S&P 500 数据的建模与预测：基于线性时间序列模型*

统计 91 董晟渤, 2193510853

西安交通大学数学与统计学院

日期：2022 年 4 月 28 日

摘　　要

S&P 500 是美国股市的一个具有代表性的指数. 本篇报告对 2019 年 1 月 2 日至 2021 年 12 月 30 日的 S&P 500 的收盘价的对数值, 首先基于 ACF、PACF 与 EACF, 筛选了合适的模型; 其次, 基于 AIC 与 BIC, 比较了不同的模型, 并为最小化 AIC, 选择模型 ARIMA$(4,1,4)$; 为了对该模型进行参数估计, 报告中比较了 ML 估计与 CSS 估计的结果, 并选用 CSS-ML 估计的结果, 从而建立了模型; 在建立了 ARIMA$(4,1,4)$ 模型之后, 报告中对残差的正态性与相关性进行了检验, 得到了残差具有正态性与平稳性的结论, 说明了模型的合理性; 最后, 基于该模型, 给出了近期的预测结果. 本篇报告实现了对真实数据进行线性时间序列建模. 报告的最后, 总结了建模的步骤.

关键词： S&P 500, 时间序列, ARMA 模型, ARIMA 模型.

*2021-2022 学年第二学期, 课程: 时间序列与金融统计, 指导老师: 惠永昌.

目录

1 概述

1.1 问题背景

标准普尔 500 指数 (Standard & Poor's 500), 简称 S&P 500, 是一个由 1957 年起记录美国股市的平均记录, 观察范围达 500 只普通股, 占总市值约 80%. 标准普尔 500 指数由标普道琼斯指数公司 (S&P Dow Jones Indices LLC, 标准普尔全球控股公司控制的合资公司) 开发并继续维持. S&P 500 的代表性够强, 甚至足以显示美国经济的兴衰. [1]

本篇报告中, 将对 S&P 500 从 2019 年 1 月 2 日至 2021 年 12 月 30 日的历史数据, 建立时间序列模型.

1.2 线性时间序列模型

常用的线性时间序列模型有 MA 模型、AR 模型、ARMA 模型和 ARIMA 模型. [2]

在介绍时间序列模型时, 首先引入的是 $\mathrm{MA}(q)$ 模型, 其中 $q \geq 0$ 是给定的整数. 其中, "MA" 的含义是滑动平均 (Moving Average).

定义 1.1 (MA 模型)**.** 设 $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$, 整数 $q \geq 0$, 若对常数 $\mu, a_1, a_2 \cdots, a_q$, 有

$$X_t = \mu + \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_q\varepsilon_{t-q},$$

则记 $X_t \sim \mathrm{MA}(q)$, 称 q 为阶数.

另外一类重要的模型是 $\mathrm{AR}(p)$ 模型, 其中 $p \geq 0$ 是给定的整数. 其中, AR 的含义是自回归 (Auto Regressive).

定义 1.2 (AR 模型)**.** 设 $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$, 整数 $q \geq 0$, 若对常数 $c, b_1, b_2, \cdots, b_p$, 有

$$X_t = c + b_1X_{t-1} + \cdots + b_pX_{t-q} + \varepsilon_t,$$

则记 $X_t \sim \mathrm{AR}(p)$, 称 p 为阶数.

基于 $\mathrm{MA}(q)$ 模型和 $\mathrm{AR}(p)$ 模型, 我们提出 $\mathrm{ARMA}(p, q)$ 模型, 其中 $p, q \geq 0$ 是给定的整数. 该模型同时包含了滑动平均和自回归的性质.

定义 1.3 (ARMA 模型)**.** 设 $\varepsilon_t \sim \mathrm{WN}(0, \sigma^2)$, 整数 $p, q \geq 0$, 若对常数 $c, a_1, a_2, \cdots, a_q$ 及 $b_1, b_2, \cdots, b_p$, 有

$$X_t = c + b_1X_{t-1} + \cdots + b_pX_{t-q} + \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_q\varepsilon_{t-q},$$

则记 $X_t \sim \mathrm{ARMA}(p, q)$, 称 p, q 为阶数.

[1] 关于 S&P 500 的介绍, 详见 [1].

[2] 关于时间序列模型的介绍, 详见 [2].

在建立时间序列模型时, 可能需要对数据进行差分, 以保证平稳性. 设进行的差分次数为正整数 $d \geq 1$, 定义差分算子 ∇, 满足

$$\nabla X_t = X_t - X_{t-1}, \quad \nabla^d X_t = \nabla\left(\nabla^{d-1} X_t\right), \quad \forall k \geq 2.$$

定义 1.4 (ARIMA 模型)**.** 设整数 $p, q \geq 0$, $d \geq 1$, 若 $\nabla^d X_t \sim \mathrm{ARMA}(p, q)$, 则记 $X_t \sim \mathrm{ARIMA}(p, d, q)$, 称 p, q 为阶数, d 为差分次数.

以上所介绍的模型, 将在建模中应用. 关于模型选择、参数估计、模型诊断和预测的方法, 均在 [2] 中介绍, 本篇报告将忽略一些繁琐的细节, 将关注点放在建模方法的应用上.

2 准备工作: 数据收集与预处理

2.1 S&P 500 历史数据的收集

本篇报告所用的数据为 S&P 500 从 2019 年 1 月 2 日至 2021 年 12 月 30 日的历史数据, 数据来自英为财情[3], 可供下载的数据有每月、每周和每日的历史数据. 为了建模的准确, 本篇报告在建立模型时, 选取的是按日的数据.

表 1: S&P 500 从 2019 年 1 月 2 日至 2021 年 12 月 30 日的历史数据 (按日)

日期	收盘价	开盘价	最高价	最低价
2019 年 1 月 2 日	2,510.03	2,476.96	2,519.49	2,467.47
2019 年 1 月 3 日	2,447.89	2,491.92	2,493.14	2,443.96
2019 年 1 月 4 日	2,531.94	2,474.33	2,538.07	2,474.33
2019 年 1 月 7 日	2,549.69	2,535.61	2,566.16	2,524.56
...	...	...	...	...
2021 年 12 月 29 日	4,793.06	4,788.64	4,804.06	4,778.08
2021 年 12 月 30 日	4,778.73	4,794.23	4,808.93	4,775.33

表 1列出了 S&P 500 的部分历史数据. 其中:

- 收盘价是当日该股最后一笔交易前一分钟所有交易的成交量加权平均价;
- 开盘价是当日该股第一笔交易的每股成交价;
- 最高价和最低价是当日该股最高的每股成交价和最低的每股成交价.

收盘价是市场参与者们所共同认可的价格, 因此本报告建模的对象是收盘价.

[3]历史数据见https://cn.investing.com/indices/us-spx-500.

2.2 报告中符号的说明

本报告中使用的符号、符号说明及定义如**表 2**所示.

表 2: 符号说明

符号	符号说明	来源或定义
t	日期, 取值从 1 到 T	**表 1**中的数据
P_t	t 时刻的收盘价	**表 1**中的数据
p_t	t 时刻的对数收盘价	$p_t := \ln P_t$
r_t	t 时刻的对数收益率	$r_t := \nabla p_t$

2.3 数据的图像与预处理

为了处理数据, 记 2019 年 1 月 2 日对应 $t = 1$, 2021 年 12 月 30 日对应 $t = T$.

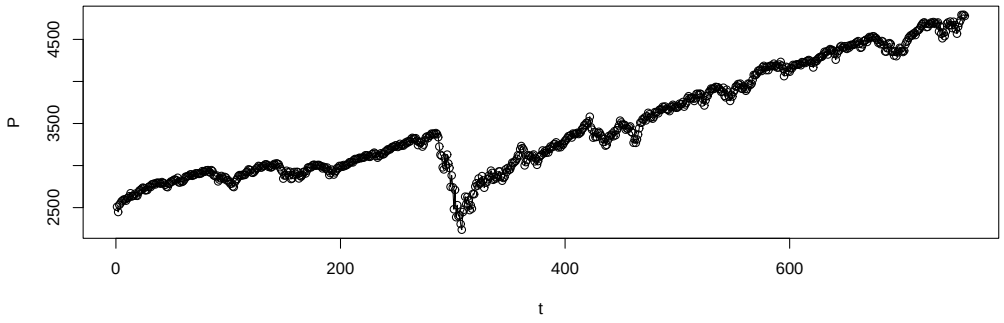


图 1: 收盘价 P_t 随时间 t 变化的图像

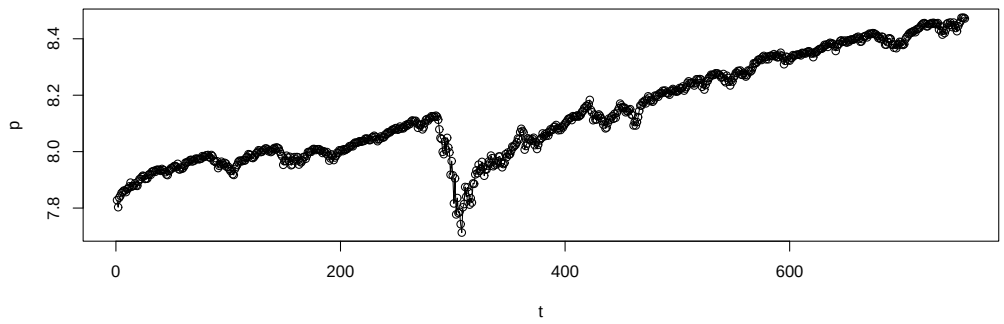


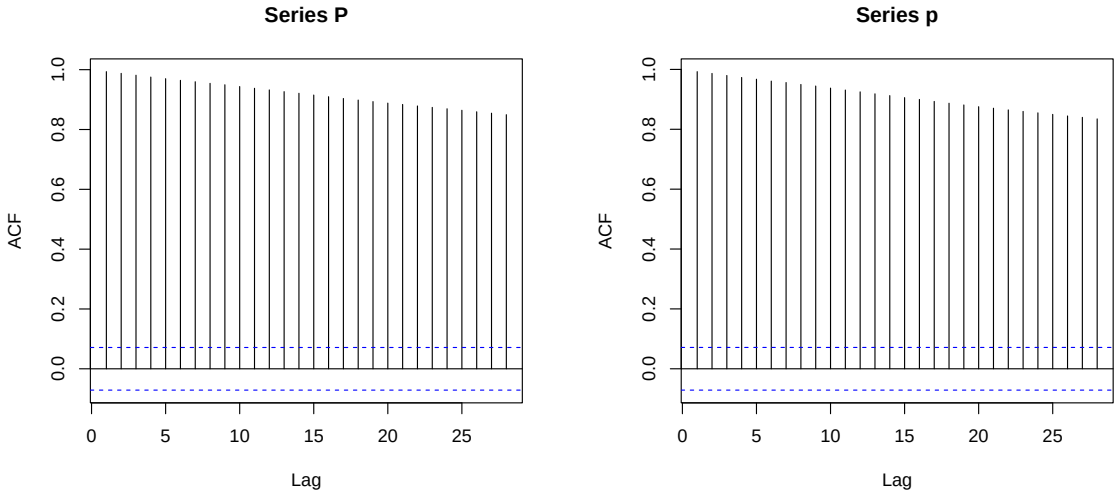
图 2: 对数收盘价 p_t 随时间 t 变化的图像

首先, 绘制出 $t - P_t$ 图像如**图 1**所示. 对于金融数据, 为了看出其规律, 常常会进行取对数的操作. 记对数收盘价

$$p_t := \ln P_t, \quad t = 1, 2, \cdots, T,$$

与**图 1**相对比, 绘制出 $t - p_t$ 图像如**图 2**所示.

接下来, 希望能对 P_t 或 p_t, 建立时间序列模型. 在此之前, 还需要考虑 P_t 和 p_t 的平稳性或相关性. 用于检验平稳性的方法有查看自相关函数 (ACF) 的图像, 以及 ADF 检验.



(a) P_t 的自相关函数图像　　(b) p_t 的自相关函数图像

图 3: P_t 和 p_t 对应的自相关函数图像

一方面, P_t 和 p_t 对应的自相关函数见**图 3**. 不管是对于 P_t 还是 p_t, 自相关函数都接近 1, 且缓慢递减, 说明该过程具有长记忆性, 从而不具有平稳性; 另外一方面, 对 P_t 和 p_t 进行 ADF 检验, 得到的 p 值分别为 0.4746 和 0.2155, 故认为 P_t 和 p_t 是不平稳的.

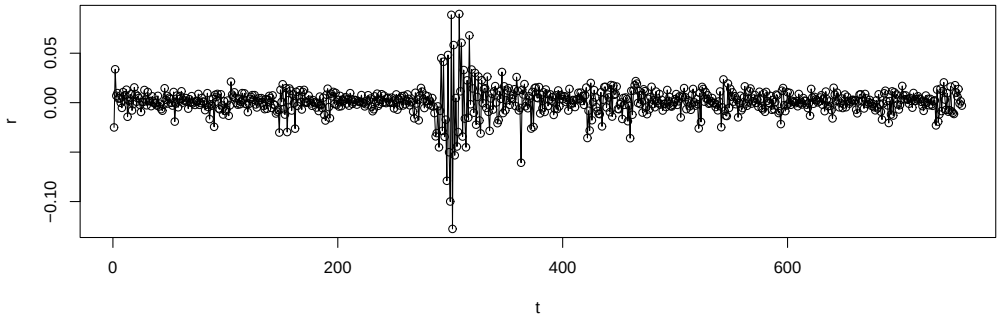


图 4: 对数收益率 r_t 随时间 t 变化的图像

为了处理不平稳的数据, 可以采用差分的方法. 考虑对对数收盘价 p_t 进行差分, 记

$$r_t := \nabla p_t = p_t - p_{t-1} = \ln \frac{P_t}{P_{t-1}}, \quad t = 2, 3, \cdots, T,$$

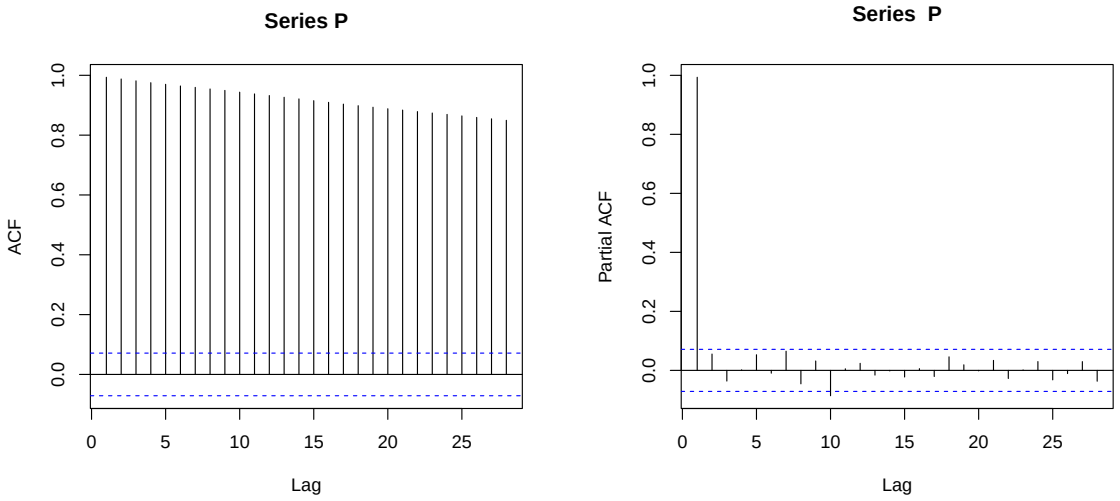
这通常被称为对数收益率. 绘制出 $t - r_t$ 图像如**图 4**所示. 对 r_t 进行 ADF 检验, 所得的 p 值为 0.01, 说明一次差分即可得到平稳的数据. 在建立模型时, 也可以考虑对 r_t 建立 ARMA(p, q) 模型, 这相当于对 p_t 建立了 ARIMA$(p, 1, q)$ 模型.

3 建立模型: 模型选择与参数估计

3.1 基于 ACF、PACF 与 EACF 的模型选择

在建立 MA 模型时, 我们可以使用自相关函数 (ACF) 的图像来估计阶数; 相应地, 在建立 AR 模型时, 我们可以使用偏自相关函数 (PACF) 的图像来估计阶数; 对于 ARMA 模型而言, 使用 ACF 和 PACF 都难以解决问题, 但可以用推广的自相关系数 (EACF) 来估计阶数.

对于 P_t, 绘制 ACF 图和 PACF 图如**图 5**所示.



(a) 自相关函数图像　　(b) 偏自相关函数图像

图 5: P_t 的自相关函数图像和偏自相关函数图像

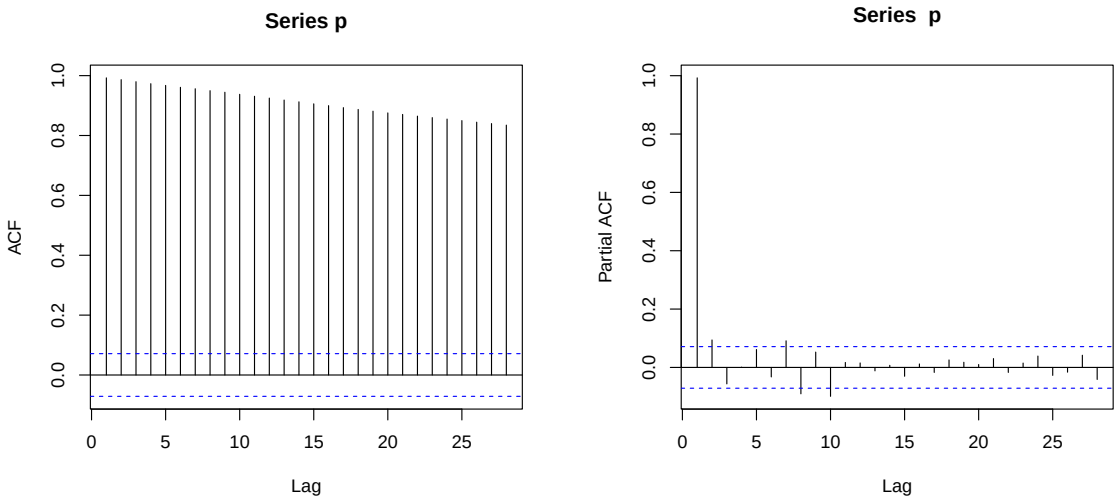
另外, 对于 P_t, 计算得到的 EACF 为:

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x o x o x x x x x o  o  x  x
2 x x o x o o x o x o o  o  o  o
```

```
3 x x o x o o x o x o o  o  o  o
4 x x o x o o o o x o o  o  o  o
5 x x x x o o o o x o o  o  o  o
6 x x x x o o o o x o o  o  o  o
7 x x x x x x o o x o o  o  o  o
```

根据上述结果, 可以对 P_t 建立模型 AR(1) 和 ARMA$(2,4)$.

对于 p_t, 绘制 ACF 图和 PACF 图如**图 6**所示.



(a) 自相关函数图像　　(b) 偏自相关函数图像

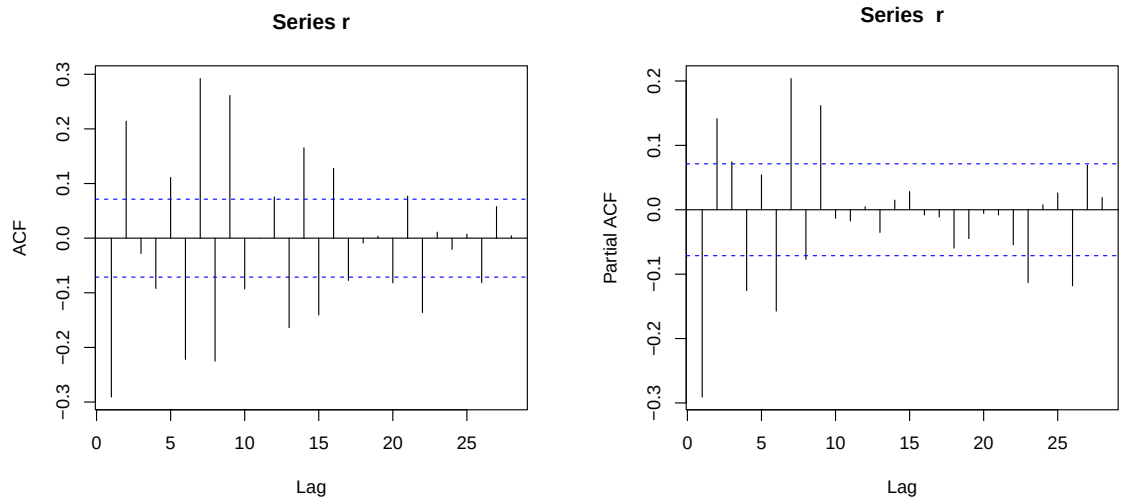
图 6: p_t 的自相关函数图像和偏自相关函数图像

另外, 对于 p_t, 计算得到的 EACF 为:

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x x x  x  x  x
1 x x o x x x x x x x o  x  x  x
2 x x o o o o x o x x o  o  o  o
3 x x o x o o x o x o o  o  o  x
4 x x o x o o x o x o o  o  o  o
5 x x x x o o o o x o o  o  o  o
6 x x x x x x o o x o o  o  o  o
7 x x x x x x o o x o o  o  o  o
```

根据上述结果, 可以对 p_t 建立模型 AR(2) 和 ARMA$(2,3)$.

对于 r_t, 绘制 ACF 图和 PACF 图如**图 7**所示.



(a) 自相关函数图像　　(b) 偏自相关函数图像

图 7: r_t 的自相关函数图像和偏自相关函数图像

另外, 对于 r_t, 计算得到的 EACF 为:

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o x x x x x x x o  x  x  x
1 x x o o o o x o x x o  o  o  o
2 x x o x o o x o x o o  o  o  o
3 x x o x o o x o x o o  x  o  o
4 x x x x o o o o x o o  o  o  o
5 x x x x o x o o x o o  o  o  o
6 x x x x x x o o x o o  o  o  o
7 x x o x o x x o o o o  o  o  o
```

根据上述结果, 可以对 r_t 建立模型 MA(2)、AR(2) 和 ARMA$(2,3)$.

对于 P_t, p_t 和 r_t, 根据 ACF、PACF 和 EACF 所得到的模型如**表 3**所示. 需要注意, r_t 的 ARMA(p,q) 模型实质上就是 p_t 的 ARIMA$(p,1,q)$ 模型, 在**表 3**中不再另外列出.

表 3: 由 ACF、PACF 和 EACF 得到的 MA 模型、AR 模型和 ARMA 模型

变量	MA 模型	AR 模型	ARMA 模型
P_t	-	AR(1)	ARMA$(2,4)$
p_t	-	AR(2)	ARMA$(2,3)$
r_t	MA(2)	AR(4)	ARMA$(1,3)$

3.2 基于 AIC 与 BIC 的模型选择

另外一种进行模型选择的方式是在假定 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 的条件下, 计算模型的 AIC 和 BIC, 并选择使得 AIC 或者 BIC 最小的模型. 在本篇报告中, 主要考虑的是最小化 AIC, 并且为了方便, 仅考虑对 p_t 和 r_t 建立模型.

对于 p_t, 各个模型的 AIC 如**表 4**所示. 表中, 空出来的部分 (例如 $\mathrm{ARMA}(3,3)$) 代表建立的模型不是平稳的, 或者进行参数估计时算法不收敛, 导致无法计算 AIC.

表 4: p_t 的 ARMA 模型的 AIC

AR/MA	0	1	2	3	4	5
0	-446.2215	-1426.307	-2104.07	-2625.097	-2931.41	-3250.085
1	-4269.619	-4314.26	-4351.963	-4352.314	-4351.536	-4349.143
2	-4331.163	-4338.882	-4338.294	-4372.177	-4379.545	-4378.7
3	-4346.388	-4346.903	-4350.897	-	-4411.69	-
4	-4349.886	-4398.016	-4412.484	-	-4409.81	-4407.952
5	-4358.985	-4401.016	-4411.971	-4410.062	-	-4418.987

对于 r_t 模型, 建立的 $\mathrm{ARMA}(p,q)$ 模型相当于是 p_t 的 $\mathrm{ARIMA}(p,1,q)$ 模型. 各个模型的 AIC 如**表 5**所示.

表 5: r_t 的 ARMA 模型 (也即 p_t 的 ARIMA 模型, $d=1$) 的 AIC

AR/MA	0	1	2	3	4	5
0	-4274.138	-4321.189	-4357.754	-4358.596	-4357.39	-4355.799
1	-4339.095	-4345.62	-4357.895	-4357.69	-4386.073	-4385.047
2	-4352.764	-4352.362	-4356.563	-4355.847	-4418.071	-4382.426
3	-4355.065	-4403.846	-4362.345	-4417.018	-4415.055	-4418.239
4	-4365.146	-4407.489	-4417.687	-4415.929	-4426.883	-4425.76
5	-4365.36	-4406.836	-4403.653	-4410.474	-4425.699	-4423.76

综合以上结果, 为最小化 AIC, 对 r_t 建立的模型为 $\mathrm{ARMA}(4,4)$, 此时对 p_t 建立的模型为 $\mathrm{ARIMA}(4,1,4)$, AIC 的值为-4426.883.

3.3 ARIMA 模型参数的估计

根据以上过程, 我们对 r_t 建立了 ARMA$(4,4)$ 模型, 也即

$$r_t = c + \sum_{i=1}^{4} b_i r_{t-i} + \varepsilon_t + \sum_{j=1}^{4} a_j \varepsilon_{t-j}.$$

然而, 模型里面涉及的参数 c, $\boldsymbol{b} = (b_1, b_2, b_3, b_4)$ 和 $\boldsymbol{a} = (a_1, a_2, a_3, a_4)$ 还没有确定下来. 为了估计 $(c, \boldsymbol{b}, \boldsymbol{a})$, 我们可以考虑的方法有极大似然估计 (ML) 和最小二乘估计 (CSS). 其中, ML 方法需要假定 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. 除此之外, R 还提供了使用最小二乘估计初值、使用极大似然估计参数的 CSS-ML 方法. 分别使用这些估计方法, 所得的参数的值见**表 6**.

表 6: ARMA$(4,4)$ 模型的参数估计

方法	$\hat{c}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$
ML	9e-04	-0.5746	0.2124	-0.5326	-0.7884	0.4126	-0.1346	0.6688	0.5890
CSS	9e-04	-0.6555	0.1600	-0.4306	-0.6854	0.4978	-0.0751	0.5961	0.5190
CSS-ML	9e-04	-0.5756	0.2117	-0.5322	-0.7883	0.4138	-0.1341	0.6686	0.5894

根据**表 6**, 使用三种方法估计出来的参数都较为接近. 考虑到在应用 AIC 选择模型时, 假定了 $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, 因此我们更倾向于使用 CSS-ML 方法估计出的参数. 从而, 我们最终建立的模型为

$$\begin{aligned} r_t = {} & 9 \times 10^{-4} - 0.5756 \times r_{t-1} + 0.2117 \times r_{t-2} - 0.5322 \times r_{t-3} - 0.7883 \times r_{t-4} \\ & + \varepsilon_t + 0.4138 \times \varepsilon_{t-1} - 0.1341 \times \varepsilon_{t-2} + 0.6686 \times \varepsilon_{t-3} + 0.5894 \times \varepsilon_{t-4}, \end{aligned} \tag{1}$$

其中对数收益率

$$r_t = \nabla p_t = p_t - p_{t-1} = \ln P_t - \ln P_{t-1} = \ln \frac{P_t}{P_{t-1}}.$$

根据模型(1), 估计得到 $\hat{\sigma}^2 = 0.0001638$. 此时, 对数似然为 2223.44, AIC 为-4426.88. $\left(\hat{c}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{a}}\right)$ 的估计值和误差如**表 7**所示.

表 7: ARMA$(4,4)$ 模型的 ML 估计

类别	$\hat{c}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$
估计值	9e-04	-0.5756	0.2117	-0.5322	-0.7883	0.4138	-0.1341	0.6686	0.5894
标准差	4e-04	0.0642	0.0693	0.0774	0.0577	0.0745	0.0601	0.0640	0.0638

4 模型诊断: 残差分析

本节中, 我们将对模型(1)的残差 ε_t 进行诊断. 根据该模型, 作出 $t-\varepsilon_t$ 图如**图 8**所示.

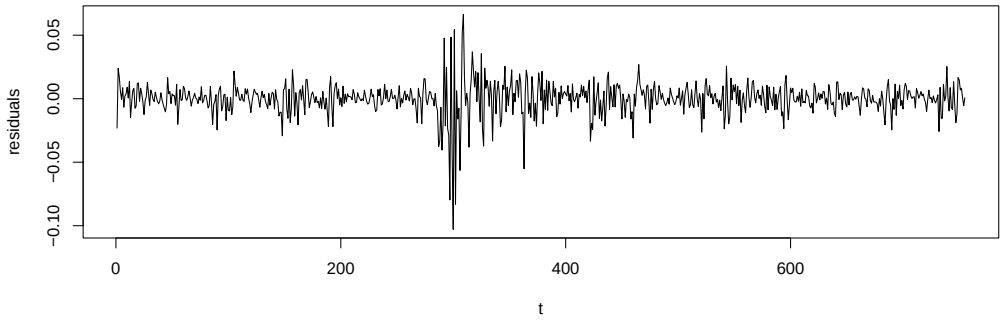


图 8: 残差 ε_t 随时间 t 变化的图像

4.1 残差的正态性检验

在建立模型时, 我们假定 $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$, 得到了模型(1). 然而, 真实数据的残差 ε_t 不一定是来自正态分布的. 为了对残差进行正态性检验, 可以作出残差的 Q-Q 图.

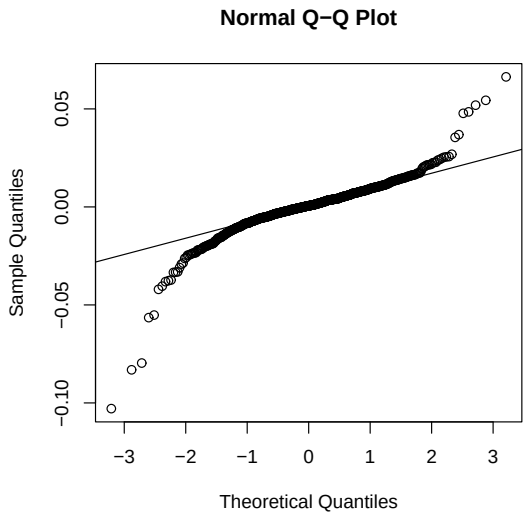


图 9: 残差 ε_t 的 Q-Q 图

作出残差的 Q-Q 图如**图 9**所示. 注意到在图中, 大部分点围绕着中间的直线, 有少部分的点离直线较远, 说明了 ε_t 可以近似为正态分布.

4.2 残差的相关性检验

在建立模型时, 我们还假定了残差至少是白噪声过程. 从而, 它们一定是不相关的.

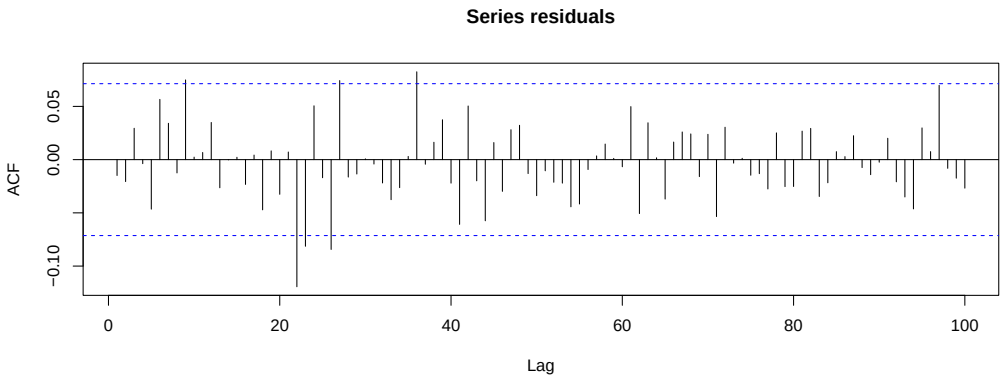


图 10: 残差 ε_t 的自相关函数图像

首先, 作出自相关函数 (ACF) 的图像如**图 10**所示. 根据图像可以看出, ACF 的取值几乎都落在置信区间内, 从而可以认为 ε_t 之间是不相关的.

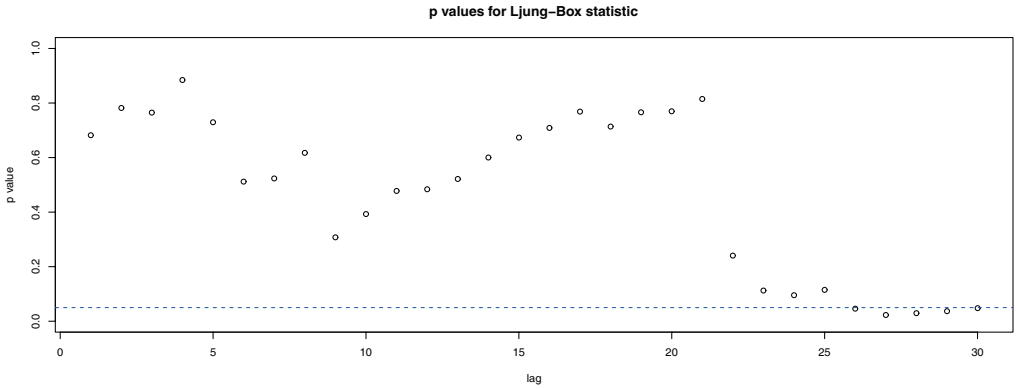


图 11: 残差 ε_t 进行 LB 检验的 p 值

为了进一步检验残差之间的相关性, 在这里进行 LB 检验, 得到的 p 值如**图 11**所示. 根据图像得知, 可以认为 ε_t 之间的相关性较弱.

5 模型应用: 未来数据的预测

在前面几节中, 我们完成了模型选择、模型参数的估计和模型诊断等工作, 成功地建立了模型(1). 基于我们所建立的模型(1), 也即对 r_t 的 ARMA$(4,4)$ 模型, 可以通过取条件期

望的方式, 对未来数据进行简单的预测. 首先, 对对数收益率 r_t 进行预测, 所得的结果如图 12所示.

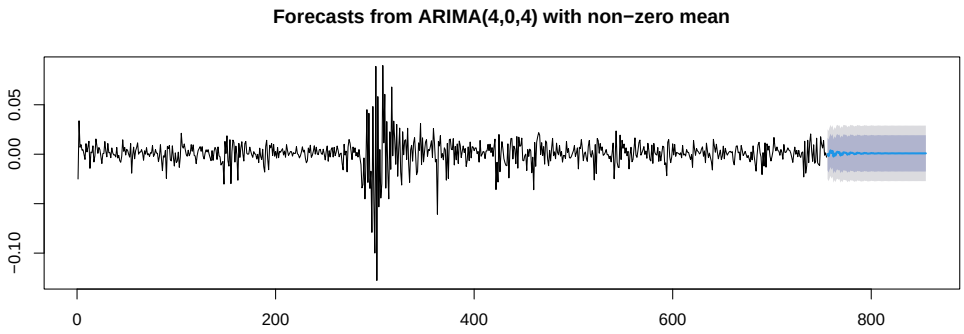


图 12: 利用 ARMA$(4,4)$ 模型对 r_t 进行预测

模型(1)对应着 p_t 的 ARIMA$(4,1,4)$ 模型. 在图 12的预测的基础上, 我们可以得到对数收盘价的预测, 结果如图 13所示.

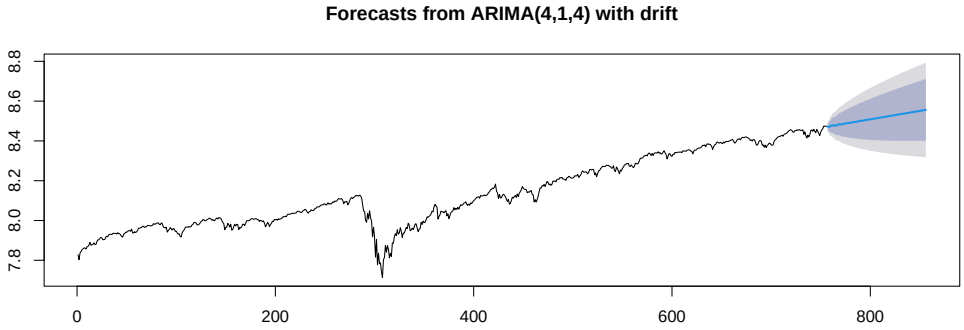


图 13: 利用 ARIMA$(4,1,4)$ 模型对 p_t 进行预测

6 总结

本篇报告实现了对真实数据的线性时间序列建模. 总而言之, 建立线性时间序列模型, 需要进行的步骤包括:

- 模型初筛, 选择合适的模型 (如查看 ACF 图、PACF 图或 EACF, 最小化 AIC 或 BIC);
- 参数估计, 对选择的模型, 选取合适的方法 (如 ML、CSS 或 CSS-ML) 估计模型的参数;
- 模型诊断, 对模型的 ε_t 进行正态性检验 (如画出 Q-Q 图) 和相关性检验 (如 LB 检验).

在建模完成后, 可以基于所建立的模型, 进行未来数据的预测.

参考文献

[1] 维基百科. 标准普尔 *500* 指数. https://zh.wikipedia.org/wiki/S%26P_500.

[2] Jianqing Fan & Qiwei Yao. *The Elements of Financial Econometrics*. Science Press.

附录

A 所用软件

在编写本篇报告的过程中, 主要使用的软件为:

- Microsoft Excel, 用于整理收集所得的数据;
- 基于 RStudio 的 R, 用于数据处理与绘制图像;
- 基于 Visual Studio Code 的 LaTeX, 用于排版报告.

B 代码

以下代码使用 R 编写.

```r
# 调用程序包

gc()
library(tseries)
library(TSA)
library(forecast)

# 原始数据

data = read.csv("Data_1.csv", header = TRUE)
P = as.numeric(data[, 2])
plot(P, xlab = "t", ylab = "P") # 图像
lines(P)
acf(P) # ACF图像
adf.test(P) # ADF检验

# 对数收盘价

p = log(P)
plot(p, xlab = "t", ylab = "p") # 图像
lines(p)
```

```
acf(p) # ACF图像
adf.test(p) # ADF检验

# 对数收益率

r = diff(p) # 差分
plot(r, xlab = "t", ylab = "r") # 图像
lines(r)
acf(r) # ACF图像
pacf(r) # PACF图像
eacf(r) # EACF表

# 建立模型

r.fit = auto.arima(r)
p.fit = auto.arima(p)

# 残差诊断

plot(r.fit $ residuals, xlab = "t", ylab = "residuals") # 残差图像
tsdiag(r.fit, gof.lag = 30) # 残差的检验
qqnorm(r.fit $ residuals) # 残差的QQ图
qqline(r.fit $ residuals)
Box.test(r.fit $ residuals, type = "Ljung-Box") # 残差的LB检验

# 预测


plot(forecast(r.fit, 100))
plot(forecast(p.fit, 100))
```