doi: 10.16862/j.cnki.issn1674-3873.2020.01.007

基于岭回归和 Lasso 回归的螺纹钢期货价格实证分析

王纯杰, 温丽男, 马元嘉

(长春工业大学 数学与统计学院, 吉林 长春 130000)

摘　要: 期货是金融市场的重要组成部分, 期货的交割通常是在一段时间后进行, 因此对于期货价格预测显得尤为重要, 其中螺纹钢期货价格预测成为提高我国钢铁产业竞争力的重要举措. 通过对螺纹钢期货结算价及其影响因素的数据进行分析, 分别利用岭回归和 Lasso 回归两种方法消除共线性的影响, 得到两种修正多元回归模型, 应用两种修正多元回归模型分别对螺纹钢未来一周的期货价格进行预测, 并与真实价格进行比较, 最终发现基于两种方法得到的模型预测准确率均高于 95% 以上, 且基于 Lasso 回归方法的拟合效果更好, 证明构建的两种回归模型对螺纹钢价格的走势与预测均有重要的参考价值.

岭回归和Lasso回归

关键词: 螺纹钢期货价格; 共线性; 岭回归; Lasso 回归

中图分类号: O212.4　　**文献标志码**: A　　**文章编号**: 1674-3873-(2020) 01-0036-06

0　引言

近年来, 国内对螺纹钢期货价格的预测与分析依然是热点, 很多学者通过不同的方法对螺纹钢期货价格进行预测与分析. 方雯等[1] 利用共同因子模型分析了国内外钢材期货市场价格规律, 结果表明, 国内期货市场的钢材价格更好的反映出市场价格波动. 靳朝翔等[2] 对焦炭、铁矿石、螺纹钢的期货价格利用单位根检验和协整的方法发现三者的期货价格相互影响. 王珂[3] 利用 BP 神经网络对螺纹钢期货价格进行预测并与小波神经网络预测结果进行对比, 证明 BP 神经网络的期货价格预测准确度高于小波神经网络模型预测的结果. 石宝峰等[4] 将剔除残差相关性的最小二乘算法引用到向量误差修正模型中, 构建了基于 PT 和 IS 共同因子的螺纹钢期货价格预测模型.

就是自己设置对照实验

作为期货市场里的黑色系商品, 螺纹钢、铁矿石、焦煤、焦炭和热轧卷板是炼钢的必要资源[5], 另外考虑到期货市场常受人民币兑美元中间价的影响[6], 故将铁矿石、焦煤、焦炭和热轧卷板及人民币兑美元中间价作为影响螺纹钢期货价格的主要因素.

本文首先阐述了多元线性回归模型的基本理论, 通过岭回归和 Lasso 回归方法消除自变量间存在共线性的问题, 得到岭回归和 Lasso 回归模型, 最后对数据进行预测并分析.

1　多元线性回归模型

多元线性回归模型探究了相关自变量直接影响因变量的问题. 模型[7] 为

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon. \tag{1}$$

其中: y 是因变量, $x_j(j=1,2,\cdots,p)$ 是 p 维自变量, $\beta_j(j=0,1,2,\cdots,p)$ 是未知参数, ε 是随机误差项.

收稿日期: 2019-12-16
基金项目: 吉林省"十三五"科学技术研究规划项目(2016315)
第一作者简介: 王纯杰(1978—) , 女, 辽宁省灯塔市人, 教授, 博士, 博士生导师. 研究方向: 生存分析与数理统计.

2　模型的建立

2.1　数据来源

本文数据源于大连商品交易所、上海期货交易所和国家外汇管理局 2018 年 1 月 16 日至 2018 年 12 月 28 日的期货交易代码为 1901 的黑色系品种的商品期货以及人民币对美元汇率中间价,共有数据 233 条,并选取 2018 年 12 月 20 日至 2018 年 12 月 28 日的 7 条数据验证模型的精确度.

文中因变量为螺纹钢结算价,自变量为 X_1—X_5,具体解释见表 1.

表 1　变量介绍

Table 1　Variable introduction

变量	含义
Y	螺纹钢(RB1901) 结算价(元/吨)
X_1	焦煤(JM1901) 结算价(元/吨)
X_2	焦炭(J1901) 结算价(元/吨)
X_3	铁矿石(I1901) 结算价(元/吨)
X_4	热扎卷板(HC1901) 结算价(元/吨)
X_5	人民币对美元汇率中间价(元/吨)

2.2　建立多元线性回归模型

2.2.1　螺纹钢结算价与各影响因素的相关性

首先,判断各影响因素是否对螺纹钢结算价存在线性关系. 表 2 为螺纹钢与各影响因素相关系数的情况.

表 2　螺纹钢与各影响因素的 Pearson 相关系数

Table 2　Pearson correlation coefficient between rebar and various influencing factors

Y	X_2	X_4	X_5	X_1	X_3
r	0. 909 77	0. 892 75	0. 768 80	0. 607 60	0. 597 55
P	<0. 000 1	<0. 000 10	<0. 000 10	<0. 000 10	<0. 000 10

Pearson 相关系数常用来衡量两个变量间线性相关强弱,用 r 表示,绝对值越大相关性越强. 从表 2 中可以看出,Y 与各变量之间都存在很强的相关关系,且 X_2 与 X_4 对其线性关系最强,接近 0. 9,其余变量对 Y 均有统计学意义,故将这 5 个变量引入模型中并进行相关分析.

2.2.2　建立多元回归方程

本文结合 SAS9. 4 版和 R3. 5. 3 版软件共同对所选变量进行统计分析,并得到以下分析结果. 首先对多元回归模型做方差分析,即 F 检验. 从表 3 可以获得 F 值为 935. 14,P 值小于显著性水平 0. 05,表明回归方程拟合效果显著. 接下来,进行了拟合优度检验,从表 4 中可以看出,回归模型的拟合度 R^2 达到 95. 4%,表示该数据适用于回归模型.

表 3　方差分析

Table 3　Variance analysis

	自由度	平方和	均方	F 值	Pr > F
模型	5	17 982 517	3 596 503	935. 14	<0. 000 1
误差	220	846 106	3 845. 936 61		
校正合计	225	18 828 623			

调整R2是为什么

表4 模型1拟合优度检验

Table 4 Model 1 goodness of fit test

R^2	调整后 R^2	均方根误差	因变量均值	变异系数
0.955 1	0.954 0	62.015 62	3 747.415 93	1.654 89

表5 模型1参数估计

Table 5 Model 1 parameter estimation

变量	参数估计	标准误差	t 值	$Pr > \|t\|$	容差	方差膨胀因子
Intercept	−2 141.590 9	267.343 8	−8.01	<0.000 1	—	0
X_1	−0.094 0	0.141 2	−0.67	0.506 3	0.108 8	9.192 0
X_2	0.233 5	0.080 4	2.90	0.004 1	0.046 3	21.578 3
X_3	1.825 5	0.274 5	6.65	<0.000 1	0.272 3	3.672 6
X_4	0.696 0	0.047 9	14.53	<0.000 1	0.148 8	6.719 2
X_5	3.025 0	0.372 2	8.13	<0.000 1	0.184 6	5.416 6

表5是通过回归模型得到的参数估计结果,在显著性水平0.05的情况下,X_1 变量的参数结果并不显著($P=0.506\ 3$),但在实际意义中,X_1(焦煤(JM1901)结算价)是影响 Y(螺纹钢(RB1901)结算价)的重要影响因素,在理论方面,可以通过Pearson相关系数检验.以上结论说明模型可能失真,考虑的主要原因是自变量间可能存在共线性的情况,接下来进行共线性诊断.

2.2.3 共线性诊断

R. Frosch[8]首次提出多重共线性的概念,即自变量间高度相关.当存在多重共线性时,常会降低模型参数估计的精度,导致参数估计的含义不合理,模型拟合达不到良好的效果.目前,传统的共线性的判定方法多采用如下几种:

(1)特征值法.比对第3和第4维度是否趋近于0.

(2)条件指数法.某特定维数条件指数是否大于30.

(3)方差比例法.是否存在方差比例趋于1的值.

(4)方差膨胀因子(VIF).方差膨胀因子值是否大于10.

(5)逐步回归方法.将引起多重共线性的自变量通过逐步回归法排除.

目前,多数学者在最小二乘的基础上引入岭回归方法和Lasso回归方法,分别运用损失部分信息、降低精度为手段和构造惩罚函数压缩回归系数的方法建立更为精炼、更具现实意义的回归模型.

表6 共线性诊断

Table 6 Collinear diagnosis

维数	特征值	条件指数	方差比例					
			Intercept	X_1	X_2	X_3	X_4	X_5
1	5.987 8	1.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
2	0.006 9	29.530 5	0.009 9	0.000 6	0.053 8	0.006 4	0.000 0	0.000 7
3	0.003 3	42.693 9	0.002 7	0.065 6	0.002 1	0.049 3	0.057 4	0.001 8
4	0.001 6	60.894 6	0.013 1	0.011 4	0.003 1	0.222 4	0.071 9	0.045 8
5	0.000 4	129.330 7	0.028 4	0.479 2	0.026 9	0.563 9	0.184 9	0.221 1
6	0.000 1	250.378 1	0.945 9	0.443 3	0.914 2	0.157 9	0.685 8	0.730 6

通过观察表6,根据共线性的判断准则,特征值在第4维时接近0,为0.001 61,条件指数最大高为250.378 14,且方差比例存在接近1的数值(0.945 88)以及表5中的方差膨胀因子(VIF)检验值最大高为21.578 31,均表明自变量间存在严重多重共线性,故需要解决这个问题.本文选取岭回归方法对自变量进行选择.

可以学习这里的岭回归模型的构建工作

2.3　通过岭回归分析建立线性模型

A. E. Hoerl 在 1962 年首次提出岭回归方法,用以控制与最小二乘估计相关的方差膨胀性和产生的不稳定性,A. E. Hoerl 和 R. W. Kennard[9] 对岭回归给出了具体的分析与证明. G. C. Mcdonald[10] 提供了岭回归方法的简要概述,证明了岭回归的相关性质. 岭回归[11] 是在普通最小二乘的参数估计 $\hat{\beta}=(X'X)^{-1}X'Y$ 中引入一个矩阵 $k\boldsymbol{I}$($k>0$,$\boldsymbol{I}$ 为单位矩阵),得到 $\hat{\beta}(k)=(X'X+k\boldsymbol{I})^{-1}X'Y$,这里 $\hat{\beta}(k)$ 为岭回归估计,k 为岭参数. 由于岭参数不唯一,故得到的 $\hat{\beta}(k)$ 是参数 β 的估计族. 本文结合回归系数的岭迹图、岭参数 k 值的基本原则[11] 以及 R 软件 ridge 包中的 linearRidge( ) 函数确定最恰当的岭参数 k.

我想到了，随机森林有文献说明可以缩小自变量之间的相关性，
也就是说尽管自变量之间是具有相关性的，但是并不会影响结果
而Lasso回归是用来选择特征的，就是特征之间是有相关性的，而通过Lasso回归之后
可以更好地来构建模型

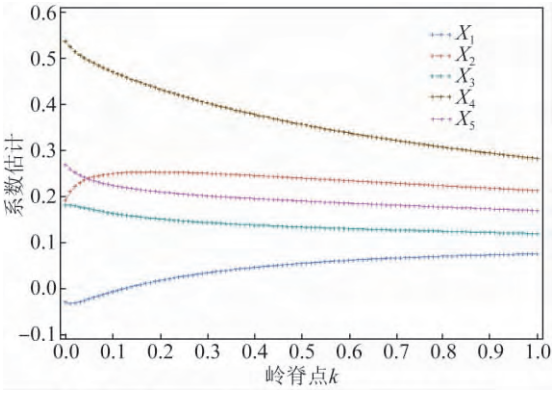

图 1　岭迹图

Fig. 1　Ridge map

表 7　选取不同岭回归系数的变化

Table 7　Select changes in regression coefficients of different ridges

k 值	均方根误差	X_1	X_2	X_3	X_4	X_5
0.0	0.214 38	−0.028 843	0.192 85	0.182 17	0.538 17	0.270 34
0.1	0.221 52	−0.005 935	0.250 39	0.164 21	0.471 36	0.224 6
0.2	0.235 37	0.018 246	0.253 75	0.151 81	0.433 83	0.210 48
0.3	0.251 74	0.034 776	0.250 86	0.144 10	0.403 80	0.202 18
0.4	0.268 88	0.046 603	0.246 06	0.138 65	0.378 72	0.195 98
0.5	0.285 99	0.055 345	0.240 60	0.134 40	0.357 32	0.190 76
0.6	0.302 70	0.061 940	0.234 96	0.130 87	0.338 76	0.186 07
0.7	0.318 86	0.066 975	0.229 35	0.127 78	0.322 45	0.181 73
0.8	0.334 40	0.070 842	0.223 88	0.125 00	0.307 97	0.177 64
0.9	0.349 31	0.073 816	0.218 58	0.122 44	0.295 00	0.173 75
1.0	0.363 60	0.076 092	0.213 48	0.120 04	0.283 28	0.170 04

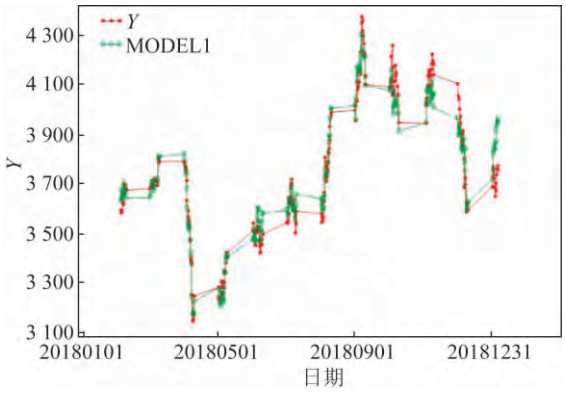

图 2　岭回归拟合值与实际值对比

Fig. 2　Ridge regression fitting value and actual value comparison

通过 linearRidge( ) 函数给出的结果来看,岭回归参数为 0.001 597 383,此时变量 X_1 的 P 值仍不显著(P =0.474 2) ,故剔除 X_1 变量,得到的岭回归参数均显著,得到的岭回归模型为

$$Y = -2236.64297 + 0.19337X_2 + 1.68306X_3 + 0.72021X_4 + 3.08815X_5. \quad (2)$$

2.4 通过 Lasso 回归分析建立线性模型

Lasso 回归最早由 R. Tibshirani[12] 提出,此后广泛应用在变量选择和参数估计中,基本思想是对参数进行压缩,进而选择重要的变量,定义为

$$\hat{\beta}(\text{Lasso}) = \underset{\beta}{\operatorname{argmin}} \left\| y - \sum_{i=1}^{p} x_i\beta_i \right\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|.$$

其中 λ 为非负正则参数, $\lambda \sum_{i=1}^{p} |\beta_i|$ 为惩罚项.

CP 统计量是选择最优子集的一种方法,其值越小表示所选子集个数最优. 表 8 为 CP 统计量的变化值,可以看出,当变量个数为 4 时,CP 值达到最小,故选择 4 个变量进入模型,这 4 个变量分别为 X_2、X_3、X_4 和 X_5,对应的回归参数估计值如表 9 所示,故得到 Lasso 回归模型为

$$Y = -2211.2475 + 0.1960X_2 + 1.6699X_3 + 0.7181X_4 + 3.0632X_5. \quad (3)$$

表 8 CP 变化值

Table 8 CP change value

个数	0	1	2	3	4	5
CP 值	4 671.719	4 074.617	427.134	402.254	4.469	6.000

表 9 Lasso 回归参数估计

Table 9 Lasso regression parameter estimation

参数	Intercept	X_2	X_3	X_4	X_5
估计值	-2 211.247 533	0.195 937	1.669 913	0.718 099	3.063 214

图 3 表示通过 Lasso 回归模型得到的参数值与实际值的拟合图,回归拟合值和实际值的走势大致相同,拟合程度较好,证明有实际应用价值. 表 10 是模型(1) 和(2) 对 2018 年 12 月 20 日至 2018 年 12 月 28 日的螺纹钢的价格进行预测估计的情况,预测的准确率均高达 95% 以上,再次验证了模型的有效性.

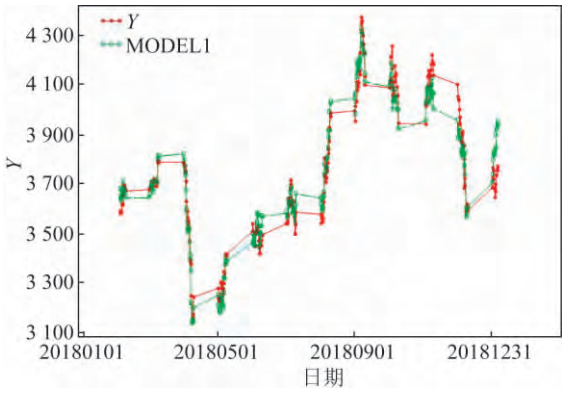

图 3 Lasso 回归拟合值与实际值对比

Fig. 3 Lasso regression fit and actual value comparison

表 10　螺纹钢价格的预测值与真实值

Table10　Predicted value and true value of rebar price

日期(年-月-日)	Lasso 预测值/(元/吨)	岭回归预测值/(元/吨)	实际值/(元/吨)	Lasso 准确率	岭回归准确率
2018-12-20	3 963.406 6	3 964.019	3 782	0.954 2	0.954 1
2018-12-21	3 980.938 6	3 981.644	3 829	0.961 8	0.961 7
2018-12-24	3 955.735 6	3 956.379	3 816	0.964 7	0.964 5
2018-12-25	3 930.710 5	3 931.303	3 805	0.968 0	0.967 9
2018-12-26	3 913.724 7	3 914.377	3 836	0.980 1	0.979 9
2018-12-27	3 925.452 1	3 925.995	3 893	0.991 7	0.991 6
2018-12-28	3 903.271 3	3 903.804	3 870	0.991 5	0.991 3

4　结论

本文通过分析螺纹钢期货价格及其影响因素,运用岭回归和 Lasso 回归分别对自变量间存在多重共线性进行处理,将变量 X_1(焦煤结算价) 剔除在模型外,再利用最小二乘回归分析方法对螺纹钢期货价格建立了岭回归模型和 Lasso 回归模型,并运用得到的模型对螺纹钢期货价格未来一周进行预测,从预测准确率来看,Lasso 回归模型的预测准确率略好一点. 但从整体来看,两种方法得到的预测效果均令人满意,拟合效果良好,具有实际应用价值.

参　考　文　献

[1]方雯,冯耕中,陆凤彬,等. 国内外钢材市场价格发现功能研究[J]. 系统工程理论与实践,2013,33(1) : 50-60.
[2]靳朝翔,梁仁方,刘建和. 基于神经网络模型的商品期货跨品种套利策略—以焦炭、铁矿石和螺纹钢为例[J]. 云南财经大学学报,2016,32(4) : 150-160.
[3]王珂. 基于神经网络模型的期货螺纹钢价格预测[D]. 太原: 中北大学,2018.
[4]石宝峰,李爱文,王静. 中国螺纹钢期货市场价格发现功能研究[J]. 运筹与管理,2018,147(6) : 166-175.
[5]姚远. 我国螺纹钢期货市场价格发现功能实证研究[D]. 重庆: 西南大学,2012.
[6]陈海鹏,卢旭旺,申铉京,等. 基于多元线性回归的螺纹钢价格分析及预测模型[J]. 计算机科学,2017,44(11A) : 61-64.
[7]王海波等. SAS 统计分析与应用从入门到精通[M]. 第 2 版. 北京: 人民邮电出版社,2013.
[8]鲁茂,贺昌政. 对多重共线性问题的探讨[J]. 统计与决策,2007(8) : 6-9.
[9]HOERL A E,KENNARD R W. Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics,1970,12(1) : 55-67.
[10]MCDONALD G C. Ridge regression[J]. Wires computational statistics,2009,1(1) : 93-100.
[11]何晓群. 应用回归分析[M]. 第 4 版. 北京: 中国人民大学出版社,2015.
[12]TIBSHIRANI R. Regression shrinkage and selection via the Lasso[J]. J R Statist Soc B,1996,58(1) : 267-288.

Empirical analysis of rebar futures price based on ridge regression and Lasso regression

WANG Chun-jie, WEN Li-nan, MA Yuan-jia

(School of Mathematics and Statistics, Changchun University of Techonolgy, Changchun 130000, China)

Abstract: Futures were an important part of the financial market. The delivery of futures was usually carried out after a period of time, so it was particularly important for futures price forecasting. The rebar forecasting of rebar futures was an important measure to improve the competitiveness of China's steel industry. The data of the rebar futures settlement price and its influencing factors were analyzed by using the two methods of ridge regression and Lasso regression to eliminate the influence of collinearity. Two modified multiple regression models were applied respectively. The futures price of rebar in the future was predicted and compared with the real price. Finally, the prediction accuracy of the model based on the two methods was higher than 95%, and the fitting effect based on the Lasso regression method was better. It was proved that the two regression models constructed have important reference value for the trend and prediction of rebar price.

Key words: rebar futures price; collinear; ridge regression; Lasso regression

(责任编辑: 孙爱慧)