Business and Globalization 商业全球化, 2016, 4(2), 47-55
Published Online April 2016 in Hans. http://www.hanspub.org/journal/bglo
http://dx.doi.org/10.12677/bglo.2016.42007

Hans汉斯

Forecast Analysis of Securities Index Based on Ridge Regression

—In Case of Shanghai Composite Index

Rengkang Wu

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: wurengkang@163.com

Received: Mar. 15th, 2016; accepted: Apr. 2nd, 2016; published: Apr. 7th, 2016

Abstract

Security market is an important indicator to measure a country's overall level of economic development. The securities index is a reflection of the overall level of the securities market, and it is an important index of the majority of investors concerned. It not only reflects the basic situation of the securities market, but also plays an important role in guiding the economic trend. Prediction of securities index and trend analysis plays an important role to stabilize market and guide investors. However, serious multi-collinearity between variables often appears in the establishment of the corresponding statistical prediction model. In this paper, the model is improved, and the problem of multi-collinearity between independent variables is solved by using ridge regression. And taking the real data of Shanghai Composite Index as an example, the predicted value of the improved ridge regression model is compared with the true value, and the fitting result is better.

Keywords

Securities Index, Ridge Regression, Multi-Collinearity

基于岭回归的证券指数的预测分析

—以上证综合指数为例

吴仍康

云南财经大学统计与数学学院，云南 昆明

文章引用: 吴仍康. 基于岭回归的证券指数的预测分析[J]. 商业全球化, 2016, 4(2): 47-55.
http://dx.doi.org/10.12677/bglo.2016.42007

Email: wurengkang@163.com

收稿日期：2016年3月15日；录用日期：2016年4月2日；发布日期：2016年4月7日

摘　要

证券市场的是衡量一个国家总体经济发展水平的重要指标。而证券指数是对各个证券市场总体水平的反映，是广大投资者关注的重要指数。它不仅反映了证券市场的基本状况，同时对经济走向也具有重要的导向作用。对证券指数的预测分析以及趋势研判对稳定市场、引导投资者具有重大意义。然而，在建立相应统计预测模型时自变量之间常常出现严重的多重共线。本文通过对模型进行改进，综合运用岭回归解决了自变量间多重共线性的问题。并且以我国上证综合指数的真实数据为例，将改进后岭回归模型的预测值与真实值进行对比，拟合结果较好。

关键词

证券指数，岭回归，多重共线性

1. 引言

1.1. 股票指数

证券市场的成熟度是衡量一个国家经济总体发展水平的重要指标。而在我国“上证综合指数”(也称“大盘指数”)是反映整个股市行情最重要的指标。投资者们时刻在关注的上证指数的波动，并试图预测其发展趋势。然而影响上证指数的因素很多，其作用机制也相当复杂，若想预测其长期走势十分困难。然而，随着计算机技术、统计理论的发展尤其是在当下的大数据时代，对于短期的股指预测却成为可能。

特别是根据股指当天的“开盘价”、“最高价”、“最低价”对其收盘价进行预测。因为不论市场是处于牛市还是熊市环境下，股票当期开盘价、最高价、最低价对收盘价的影响程度均远远的超过历史期收盘收益的影响，这可能是由于开盘价、最高价、最低价与收盘价是同期的原因。

因此，如何判断或预测股票指数走势引起了众多研究者和市场分析人员的极大兴趣，各种预测方法相继涌现。其中邱剑和艾立翔(2011)基于多元线性模型和层次分析法对上证指数建立了预测模型，确定了各个参数的权重，克服了传统定性分析方法不准确的缺点；吴小强和吕文龙(2012)运用时间序列分析对上证指数进行了趋势预测，由于金融时间序列的复杂性模型仅适用于特定的假设下；石鸿雁、尤作军等(2014)基于小波分析的 ARIMA 模型对上证指数进行了分析与预测。然而，在各种模型的预测分析中都无可避免的存在着自变量之间的自相关存在。若能减弱或消除自变量之间的多重共线性，那么在一定程度上则能提高对指数预测的精度以及可靠性。因此，本文将运用岭回归分析方法对模型进行改进解决这一问题。

1.2. 收盘价与开盘价、最高价、最低价之间的关系及其意义

开盘价、收盘价是股票日交易行情中两个比较重要的分析工具。开盘是一天交易的开始，而收盘则标志着一天交易的结束。开盘价是市场各方对当日股价的一个预期，虽然开盘价不能作为判断股价走势的唯一依据，但却可以作为一种参考，特别是一些特殊的开盘价往往能预示当日全天的走势。而收盘价是当日行情的标准，如隔夜没有显著的信息变化，它又是下一个交易日开盘价的依据，可以用来预测下一成交日的股票市场行情。

最高价为当日交易过程中产生的最高价位。最低价为当日交易过程中产生的最低价位。由于价格反应了在交易过程中供给双方的博弈结果，因此最高价与最低价也是投资者十分关注的重要指标，进而影响着收盘时的价位。

综上所述，我们有理由相信收盘价与开盘价、最高价、最低价之间具备着一定的关系。因此，我们可以运用线性模型的相关知识去探寻他们之间的详细的数量关系[1] [2]。

2. 岭回归

岭回归法是 A.E.Horel 在 1962 年提出的一种能统一诊断和处理多重共线性问题的特殊方法， 在多重共线性十分严重的情况下，两个共线变量的系数之间的二维联合分布是一个山岭状曲面，曲面上的每一个点均对应一个残差平方和，点的位置越高，相应的残差平方和越小。因此，山岭的最高点和残差平方和的极小值相对应，相应的参数值便是参数的 OLS 估计值。由于有多重共线形存在时 OLS 估计量已不适用，一个自然的想法就是应寻找别的更合适的估计量。这种估计量既要具有最小的方差，又不能使残差平方和过分的偏离其极小值。在参数的联合分布曲面上，能满足这种要求的点只能沿着山岭寻找，这就是岭回归法。

岭估计方法：

若线性回归模型为：$Y = X\beta + e$

则参数的最小二乘估计为：$\hat{\beta} = (X'X)^{-1}X'Y$

而回归系数 β 的岭估计定义为：$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y$

这里的 $k > 0$ 为可选择参数，称为岭参数或偏参数。因次，对一切 $k \neq 0$，岭估计是有偏估计。它实际上是一种改良的最小二乘法，是以放弃最小二乘的无偏性，放弃部分精确度为代价来寻求效果稍差但更符合实际的回归过程。虽然岭回归所得残差平方和比最小二乘回归要大，但这样一来，它对病态数据的耐受性就远远强于最小二乘法。岭回归方法也非常灵活，它的使用存在着一定的主观人为性，但这种人为性正好是发挥定性分析与定量分析有机结合之处，在解决多重共线性问题中有着独特作用。

岭迹法——一种求 k 值的方法

岭估计 $\hat{\beta}(k) = (X'X + kI)^{-1}X'Y$ 是随着 k 值的改变而变化。

若记 $\hat{\beta}_i(k)$ 为 $\hat{\beta}(k)$ 的第 i 个分量，则它是 k 的一元函数，当 k 在 $[0,+\infty)$ 上变化时，$\hat{\beta}_i(k)$ 的图形称为“岭迹”。

选择岭迹的方法：将 $\hat{\beta}_i(k)$ 的岭迹画在同一个图上，根据岭迹的变化趋势选择 k 值，使得各个回归系数的岭估计大体上稳定，并且各个回归系数的岭估计的符号比较合理[3]-[5]。

3. 岭回归上证综合指数的岭回归分析

3.1. 对数据进行多元线性回归

现在对 1990 年 12 月 19 日~2013 年 12 月 31 日上海证券综合指数的日 K 线图的数据进行分析。由于全部数据量太大，故仅将其中部分数据展示见表 1。

因此，根据数据建立收盘指数与开盘指数、最高指数、最低指数之间的多元线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \mu \quad (1)$$

对(1)中所建立的模型直接用最小二乘法运用 SPSS 软件得出相应的分析结果如图 1。

由结果分析可知：

1) 多元线性回归方程的可决系数 R^2 接近于 1，调整后的可决系数也接近于 1。这说明模型的拟合优度非常好。

Table 1. The Shanghai Composite Index on K-line part data
表 1. 上证综合指数日 K 线部分数据

时间	开盘指数	最高指数	最低指数	收盘指数
2013/01/04	2289.51	2296.11	2256.56	2276.99
2013/01/07	2271.66	2293.32	2266.86	2285.36
2013/01/08	2284.65	2289.14	2262.98	2276.07
2013/01/09	2271.3	2283.9	2259.05	2275.34
2013/01/10	2274.38	2295.48	2268.65	2283.66
2013/01/11	2285.19	2290.21	2234.95	2243
2013/01/14	2236.29	2317.62	2235.11	2311.74
2013/01/15	2312.47	2332.78	2309.32	2325.68
2013/01/16	2322.16	2326.76	2279.51	2309.5
2013/01/17	2305.14	2305.6	2275.88	2284.91
2013/01/18	2295.81	2324.51	2285.97	2317.07
2013/01/21	2321.49	2329.58	2305.1	2328.22
2013/01/22	2326.49	2335.81	2301.3	2315.14
2013/01/23	2308.52	2325.12	2296.49	2320.91
2013/01/24	2320.26	2362.94	2287.3	2302.6
2013/01/25	2300	2308.38	2288.26	2291.3
2013/01/28	2295.35	2346.92	2295.35	2346.5
2013/01/29	2347.22	2363.8	2337.35	2358.98
2013/01/30	2360.75	2383.76	2347.89	2382.48
2013/01/31	2383.43	2391.82	2371.23	2385.42

2) 多元线性回归方程的 F 检验的 P 值远小于 0.05。因此，对方程的检验是显著的。

3) 对三个自变量回归系数的 t 检验的 P 值均远小于 0.05。因此，对三个变量的回归系数是显著的。

4) 由 D.W 检验法可知，D.W 值接近于 2。因此，模型不存在序列相关。

5) 然而，由变量之间的相关矩阵可以看出，变量之间的相关系数很高，部分甚至接近于 1。因此，有理由怀疑模型的自变量之间存在严重的多重共线性。

因此，下面通过计算条件数进一步判定模型中是否存在严重的多重共线性。

通过 MATLAB 软件，先将原始数据中心化和标准化，再计算 $X'X$ 得：

$$X'X = \begin{bmatrix} 1.0000 & 0.9998 & 0.9997 \\ 0.9998 & 1.0000 & 0.9995 \\ 0.9997 & 0.9995 & 1.0000 \end{bmatrix}$$

在计算其三个特征值，分别：$\lambda_1 = 0.0002, \lambda_2 = 0.0005, \lambda_3 = 2.9993$

因此根据条件数的定义：

描述性统计量

	均值	标准 偏差	N
收盘	1678.3482	1001.44111	5618
开盘	1677.8994	1000.67552	5618
最高	1694.8395	1011.32039	5618
最低	1658.7885	988.05842	5618

相关性

		收盘	开盘	最高	最低
Pearson 相关性	收盘	1.000	.999	1.000	1.000
	开盘	.999	1.000	1.000	1.000
	最高	1.000	1.000	1.000	1.000
	最低	1.000	1.000	1.000	1.000
Sig.（单侧）	收盘	.	.000	.000	.000
	开盘	.000	.	.000	.000
	最高	.000	.000	.	.000
	最低	.000	.000	.000	.
N	收盘	5618	5618	5618	5618
	开盘	5618	5618	5618	5618
	最高	5618	5618	5618	5618
	最低	5618	5618	5618	5618

模型汇总[b]

模型	R	R 方	调整 R 方	标准 估计的误差	更改统计量					Durbin-Watson
					R 方更改	F 更改	df1	df2	Sig. F 更改	
1	1.000[a]	1.000	1.000	15.05876	1.000	8278610.086	3	5614	.000	2.159

a. 预测变量: (常量), 最低, 最高, 开盘。

b. 因变量: 收盘

系数[a]

模型		非标准化系数		标准系数	t	Sig.	相关性			共线性统计量	
		B	标准 误差	试用版			零阶	偏	部分	容差	VIF
1	(常量)	-1.249	.393		-3.178	.001					
	开盘	-.709	.011	-.708	-61.856	.000	.999	-.637	-.012	.000	3254.335
	最高	.982	.010	.992	100.393	.000	1.000	.801	.020	.000	2423.976
	最低	.726	.008	.716	92.952	.000	1.000	.779	.019	.001	1474.719

a. 因变量: 收盘

Figure 1. Results of regression analysis
图 1. 回归分析结果

$$k=\frac{\lambda_3}{\lambda_1}=\frac{2.9993}{0.0002}=14996$$

由此可知，条件数非常大，因此可以判定在模型中存在严重的多重共线性。

3.2. 运用岭回归法对模型进行改进

由于回归系数 β 的岭估计为：$\hat{\beta}(k)=(X'X+kI)^{-1}X'Y$

其关键在于确定岭参数 k 的值，有前面所介绍可知运用“岭迹法”可确定岭参数 k 的值。

运用 MATLAB 软件在给定 k 值范围在区间[0,30]内，对中心化和标准化后的数据进行岭回归分析。可以得到如下结果见表 2。

岭迹图如图 2。

因此，从图像中以及表格中我们不难发现，三个变量的回归系数在 $k=10$ 后开始收敛，故总体上看大致我们可以取 $k=10$。

带入原模型后得如下岭回归方程：$\hat{Y}=-1.249+0.2257X_1+0.3871X_2+0.38645X_3$

4. 预测

通过岭回归所得到的线性模型，对 2014 年 11 月 3 日~12 月 26 日每日的收盘价进行预测，并与实际的指数进行对比，见表 3。

Table 2. Results of ridge regression analysis of Shanghai Composite Index
表 2. 上证综合指数岭回归分析结果

k	B1	B2	B3
0.00	−0.70799	0.9917	0.71618
1.00	−0.21948	0.65785	0.56144
2.00	−0.04417	0.54522	0.49868
3.00	0.046385	0.48961	0.46368
4.00	0.10177	0.45673	0.44112
5.00	0.13916	0.43509	0.42529
5.10	0.14225	0.43333	0.42397
5.50	0.15365	0.42685	0.41902
6.00	0.16612	0.41982	0.41355
6.10	0.16841	0.41853	0.41254
6.50	0.17697	0.41375	0.40874
7.00	0.18648	0.40847	0.40448
7.10	0.18825	0.40749	0.40368
7.50	0.1949	0.40383	0.40067
8.00	0.2024	0.39971	0.39725
8.10	0.20381	0.39895	0.39661
8.50	0.20913	0.39605	0.39416
9.00	0.2152	0.39276	0.39135
9.10	0.21634	0.39214	0.39082
9.20	0.21746	0.39154	0.3903
9.30	0.21856	0.39094	0.38979
9.40	0.21964	0.39036	0.38929
9.50	0.22069	0.38979	0.38879
9.60	0.22173	0.38924	0.38831
9.70	0.22275	0.38869	0.38783
9.80	0.22375	0.38815	0.38736
9.90	0.22473	0.38762	0.3869
10.0	0.2257	0.3871	0.38645
11.0	0.23447	0.38241	0.3823
12.0	0.24192	0.37846	0.37875
13.0	0.24831	0.37509	0.37567
14.0	0.25386	0.37218	0.37298
15.0	0.25872	0.36963	0.3706
16.0	0.26301	0.36739	0.36848
17.0	0.26683	0.36541	0.36659
18.0	0.27025	0.36363	0.36489
19.0	0.27334	0.36203	0.36334
20.0	0.27612	0.36059	0.36194

Table 3. Comparison of the predictive value and the real value of Shanghai Composite Index
表 3. 上证综合指数预测值与真实值的对比

时间	回归预测值	真值
2014/11/03	2423.9	2430.03
2014/11/04	2423.6	2430.679
2014/11/05	2423.5	2419.25
2014/11/06	2412.5	2425.86
2014/11/07	2427.2	2418.17
2014/11/10	2444.8	2473.67
2014/11/11	2475.6	2469.67
2014/11/12	2463.3	2494.479
2014/11/13	2487.7	2485.61
2014/11/14	2468	2478.82
2014/11/17	2491.2	2474.01
2014/11/18	2462.8	2456.37
2014/11/19	2449	2450.99
2014/11/20	2443.8	2452.66
2014/11/21	2461	2486.79
2014/11/24	2514.5	2532.88
2014/11/25	2541	2567.6
2014/11/26	2581.2	2604.35
2014/11/27	2612.1	2630.49
2014/11/28	2644.2	2682.83
2014/12/01	2690.8	2680.16
2014/12/02	2706.2	2763.54
2014/12/03	2773.4	2779.52
2014/12/04	2821.2	2899.459
2014/12/05	2899.2	2937.65
2014/12/08	2945.4	3020.26
2014/12/09	2966.2	2856.27
2014/12/10	2869	2940.01
2014/12/11	2921.9	2925.74
2014/12/12	2933.2	2938.17
2014/12/15	2921.2	2953.42
2014/12/16	2972.9	3021.52
2014/12/17	3030.8	3061.02
2014/12/18	3057.1	3057.52
2014/12/19	3061.1	3108.6
2014/12/22	3134.2	3127.449
2014/12/23	3078.6	3032.609
2014/12/24	2999.7	2972.53
2014/12/25	3011.5	3072.54
2014/12/26	3102.5	3157.6

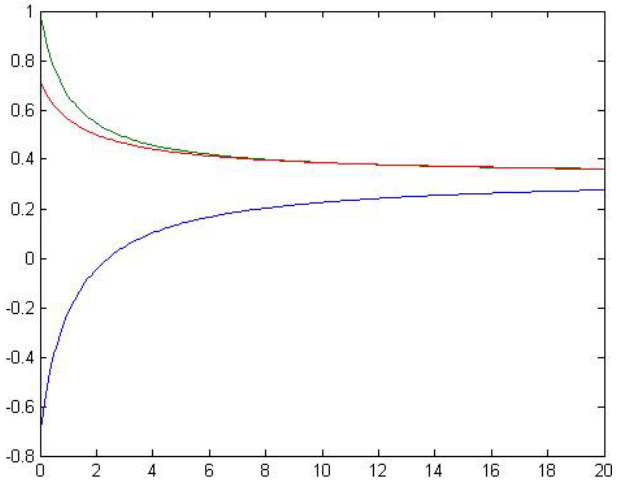

Figure 2. Results of regression analysis
图 2. 回归分析结果

由表分析可知，该模型的拟合效果较好。

5. 结论

本文通过对 1990 年 12 月 19 日~2013 年 12 月 31 日的上证综合指数日 K 线数据进行研究，通过岭回归解决了变量之间的严重的多重共线性问题。最后得到线性回归模型，并通过对比 2014 年 11 月 3 日~12 月 26 上证综合指数回归预测值与真值，说明了该模型的拟合效果较好。

因此，我们有一下结论：

$\hat{\beta}_1 = 0.2257$ 表示每日开盘指数每增加一个单位，当天收盘指数则增加 0.2257 个单位。

$\hat{\beta}_2 = 0.3871$ 表示每日最高指数每增加一个单位，当天收盘指数则增加 0.3871 个单位。

$\hat{\beta}_3 = 0.38645$ 表示每日最低指数每增加一个单位，当天收盘指数则增加 0.38645 个单位。

收盘指数变化可由开盘指数，、最高指数和最低日指数的变化(线性)来解释。

参考文献 (References)

[1] 陈怡玲, 宋逢明. 中国股市价格变动与交易量关系的实证研究[J]. 管理科学学报, 2000, 3(2): 62-68.

[2] 赵传刚. 我国 A 股市场量价关系的实证分析[D]: [硕士学位论文]. 南昌: 江西财经大学, 2007: 20-22.

[3] 王松桂, 史建红, 等. 线性模型引论[M]. 北京: 科学出版社, 2004.

[4] 张尧庭, 方开泰. 多元统计分析引论[M]. 北京: 科学出版社, 1982.

[5] 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.

附录

主要的 MATLAB 的程序：

标准化数据并求出矩阵 X'X 以及其特征值：

x11 = zscore(x1)

x22 = zscore(x2)

x33 = zscore(x3)

X = [x11，x22，x33]

X'X

eig(X'X)

求出在不同 K 值的情况下的岭回归系数并画出岭迹图

K = 0:0.01:30

B0 = ridge(Y，X，K，0)

plot(K，B0')

3、通过所求模型对数据进行预测

B1 = 0.2257

B2 = 0.3871

B3 = 0.38645

U = −1.249

Y = U*ones(20，1)+B1*X1+B2*X2+B3*X3