时间序列分析实验报告书

班级：统计 2001 姓名：张逸敏　　实验日期：2023.4.18

实验一　ARMA 模型的建模与预测

1 实验目的

R 语言的基本操作，针对平稳序列建立 ARMA 模型并做预测。

2 实验条件

PC 机，R 语言

3 实验原理

3.1 AR 模型

如果误差项$\{\varepsilon_t\}$是白噪声$WN(0,\sigma^2)$，实数$a_1,a_2,\ldots,a_p\,(a_p\neq 0)$使得多项式的零点都在单位圆外:

$$A(z)=1-\sum_{j=1}^{p}a_jz^j\neq 0,\ |z|\leq 1$$

则称p阶差分方程

$$X_t=\sum_{j=1}^{p}a_jX_{t-j}+\varepsilon_t,\ \ t\in\mathbb{Z}$$

是一个阶自回归模型，简称为$AR(p)$模型.

满足 AR 模型的平稳时间序列称为的平稳解，也称作 AR(p)序列.称$\boldsymbol{a}=(a_1,a_2,\cdots,a_p)^T$是 AR(p)模型的自回归系数. $A(z)$称为模型的特征多项式。模型可用推移算子写成

$$A(\mathscr{B})X_t=\varepsilon_t,\quad t\in\mathbb{Z}$$

3.2 MA 模型

设$\{\varepsilon_t\}$是 WN$(0,\sigma^2)$，如果实数$b_1,b_2,\ldots,b_q\,(b_q\neq 0)$使得

$$B(z)=1+\sum_{j=1}^{q}b_jz^j\neq 0,\ |z|<1,$$

则称

$$X_t=\varepsilon_t+\sum_{j=1}^{q}b_j\varepsilon_{t-j},\quad t\in\mathbb{Z}$$

是q阶滑动平均模型，简称为 MA(q)模型。

模型中的$\{X_t\}$显然是平稳列。称平稳序列$\{X_t\}$是滑动平均序列，简称为MA(q)序列.

3.3 ARMA 模型

设$\{\varepsilon_t\}$是 WN$(0,\sigma^2)$，实系数多项式$A(z)$和$B(z)$没有公共根，满足$b_0=1$，$a_p b_q \neq 0$和

$$A(z)=1-\sum_{j=1}^{p} a_j z^j \neq 0, |z| \leq 1,$$

$$B(z)=\sum_{j=0}^{q} b_j z^j \neq 0, \ |z|<1$$

就称差分方程：

$$X_t=\sum_{j=1}^{p} a_j X_{t-j}+\sum_{j=0}^{q} b_j \varepsilon_{t-j}, \quad t \in \mathbb{Z}$$

是一个自回归滑动平均模型，简称为 ARMA(p,q)模型. 称满足上式的平稳序列$\{X_t\}$为平稳解或 ARMA(p,q)序列.

模型写成

$$A(\mathscr{B}) X_t=B(\mathscr{B}) \varepsilon_t, \quad t \in \mathbb{Z}$$

4 实验过程与结果

4.1 实验案例表述

18. 某地区连续 74 年的谷物产量（单位：千吨）如表 3—21 所示（行数据）。

表 3—21

0.97	0.45	1.61	1.26	1.37	1.43	1.32	1.23	0.84	0.89	1.18
1.33	1.21	0.98	0.91	0.61	1.23	0.97	1.10	0.74	0.80	0.81
0.80	0.60	0.59	0.63	0.87	0.36	0.81	0.91	0.77	0.96	0.93
0.95	0.65	0.98	0.70	0.86	1.32	0.88	0.68	0.78	1.25	0.79
1.19	0.69	0.92	0.86	0.86	0.85	0.90	0.54	0.32	1.40	1.14
0.69	0.91	0.68	0.57	0.94	0.35	0.39	0.45	0.99	0.84	0.62
0.85	0.73	0.66	0.76	0.63	0.32	0.17	0.46			

（1）判断该序列的平稳性与纯随机性。

（2）选择适当模型拟合该序列的发展。

（3）利用拟合模型，预测该地区未来 5 年的谷物产量。

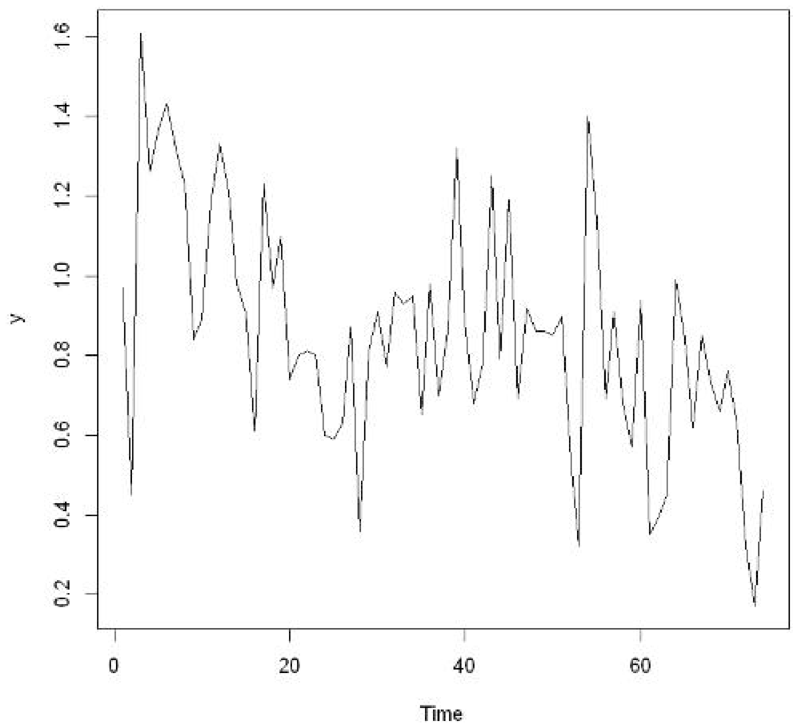
4.2 实验过程与代码

4.2.1 判断$\{x_t\}$的平稳性和纯随机性

首先读入数据，创建时间序列对象。

```
y = ts(c(0.97, 0.45, 1.61, 1.26, 1.37, 1.43, 1.32, 1.23, 0.84, 0.89, 1.18, 1.33, 1.21, 0.98, 0.91, 0.61, 1.23, 0.97, 1.10, 0.74, 0.80, 0.
plot(y)
```

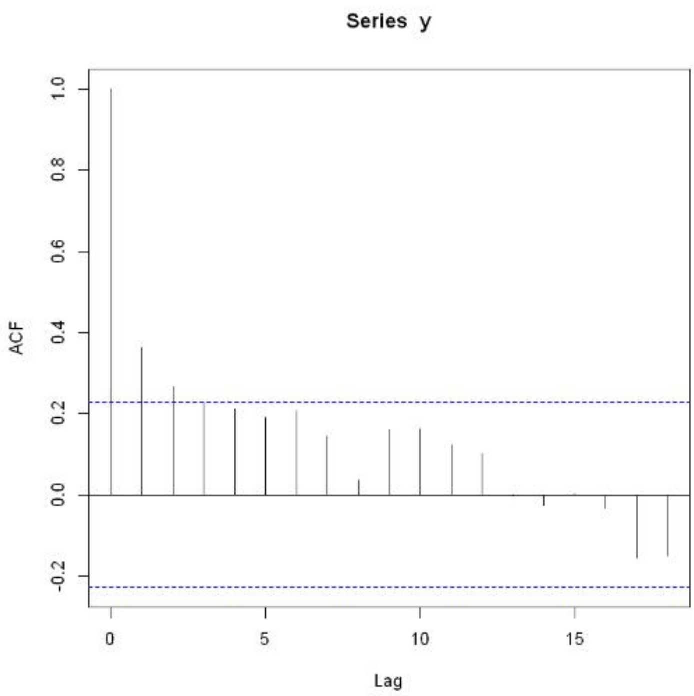
接下来绘制时序图。



观察时序图，没有发现明显的趋势或者周期，基本上可以视为平稳序列。

为了稳妥起见，还需要进行自相关图检验。

```
acf(y)
```



自相关图显示大多数自相关系数在 2 倍标准差以外，说明序列具有短期相关性，是平稳序列。

利用 LB 统计量进行纯随机性检验，结果如下。

```
Box.test(y, type = "Ljung-Box", lag=6)
Box.test(y, type = "Ljung-Box", lag=12)
```

```
        Box-Ljung test

data:  y
X-squared = 29.872, df = 6, p-value = 4.156e-05


        Box-Ljung test

data:  y
X-squared = 38.58, df = 12, p-value = 0.0001234
```

延迟 6 期的 LB 统计量和延迟 12 期的 LB 统计量的 p 值均<0.05，表示延迟期数小于等于 6、12 的序列值之间存在相关性，时间序列不具有纯随机性，值得继续进行建模。

4.2.2 选择适当模型拟合该序列的发展

选择$AR(1)$模型进行建模。

```
library(forecast)
fit = arima(y, order=c(1,0,0), include.mean = T)
fit
```

```
Call:
arima(x = y, order = c(1, 0, 0), include.mean = T)

Coefficients:
         ar1  intercept
      0.3681     0.8491
s.e.  0.1085     0.0499

sigma^2 estimated as 0.07467:  log likelihood = -9.07,  aic = 24.14
```

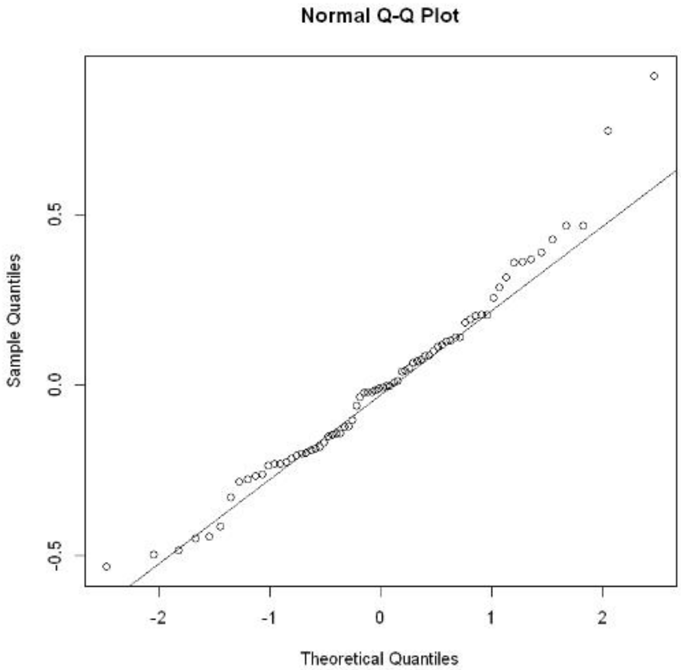
一般来说，一个模型如果合适，那模型的残差应该满足均值为 0 的正态分布，并且对于任意的滞后阶数，残差自相关系数都应该为零。换句话说，模型的残差应该满足独立正态分布（即残差间没有关联）。

下面进行对模型残差进行检验。

```
qqnorm(fit$residuals)
qqline(fit$residuals)
Box.test(fit$residuals, type="Ljung-Box")
```

```
	Box-Ljung test

data:  fit$residuals
X-squared = 0.25563, df = 1, p-value = 0.6131
```



如果数据呈现正态分布，那么数据点会落在图中直线上。显然，本模型的残差大致符合正态分布，模型效果不错。

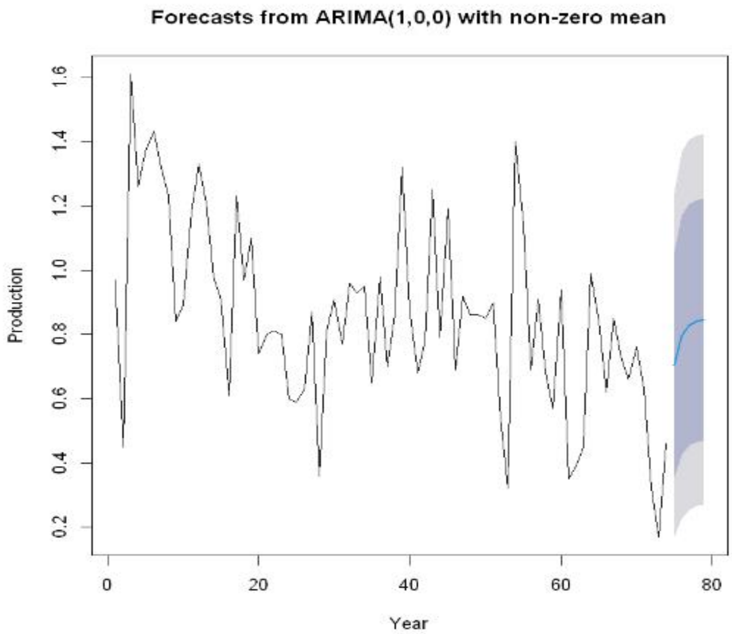
对残差的自相关检验，$p=0.6131 > 0.05$，接受原假设，即认为残差之间的自相关系数为 0，符合残差独立的假设。

4.2.3 利用拟合模型预测该地区未来 5 年的谷物产量

利用 forecast()函数，并画出 80%和 95%的置信区间。

```
forecast(fit, 5)
plot(forecast(fit, 5), xlab="Year", ylab="Production")
```

```
   Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
75      0.7058632 0.3556779 1.056049 0.1703010 1.241425
76      0.7963730 0.4232128 1.169533 0.2256736 1.367072
77      0.8296925 0.4535266 1.205858 0.2543964 1.404989
78      0.8419584 0.4653870 1.218530 0.2660422 1.417875
79      0.8464738 0.4698476 1.223100 0.2704737 1.422474
```



4.2.4 采用自动定阶函数建立模型

```
# 采用自动定阶
library(zoo)
library(forecast)
autofit = auto.arima(y)
autofit
```

```
Series: y
ARIMA(0,1,1) with drift

Coefficients:
          ma1    drift
      -0.8695  -0.0085
s.e.   0.0873   0.0045

sigma^2 = 0.06925:  log likelihood = -5.82
AIC=17.64   AICc=17.98   BIC=24.51
```

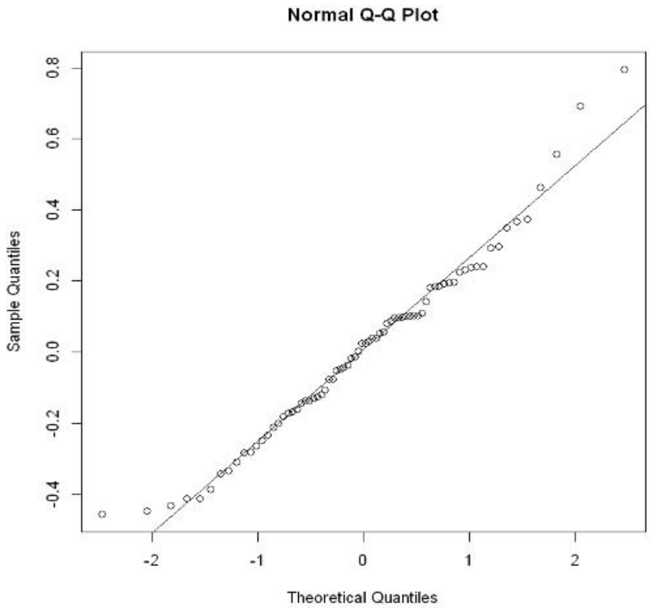
利用自动定阶函数 auto.arima()建立 $ARIMA(0,1,1)$ 模型，模型 $AIC=17.64$，小于 $AR(1)$ 模型的 24.14，$ARIMA(0,1,1)$ 模型更优。

对 ARIMA 模型进行残差检验：

```
: qqnorm(autofit$residuals)
  qqline(autofit$residuals)
  Box.test(autofit$residuals, type="Ljung-Box")
```

```
	Box-Ljung test

data:  autofit$residuals
X-squared = 0.62448, df = 1, p-value = 0.4294
```

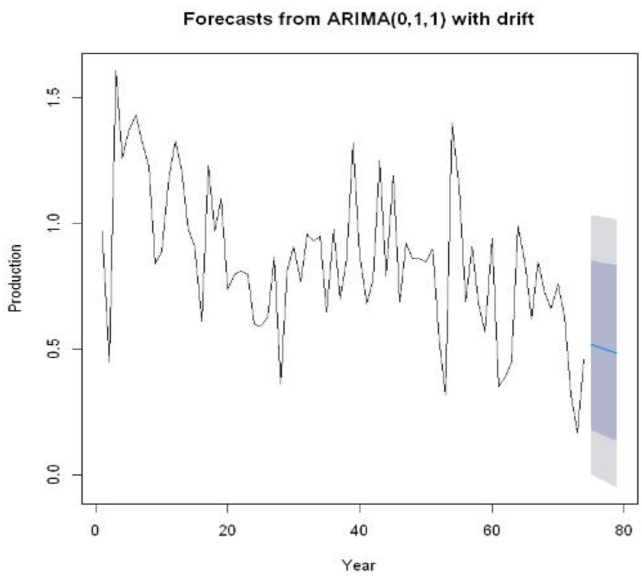


符合残差独立同分布于正态分布的假设。

模型预测结果如下：

```
: forecast(autofit, 5)
  plot(forecast(autofit, 5), xlab="Year", ylab="Production")
```

```
   Point Forecast     Lo 80     Hi 80        Lo 95    Hi 95
75      0.5181957 0.1809510 0.8554403  0.002424464 1.033967
76      0.5097196 0.1696165 0.8498226 -0.010423137 1.029862
77      0.5012435 0.1583059 0.8441811 -0.023234301 1.025721
78      0.4927674 0.1470185 0.8385164 -0.036009927 1.021545
79      0.4842914 0.1357538 0.8328290 -0.048750873 1.017334
```



5 实验分析与总结

5.1 实验分析

（1）时序图显示该地区连续 74 年的谷物产量的波动是比较平稳的。
（2）自相关图检验，考察该样本序列的自相关图，可以发现延迟 2 阶以后，自相关系数都在两倍标准差范围左右，没有明显超出两倍标准差。可以看出，这是一个典型的短期相关的样本自相关图。由时序图和样本自相关图的性质，可以认为该序列平稳。
（3）纯随机性检验。利用 LB 统计量进行纯随机性检验，计算在 6 延迟阶数和 12 延迟阶数下的 LB 统计量的 p 值。在 6 和 12 延迟阶数下的 LB 统计量的 p 值都为 0(<0.001)，所以有大于99.99%的置信水平认为该地区连续 74 年的谷物产量属于非白噪声序列。

总结（1）、（2）、（3）的结果可以发现，这个序列是一列平稳的、蕴含相关信息的序列，值得去分析和研究。

5.2 实验总结

（1）建立 ARIMA 模型的步骤

1.1 确保时序是平稳的

1.2 找到一个（或几个）合理的模型（即选定可能的 p 值和 q 值）

1.3 拟合模型

1.4 从统计假设和预测准确性等角度评估模型

1.5 预测

（2）判断不同模型效果的准则

本文采用 AIC 准则来评价不同模型的效果好坏。AIC 是衡量统计模型拟合优良性的一种标准，由于它为日本统计学家赤池弘次创立和发展的，因此又称赤池信息量准则。它建立在熵的概念基础上，可以权衡所估计模型的复杂度和此模型拟合数据的优良性。

AIC 的假设条件是模型的误差服从独立正态分布，其公式为

$$AIC=2k+n\ln(SSR/n)$$

其中 k 是参数的数量 n 为样本数，SSR 为残差平方和。