选样空间，按近半年日均总市值从高到低排序，筛选排名前 50 名的股票构成样本股[9]。根据指数的编制方法，易知区块链 50 指数是 50 只成分股股价的加权平均，权重与成分股的股本有关。其成分股及代码如表 1 所示：

Table 1. List of blockchain 50 components
表 1. 区块链 50 成分股列表

序号	代码	名称	序号	代码	名称
1	000001.SZ	平安银行	26	002537.SZ	海联金汇
2	000100.SZ	TCL 科技	27	002558.SZ	巨人网络
3	000158.SZ	常山北明	28	002602.SZ	世纪华通
4	000333.SZ	美的集团	29	002610.SZ	爱康科技
5	000555.SZ	神州信息	30	002727.SZ	一心堂
6	000676.SZ	智度股份	31	300007.SZ	汉威科技

Continued

7	000681.SZ	视觉中国	32	300033.SZ	同花顺
8	000776.SZ	广发证券	33	300059.SZ	东方财富
9	000997.SZ	新大陆	34	300099.SZ	精准信息
10	002010.SZ	传化智联	35	300113.SZ	顺网科技
11	002063.SZ	远光软件	36	300130.SZ	新国都
12	002065.SZ	东华软件	37	300170.SZ	汉得信息
13	002104.SZ	恒宝股份	38	300212.SZ	易华录
14	002117.SZ	东港股份	39	300271.SZ	华宇软件
15	002123.SZ	梦网科技	40	300339.SZ	润和软件
16	002131.SZ	利欧股份	41	300352.SZ	北信源
17	002152.SZ	广电运通	42	300377.SZ	赢时胜
18	002195.SZ	二三四五	43	300379.SZ	东方通
19	002230.SZ	科大讯飞	44	300386.SZ	飞天诚信
20	002268.SZ	卫士通	45	300465.SZ	高伟达
21	002352.SZ	顺丰控股	46	300468.SZ	四方精创
22	002400.SZ	省广集团	47	300525.SZ	博思软件
23	002410.SZ	广联达	48	300579.SZ	数字认证
24	002517.SZ	恺英网络	49	300663.SZ	科蓝软件
25	002530.SZ	金财互联	50	300676.SZ	华大基因

收盘价指股市收盘价，为当日该证券最后一笔交易前一分钟所有交易的成交量加权平均价(含最后一笔交易)。收盘价计算方式：下午 3 时收盘前的 3 分钟将实施收盘集合竞价的方式，用以确定收盘价，收盘集合竞价不能产生收盘价的，以最后一笔成交价为当日收盘价。本文选用 2020 年 1 月 2 日至 2022 年 7 月 29 日的区块链 50 指数及其成分股的日 K 线的收盘价，含 50 个自变量，1 个因变量，共有 51 列 624 行，共计 31,824 个样本数据。按照训练集:测试集 = 2:1 的原则来划分，样本数据追踪期间为 2020 年 1 月 2 日至 2021 年 9 月 15 日，检验期为 2021 年 9 月 16 日至 2022 年 7 月 29 日，数据示例如表 2 所示。本文数据来源于 Choice 金融终端。

Table 2. Table of data
表 2. 数据示例表

平安银行	TCL 科技	常山北明	美的集团	…	数字认证	科蓝软件	华大基因	Y
16.87	4.57	6.92	59.75	…	39.33	26.31	68.5	3273.823
17.18	4.65	6.88	58.26	…	39.72	26.29	68.7	3293.014
17.07	4.61	6.72	57.2	…	38.9	26.39	67.16	3296.561
…	…	…			…	…	…	…
12.79	4.28	6.63	55.92	…	25.18	16	63.98	2923.4903
12.88	4.43	6.6	55.75	…	24.91	15.94	64.02	2937.6198
12.68	4.44	6.42	54.98	…	24.7	15.97	62.02	2894.3303

区块链 50 指数收盘价从 2020 年 1 月 2 日至 2022 年 7 月 29 日的走势图如图 1 所示。本文旨在通过收集到 s 的数据建立区块链 50 指数与各个成分股的线性回归方程，用于描述区块链 50 指数的跟踪效果。

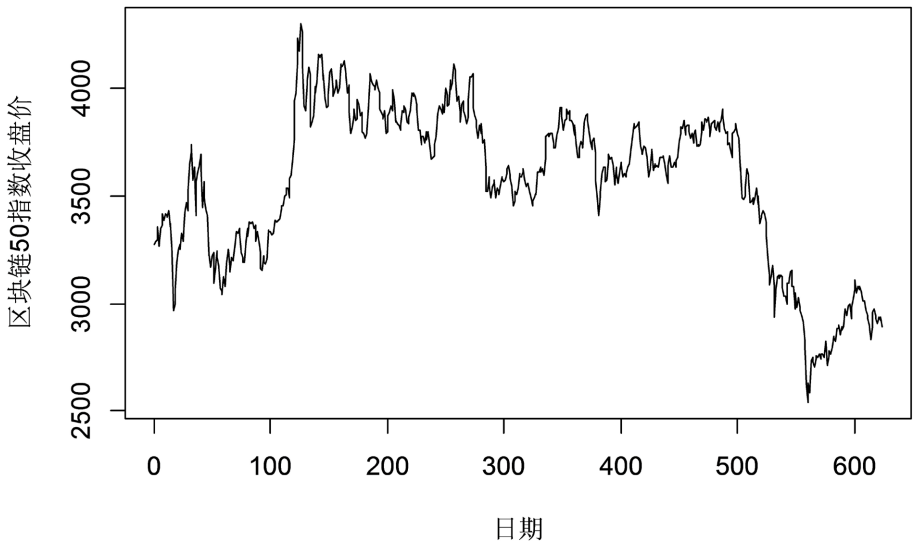

Figure 1. Blockchain 50 index chart
图 1. 区块链 50 指数走势图

2.2. 描述性统计分析

首先检验区块链 50 指数收盘价(Y)的分布，并进行描述性统计分析，便于把握该数据的总体特征。

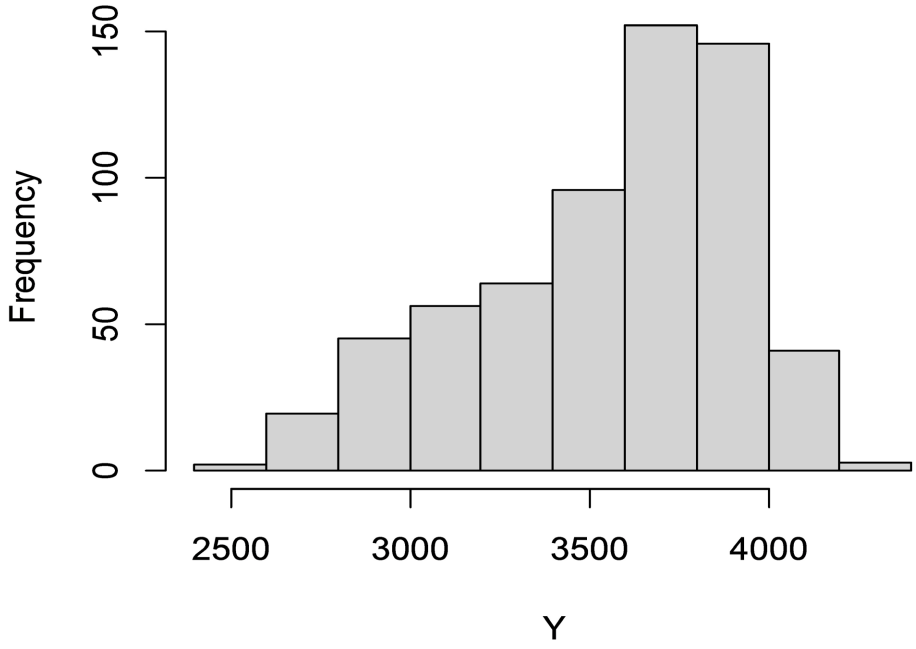

Figure 2. Block chain 50 index histogram
图 2. 区块链 50 指数直方图

由图 2 可知，区块链 50 指数收盘价的分布呈左偏、高峰的特征。表 3 给出了区块链 50 指数收盘价的描述性统计分析结果。

Table 3. Descriptive statistics of the blockchain 50 index closing price
表 3. 区块链 50 指数收盘价的描述性统计

目标指数	平均值	标准差	中位数	最小值	最大值	偏度	峰度
区块链 50 指数	3560.37	362.49	3635.55	2539.75	4304.62	−0.63	−0.44

3. 模型介绍

3.1. 最小二乘模型

对于线性模型：

$$\begin{cases} Y = X\beta + \varepsilon \\ E\varepsilon = 0, \operatorname{cov}(\varepsilon) = \sigma^2 I_n \end{cases} \tag{3.1}$$

来说，回归系数 β 的最小二乘估计为 $\hat{\beta}_{lse} = (X'X)^{-1} X'Y$。最小二乘估计是一个无偏估计，它对数据的分布假设没有要求，同时在无偏估计类中，最小二乘估计可得出残差平方和最小的回归模型，因此是回归分析中最为常用的方法之一。

在参数估计理论中，虽然最小二乘估计在所有的线性无偏估计中具有最小方差，但是当数据之间存在非常严重的多重共线性时，设计阵呈病态，此时其方差在线性无偏估计中最小，但是其值却很大，一般认为它不再是一个良好的估计。有偏估计是目前改善最小二乘估计的一种重要方法，它以牺牲估计量的无偏性代价来提高估计量稳定性[10]。

3.2. 岭估计

传统回归模型在变量间存在多重共线性时不再适用，最小二乘估计由于结构问题会导致估计的均方误差增大，此时考虑用有偏估计替代最小二乘估计。Horel 和 Kennard [11]在 1970 年提出岭估计，可解决条件极值问题获得

$$(y - X\beta)'(y - X\beta) + k\beta'\beta \tag{3.2}$$

其中，k 是拉格朗日乘数(Lagrangian Multipliers)，岭估计有如下表达式

$$\hat{\beta}(k) = (X'X + kI)^{-1} X'y \tag{3.3}$$

其中，$k \geq 0$ 是岭参数。通过对 k 值的选择，可以减少多重共线性的影响，取不同的 k 值，可以得到不同的估计，因此岭估计 $\hat{\beta}(k)$ 是一个估计类。当 $k = 0$，$\hat{\beta}(0) = (X'X)^{-1} X'y$ 就是常用的最小二乘估计[12]。

3.3. 绝对约束回归(Lasso)

Tibshirani [13]提出了一种解决高维变量选择的正则化方法——Lasso，该方法是在最小二乘估计基础上对回归数施加 L1 范数约束：

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.4}$$

其中，$\lambda > 0$ 为惩罚参数，取值越大，惩罚力度越强[14]。随着 λ 的增加，L1 惩罚项不但可以压缩回归系数趋于 0，而且当 λ 充分大时，可以使一些不重要的变量系数为 0，同时完成变量选择和参数估计[15]。因此，惩罚参数的选择至关重要，一般可通过 AIC 准则、BIC 准则、CV 交叉验证等准则选取[16]。

Efron [17]提出最小角回归(LARS)方法，这种方法既可以进行变量选择，可以用来解决 Lasso 问题，

并且可以提高计算效率。LARS 算法的基本思想是：首先选择一个与因变量相关性最大的协变量，然后沿这个方向走一定长度，知道出现第二个协变量，这两个协变量与残差的相关性相同，就沿着与这两个变量等角度的方向继续走，以此类推，选择出需要的协变量。LARS 算法的数学描述如下：由于 LARS 算法中，要选择多个变量等角度的方向，因此首先介绍如何来选择等角度的方向，设第 k 步时，前 k 个自变量被选择进来，记它们的集合为 A。由前 k-1 步得到的对响应变量的拟合为 u_{k-1}，定义矩阵

$$X_A = \left(S_1 X_1, S_2 X_2, \cdots, S_k X_k\right) \tag{3.5}$$

其中，$S_j = sign\left(\left(Y - u_{k-1}\right)' X_j\right)$。记 $G_A = X_A' X_A$，$C_A = \left(I_A' G_A^{-1} I_A\right)^{-\frac{1}{2}}$ 则下一步的搜索方向定义为

$$\mu_A = X_A \omega_A,\ \omega_A = C_A G_A' I_A \tag{3.6}$$

可以验证，它满足

$$X_A' \mu_A = C_A I_A,\ \left(\mu_A\right)^2 = 1.$$

因此，μ_A 是一个与所有已选入自变量方向成相同夹角的方向，在该方向上前进会导致残差与各自变量方向与各自变量内积等量递减[18] [19]。

3.4. 刘估计

1993 年 Liu Ke-jian [20] [21]借助岭回归的思想，对线性模型(4.1)，参数 β 的估计：

$$\beta(d) = \left(X'X + I\right)^{-1}\left(X'Y + d\hat{\beta}_{lse}\right) \tag{3.7}$$

为刘估计的待估回归系数，其中 $\hat{\beta}_{lse}$ 为最小二乘估计，$0 < d < 1$ 是参数。

3.5. 弹性约束估计

2005 年 Zou 与 Hastie [22]综合考虑岭回归和 Lasso 的约束方式，提出了弹性约束估计。弹性约束估计融合了 Lasso 估计和岭估计的特点，能处理高维数据，而且一般能挑选出相对于 Lasso 估计较少的变量[22]。弹性约束估计定义如下

$$\tilde{\beta} = \arg\min_{\beta}\left(\sum_{i}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right) \tag{3.8}$$

等价找到使

$$\sum_{i}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{3.9}$$

满足

$$(1-\lambda)\sum_{j=1}^{p}\left|\beta_j\right| + \lambda\sum_{j=1}^{p}\beta_j^2 \le t \tag{3.10}$$

达到最小的 $\beta_j, j = 1,2,3,\cdots,p$。

4. 实证分析

4.1. 最小二乘估计

首先用最小二乘估计建立区块链 50 指数及其成分股的线性回归方程，结果如表 4 所示。

Table 4. Least squares estimation regression results
表 4. 最小二乘估计回归结果

变量	成分股	Estimate	Std.Error	t-value	Pr(>|t|)	显著性
(Intercept)		−18.93860	30.63242	−0.618	0.536794	
X_1	平安银行	18.63554	0.91056	20.466	<2e − 16	***
X_2	TCL 科技	−1.30753	1.79949	−0.727	0.467927	
X_3	常山北明	4.96318	1.40860	3.523	0.000480	***
X_4	美的集团	3.97739	0.23642	16.823	<2e − 16	***
X_5	神州信息	10.51254	1.58910	6.615	1.32e − 10	***
X_6	智度股份	0.40455	1.74288	0.232	0.816577	
X_7	视觉中国	−0.28263	1.00863	−0.280	0.779477	
X_8	广发证券	2.31175	1.49571	1.546	0.123070	
X_9	新大陆	−3.40366	1.58981	−2.141	0.032943	*
X_{10}	传化智联	9.37521	2.78003	3.372	0.000825	***
X_{11}	远光软件	−1.04831	1.42894	−0.734	0.463647	
X_{12}	东华软件	12.46838	1.87658	6.644	1.11e − 10	***
X_{13}	恒宝股份	13.82414	2.74010	5.045	7.16e − 07	***
X_{14}	东港股份	5.22339	2.21317	2.360	0.018794	*
X_{15}	梦网科技	2.40458	0.84760	2.837	0.004809	**
X_{16}	利欧股份	28.17065	4.10101	6.869	2.79e − 11	***
X_{17}	广电运通	6.28846	1.72754	3.640	0.000312	***
X_{18}	二三四五	84.83666	10.57142	8.025	1.40e − 14	***
X_{19}	科大讯飞	1.47051	0.37950	3.875	0.000126	***
X_{20}	卫士通	5.68411	0.52546	10.817	<2e − 16	***
X_{21}	顺丰控股	3.39530	0.19705	17.231	<2e − 16	***
X_{22}	省广集团	8.83649	1.32320	6.678	9.04e − 11	***
X_{23}	广联达	2.58247	0.27996	9.225	<2e − 16	***
X_{24}	恺英网络	26.12347	3.25861	8.017	1.49e − 14	***
X_{25}	金财互联	−5.75406	1.48189	−3.883	0.000122	***
X_{26}	海联金汇	4.63717	2.34821	1.975	0.049048	*
X_{27}	巨人网络	11.66566	1.39395	8.369	1.26e − 15	***
X_{28}	世纪华通	−4.73219	1.66271	−2.846	0.004676	**
X_{29}	爱康科技	22.20637	3.91331	5.675	2.84e − 08	***
X_{30}	一心堂	0.56627	0.41191	1.375	0.170055	
X_{31}	汉威科技	4.32893	0.87786	4.931	1.24e − 06	***
X_{32}	同花顺	0.95274	0.13999	6.806	4.14e − 11	***
X_{33}	东方财富	12.50353	0.66244	18.875	<2e − 16	***

Continued

X_{34}	精准信息	−1.11621	1.51867	−0.735	0.462815	
X_{35}	顺网科技	2.89824	0.69815	4.151	4.12e − 05	***
X_{36}	新国都	−0.71194	1.60362	−0.444	0.657336	
X_{37}	汉得信息	11.33319	2.66815	4.248	2.75e − 05	***
X_{38}	易华录	0.86292	0.48338	1.785	0.075060	.
X_{39}	华宇软件	7.22730	0.78422	9.216	<2e − 16	***
X_{40}	润和软件	5.85516	0.40007	14.635	<2e − 16	***
X_{41}	北信源	7.21819	2.80058	2.577	0.010347	*
X_{42}	赢时胜	−5.86753	1.40887	−4.165	3.89e − 05	***
X_{43}	东方通	2.06954	0.28649	7.224	2.97e − 12	***
X_{44}	飞天诚信	0.83569	1.06822	0.782	0.434530	
X_{45}	高伟达	2.62079	1.23196	2.127	0.034063	*
X_{46}	四方精创	−0.03147	0.44981	−0.070	0.944261	
X_{47}	博思软件	−1.03096	0.42799	−2.409	0.016498	*
X_{48}	数字认证	−0.21611	0.42173	−0.512	0.608658	
X_{49}	科蓝软件	0.98143	0.42106	2.331	0.020304	*
X_{50}	华大基因	1.32094	0.12637	10.453	<2e − 16	***

Table 5. Model test table
表 5. 模型检验表

模型显著性检验 P 值	R^2	残差平方和	最大特征值	最小特征值	条件数
<2.2e − 16	0.9986	45592	23.1686	0.0053	4346.493

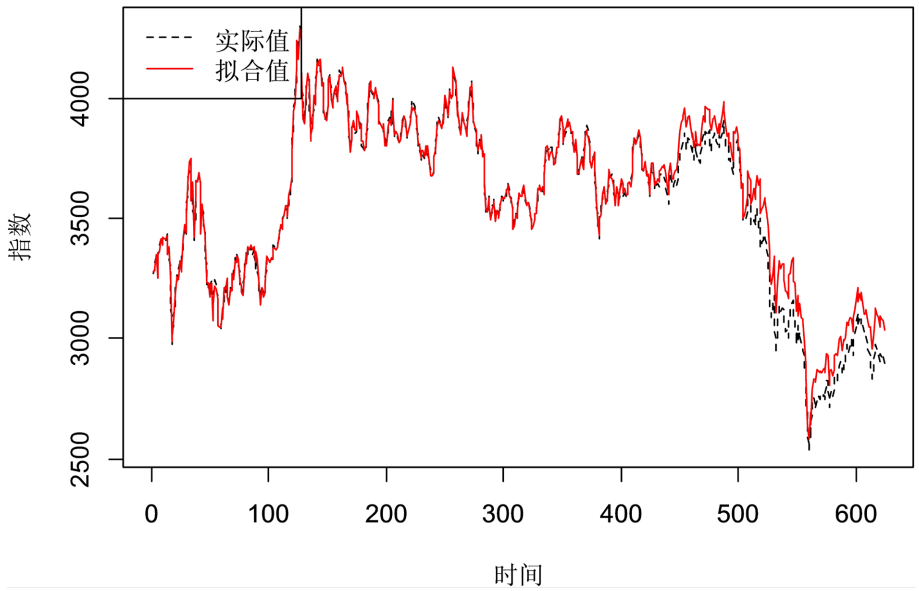

Figure 3. Blockchain 50 Index Tracking (olse)
图 3. 区块链 50 指数追踪(olse)

得到如下经验回归方程如下：

$$\hat{y} = -18.94 + 18.64x_1 - 1.31x_2 + 4.96x_3 + 3.98x_4 + 10.51x_5 + 0.40x_6 - 0.28x_7 + \cdots + 2.62x_{45} - 0.03x_{46} - 1.03x_{47} - 0.22x_{48} + 0.98x_{49} + 1.32x_{50} \quad (4.1)$$

TCL 科技、智度股份、视觉中国、广发证券、远光软件、一心堂、精准信息、新国都、飞天诚信、四方精创、数字认证这 11 只成分股的系数没有通过显著性检验，有 12 只成分股的系数为负数。

如表 5 所示的模型检验表明：R^2 为 0.9986，说明拟合效果很好，且模型通过显著性检验。预测指数跟踪如图 3 所示，指数走势跟实际指数的走势基本一致，说明通过回归模型跟踪区块链 50 指数的走势非常成功。但由于条件数为 4346.493，说明存在严重的多重共线性，因此需改进方法。

4.2. Cp 准则

4.2.1. Cp 准则下的岭估计

首先通过岭迹法选择参数 k，绘制岭迹图如图 4 所示。

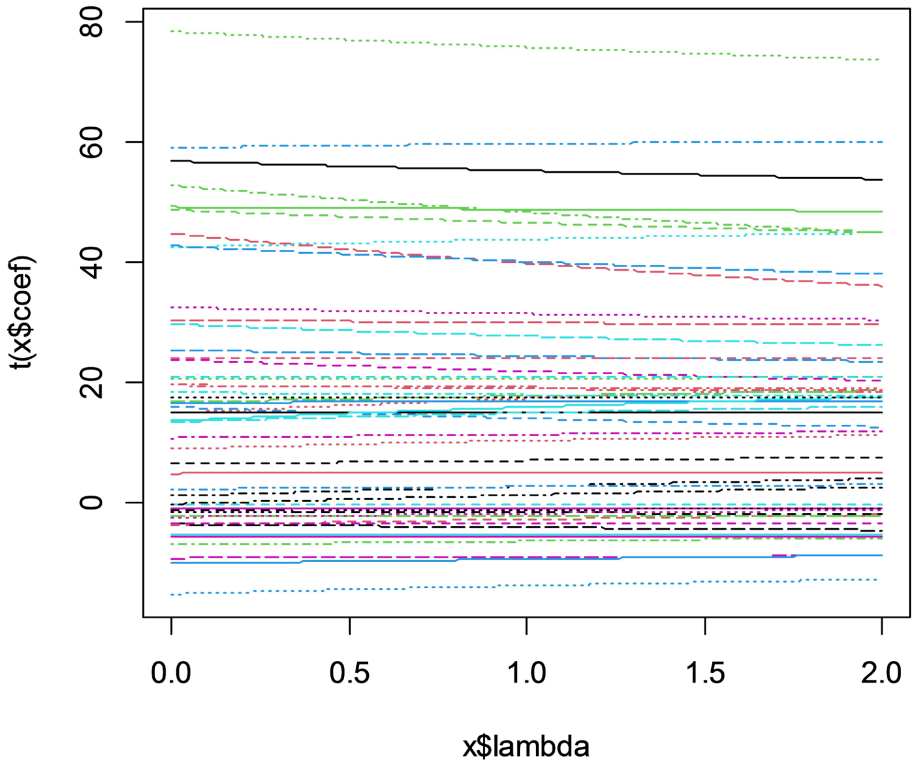

Figure 4. Ridge map
图 4. 岭迹图

Table 6. Ridge parameter value table
表 6. 岭参数取值表

岭参数	取值
ModifiedHKBestimator	0.2266915
ModifiedL-Westimator	0.06134589
SmallestvalueofGCV	0.405

结合图 4 和表 6 可知，选择最小的 k 值 0.06，得到岭回归方程：

$$\hat{y}=39.66+17.10x_1-2.40x_2+10.83x_3+4.51x_4+13.00x_5-0.56x_6-0.43x_7 \\ +\cdots-1.30x_{46}-0.03x_{47}-0.30x_{48}-0.01x_{49}+1.02x_{50} \tag{4.2}$$

具体系数见表 7：

Table 7. Ridge-estimated variable coefficients
表 7. 岭估计变量系数

变量	系数	变量	系数	变量	系数
截距项	39.6598	X_{17}	13.1572	X_{34}	1.7858
X_1	17.1006	X_{18}	72.4284	X_{35}	2.5800
X_2	−2.3955	X_{19}	0.1481	X_{36}	−2.1358
X_3	10.8281	X_{20}	3.6133	X_{37}	5.0385
X_4	4.5072	X_{21}	3.0759	X_{38}	0.8731
X_5	12.9969	X_{22}	8.5396	X_{39}	7.1460
X_6	−0.5622	X_{23}	3.6327	X_{40}	4.5601
X_7	−0.4330	X_{24}	11.5470	X_{41}	10.3707
X_8	5.2800	X_{25}	−1.3717	X_{42}	−2.0613
X_9	−3.9316	X_{26}	3.6434	X_{43}	1.7899
X_{10}	21.8649	X_{27}	11.9289	X_{44}	4.3490
X_{11}	−2.5043	X_{28}	−4.7293	X_{45}	−0.7980
X_{12}	8.8000	X_{29}	15.9100	X_{46}	−1.3008
X_{13}	−0.7897	X_{30}	−1.2858	X_{47}	−0.0304
X_{14}	10.1209	X_{31}	3.5333	X_{48}	−0.3004
X_{15}	5.2571	X_{32}	1.1599	X_{49}	−0.0105
X_{16}	23.7213	X_{33}	12.1031	X_{50}	1.0247

普通残差图如图 5 所示。由此可见岭估计给出的岭回归方程较好地刻画了资源 50 的趋势，如图 6 所示。

4.2.2. Cp 准则下的绝对约束估计(Lasso)

通过 LARS 进行变量选择，其系数图如图 7 所示。在 Cp 准则下，选择最小的 Cp 值对应的变量集。结果显示，最小值 Cp = 44.20989 对应的变量集包含 46 个变量，即通过变量选择，保留原始 46 个变量进行指数追踪。

对应的线性回归方程为

$$\hat{y}=18.36x_1-1.38x_2+5.45x_3+3.99x_4+\cdots-0.91x_{47}-0.28x_{48}+1.02x_{49}+1.33x_{50} \tag{4.3}$$

由表 8 可知，从回归系数上看，智度股份、视觉中国、新国都、四方精创这 4 只股票回归系数为 0，说明予以剔除是合理的，余下 46 个变量的最优子集。

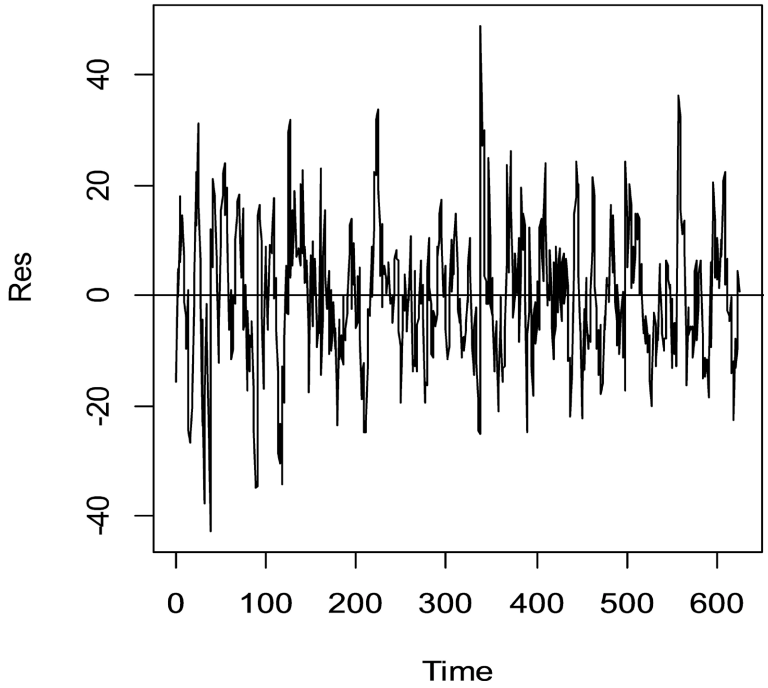

Figure 5. Ordinary residual plot of the ridge regression
图 5. 岭回归的普通残差图

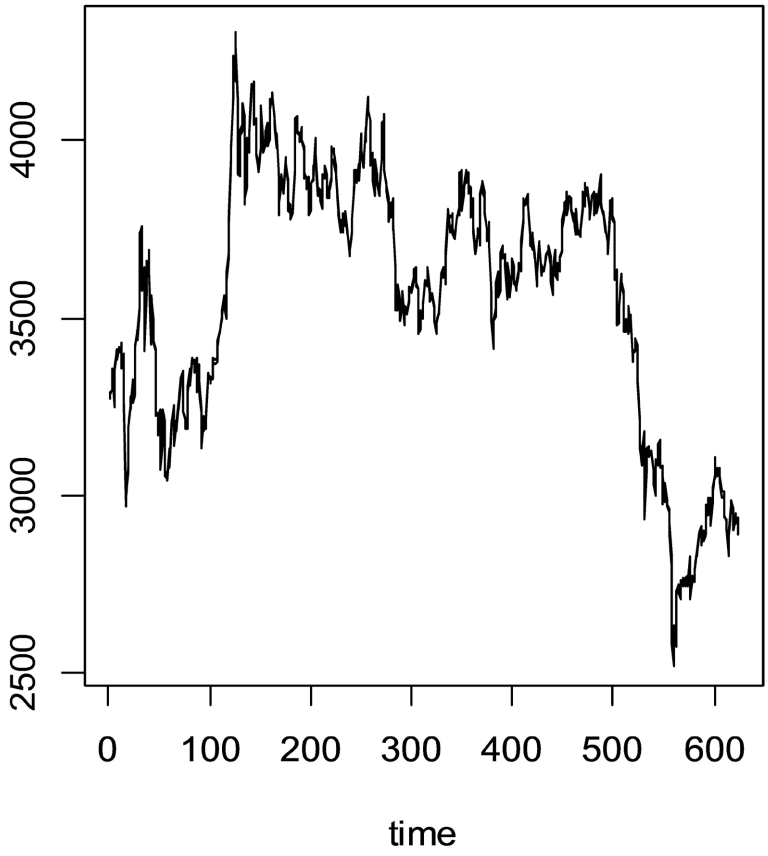

Figure 6. Fit plots of the dependent variable and predictive values
图 6. 因变量和预测值的拟合图

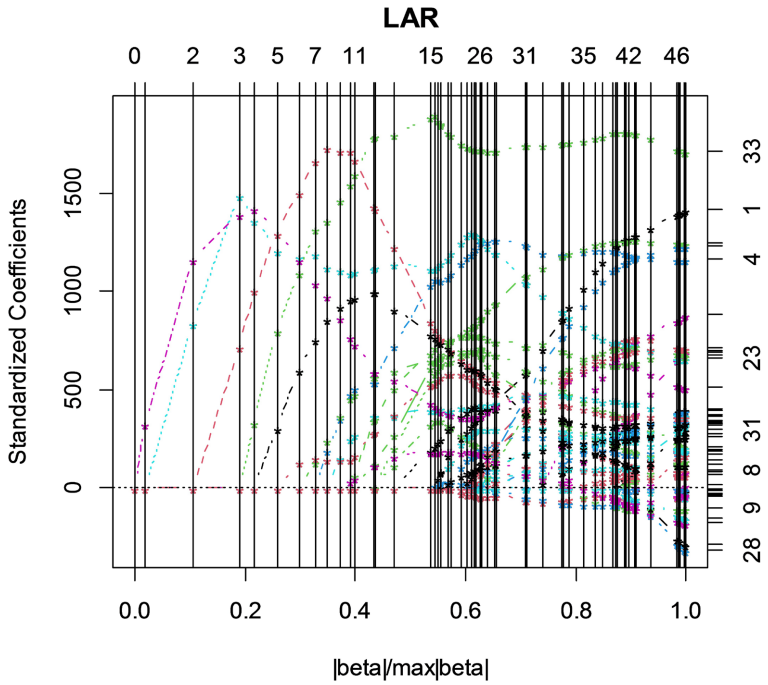

Figure 7. Selection coefficient plot of the LARS variables
图 7. LARS 变量选择系数图

Table 8. The Lasso parameter estimation table
表 8. Lasso 参数估计表

变量	系数	变量	系数	变量	系数
X_1	18.3597	X_{18}	81.7613	X_{35}	2.5350
X_2	−1.3841	X_{19}	1.3286	$\boldsymbol{X_{36}}$	**0.0000**
X_3	5.4526	X_{20}	5.7393	X_{37}	10.8152
X_4	3.9868	X_{21}	3.4184	X_{38}	0.7463
X_5	10.5789	X_{22}	8.6153	X_{39}	7.4298
$\boldsymbol{X_6}$	**0.0000**	X_{23}	2.5790	X_{40}	5.8273
$\boldsymbol{X_7}$	**0.0000**	X_{24}	27.0778	X_{41}	7.3843
X_8	2.4067	X_{25}	−5.1376	X_{42}	−5.4299
X_9	−3.5532	X_{26}	4.3028	X_{43}	2.0896
X_{10}	9.3258	X_{27}	11.4822	X_{44}	0.7904
X_{11}	−0.8940	X_{28}	−4.2282	X_{45}	2.4266
X_{12}	11.8088	X_{29}	22.3006	$\boldsymbol{X_{46}}$	**0.0000**
X_{13}	13.4136	X_{30}	0.4877	X_{47}	−0.9154
X_{14}	4.8950	X_{31}	4.2527	X_{48}	−0.2817
X_{15}	2.6697	X_{32}	0.9605	X_{49}	1.0159
X_{16}	27.8514	X_{33}	12.6538	X_{50}	1.3287
X_{17}	6.4466	X_{34}	−1.2952		

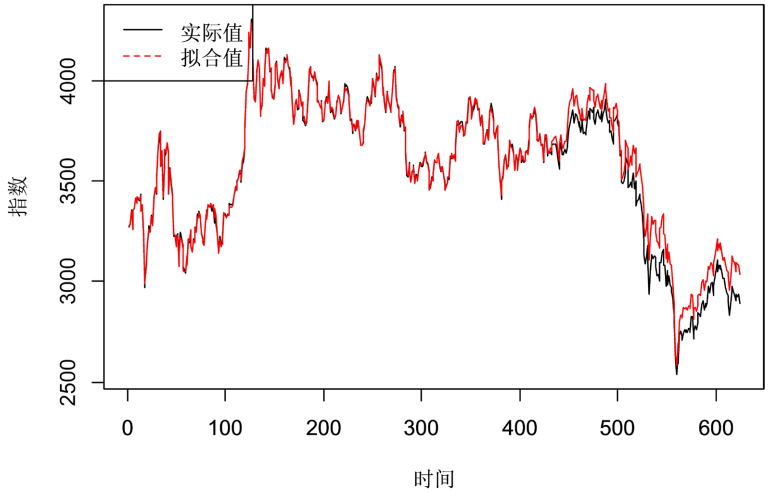

Figure 8. Blockchain 50 index tracking (Ridge Estimates)
图 8. 区块链 50 指数追踪(岭估计)

由图 8 可知，指数走势跟实际指数的走势基本一致，说明通过 Lasso 回归模型跟踪区块链 50 指数的走势较为成功。

4.3. 弹性约束估计

4.3.1. 岭估计交叉验证法

通过 CV 交叉验证，确定 $\lambda_{min} = 22.69$。由图 9、图 10 可知，保留变量个数是 50，其系数表如表 9 所示。

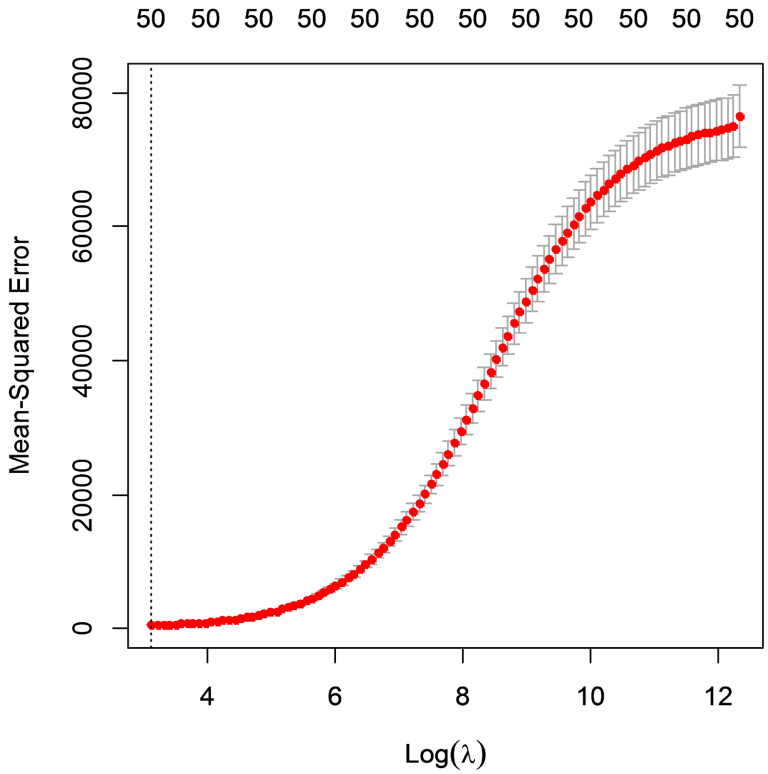

Figure 9. λ Select the graph
图 9. λ 选择图

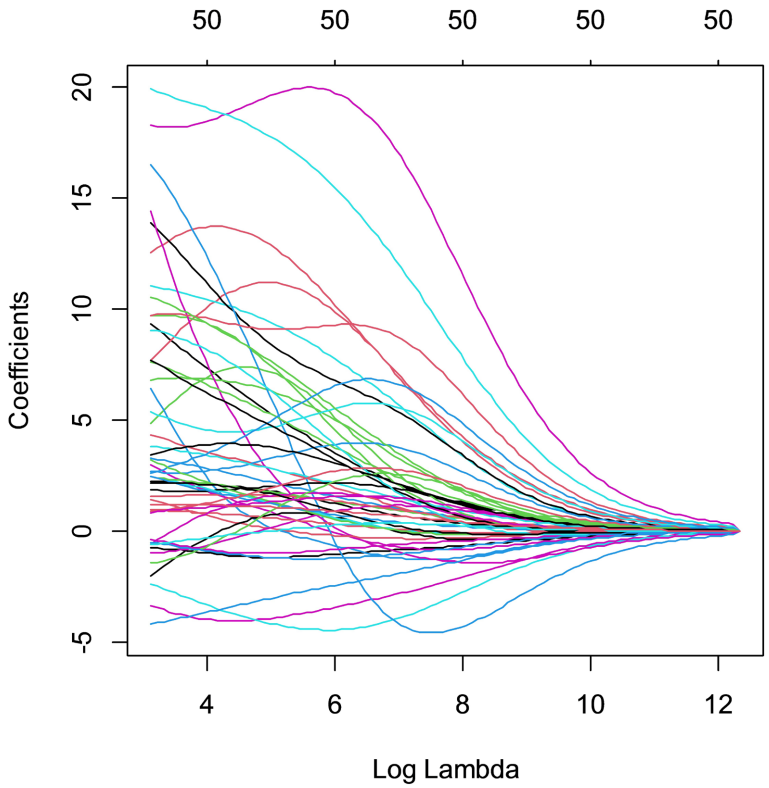

Figure 10. Ridge regression coefficient diagram
图 10. 岭回归系数图

Table 9. Ridge regression parameter estimation table (retain all variables)
表 9. 岭回归参数估计表(保留全部变量)

变量	系数	变量	系数	变量	系数
(Intercept)	1.07E − 12	X_{17}	3.59E + 01	X_{34}	4.44E + 01
X_1	2.41E + 01	X_{18}	4.99E + 00	X_{35}	1.91E + 00
X_2	−1.46E + 01	X_{19}	2.36E + 01	X_{36}	2.64E + 00
X_3	−6.05E + 00	X_{20}	1.00E + 01	X_{37}	2.61E + 01
X_4	−1.60E − 01	X_{21}	3.57E + 01	X_{38}	1.72E + 01
X_5	3.15E + 01	X_{22}	−1.86E + 01	X_{39}	3.40E + 00
X_6	−8.93E + 00	X_{23}	1.64E + 01	X_{40}	3.03E + 01
X_7	3.76E + 00	X_{24}	−7.37E + 00	X_{41}	1.17E + 02
X_8	1.01E + 01	X_{25}	5.66E + 00	X_{42}	2.13E + 00
X_9	2.32E + 00	X_{26}	1.53E + 01	X_{43}	2.63E + 00
X_{10}	3.34E + 01	X_{27}	3.72E + 00	X_{44}	9.65E + 00
X_{11}	−3.31E + 00	X_{28}	3.85E + 01	X_{45}	9.62E + 00
X_{12}	9.26E + 00	X_{29}	3.19E + 01	X_{46}	2.50E + 01
X_{13}	8.07E + 00	X_{30}	1.52E + 02	X_{47}	2.37E + 01
X_{14}	1.19E + 00	X_{31}	4.38E + 01	X_{48}	1.96E + 02
X_{15}	2.55E + 01	X_{32}	2.30E + 02	X_{49}	6.28E + 02
X_{16}	1.77E + 01	X_{33}	−2.88E + 00	X_{50}	5.49E + 00

从表 9 可以看出，与 Lasso 相比，岭估计得到的模型一直都是 50 个变量，因此岭估计没有变量筛选的功能。区块链 50 指数追踪图如图 11 所示，可知追踪效果较好。

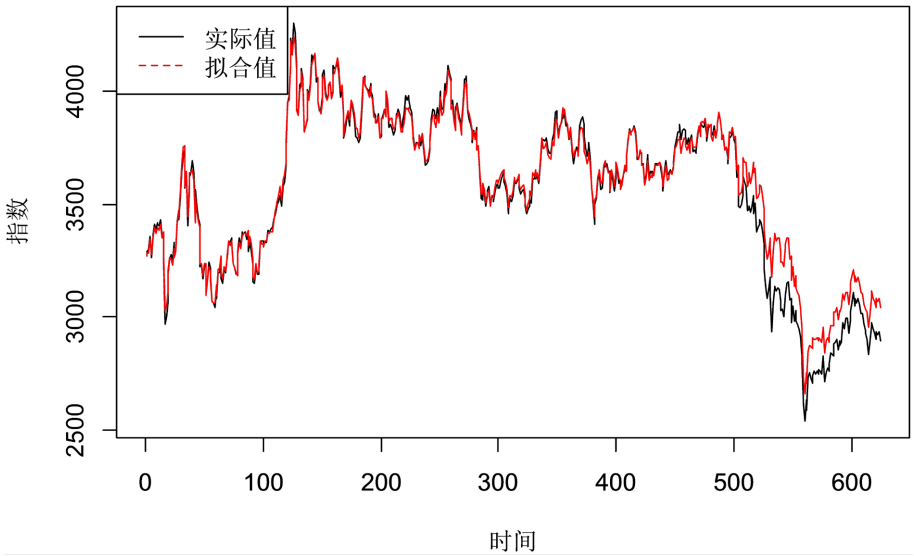

Figure 11. Blockchain 50 index tracking (Ridge estimation cross-validation method)
图 11. 区块链 50 指数追踪(岭估计交叉验证法)

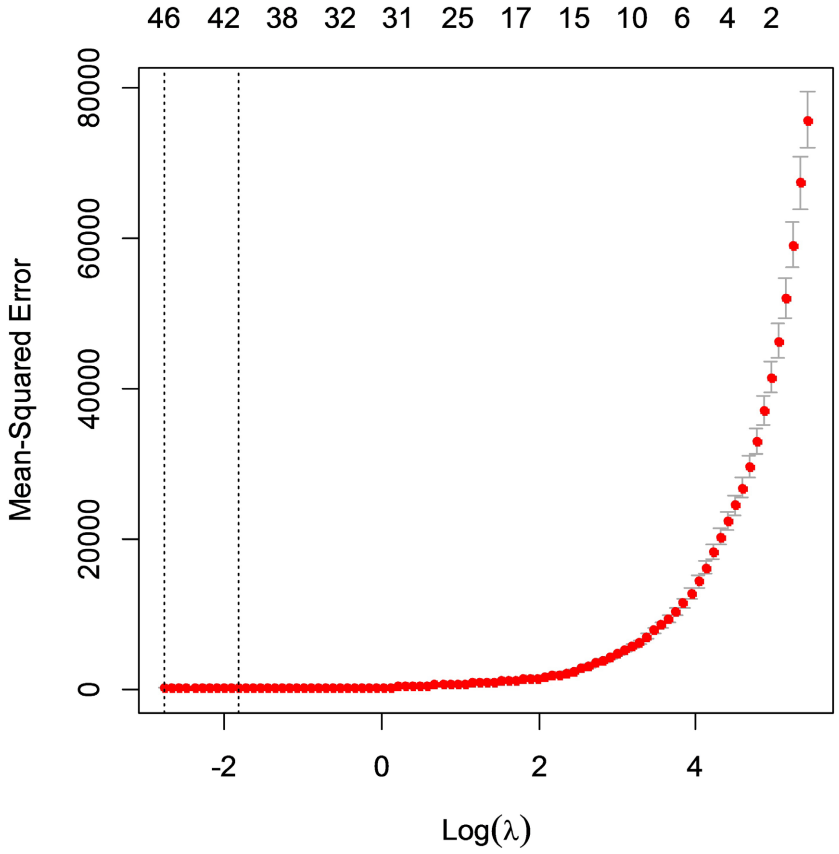

Figure 12. λ Select the graph
图 12. λ 选择图

4.3.2. Lasso 交叉验证法

通过 CV 交叉验证，确定 $\lambda_{min} = 0.06313$。由图 12、图 13 可知，保留变量个数是 46，其系数表如表 10 所示。

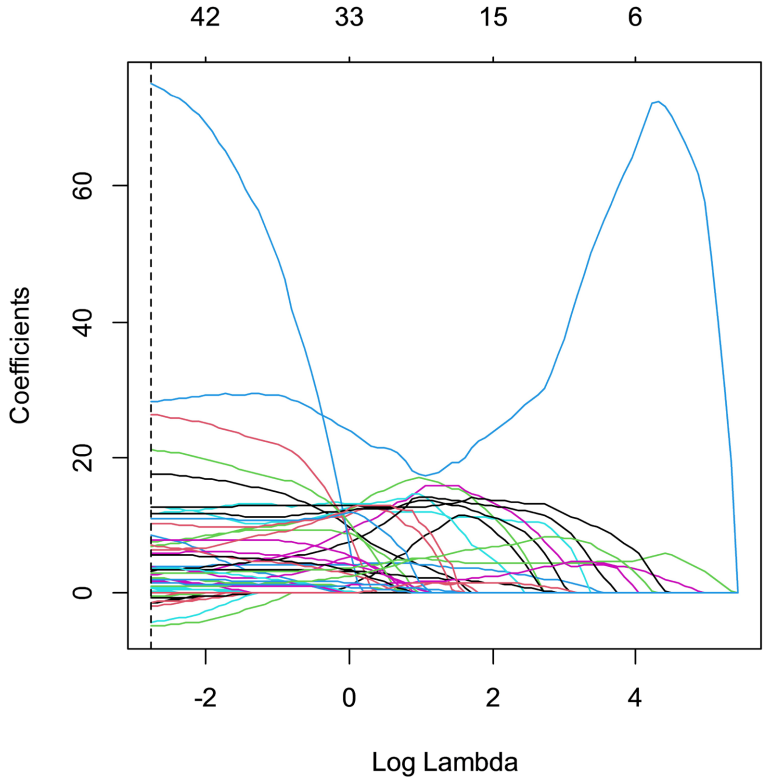

Figure 13. Lasso regression coefficient Fig
图 13. Lasso 回归系数图

Table 10. Lasso parameter estimation table (46 variables retained)
表 10. Lasso 参数估计表(保留 46 个变量)

变量	系数	变量	系数	变量	系数
截距项	1.42E − 13	X_{17}	7.53E + 00	X_{34}	−1.33E + 00
X_1	1.76E + 01	X_{18}	7.51E + 01	X_{35}	2.17E + 00
X_2	−5.54E − 01	X_{19}	8.81E − 01	X_{36}	.
X_3	6.93E + 00	X_{20}	5.94E + 00	X_{37}	1.11E + 01
X_4	4.00E + 00	X_{21}	3.39E + 00	X_{38}	2.34E − 01
X_5	1.18E + 01	X_{22}	7.01E + 00	X_{39}	7.68E + 00
$\boldsymbol{X_6}$	.	X_{23}	2.83E + 00	X_{40}	5.59E + 00
$\boldsymbol{X_7}$	.	X_{24}	2.82E + 01	X_{41}	6.40E + 00
X_8	2.70E + 00	X_{25}	−4.27E + 00	X_{42}	−4.95E + 00
X_9	−1.47E + 00	X_{26}	2.30E + 00	X_{43}	2.07E + 00
X_{10}	1.03E + 01	X_{27}	1.16E + 01	X_{44}	6.22E − 01
X_{11}	−5.20E − 01	X_{28}	−2.03E + 00	X_{45}	1.73E + 00

Continued

X_{12}	8.60E + 00	X_{29}	2.12E + 01	X_{46}	.
X_{13}	1.27E + 01	X_{30}	1.22E − 01	X_{47}	−7.51E − 01
X_{14}	3.80E + 00	X_{31}	3.78E + 00	X_{48}	−4.16E − 02
X_{15}	3.28E + 00	X_{32}	9.95E − 01	X_{49}	1.03E + 00
X_{16}	2.63E + 01	X_{33}	1.25E + 01	X_{50}	1.36E + 00

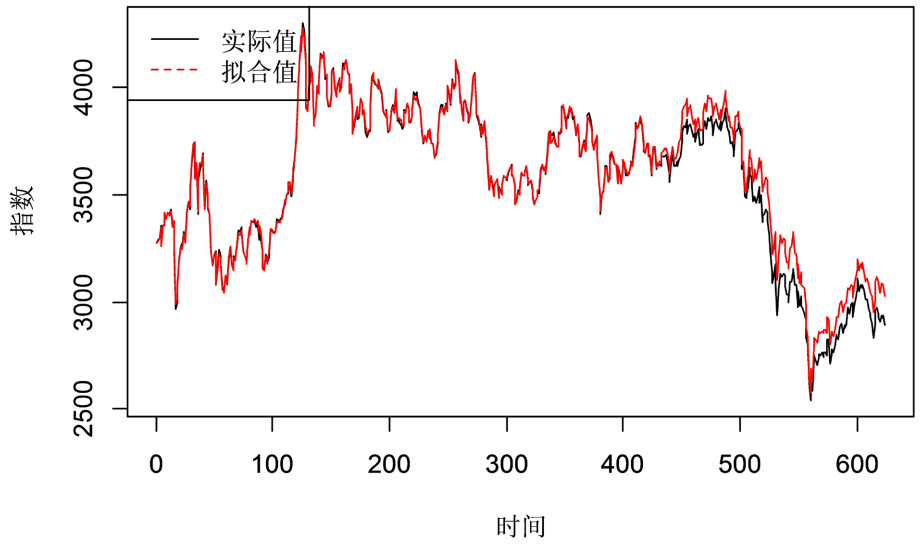

Figure 14. Blockchain 50 index tracking (Lasso cross-validation method)
图 14. 区块链 50 指数追踪(Lasso 交叉验证法)

由图 14 可知，指数走势跟实际指数的走势基本一致，说明通过 Lasso 交叉验证的弹性约束估计回归模型跟踪资源 50 指数的走势较成功。残差平方和为 47,983.53。

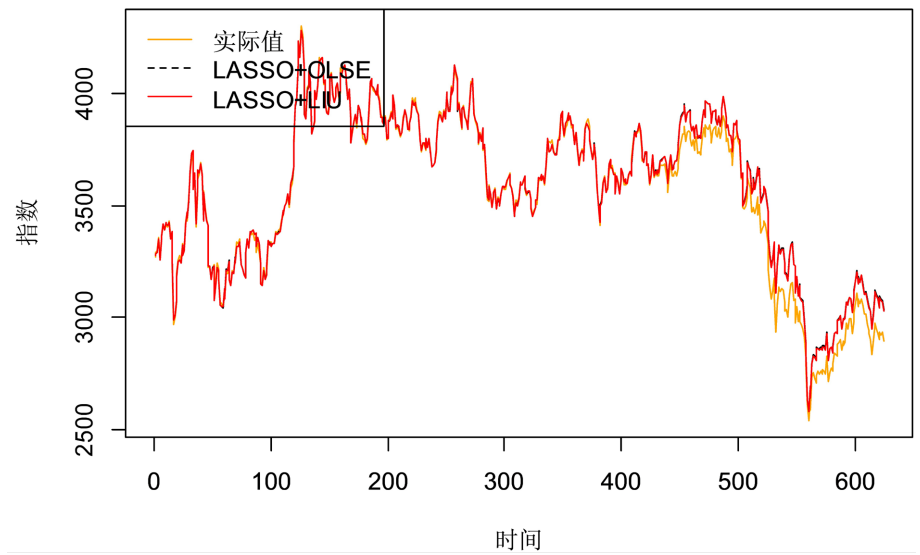

Figure 15. Blockchain 50 index tracking (two-step estimate)
图 15. 区块链 50 指数追踪(两步估计)

4.4. 两步估计

现采用两步估计的方法，由 Lasso 筛选出 46 只成分股，再通过最小二乘估计、岭估计、刘估计等回归方法建立模型，从而进行指数追踪，结果如图 15 所示，可以看出 LASSO + OLSE 和 LASSO + LIU 这两个组合的预测收盘价都能很好地跟踪到区块链 50 指数收盘价的运行趋势。

5. 结论

将上述方法对区块链 50 指数进行追踪的效果进行对比，如表 11 所示。在 Cp 准则下，Lasso 在测试集上的残差标准差(SD)优于岭估计，但在测试集上的平均残差平方和(RMS)不如岭估计；结合残差图(图 3)来看，可以认为 Cp 准则下的岭估计优于 Lasso；在 CV 准则下，Lasso 在测试集上的平均残差平方和(RMS)和残差标准差(SD)两种指标优于岭估计；在两步估计(Lasso 变量选择)方法下，进一步运用刘估计进行回归，即 Lasso + Liu，测试集上的平均残差平方和(RMS)和残差标准差(SD)两种指标优于 Lasso + OLSE、Lasso + 岭估计，得到较好的外预测效果。

Table 11. Track acking effect of each method
表 11. 各方法追踪效果对比

方法		训练集		测试集	
		RMS	SD	RMS	SD
Cp 准则	Lasso	123.965	10.49877	13897.34	47.19733
	岭估计	290.4769	10.4829	**13687.34**	**47.88227**
CV 准则	Lasso	129.3357	10.75281	**12348.29**	**44.19456**
	岭估计	508.5015	21.14796	17101.87	79.89194
两步估计	Lasso + OLSE	123.6443	10.48519	13740.74	47.82017
	Lasso + 岭估计	283.4467	10.48636	13595.95	47.84592
	Lasso + Liu	124.2304	10.51001	**13161.81**	**46.89458**

本文以区块链 50 指数及其成分股的日线收盘价数据为研究对象，不断修正回归模型，得到了效果较好的区块链 50 指数回归模型，对投资者有一定的参考价值。但由于数据、估计方法的一定的改进空间，还应结合市场特点对股票指数趋势进行分析。

参考文献

[1] 杨楠. 岭回归分析在解决多重共线性问题中的独特作用[J]. 统计与决策, 2004(3): 14-15.

[2] 薛宏刚, 张锐敏, 胡春萍, 李乃成. 基于岭回归的套期保值方法[J]. 统计与决策, 2012(5): 77-79.

[3] 张家茂, 杨思思. 房地产股价线性模型的变量选择实证研究[J]. 重庆工商大学学报(自然科学版), 2017, 34(4): 35-40.

[4] 张慧伟. 基于弹性估计筛选部分成分股追踪股指变化[J]. 广西质量监督导报, 2018(12): 83-84.

[5] 杨思思. 中证 100 股票指数回归模型的实证分析[J]. 重庆文理学院学报(社会科学版), 2018, 37(2): 121-126.

[6] 王琪, 冷林峰, 常永莲. 改进岭回归与主成分回归的股指跟踪研究[J]. 重庆理工大学学报(自然科学), 2018, 32(1): 212-221.

[7] Ranstam, J. and Cook, J.A. (2018) LASSO Regression. *Journal of British Surgery*, **105**, 1348-1348. https://doi.org/10.1002/bjs.10895

[8] 韩笑, 滕兴虎, 窦婷. 基于银行类指数及其成分股的分析和预测[J]. 统计学与应用, 2020, 9(4): 506-514.

[9] 深圳证券交易所. 关于发布深证区块链 50 指数的公告[EB/OL].

http://www.szse.cn/disclosure/notice/general/t20191224_572813.html, 2019-12-24.

[10] 太思梦. 两类改进 LIU 估计在股指追踪中的应用[D]: [硕士学位论文]. 重庆: 重庆大学, 2019.

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

[12] 杨虎, 杨玥含. 金融大数据统计方法与实证[M]. 北京: 科学出版社, 2016: 122-123.

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

[14] 蒋翠侠, 刘玉叶, 许启发. 基于 LASSO 分位数回归的对冲基金投资策略研究[J]. 管理科学学报, 2016, 19(3): 107-126.

[15] 张靖, 胡学钢, 李培培, 张玉红. 基于迭代 Lasso 的肿瘤分类信息基因选择方法研究[J]. 模式识别与人工智能, 2014, 27(1): 49-59. https://doi.org/10.16451/j.cnki.issn1003-6059.2014.01.001

[16] 彭胜银. 基于 Lasso 分位数的非负两阶段方法及在标普 500 指数追踪的应用[D]: [硕士学位论文]. 重庆: 重庆大学, 2019.

[17] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[18] 梁斌, 陈敏, 缪柏其, 黄意球, 陈钊. 基于 LARS-Lasso 的指数跟踪及其在股指期货套利策略中的应用[J]. 数理统计与管理, 2011, 30(6): 1104-1113.

[19] 梁斌. 股指期货套期保值和套利策略研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2010.

[20] Liu, K. (1993) A New Class of Blased Estimate in Linear Regression. *Communications in Statistics—Theory and Methods*, **22**, 393-402. https://doi.org/10.1080/03610929308831027

[21] Liu, K.J. (2003) Using Liu-Type Estimator to Combat Collinearity. *Communications in Statistics Theory & Methods*, **32**, 1009-1020. https://doi.org/10.1081/STA-120019959

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