

ARMA模型在股票预测中的应用

陶友凤, 张欣, 张德飞*

红河学院数学与统计学院, 云南 蒙自

收稿日期: 2021年12月26日; 录用日期: 2022年1月16日; 发布日期: 2022年1月28日

摘要

通过R软件研究ARMA模型在平稳时间序列数据预测方面的优势,对太平洋股份和云铝股份的股票日收盘价进行实证分析,并作出试探性的预测分析。

关键词

R软件, ARMA模型, 时间序列

Application of ARMA Model in Stock Forecasting

Youfeng Tao, Xin Zhang, Defei Zhang*

School of Mathematics and Statistics, Honghe University, Mengzi Yunnan

Received: Dec. 26th, 2021; accepted: Jan. 16th, 2022; published: Jan. 28th, 2022

Abstract

This paper studies the advantages of ARMA model in the prediction of stationary time series data by R software, makes an empirical analysis on the daily closing prices of Pacific Securities stock and Yunnan Aluminum stock, and makes a tentative prediction analysis.

Keywords

R Software, ARMA Model, Time Series

*通讯作者。



1. 引言

中国证券市场规模不断扩大, 规范化程度不断提高, 市场功能和作用逐步显现, 在国民经济运行中发挥着越来越重要的作用。影响股票市场涨跌的因素多种多样, 例如宏观经济的发展趋势、行业政策、资金供求格局、利率的走势、政治因素、突发事件, 证监会的重大政策信息等因素的影响。对于投资者来说, 进行股票投资时, 应了解宏观大环境, 正确解读其中的异动, 把握涨跌的趋势, 寻找合适的买卖点。中国的股票市场是在社会主义市场经济条件下的资本流动市场, 起步较晚, 且不够完善, 国内有很多学者对我国股票市场做了大量研究, 结果表明我国的股市是可以进行预测研究的。

2. 文献综述

在对股票收盘价分析过程中如果使用时间序列分析方法对其收盘价进行分析, 那么得到的结果不是很理想, 然而对于具有很大波动性和聚集性的股票市场而言, 使用较为一般的方法去描述是不够准确的。许多人在研究过程中采用自回归移动平均模型(ARMA), 通过利用这个模型较为集中的体现方差的性质以及对于股票价格的短期预测效果良好。目前有很多学者对于不平稳的时间序列数据中得到了很多启发。例如: 高远利用对乐视网股票收盘价格建立 ARMA 模型, 得出 ARMA 模型对投资者具有一定的指导作用[1]; 谢海宁、尚园利用 ARMA 模型对中信银行开盘价进行预测, 得出 ARMA 模型能够对短期内的股票价格进行很好的预测[2]; 彭紫君以上证指数为例建立 ARMA-GARCH 模型, 该模型可以推广到金融风险管理中[3]。

3. 建立 ARMA(p, q)模型方法

3.1. ARMA 模型简介

时间序列计量经济学中一个基本的模型是 ARMA 模型[4]。该模型遵循博克斯-詹金斯(Box-Jenkins)方法, 该方法的目的是找到数据生成过程的最简约模型, 首先通过确定数据生成过程, 其次估计这一过程的参数, 最后判断估计模型的残差, 这是一套完整的方法。通过诸多方法的对比, 后来人们在时间序列的预测中也开始普遍使用 ARMA 模型。例如在当今激烈的股票市场中, 投资者可以根据之前的股票价格数据, 估计出未来时间里的股票价格趋势, 由此更方便投资者根据预测数据做出合理的投资决策。

ARMA 模型的全称是自回归移动平均模型, 它是目前最常用的拟合平稳序列的模型。它又可以细分为 AR 模、MA 模型和 ARMA 模型三大类, 是研究平稳随机过程有理有谱的典型方法。模型中具有(p, q)阶自回归移动平均模型通常记为 ARMA(p, q), 其数学表达公式(1)如下, 其中: p 为 AR 模型的阶数; q 为 MA 模型的阶数; $\phi_i (i=1, 2, \dots, p)$; $\theta_j (j=1, 2, \dots, q)$ 为模型的待定系数; ε_t 为残差; x_t 为观测值。

$$\begin{cases} x_t = \phi_0 + \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q} \\ \phi_p \neq 0, \theta_q \neq 0 \\ E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) = 0, \forall s < t \end{cases} \quad (1)$$

3.2. 建模流程

- 第一、进行模型的参数估计，确定模型阶数。
- 第二、模型的拟合检验。
- 第三、使用建立的 ARMA 模型对研究的股票序列进行预测。

4. 实证分析

4.1. 数据来源

本文的数据来源于大智慧软件中选取云铝股份(本文称作股份 A)、太平洋股份(本文称作股份 B)两只股票在 2010 年 1 月至 2019 年 12 月之间每天的收盘价作为样本数据,样本个数分别 $n_1 = 2076$, $n_2 = 2445$ 。在数据收集以及整理的过程中,要保证两支股票收盘价时间的统一性,首先是以规定好的时间为对照,这样对照后,则可以对多出来的数据进行删减。表 1 为 A 股份部分样本数据。

Table 1. A share part data
表 1. A 股份部分数据

日期	收盘(云铝股份)
2010/1/4	13.61
2010/1/5	14.19
2010/1/6	14.55
2010/1/7	14.89
2010/1/8	14.66
2010/1/11	14.8
2010/1/12	14.93
2010/1/13	13.67
2010/1/14	13.84
2010/1/15	13.86

4.2. 建模与分析

4.2.1. 对股票开盘价时间序列研究

首先假如时序图始终围绕一个常数值较小幅度的上下波动时,则序列是平稳序列;其次假如时间序列有明显的大幅度波动趋势或周期性特征时,则序列是不平稳的时间序列。截取 A 股份从 2010 年 1 月 4 日至 2019 年 12 月 31 日 2076 个交易日和 B 股份的 2445 个交易日的开盘价作为样本数据,借助 R 软件,分别对样本数据分析,绘制出时序图 1、图 2,从图中可判断出,这两组序列是不平衡的。

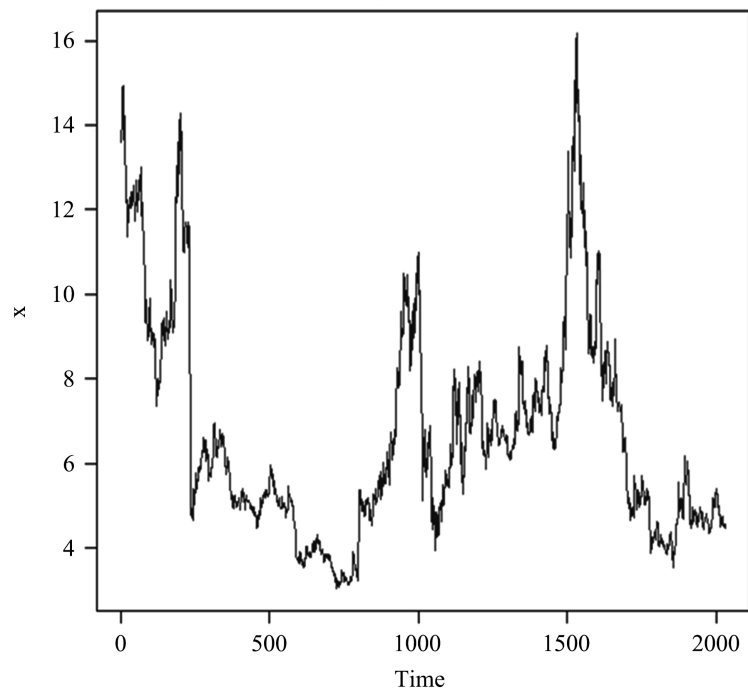


Figure 1. A share timing chart

图 1. A 股份时序图

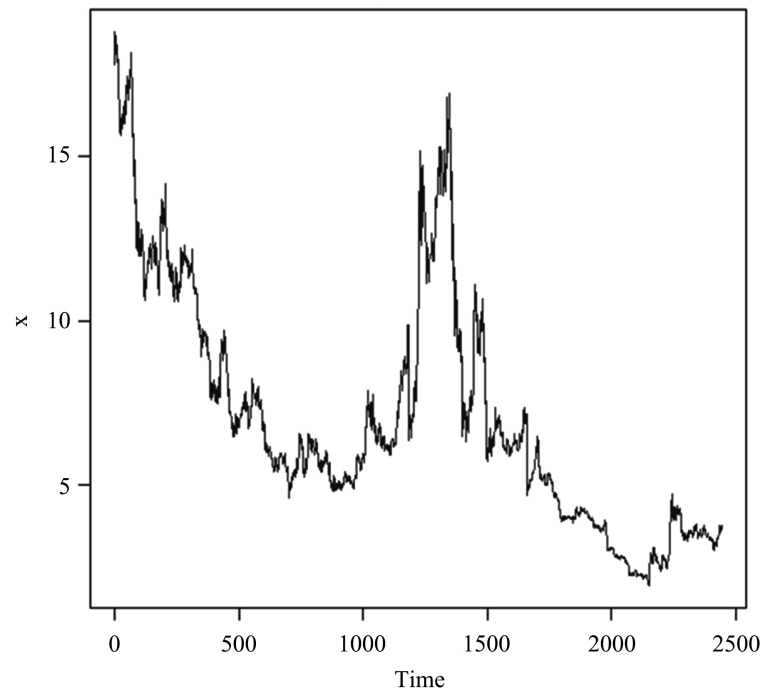


Figure 2. B share timing chart

图 2. B 股份时序图

由时序图可知，两组序列不平衡，进而采用 ADF 单位根检验方法。可知存在异方差，为消除异方差，主要对 A 股份的每日收盘价进行平稳性检验。见表 2：

Table 2. ADF test
表 2. ADF 检验

lag	ADF	P 值
[1]	-1.79	0.0736
[2]	-1.89	0.0597
[3]	-1.94	0.0521
[4]	-1.99	0.0460
[5]	-1.96	0.0490
[6]	-1.98	0.0474
[7]	-1.99	0.0461
[8]	-1.80	0.0723

$P \leq 0.01$ 。

通过 R 软件得出检验结果，由表 2 可以发现有多种模型的统计量的 P 值大于显著水平 0.01，根据 ADF 检验方法可知该序列不平稳。为了可以充分提取确定性信息，下一步做一阶差分。如图 3：

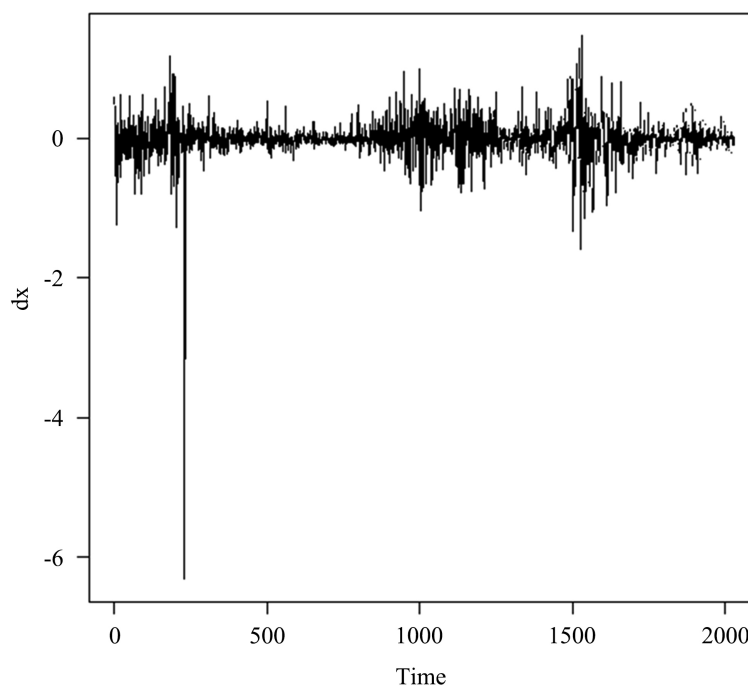


Figure 3. Differential sequence
图 3. 差分序列

从图 3 上看出对数据进行一阶差分处理后实现趋势平稳，进一步做 ADF 检验。见表 3：

Table 3. ADF test
表 3. ADF 检验

lag	ADF	p.value
[1]	-42.5	0.01
[2]	-32.3	0.01
[3]	-25.6	0.01
[4]	-21.0	0.01
[5]	-19.5	0.01
[6]	-18.4	0.01
[7]	-16.1	0.01
[8]	-15.0	0.01

通过上面的检验结果可以知道：因为有多种模型统计量的 P 值小于且等于显著水平，所以可以认为该序列是平稳的，接下来进一步做纯随机检验。

4.2.2. 纯随机检验结果(表 4)

Table 4. Pure random test
表 4. 纯随机检验

延迟	统计量检验	
	LB 统计量值	p 值
延迟 6 阶	21.285	0.00163
延迟 12 阶	34.403	0.0005826

因为纯随机检验统计量的 P 值是非常明显的小于显著性水平 α 的，所以该序列应该拒绝 LB 的原假设。换句话说，我们可以判断该序列是平稳非白噪声序列，认为该序列的波动是有统计规律可循的。

4.2.3. 模型定阶

利用序列的 acf 和 pacf 条形图来判断模型阶数。pacf 是用最小二乘法估计出来的偏自相关函数。根据拖尾指条形图以指数形式或周期形式衰减，截尾指图形在若干期之后变得很小而且没有什么模式。如果偏自相关函数是截尾的，而自相关函数是拖尾的，则我们可以使用 AR 模型；如果偏自相关函数是拖尾的，而自相关函数是截尾的，则我们可以使用 MA 模型；如果偏自相关函数和自相关函数都是拖尾的，则我们可以使用 ARMA 模型。通过利用 R 可以得到一阶差分后的相关图。观察自相关图(图 4)、偏自相关系数图(图 5)可以得出它们是拖尾，因此初步判定认为模型 ARMA 模型。

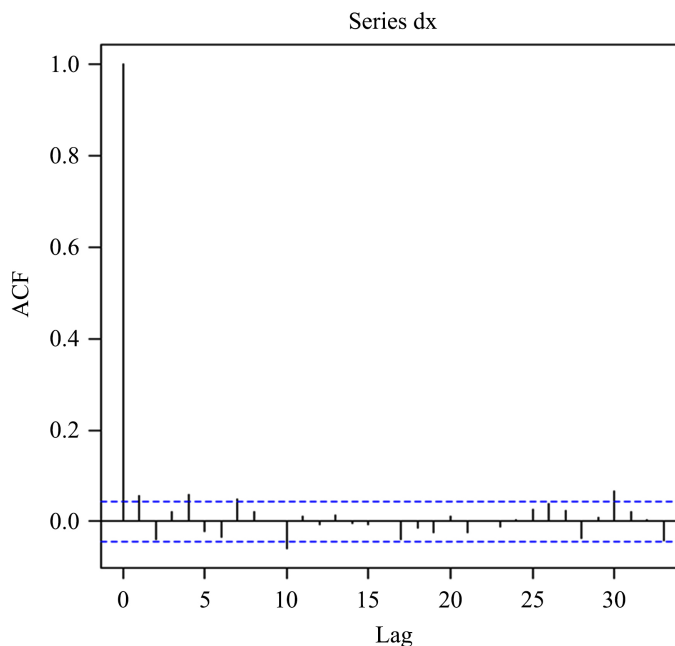


Figure 4. A-share autocorrelation diagram

图 4. A 股自相关图

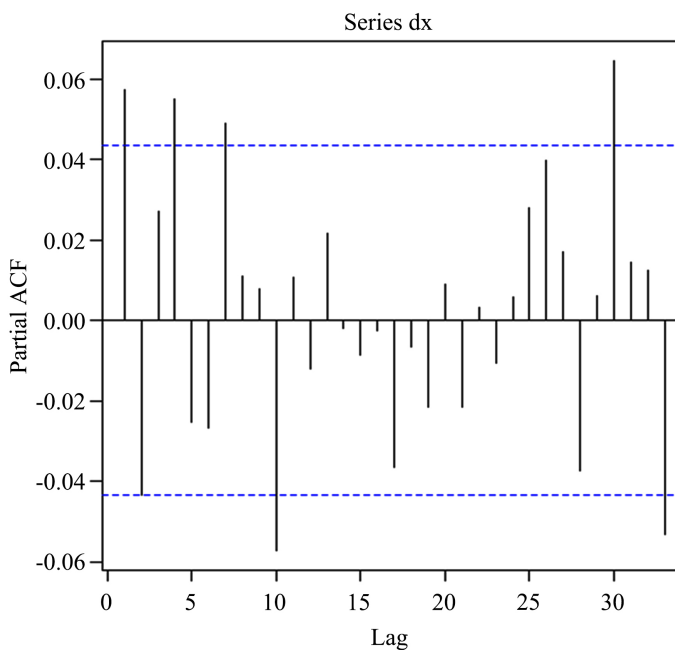


Figure 5. A-share partial autocorrelation diagram

图 5. A 股偏自相关图

以 A 股份为研究对象, 尝试使用 ARMA(2, 2)、ARMA(2, 3)、ARMA(3, 2)、ARMA(3, 3)等模型通过对时间序列进行了模拟, 根据不同组合的试验以及 AIC 准则并且综合各方面的参数可以得到参数估计值、模型的系数均为显著的, 当 $p = 2$, $q = 3$ 时, 对应的 AIC 值相对比较小, 为 593.7118; BIC 信息量的值同样较小, 为 633.1758。最终确定模型的阶数为 $p = 2$, $q = 3$, 即 ARMA(2, 3)见表 5。

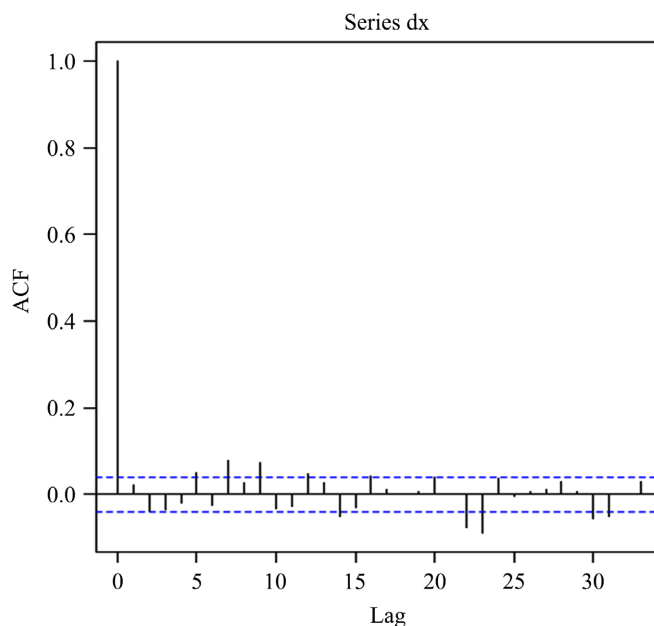


Figure 6. B-share autocorrelation diagram

图 6. B 股自相关图

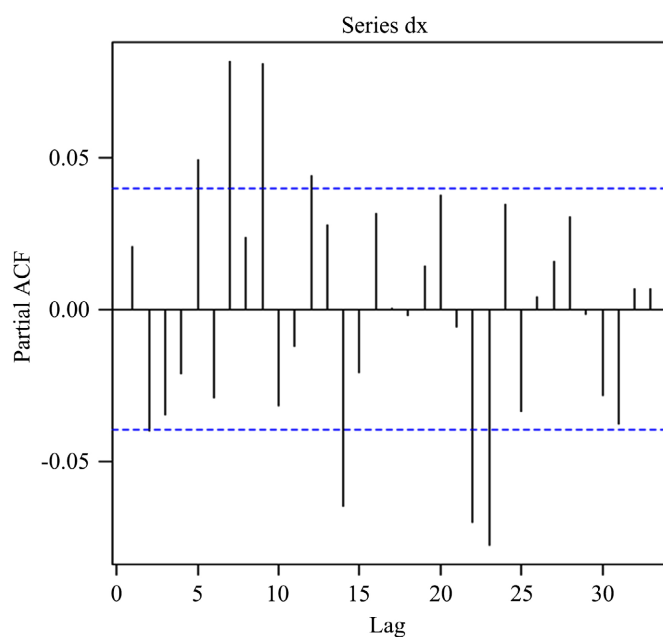


Figure 7. B-share partial autocorrelation diagram

图 7. B 股偏自相关图

以 B 股份为研究对象，自相关图和偏自相关图见图 6 和图 7，尝试使用 ARMA(1, 22)、ARMA(1, 23)、ARMA(2, 25)、ARMA(2, 26)等模型通过对时间序列进行了模拟，根据不同组合的试验以及 AIC 准则并且综合各方面的参数可以得到参数估计值、模型的系数均为显著的，当 $p=2$ ， $q=25$ 时，对应的 AIC 值相对比较小，为 31.80462；BIC 信息量的值同样较小，为 200.045。运用同样的定阶原理可得 B 股份最终确定模型的阶数为 $p=2$ ， $q=25$ ，即 ARMA(2, 25)。

Table 5. AIC and BIC values
表 5. A 股 AIC、BIC 值

	AIC	BIC
ARMA(2, 2)	601.9779	635.8042
ARMA(2, 3)	593.7118	633.1758
ARMA(3, 2)	599.036	638.5
ARMA(3, 3)	600.2445	645.3462

4.2.4. 模型口径

根据 R 软件编写程序输出结果，可以看出 A 股份 ARMA 模型的口径为[5]：

$$X_t = -0.00476121 + 0.1647X_{t-1} + \varepsilon_t - 0.1428\varepsilon_{t-1}$$

$$\text{Var}(\varepsilon_t) = 0.07741$$

根据以上的建模过程同理可得 B 的模型口径为：

$$X_t = 0.2893X_{t-1} - 0.9442X_{t-2} - 0.00678509 + \varepsilon_t - 0.2281\varepsilon_{t-1} + 0.9159\varepsilon_{t-2} + 0.092\varepsilon_{t-3}$$

$$\text{Var}(\varepsilon_t) = 0.07741$$

从图 8、图 9 可以看得出各阶延迟下的白噪声检验统计量的 p 值都显著大于 0.05，我们可以认为这个拟合模型的残差序列是属于白噪声序列，即该拟合模型是显著成立的。结合模型的诊断图，说明拟合的 ARMA(2, 3)模型和 ARMA(2, 25)的显著性检验通过。

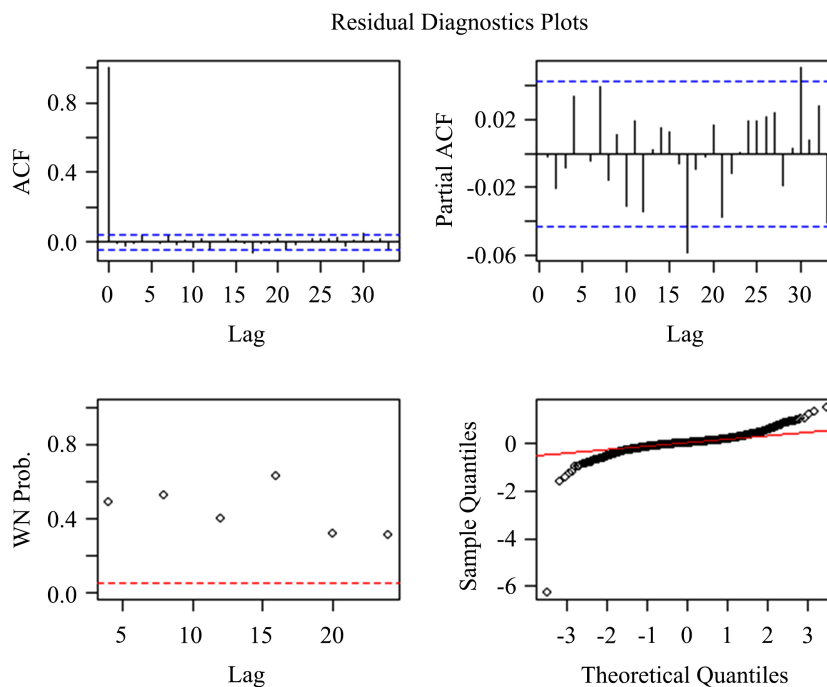


Figure 8. A shares

图 8. A 股份

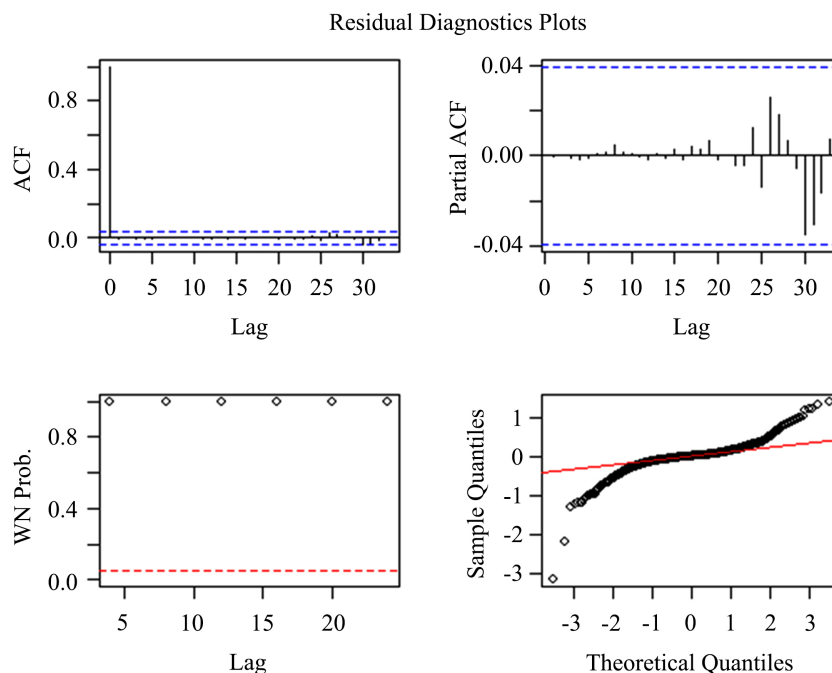


Figure 9. A shares
图 9. A 股份

4.2.5. 拟合效果与预测

为了检验模型优度，采取数据最后十个数据进行模型拟合以及相应检验，根据预测判断模型优度。不带漂移项的预测可以看出后期的预测变化大小趋于常数，比较接近实际值的。带漂移项模型的预测前期接近实际值，后期波动也大，有变低的趋势，不符合实际情况。总体而言，不带漂移项的预测效果较优于带漂移项模型的预测效果。见表 6、表 7：

Table 6. ARMA (2, 3) model fitting without drift term
表 6. 不带漂移项的 ARMA(2, 3)模型拟合

日期	预测值	真实值	误差比
2019/12/18	5.005	5.090	0.085
2019/12/19	5.009	5.140	0.131
2019/12/20	5.010	5.120	0.110
2019/12/23	5.007	5.010	0.003
2019/12/24	5.005	5.250	0.245
2019/12/25	5.007	5.180	0.173
2019/12/26	5.010	5.150	0.140
2019/12/27	5.009	5.250	0.241
2019/12/30	5.006	5.190	0.184
2019/12/31	5.006	5.140	0.134

Table 7. ARMA(2, 3) model fitting with drift term
表 7. 带漂移项的 ARMA(2, 3)模型拟合

日期	预测值	真实值	误差比
2019/12/18	5.001	5.090	0.089
2019/12/19	5.001	5.140	0.139
2019/12/20	4.998	5.120	0.122
2019/12/23	4.990	5.010	0.020
2019/12/24	4.984	5.250	0.266
2019/12/25	4.982	5.180	0.198
2019/12/26	4.980	5.150	0.170
2019/12/27	4.975	5.250	0.275
2019/12/30	4.968	5.190	0.222
2019/12/31	4.964	5.140	0.176

图 9 中：虚线为序列观察值，实线为模型拟合值与预测值，深色阴影部分为预测值的 80%置信区间和浅色阴影部分为预测值的 95%置信区间。输出结果见表 8，预测效果图见图 10：

Table 8. Model prediction results of A share without drift term
表 8. A 股份不带漂移项的模型预结果

预测日期	预测值	80%置信区间	95%置信区间
1	5.135	(4.778, 5.491)	(4.589, 5.68)
2	5.137	(4.617, 5.657)	(4.342, 5.932)
3	5.139	(4.498, 5.78)	(4.159, 6.119)
4	5.137	(4.389, 5.886)	(3.993, 6.282)
5	5.135	(4.29, 5.98)	(3.843, 6.427)
6	5.136	(4.208, 6.064)	(3.717, 6.555)
7	5.138	(4.137, 6.139)	(3.608, 6.669)
8	5.138	(4.067, 6.209)	(3.50, 6.776)
9	5.136	(3.996, 6.276)	(3.392, 6.879)
10	5.135	(3.931, 6.34)	(3.293, 6.977)

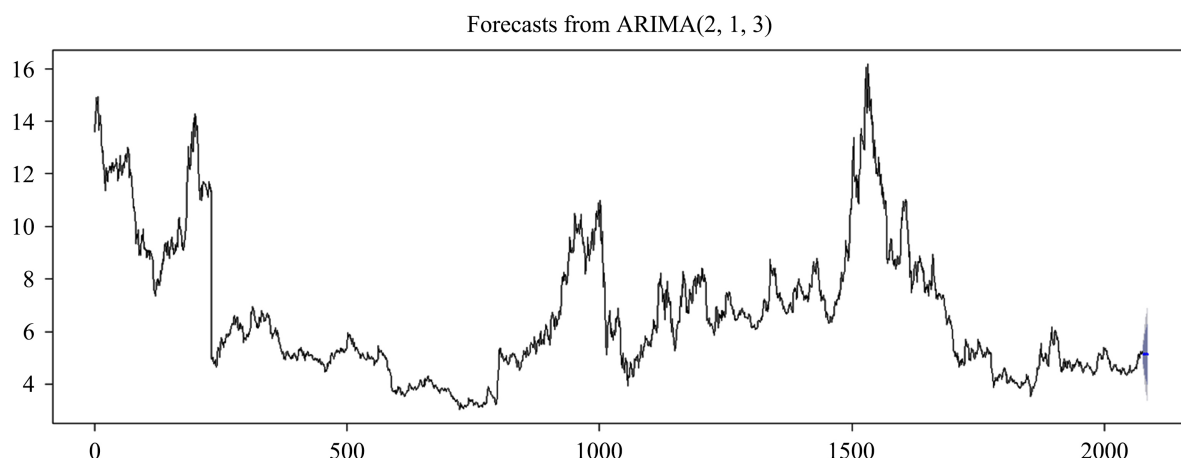


Figure 10. Effect diagram of model prediction without drift term

图 10. 不带漂移项的模型预测效果图

根据这个方法可得太平洋不带漂移项的预测效果较优于带漂移项模型的预测效果。预测结果见表 9:

Table 9. Model prediction results of B share without drift term

表 9. B 股份不带漂移项的模型预结果

预测日期	预测值	80%置信区间	95%置信区间
1	3.759	(3.451, 4.067)	(3.287, 4.231)
2	3.782	(3.341, 4.223)	(3.107, 4.457)
3	3.758	(3.222, 4.295)	(2.938, 4.579)
4	3.768	(3.155, 4.381)	(2.83, 4.706)
5	3.745	(3.069, 4.422)	(2.71, 4.781)
6	3.773	(3.03, 4.516)	(2.637, 4.909)
7	3.763	(2.961, 4.565)	(2.537, 4.99)
8	3.771	(2.905, 4.637)	(2.446, 5.095)
9	3.770	(2.844, 4.696)	(2.354, 5.187)
10	3.761	(2.771, 4.75)	(2.247, 5.274)

5. 模型优缺点

5.1. 优点

收集 A 股份、B 股份两只股票的交易日收盘价数据, 并进行整理分析, 建立了拟合效果最优的 $ARMA(p, q)$ 模型。从不带漂移项模型预测未来 10 日的股票价格走势, 预测结果总体较好。因此从这些数据中可以得出这种方法对股票的预测具有较好的效果。股票市场变化多端, 其趋势一般是不稳定的, 通过 ARMA 模型预测短期内的股票价格具有一定的实践意义, 投资者可以根据模型预测结果判断某只股票

短期内的盈亏变化，有助于投资者理性理财。

5.2. 缺点

ARMA 模型主要用于平稳序列，首先在数据方面的处理，可能会产生较小的误差，因此在股票每日收盘价的时候，就会存在有一定的缺失值，与此同时也会导致模型拟合的时候存在有基本的误差；股票市场是一个不稳定且特别容易受到外界诸多因素影响，其中最容易发现的则是价格的变化，它会受到开盘价、成交量，乃至大盘的影响；除此之外，国家的各种政策、国际经济贸易所产生的变化在一定因素上都会对股票市场产生一定影响，而以上说到的这些因素都是不可预测。在现实中也会有很多客观的因素可能会导致股票价格变化，然而 ARMA 模型针对短期预测具有较好的效果，而对股票长期走势进行的预测可能就不会那么理想。

6. 研究结论

本文根据平稳序列得到的样本自相关图和样本偏自相关图判断模型为 $ARMA(p, q)$ 模型；其次利用条件最小二乘估计方法，建立起 ARMA 拟合模型口径；最后通过比较模型的相关检验结果，选出了效果最好的 $ARIMA(p, d, q)$ 模型，云铝股份模型是 $ARIMA(2, 1, 3)$ 、太平洋股份模型是 $ARIMA(2, 1, 25)$ ；从不带漂移项的预测结果来看，可以知道 ARMA 模型对股票价格进行短期预测的效果较好。

基金项目

红河学院大学生科技创新基金项目(SZ2019)。

参考文献

- [1] 高远. 基于 ARMA 模型对乐视网股价预测的实证研究[J]. 现代经济信息, 2015(7): 352.
- [2] 谢海宁, 尚园. 基于 ARMA 模型分析的股价预测以中信银行为例[J]. 经济与管理科学, 2020(12): 122-124.
- [3] 彭紫君. 基于 ARMA-GARCH 模型的上证指数收益率分析[J]. 中南财经政法大学研究生学报, 2017(4): 48-55.
- [4] 陈瑶. 考虑流动性风险的 ARMA 模型在股票收益率中的应用研究[J]. 河南科学, 2014(4): 655-659.
- [5] 马紫璇. ARMA 模型在股票短期预测中的应用[J]. 广西质量监督导报, 2020(5): 152-153.