Advances in Social Sciences 社会科学前沿, 2016, 5(4), 630-637
Published Online September 2016 in Hans. http://www.hanspub.org/journal/ass
http://dx.doi.org/10.12677/ass.2016.54088

Hans汉斯

Short Term Forecast of Daily Closing Price of Shanghai Composite Index—Empirical Analysis Based on ARIMA Model

Xiaoli Zhou, Zhixiong Wu

Shanghai Maritime University, Shanghai
Email: zhouxiaoli_2016@163.com, zxwu@shmtu.edu.cn

Received: Sep. 1st, 2016; accepted: Sep. 15th, 2016; published: Sep. 22nd, 2016

Abstract

Shanghai stock exchange is one of the two major exchanges, its establishment of the Shanghai Composite Index can effectively respond to changes in the stock price listed on the exchange. Through the establishment of a suitable time series model, it is important to predict and analyze the trend of the index change. By dealing with the historical data and using R to select the model and estimate the parameters, this article achieved the short-term closing price forecasting of the Shanghai Composite Index, and this will provide the relevant reference for the trend analysis of stock market.

Keywords

Shanghai Composite Index, Times Series, ARIMA Model, Predict

上证综合指数日收盘价的短期预测——基于ARIMA模型的实证分析

周小丽，吴志雄

上海海事大学，上海
Email: zhouxiaoli_2016@163.com, zxwu@shmtu.edu.cn

文章引用: 周小丽, 吴志雄. 上证综合指数日收盘价的短期预测——基于 ARIMA 模型的实证分析[J]. 社会科学前沿, 2016, 5(4): 630-637. http://dx.doi.org/10.12677/ass.2016.54088

收稿日期：2016年9月1日；录用日期：2016年9月15日；发布日期：2016年9月22日

摘　要

上海证券交易所作为国内两大交易所之一，它编制的上证综合指数能有效反映在该交易所上市的股票价格的变动情况。通过建立合适的时间序列模型，对该指数变化趋势进行预测分析具有重要意义。文章对历史数据进行处理，通过R语言进行模型选择及估计，实现上证综合指数日收盘价格的短期预测，为股票市场投资者提供相关参考。

关键词

上证综合指数，时间序列，ARIMA模型，预测

1. 引言

上海证券综合指数(简称“上证综指”)反映了在上海证券交易所上市股票价格的变动情况，在我国，股票市场上最具代表性的指数之一就是上证综指，它的波动能及时准确的反映股票市场的运行情况[1]。文献[2]中将非平稳时间序列进行差分，并通过自相关系数及偏自相关系数来对模型进行识别，提出了完整的建模、估计、参数检验的方法。时间序列反映了对象的发展趋势，文献[3]-[6]提出了针对不同对象的建模方法。文献[3]以指数为对象，研究了上证综指中的 ARCH 现象。文献[4]通过建立对应的时间序列模型来研究风险溢价的时变性。文献[5]研究了用 ARIMA 模型来探究上证 A 股指数的变化规律。文献[6]提出了 ARIMA 模型以及相关干预模型，并结合政府对股市的重大干预，来对深圳证券交易市场成分指数进行建模。ARIMA 模型具有广阔的应用空间，适用于多种领域内的预测研究，文献[7]-[10]研究了 ARIMA 模型的应用实例。文献[7]将生态承载能力进行量化，从而获取时间序列并通过 ARIMA 模型进行动态模拟和预测。文献[8]研究了 ARIMA 模型在信用风险波动性上的应用。文献[9] [10]研究了 ARIMA 与神经网络组合模型，文献[9]将该模型用于 GDP 预测，文献[10]则将该模型用于碳排放预测，从中可以看到，没有任何一种模型是万能通用的，不同的时间序列，甚至同一序列选择不同时间段所适合的模型是不同的。本文研究股票市场的短期预测，考虑到上证综指收盘价受很多随机因素的干扰，往往是非平稳序列，因此选择 ARIMA 模型对上证综指 2014.5.12~2015.5.22 日的收盘价格进行有效建模，再利用该模型进行实证并判断预测的好坏。

2. 建模

2.1. ARIMA 模型

金融时间序列模型大都是非平稳的序列，但对于短期数据，若通过适当的差分往往能获得平稳的序列，因此本文选择常用的非平稳时间序列模型自回归求和移动平均模型即 ARIMA(p,d,q)来进行预测，该模型的本质实际上是差分运算和 ARMA(p,q)模型的结合。

2.1.1. 自回归模型(AR 模型)

若平稳序列在 t 时刻的值为 Z_t，则这个值可以表示成过去 p 个时刻 a_t 的值 $Z_{t-1,}Z_{t-2,}\cdots Z_{t-p}$ 的线性组合加上 t 时刻所对应的残差 a_t，该残差为白噪声。这个序列可以用 AR 模型表示为：

$$Z_t = c+\varphi_1 Z_{t-1}+\varphi_2 Z_{t-2}+\cdots+\varphi_p Z_{t-p}+a_t, t=1,2,\cdots,T \quad (1)$$

其中，c 为常数，ϕ_1，ϕ_2,…，ϕ_p 是自回归系数，p 为自回归阶数，a_t 为白噪声序列。

2.1.2. 移动平均模型(MA 模型)

若平稳序列在 t 时刻的值为 Z_t，则这个值可以表示成过去 q 个时刻残差序列 $\{a_t\}$ 的加权平均值和 a_t 的和，且 $\{a_t\}$ 为白噪声序列，则该序列可以用移动平均模型如下表示：

$$Z_t = u + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q}, t = 1, 2, \cdots, T \tag{2}$$

其中，u 为常数，θ_1, θ_2,…θ_q 是自回归系数，q 为移动平均模型阶数，a_t 为白噪声序列。

2.1.3. 自回归移动平均模型(ARMA 模型)

自回归移动平均过程是自回归过程和移动平均过程的组合。该模型矩阵形式如下：

$$\phi_p(B)(1-B)^d Z_t = \theta_0 + \theta_q(B) a_t \tag{3}$$

其中，p 为自回归阶数，q 为移动平均模型阶数，d 为差分次数。

2.2. 模型选择

分析采用的数据来源于 Yahoo Finance，具体为上证综指 2014.5.12~2015.5.22 日的收盘价格，共 254 个样本数据，通过分析数据特征，找到适宜模型，进行估计并预测接下来 5 天的收盘价格。

2.2.1. 平稳性检验

通过该指数时间序列图(图 1)可以看出该时间段的序列有显著向上的趋势并伴随较小的上下波动，为非平稳序列，而从 ACF 图(图 2)衰减缓慢可以得出，为了消除趋势，使该序列成为平稳的序列，需要对原序列进行一定阶数的差分，通过对此原序列进行平稳性检验，利用函数 adf.test()检验，得到 p-value = 0.9349。接下来，尝试对序列进行差分，发现 2 阶差分后序列的平稳性检验的 p-value = 0.01，在 5%的显著性水平下拒绝原假设，序列平稳。因为过度差分会损失原序列所蕴含的信息，所以，为得到平稳序列，模型取 2 阶差分即可，即取 $d = 2$。

2.2.2. 白噪声检验

如果差分后的序列属于白噪声，则说明序列包含的信息中只有较少对相关研究有价值的信息，所以

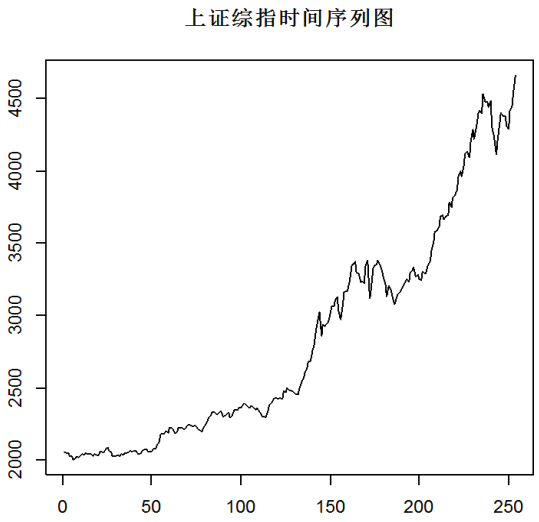

Figure 1. The time series of Shanghai composite index
图 1. 上证综指时间序列图

有必要对差分后的序列进行白噪声检验。本文中，我们选择 Q 检验，通过程序运算可得差分后的序列滞后期为 10 时 Q 检验统计量的值。所对应的 p-value = 1.166e−14，远远小于 0.05，表明差分后的序列包含较多的信息，为非白噪声序列，对该差分后的序列建模能提取到有价值的信息。

2.2.3. 模型定阶

选择合适的 ARIMA 模型，就意味着要找到合适的 p，d，q，由于上面已经判断得到二阶差分后的序列为平稳序列，所以，取定 d = 2。此外，还需要识别并找出合适的 p 和 q，识别 p 和 q 阶数的方法是通过使样本 ACF 和 PACF 与已知模型的理论形态相匹配。已知模型的理论形态如表 1 所示。

二阶查分后差分后上证综指收盘价的 ACF 和 PACF 图如图 3 和图 4 所示。

由图 3 可以看出，滞后 2 阶自相关值基本不再超过边界值，虽然隔两个滞后期后又有 2 个自相关值超过边界，但是这很可能属于偶然出现的，因为理论上我们可以期望 1 到 20 之间的会偶尔超出 95%的置信边界，所以，自相关值选 2 阶，即 p = 2。另外，图 4 显示滞后 6 阶偏自相关值不再显著超过边界，虽然其中滞后 4 期后有 1 期未超过边界但之后又连续出现 2 次 PACF 值显著超过边界，所以偏自相关值选 6 阶，即 q = 6。

综上所述，我们选定的 ARIMA 模型为 ARIMA(2,2,6)。

3. 参数估计与检验

3.1. 参数估计

在 R 语言中使用 arima()函数来估计模型参数，通过代码 arima(data,order = c(2,2,6))可得参数估计如

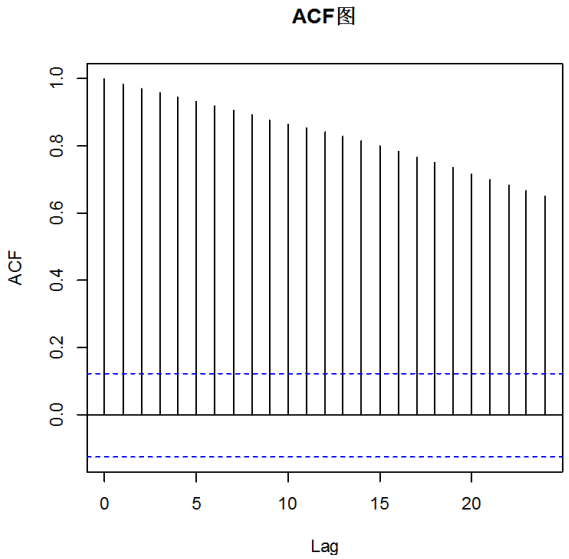

Figure 2. ACF figure
图 2. ACF 图

Table 1. The theoretical form of model
表 1. 模型的理论形态

模型	ACF	PACF
AR(p)	按指数衰减或阻尼正弦波动拖尾	p 步截尾
MA(q)	q 步截尾	按指数衰减或阻尼正弦波动拖尾
ARMA(p,q)	q-p 步截尾	p-q 步截尾

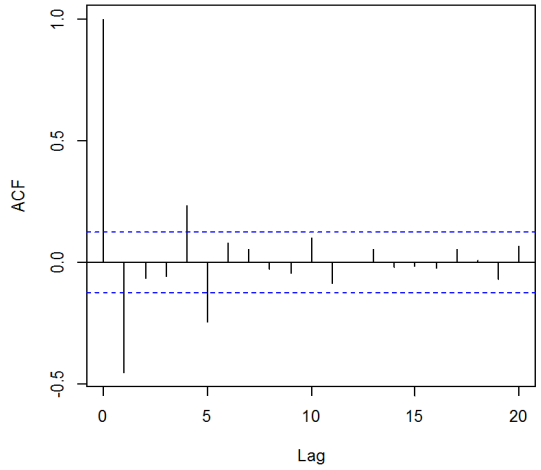

Figure 3. ACF figure of differential series
图 3. 差分后序列 ACF 图

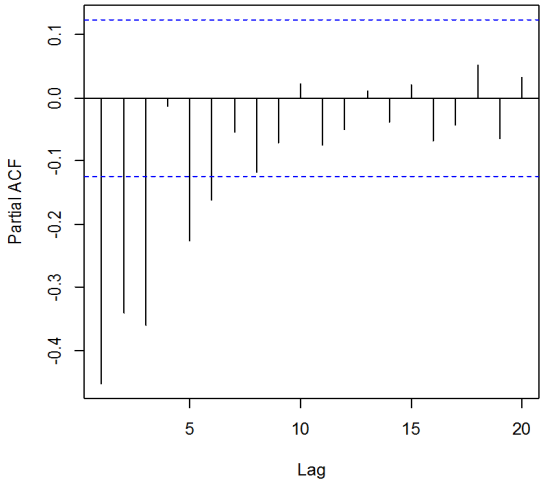

Figure 4. PACF figure of differential series
图 4. 差分后序列 PACF 图

下表 2。

所得模型如下：

$$\left(1+0.644B+0.612B^2\right)Z_t=\left(1-0.261B-0.098B^2-0.649B^3+0.139B^4-0.175B^5+0.087B^6\right)a_t$$

即：

$$Z_t=-0.644Z_{t-1}-0.612Z_{t-2}+a_t-0.261a_{t-1}-0.098a_{t-2}-0.649a_{t-3}+0.139a_{t-4}-0.175a_{t-5}+0.087a_{t-6}$$

3.2. 模型的有效性检验

判断模型是否有效，需要检验模型是否较充分的从原数据中提取相关的信息，一个有效的模型应该几乎提取了原数据中所包含的所有信息，使得剩下的残差序列中不再蕴含其他相关信息。这意味着残差序列应该是一个纯随机的序列，即白噪声序列。满足这样条件的模型才算是显著的有效的模型。

但需要注意的是，在检验残差时，若残差不是纯随机序列，而是由非独立的纯随机序列组合的序列，那么仅仅检验序列的纯随机性不能完全证明模型的有效性。所以，我们还需检查残差的分布，观察 ARIMA

模型的预测误差是否服从均值为 0 且方差为常数的正态分布。与此同时，也要观察连续预测误差是否(自)相关，这主要通过直方图和 QQ 图来检验。

首先，通过残差 ACF 图来判断残差是否为白噪声序列。

如图 5 所示，在滞后 1-20 阶(lags 1 - 20)的范围内，样本自相关值都没有超出显著(置信)边界。利用 R 软件进行 Ljung-Box 检验(即 Q 检验)，得到的 p-value = 0.9978，所以可以推断在滞后 1~20 阶(lags 1 - 20)范围内，没明显证据说明预测误差是非零自相关的。

要判断残差是否服从均值为 0，方差不变的正态分布，需要计算残差，残差直方图和 QQ 图如下图 6 和图 7 所示。

从图 6 和图 7 可以看出，该模型残差服从均值为 0，方差恒定的正态分布。

通过上述分析，此模型的残差序列为纯随机序列，可视为白噪声，且服从均值为 0，方差不变的正态分布，建立的该 ARIMA (2,2,6)为有效的。

4. 模型预测与拟合

通过上述模型的检验，可以发现 ARIMA(2,2,6)模型最适合样本数据的分析和预测，利用 R 语言中的 forecast.Arima()函数对模型序列进行预测，后 5 个交易日上证指数的收盘价格预测如下。

由表 3 预测结果和真实值的比较可知，该模型具有较好的预测效果。

5. 总结

本文在历史数据的基础上建立 ARIMA(2,2,6)模型并进行了上证综指日收盘价的短期预测。通过预测结果，可以看出，模型效果较好，预测结果与真实值相差较小，本文建立的模型对股票投资有着一定的建议作用。但是由于股市走势本生受很多复杂因素的干扰，且其还具有时变性、随机性和非线性等特点，

Table 2. Results of parameter estimation
表 2. 参数估计结果

	ar1	ar2	ma1	ma2	ma3	ma4	ma5	ma6
Coefficients	−0.644	−0.612	−0.261	−0.098	−0.649	0.139	−0.175	0.087
s.e.	0.246	0.161	0.254	0.204	0.1637	0.084	0.104	0.091

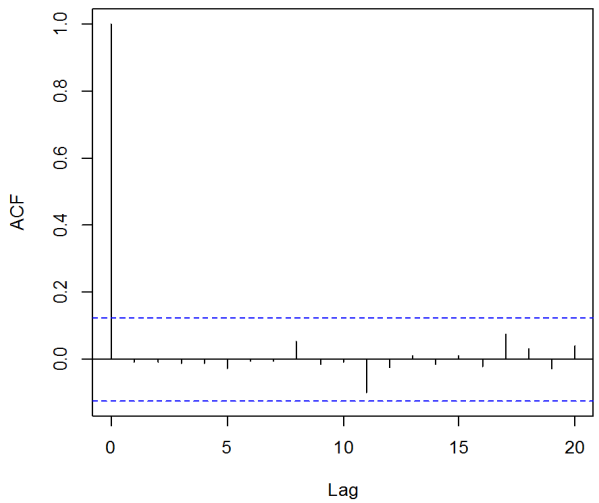

Figure 5. ACF figure of residuals
图 5. 残差 ACF 图

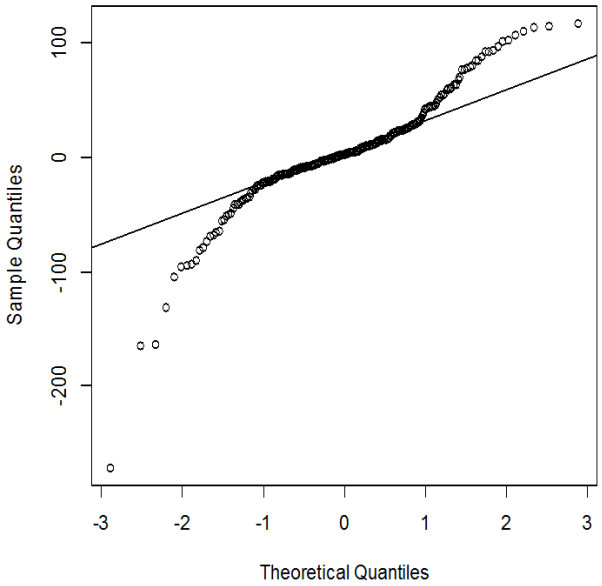

Figure 6. Histogram of residuals
图 6. 残差直方图

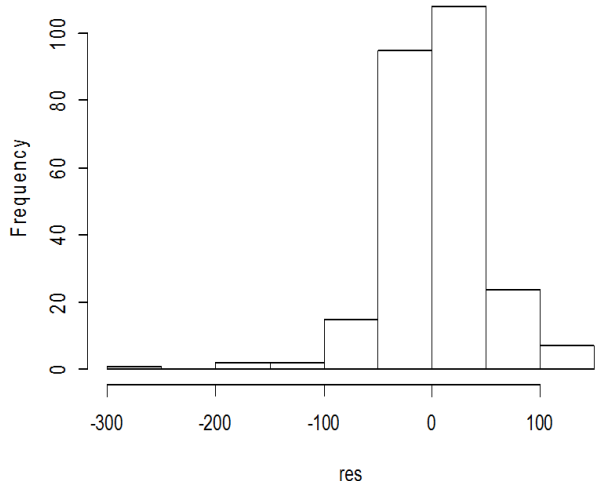

Figure 7. QQ figure of residuals
图 7. 残差 QQ 图

Table 3. The prediction of closing price
表 3. 收盘价格预测

日期	预测值	真实值
2015/5/25	4804.664	4813.8
2015/5/26	4883.443	4910.9
2015/5/27	4914.416	4941.71
2015/5/28	4650.610	4620.27
2015/5/29	4641.907	4611.74

本文建立的模型仅对短期预测有效，核心在于研究特定股票大概走势。投资者可以通过走势的判断并结合对股票基本面和技术面的分析，对投资提供参考，而关于股票市场里中长期趋势的预测还有待进一步研究。

参考文献 (References)

[1] 李文君, 尹康. 多元 GARCH 模型研究述评[J]. 数量经济技术经济研究, 2009, 26(10): 138-147.

[2] Box, G.E.P. and Jenkis. G.M. (1978) Time Series Analysis: Forecasting and Control. Revised Edition, Holden Day, San Francisco.

[3] 丁华. 股价指数波动中的 ARCH 现象[J]. 数量经济技术经济研究, 1999, 16(9): 22-25.

[4] 张思奇, 马刚, 冉华. 股票市场风险、收益与市场效率——ARMA-ARCH-M 模型[J]. 世界经济, 2000, 23(5): 19-28.

[5] 查正洪. 上证综合指数的统计分析与预测[J]. 上海海运学院报, 1999, 20(4): 80-87.

[6] 赵志峰. 对建立中国股票价格指数时间序列模型的探讨[J]. 统计与信息论坛, 2003, 18(1):.66-69.

[7] 王耕, 王嘉丽, 苏柏灵. 基于 ARIMA 模型的辽河流域生态足迹动态模拟与预测[J]. 生态环境学报, 2013, 22(4): 632-638.

[8] 丁彦皓. 中国住房抵押贷款证券化信用风险的波动特征检验——基于 ARIMA-ARCH 模型的论证[J]. 金融经济学研究, 2013, 28(4): 117-128.

[9] 熊志斌. 基于 ARIMA 与神经网络集成的 GDP 时间序列预测研究[J]. 数理统计与管理, 2011, 30(2): 306-314.

[10] 赵成柏, 毛春梅. 基于 ARIMA 和 BP 神经网络组合模型的我国碳排放强度预测[J]. 长江流域资源与环境, 2012, 21(6): 665-671.