第五章　非平稳时间序列模型

zhaixbh@126.com

学习目标与要求

- 理解非平稳时间序列的概念
- 理解确定性趋势的消除过程
- 掌握求和自回归移动平均模型的定义、性质、建模和预测
- 了解残差自回归模型的建模

本章结构

1. 非平稳序列的概念
2. 趋势的消除
3. 求和自回归移动平均模型
4. 残差自回归模型

非平稳序列的定义

- 在前面章节中，我们主要讨论了平稳时间序列，但事实上，在自然科学和经济现象中绝大部分时间序列数据都是非平稳的. 这些非平稳时间序列表现形式多样，不过我们分析的基本手段是想办法将其转化为平稳序列，然后再进一步分析. 从本节开始，我们来介绍非平稳时间序列模型及其建模过程
- 非平稳序列的定义

  所谓平稳时间序列，也即宽平稳时间序列，其实就是指时间序列的均值、方差和协方差等一、二阶矩存在但不随时间改变，表现为时间的常数.

非平稳序列的定义

因而，要判断一个序列是否平稳，只需判断下列三个条件是否同时成立:

$$\mathrm{E}(Y_t)=\mu;\ \mathrm{Var}(Y_t)=\sigma^2;\ \mathrm{Cov}(Y_t,Y_s)=\gamma(t-s) \qquad (5.1)$$

- 一般地，只要上述三个条件有一个不成立，那么我们就称该序列是**非平稳时间序列** (nonstationary time series). 进一步，根据不满足 (5.1) 式的情况，我们可归纳出非平稳时间序列数据具有如下两种形式:
- 确定性趋势时间序列;
- 随机性趋势时间序列.

确定性趋势

- 一般地，**确定性趋势 (deterministic trend) 时间序列**是指序列的期望随着时间而变化,而协方差却平稳的非平稳时间序列，其生成过程为

$$x_t = \mu_t + y_t \tag{5.2}$$

其中，y_t 是一个平稳可逆的 ARMA(p,q) 过程，期望为0，即$\Phi(B)x_t = \Theta(B)\varepsilon_t$. 由 (5.4)式显然可得，

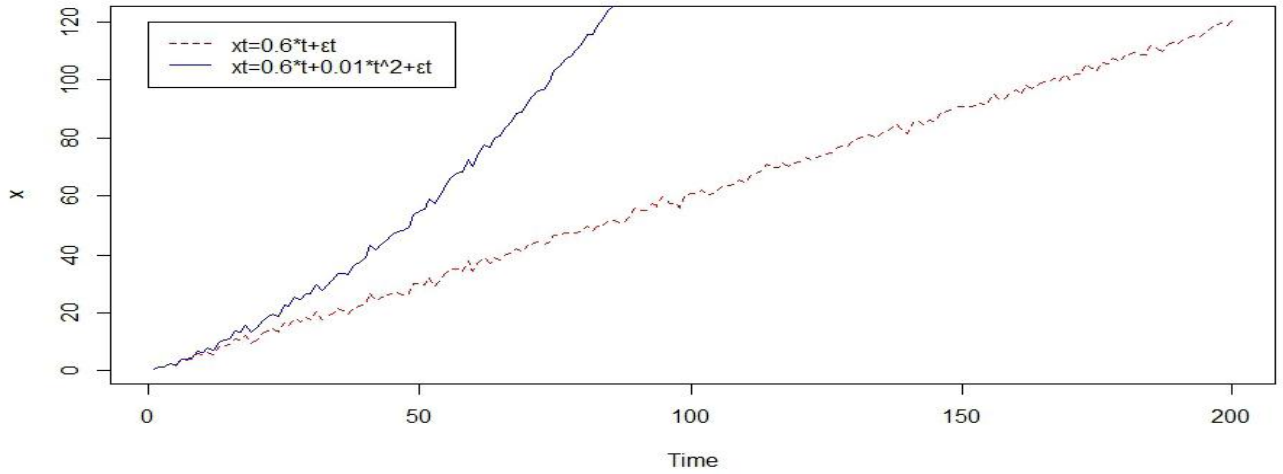
$$\mathrm{E}(x_t) = \mu_t;$$

$$\mathrm{E}[(x_t - \mu_t)(x_{t+k} - \mu_{t+k})] = \gamma(k).$$

- 由于上述序列$\{x_t\}$的方差是常数，所以它的观测值总是围绕着一个确定的趋势在有限的幅度内做波动.

确定性趋势

- 下图是由一个线性趋势和一个二次趋势分别加上一个纯随机序列形成的两个序列的时序图. 从图中可以看出, 确定性趋势时间序列的偏离是暂时的. 如果对具有确定性趋势的时间序列进行长期预测, 那么只要考虑期望函数就可以了, 这是因为不论对多长时间的序列值进行预测, 误差都是有界的. 然而, 这种预测的精度不会令人满意, 实际意义不是太大.



随机性趋势

- 通常，我们把不具有确定性趋势的非平稳时间序列称为**随机性趋势（stochastic trend）时间序列**，其一般具有自回归的形式:

$$x_t = \mu_t + x_{t-1} + y_t \qquad (5.3)$$

- 例如给定初始值 x_0 的 AR(1) 模型: $x_t = \phi x_{t-1} + \varepsilon_t$，$\phi > 1$. 经过迭代，可得

$$x_t = \phi^t x_0 + \sum_{i=0}^{t-1} \phi^i \varepsilon_{t-i}$$

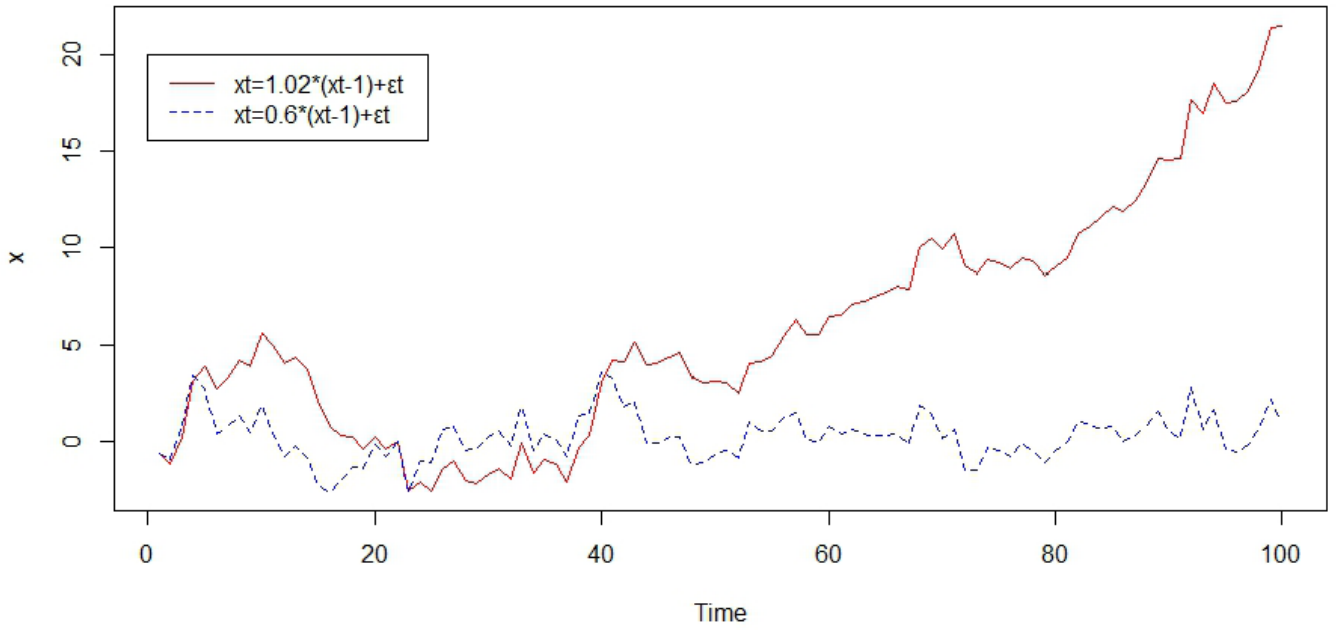
- 所以

$$\mathrm{E}(x_t) = \phi^t x_0;\ \mathrm{Var}(x_t) = \frac{\phi^{2t} - 1}{\phi^2 - 1}\sigma^2$$

- 故当 $|\phi| > 1$ 时，该序列的期望函数和方差函数都呈现指数型增长,因此该序列呈现扩散式增长,是典型的随机性趋势

随机性趋势

- 时间序列; 当$|\phi| < 1$时, 由前面章节的知识得到, 该序列是平稳的. 下图展示了平稳的 AR(1) 序列与非平稳的 AR(1) 的时序图.



本章结构

1. 非平稳序列的概念
2. 趋势的消除
3. 求和自回归移动平均模型
4. 残差自回归模型

趋势的消除

- 非平稳时间序列典型的特征是含有趋势: 确定性趋势和随机性趋势. 要想把非平稳时间序列转化成平稳的时间序列来分析, 就需要通过去趋势和差分方法消除确定性趋势和随机性趋势.在第 4 章中, 我们曾学习了一些提取确定性趋势的方法, 如: 线性拟合法、移动平均法、指数平滑法以及通过构造季节指数处理具有季节效应的序列.
- 一般来讲, 用确定性趋势时间序列减去确定性趋势部分就会得到一个平稳序列, 但是由于上述方法并不能保证趋势信息提取的充分性, 因而剩余部分不能保证平稳.对于随机性趋势的处理, 一般是通过差分运算提取趋势信息的, 但是需要特别小心过差分现象的出现. 在本节中, 我们讨论去趋势的方法.

差分运算的本质

- 在第 2 章中，我们曾学习了差分运算. 熟悉了差分运算之后，我们很容易发现，一个序列的 m 阶差分就类似于连续变量的 m 阶求导. 比如:

$$\nabla c = 0,\ \nabla t = 1,\ \nabla^2 t^2 = 2,\ \nabla^3 t^3 = 6,\ \nabla^4 t^4 = 24, \cdots$$

- 一般地，我们有

$$\nabla^m (\alpha_0 + \alpha_1 t + \cdots + \alpha_m t^m) = c,\ c \text{ 为某一常数.}$$

- 设 $\{x_t\}$ 为一个时间序列，根据 1 阶向后差分运算得

$$\nabla x_t = x_t - x_{t-1}$$

也即

$$x_t = x_{t-1} + \nabla x_t \tag{5.4}$$

差分运算的本质

- 可见，1 阶差分本质上是一个自回归过程. (5.4) 式可视为用延迟 1 期的历史数据 x_{t-1} 作为自变量来解释当期序列值 x_t 的变动情况，差分序列 $\{\nabla x_t\}$ 可视为 x_t 的1 阶自回归过程中产生的随机误差的大小.
- 一般地，对序列 $\{x_t\}$ 作 m 阶差分得

$$\nabla^m x_t = (1-B)^m x_t = \sum_{k=0}^{m} (-1)^k C_m^k x_{t-k}$$

等价于

$$x_t = \sum_{k=1}^{m} (-1)^{k+1} C_m^k x_{t-k} + \nabla^m x_t \qquad (5.5)$$

可见，(5.5) 式本质上也是一个 m 阶自回归过程.

差分运算的本质

- 借助于差分运算可以提取趋势信息.在 R 语言中, 使用函数 diff( ) 来进行差分运算. diff( ) 函数的命令格式如下:
- diff(x, lag= , differences=)

  该函数的参数说明:

  - x: 需要进行差分的序列名.
  - lag: 差分的步长, 不特意指定, 系统默认 lag=1.
  - differences: 差分次数, 不特意指定, 系统默认 differences=1.
- 根据 diff( ) 函数的参数含义, 差分命令 diff(x,d,k) 的意思是进行 k 次 d 步差分. 常用的差分运算为:

差分运算的本质

- 根据 diff( ) 函数的参数含义, 差分命令 diff(x,d,k) 的意思是进行 k 次 d 步差分. 常用的差分运算为:

  1 阶差分: diff(x);

  2 阶差分: diff(x,1,2);

  k 阶差分: diff(x,1,k);

  d 步差分: diff(x,d,1) 或简写为 diff(x,d);

  1 阶差分后再进行 d 步差分: diff(diff(x),d).

趋势信息的提取

- 一般来讲, 经过有限阶的差分运算可以提取趋势信息, 但是有过差分的风险, 过差分现象将在下一节讨论. 在本小节中, 讨论如何从实际数据出发, 通过差分运算初步提取趋势信息. 这里需要注意的是, 我们没有对差分结果的合理性进行深入研究.
- 在实际数据分析中, 通常用 1 阶差分可提取线性趋势, 2 阶或 3 阶等低阶差分可提取曲线趋势, 而对于含有季节趋势的数据, 通常选取差分的步长等于季节的周期可较好地提取季节信息.
- **例 5.1** 在例 4.2 的分析中, 我们得到美国 1974 年至 2006 年月度发电量序列蕴含一个近似线性的递增趋势.

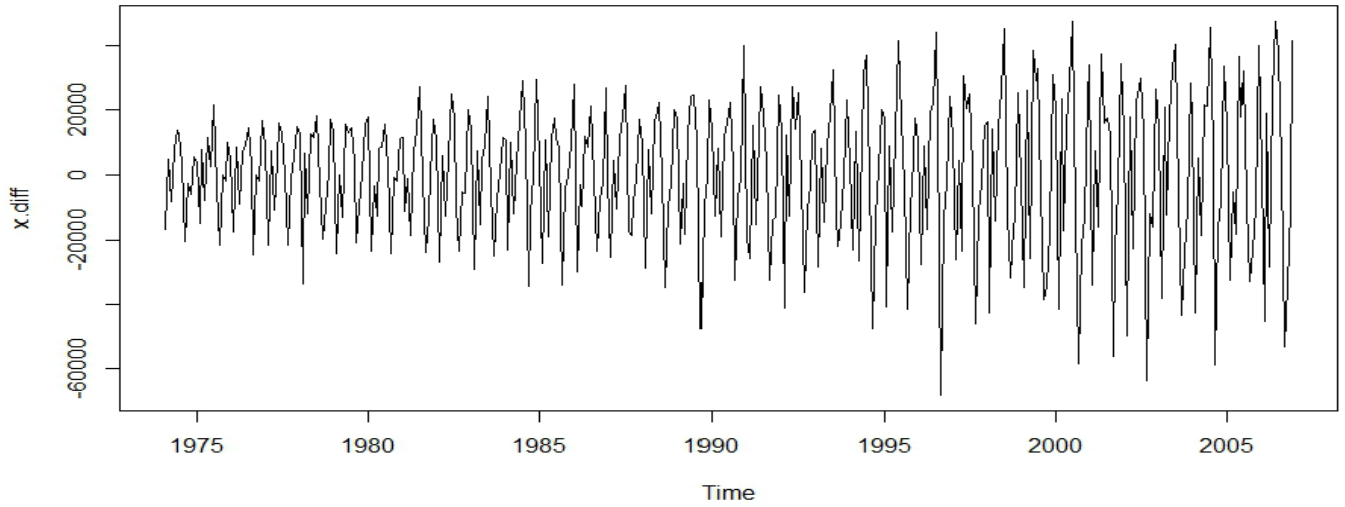
趋势信息的提取

- 现对该序列进行 1 阶差分运算, 考察差分运算对该序列线性趋势的提取作用.
- **解** :通过下述语句实现差分运算, 并对差分序列绘制时序图

```
> electricity <- scan("E:/DATA/CHAP4/1.txt")

Read 396 items

> electricity <- ts(electricity, start=c(1974,1), frequency = 12)

> x.diff <- diff(electricity)

> plot(x.diff)
```

趋势信息的提取

```
> electricity <- scan("E:/DATA/CHAP4/1.txt")
Read 396 items
> electricity <- ts(electricity, start=c(1974,1), frequency = 12)
> x.diff <- diff(electricity)
> plot(x.diff)
```



趋势信息的提取

- 上图表明: 1 阶差分运算成功地从原序列中提取出了线性趋势. 不过, 差分序列的平稳性还需进一步考察, 因为时序图也表明差分序列的方差在逐步增大.
- **例 5.2** 从例 1.13 的分析中, 我们得到 1996 年至 2015 年宁夏回族自治区地区生产总值序列蕴含一个近似二次曲线的趋势. 对该序列进行 2 阶差分运算, 考察差分运算对曲线趋势的提取作用.
- **解** 用下述语句实现差分运算, 并对差分序列绘制时序图

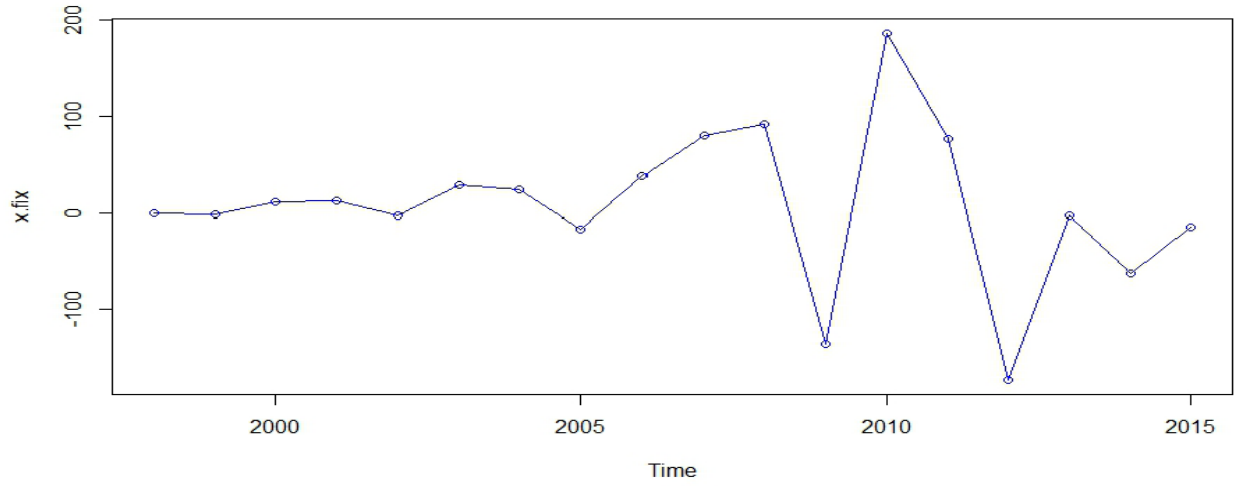
```
> a <- read.table(file="E:/DATA/CHAP1/SMGDP.csv",sep=",",header=T)

> NXGDP <- ts(a$NX, start=1996)

> x.fix <- diff(NXGDP,1,2);  plot(x.fix, type="o", col="blue")
```

趋势信息的提取

- **解** 用下述语句实现差分运算, 并对差分序列绘制时序图

```
> a <- read.table(file="E:/DATA/CHAP1/SMGDP.csv",sep=",",header=T)
> NXGDP <- ts(a$NX, start=1996)
> x.fix <- diff(NXGDP,1,2);  plot(x.fix, type="o", col="blue")
```

- 从下图可见, 经过 2 次差分之后, 差分序列的曲线趋势被提取.

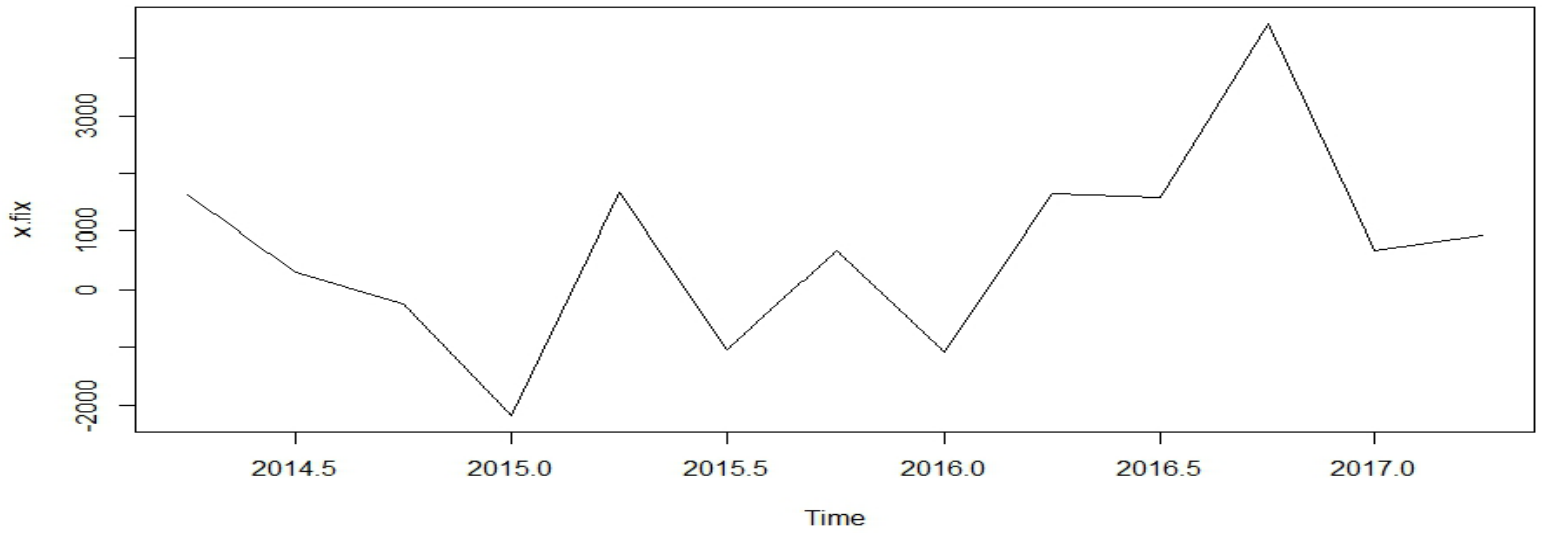


趋势信息的提取

- **例 5.3** 分析例 5.4 中 2013 年第一季度至 2017 年第二季度我国季度 GDP 数据序列, 并提取其确定性趋势信息.
- **解** 从图 5.4 中, 我们发现该序列具有线性趋势和以季度为周期的季节效应, 所以我们首先做 1 阶差分取消线性趋势, 然后做 4 步差分提取季节趋势.

```
> a <- read.table(file="E:/DATA/CHAP4/JDGDP.csv",sep=",",header=T)
> JDGDP <- ts(a$JDGDP,start=2013,frequency = 4)
> x.fix <- diff(diff(JDGDP),4)
> plot(x.fix)
```

趋势信息的提取



- 从上图可见, 经过 1 次差分和 1 次 4 步差分之后, 所得差分序列已无明显趋势.

过差分现象

- 所谓**过差分现象**，就是指由于对序列不恰当地使用差分运算而导致有效信息浪费，估计精度下降的现象. 例如: 考察随机性趋势模型 $x_t = \mu + x_{t-1} + y_t$，其中，$y_t$ 为平稳序列. 我们知道通过对序列 $\{x_t\}$ 的1 阶差分就可以消除非平稳性. 而对于线性趋势模型 $x_t = \mu + \phi t + y_t$，应用 1 阶差分，得到

$$x_t - x_{t-1} = \phi + y_t - y_{t-1}$$

- 因为 $\{y_t\}$ 为ARMA(p,q) 序列，满足形式 $\Phi(B)y_t = \Theta(B)\varepsilon_t$，所以我们有

$$\Phi(B)\nabla x_t = \Phi(1)\phi + (1-B)\Theta(B)\varepsilon_t$$

- 可见，∇x_t 是一个平稳的 ARMA(p,q + 1) 序列. 由于 MA 部分存在一个单位根，所以它是一个不可逆的序列.

过差分现象

- 这个序列是一个新的平稳序列, 而不是原来的平稳的 ARMA 序列$\{y_t\}$,这就导致出现过差分.
- 再举一例. 一个可逆的 MA(1) 模型 $y_t = \varepsilon_t + \theta\varepsilon_{t-1}$, y_t 的方差函数和自相关函数为

$$\mathrm{Var}(y_t) = (1+\theta^2)\sigma^2;\ \rho_k = \begin{cases} \dfrac{\theta}{1+\theta^2}, & k=1; \\ 0, & k>1. \end{cases}$$

- 对 MA(1) 模型做 1 阶差分, 得到

$$\nabla y_t = [1+(\theta-1)B-\theta B^2]\varepsilon_t$$

过差分现象

- ∇y_t 的方差函数和自相关函数分别为

$$\mathrm{Var}(\nabla y_t) = 2(1-\theta+\theta^2)\sigma^2;\ \rho_k = \begin{cases} -\dfrac{(\theta-1)^2}{2(1-\theta+\theta^2)}, & k=1; \\ -\dfrac{\theta}{2(1-\theta+\theta^2)}, & k=2; \\ 0, \qquad k>2. \end{cases}$$

- 因此，$\mathrm{Var}(\nabla y_t) > \mathrm{Var}(y_t)$. 可见，对于 MA(1) 序列$\{y_t\}$而言，其差分序列$\{\nabla y_t\}$是一个不可逆的 MA(1) 序列，且方差变大. 这就意味着对 MA(1) 序列差分会导致过差分现象.

过差分现象

- 实践经验表明，处理确定性趋势时间序列最好采用减去趋势部分的方法来去趋势，特别是对于曲线趋势明显的方差齐性序列来讲，采用此方法更好；而处理随机性趋势的时间序列最好使用差分运算来去趋势.
- **例 5.4** 设时间序列$\{x_t\}$的观测值满足随机游走模型：$x_t = 2.5 + x_{t-1} + \varepsilon_t$，其中，$\{\varepsilon_t\}$是标准正态白噪声序列. 很明显，如果对$\{x_t\}$作1 阶差分，那么就会得到平稳的差分序列：$\nabla x_t = 2.5 + \varepsilon_t$. 如果采用减去确定性部分来去趋势的方法，那么分析其残差序列的特征.

  解 将序列$\{x_t\}$关于时间 t 作趋势回归. 具体命令及运行结果如下：

过差分现象

```
> set.seed(10) #设定生成随机数的种子，种子是为了让结果可重复
> x0 <- 2.5+rnorm(100)
> x <- cumsum(x0) #累积和
> t <- 1:100
> x.lm <- lm(x~t)
> summary(x.lm)
Call:
lm(formula = x ~ t)
Residuals:
     Min       1Q   Median       3Q      Max
-6.1315  -2.2990  -0.2992   3.0378   5.9587
```

过差分现象

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-4.11652	0.71144	-5.786	8.64e-08	***
t	2.34736	0.01223	191.922	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.531 on 98 degrees of freedom

Multiple R-squared: 0.9973, Adjusted R-squared: 0.9973

F-statistic: 3.683e+04 on 1 and 98 DF, p-value: < 2.2e-16

过差分现象

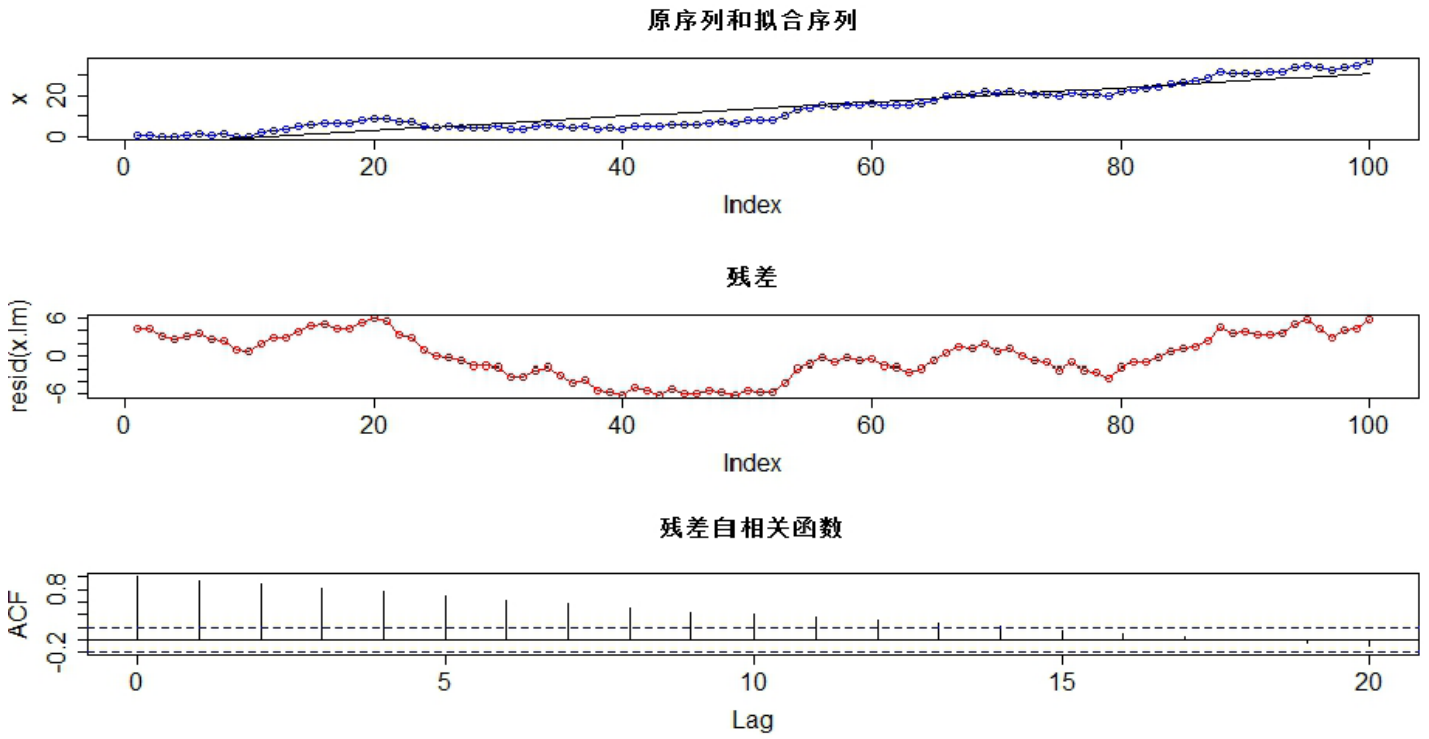
- 估计得到的回归模型为

$$x_t = -4.11652 + 2.34736t + \varepsilon_t$$

- 估计结果表明，常数项和时间趋势系数均显著异于零，这是由于随机游走模型中暗含了一个线性趋势. 下面分析回归残差项 ε_t .

```
> par(mfrow=c(3,1))
> plot(x,type="o",col="blue",main="原序列和拟合序列")
> lines(fitted.values(x.lm))
> plot(resid(x.lm),type="o",col="red",main="残差")
> acf(resid(x.lm),main="残差自相关函数")
```

过差分现象



- 上图表明尽管随机游走$\{x_t\}$具有线性趋势，但是去趋势之后的残差仍然具有很强的自相关性质. 因此，对该序列采用减去趋势部分的去趋势法是不合适的.

过差分现象

- 上图表明尽管随机游走$\{x_t\}$具有线性趋势，但是去趋势之后的残差仍然具有很强的自相关性质. 因此，对该序列采用减去趋势部分的去趋势法是不合适的.
- 上面这些例子表明，分析非平稳序列时，需要对数据所表现出来的趋势进行严谨细致的研究，否则，很容易产生人为的波动和自相关性.

本章结构

1. 非平稳序列的概念
2. 趋势的消除
3. 求和自回归移动平均模型
4. 残差自回归模型

求和自回归移动平均模型的定义

- 一般来讲，具有随机性趋势的非平稳时间序列在经过适当差分之后就会变成一个平稳时间序列. 此时，我们称这个非平稳序列为差分平稳序列. 对差分平稳序列可以使用求和自回归移动平均模型进行拟合.
- 求和自回归移动平均模型的定义

  设$\{x_t, t \in T\}$为一个序列，则我们称满足如下结构的模型为**求和自回归移动平均** (ARIMA)模型，简记为ARIMA(p,d,q)，

$$\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t$$

  其中，ε_t 为均值为零，方差为σ_ε^2 的白噪声，且

$$\mathrm{E}(x_s x_t)=0,\ \forall s < t;\ \nabla^d = (1-B)^d;\ \Phi(B) = 1-\phi_1 B-\cdots-\phi_p B^p$$

求和自回归移动平均模型的定义

- 求和自回归移动平均模型的定义

设$\{x_t, t \in T\}$为一个序列, 则我们称满足如下结构的模型为**求和自回归移动平均** (ARIMA)模型, 简记为ARIMA(p,d,q),

$$\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \qquad (5.8)$$

其中, ε_t为均值为零, 方差为σ_ε^2的白噪声, 且

$\mathrm{E}(x_s x_t)=0, \forall s < t$; $\nabla^d = (1-B)^d$; $\Phi(B) = 1-\phi_1 B - \cdots - \phi_p B^p$为平稳可逆的 ARMA(p,q) 模型的自回归系数多项式;

$\Theta(B) = 1-\theta_1 B - \cdots - \theta_q B^q$为平稳可逆的 ARMA(p,q) 模型的移动平滑系数多项式.

求和自回归移动平均模型的定义

- 注:

(1) 从ARIMA(p,d,q) 模型的定义可以看出, 该模型实质上就是$\{x_t, t \in T\}$的d 阶差分序列是一个平稳可逆的ARMA(p,q) 模型. (5.8) 也可简单记作

$$\nabla^d x_t = \frac{\Theta(B)}{\Phi(B)} \varepsilon_t \qquad (5.9)$$

式中, ε_t 为零均值白噪声序列. (5.9) 说明, 一个非平稳时间序列如果d 阶差分之后成为平稳序列了, 那么我们就可以用较为成熟可靠的 ARMA(p,q) 模型拟合其 d 阶差分序列了.

求和自回归移动平均模型的定义

- 注:

(2) ARIMA(p,d,q) 模型是比较综合的模型, 它有以下几种要的特殊形式: 当 d = 0 时, ARIMA(p,d,q) 模型就是 ARMA(p,q) 模型; 当 p = 0 时, ARIMA(p,d,q) 模型简记为 IMA(d,q) 模型; 当 q = 0 时, ARIMA(p,d,q) 模型简记为 ARI(p,d) 模型; 当 d = 1,p = q = 0 时, ARIMA(p,d,q) 模型为 $x_t = x_{t-1} + \varepsilon_t$, 这是著名的随机游走模型.

求和自回归移动平均模型的性质

- 设时间序列$\{x_t, t\in T\}$服从ARIMA(p,d,q) 模型

$$\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t$$

记 $\varphi(B)=\Phi(B)\nabla^d$，称 $\varphi(B)$ 为**广义自回归系数多项式**.

- 显然，$\{x_t, t\in T\}$的平稳性取决于$\varphi(B)=0$ 的根的分布. 由于$\{x_t, t\in T\}$ 的d 阶差分序列是平稳可逆的 ARMA(p,q) 模型，所以不妨设

$$\Phi(B)=\prod_{k=1}^{p}(1-\lambda_k B),\ |\lambda_k|<1,\ k=1,2,\cdots p$$

因而

$$\varphi(B)=\Phi(B)\nabla^d=\left[\prod_{k=1}^{p}(1-\lambda_k B)\right](1-B)^d$$

求和自回归移动平均模型的性质

$$\varphi(B)=\Phi(B)\nabla^d=\left[\prod_{k=1}^{p}(1-\lambda_k B)\right](1-B)^d$$

- 由上式容易判断, ARIMA(p,d,q) 模型的广义自回归系数多项式共有 p + d 个根, 其中 p 个根 $1/\lambda_1, 1/\lambda_2, \cdots, 1/\lambda_p$ 在单位圆外, d 个根在单位圆上. 从而 ARIMA(p,d,q) 模型有 p + q 个特征根, 其中, p 个在单位圆内, d个在单位圆上. 因为有 d 个特征根在单位圆上而非单位圆内, 所以当 $d \neq 0$ 时, ARIMA(p,d,q) 模型非平稳.
- 对于 ARIMA(p,d,q) 模型来讲,当 $d \neq 0$ 时, 均值和方差都不具有齐性. 方差不具有齐性的最简单的例子是随机游走模型 ARIMA(0,1,0): $x_t = x_{t-1} + \varepsilon_t$.

求和自回归移动平均模型的性质

- 对于 ARIMA(p,d,q) 模型来讲,当 d ≠ 0 时, 均值和方差都不具有齐性. 方差不具有齐性的最简单的例子是随机游走模型 ARIMA(0,1,0): $x_t = x_{t-1} + \varepsilon_t$.
- 这是因为根据上述递推关系可得

$$\begin{aligned} x_t &= x_{t-1} + \varepsilon_t \\ &= x_{t-2} + \varepsilon_t + \varepsilon_{t-1} \\ &\vdots \\ &= x_0 + \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1. \end{aligned}$$

从而, $\mathrm{Var}(x_t) = t\sigma_\varepsilon^2$, 这是随时间递增的函数, 当时间趋于无穷时, x_t 的方差也趋于无穷.

求和自回归移动平均模型的建模

- 正如前面所述，对于非平稳时间序列的建模，我们的策略是将其设法转化为平稳序列，然后用平稳序列建模的方法来建模.
- 对于 ARIMA 模型的建模，我们首先对观测值序列进行平稳性检验，如果检验是非平稳的序列，那么对其进行差分运算，直至检验是平稳的；如果检验是平稳的，那么转入 ARMA 模型的建模步骤. 下面举例说明.
- **例 5.5** 分析 1996 年至 2015 年，我国第三产业增加值序列，建立 ARIMA 模型，并预测2016 年的增加值.
- **解:** 读取数据，并作第三产业增加值序列的时序图. 具体命令如下:

求和自回归移动平均模型的建模

- > x <- read.table(file="E:/DATA/CHAP5/1.csv",sep=",",header=T)
- > x <- ts(x$thr,start=1996);par(mfrow=c(2,2))
- > plot(x,type="o",col="blue",sub= "图 5.7 第三产业增加值序列的时序图")

观察时序图很容易发现，该序列具有明显的增长趋势，因此可判定非平稳. 对该序列作1 阶差分运算，并对所得差分序列作出时序图、自相关图和偏自相关图. 具体命令如下，运行结果见下图.

- > x.fix <- diff(x)
- > plot(x.fix,col="blue",type="o",sub="图 5.8 第三产业增加值差分序列的
- + 时序图")

求和自回归移动平均模型的建模

- > acf(x.fix,sub="图 5.9 第三产业增加值差分序列的自相关图")
- > pacf(x.fix, sub="图 5.10 第三产业增加值差分序列的偏自相关图")

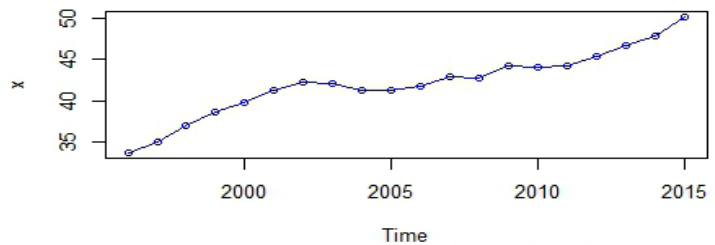


图 5.7 第三产业增加值序列的时序图

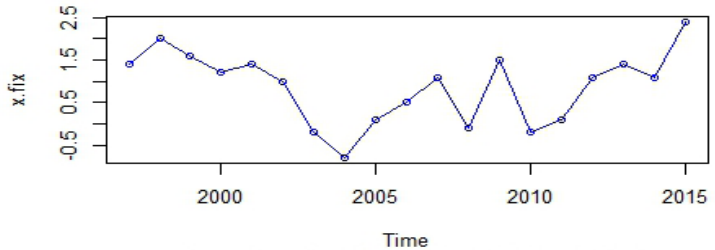


图 5.8 第三产业增加值差分序列的时序图

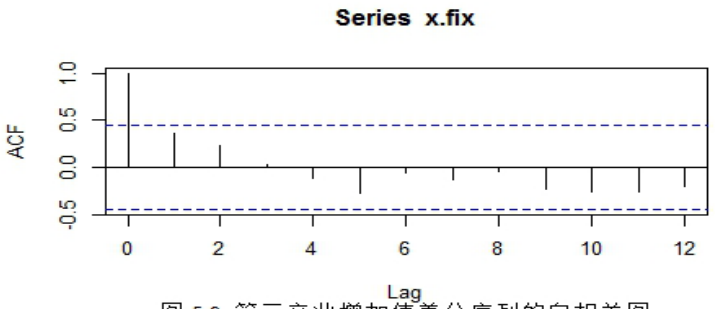


图 5.9 第三产业增加值差分序列的自相关图

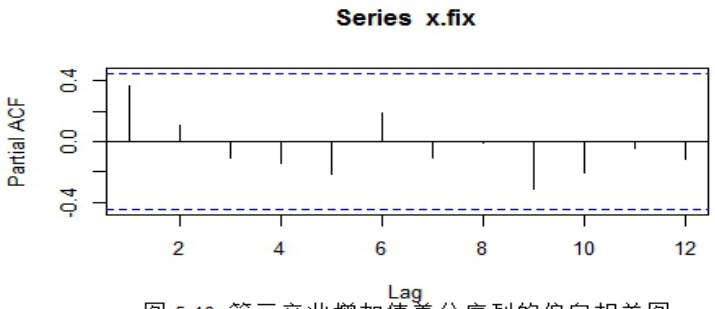


图 5.10 第三产业增加值差分序列的偏自相关图

求和自回归移动平均模型的建模

- 1 阶差分序列的时序图、自相关函数图和偏自相关函数图都表明, 差分序列具有平稳性. 而且图 5.9 和图 5.10 表明自相关函数具有延迟 1 阶的截尾特征, 而偏自相关函数具有拖尾性, 故我们选用 ARIMA(0,1,1) 模型来拟合所给数据. 具体的命令及运行结果如下:
- > x.fix <- arima(x,order=c(0,1,1))
- > x.fix
- Call:
- arima(x = x, order = c(0, 1, 1))
- Coefficients:
- ma1

求和自回归移动平均模型的建模

- > x.fix <- arima(x,order=c(0,1,1))
- > x.fix
- Call:
- arima(x = x, order = c(0, 1, 1))
- Coefficients:
- ma1
- 0.4823
- s.e. 0.1541
- sigma^2 estimated as 1.023: log likelihood = -27.31, aic = 58.62

求和自回归移动平均模型的建模

- 拟合结果为

$$x_t = x_{t-1} + \varepsilon_t + 0.4823\varepsilon_{t-1}, \quad \varepsilon_t \sim N(0,1.023).$$

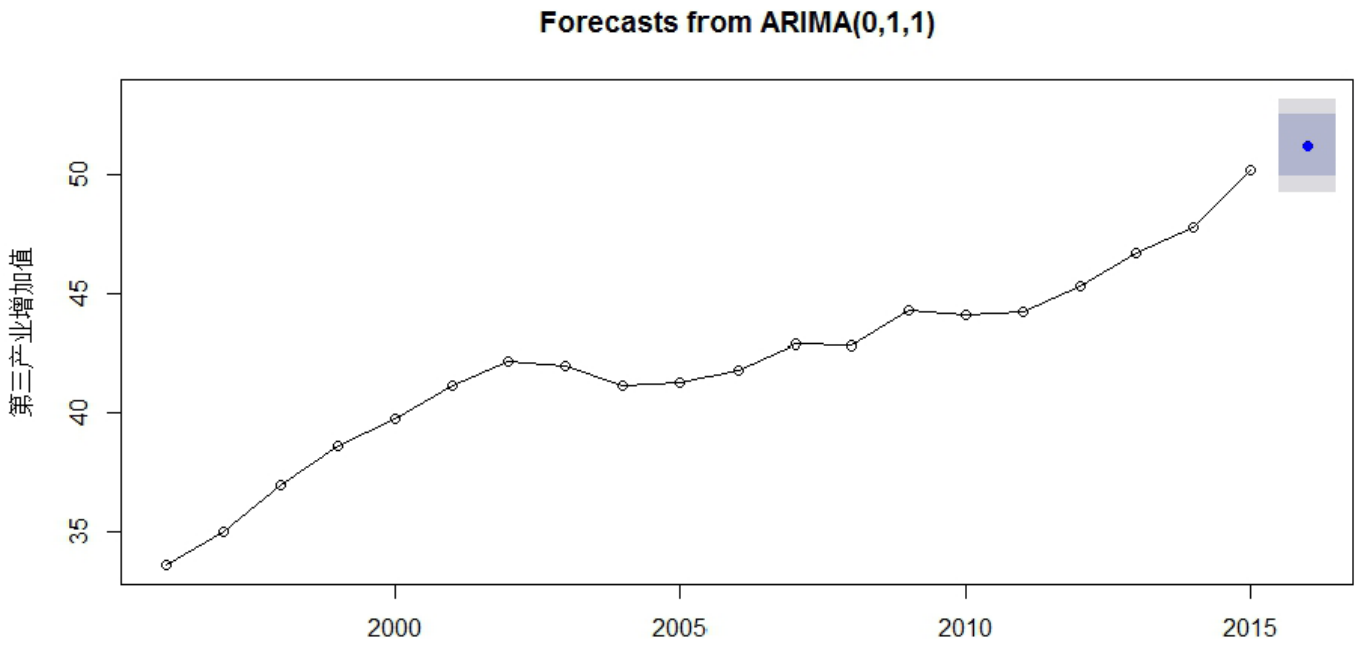
- 再对残差序列作白噪声检验, 具体命令及运行结果如下:
- > for(i in 1:2) print(Box.test(x.fix$residuals,lag=6*i))
- Box-Pierce test
- data: x.fix$residuals
- X-squared = 4.6456, df = 6, p-value = 0.59
- Box-Pierce test
- data: x.fix$residuals
- X-squared = 6.5127, df = 12, p-value = 0.8881

求和自回归移动平均模型的建模

- 白噪声检验表明, 延迟 6 阶的白噪声和延迟 12 阶的白噪声检验的 p 值都远远大于 0.05, 因此该模型显著成立, 即 ARIMA(0,1,1) 模型对该序列拟合成功. 用该模型预测 2016 年我国第三产业增加值. 具体命令及运行结果如下, 预测图如图 5.11 所示.
- > library(forecast)
- > x.fore <- forecast(x.fix,h=1)
- > plot(x.fore,type="o",ylab="第三产业增加值")
- > x.fore
- Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
- 2016 51.21497 49.91872 52.51121 49.23252 53.19741

求和自回归移动平均模型的建模

- 预测表明, 2016 年我国第三产业增加值为 51.21497%, 这个预测的 95% 的置信区间为(49.23252, 53.19741).



求和自回归移动平均模型的建模

- 在对时间序列数据进行 ARIMA(p,d,q) 建模时，有时会遇到所谓的缺省自回归系数或移动平均系数的情况.
- 一般来讲，ARIMA(p,d,q) 模型是指序列进行 d 阶差分之后会得到一个自回归阶数为 p，移动平均阶数为 q 的自回归移动平均模型，它包含了 p+q 个未知参数:

$$\phi_1,\ \phi_2,\cdots,\phi_p,\theta_1,\theta_2,\cdots,\theta_q$$

- 如果这 p+q 个参数中有部分为 0，那么称原 ARIMA(p,d,q) 模型为**疏系数模型**.
- 如果只是自回归系数中有部分缺省，那么该疏系数模型简记为 ARIMA(($p_1, p_2,\cdots, p_l$),d,q)，其中 $p_1, p_2,\cdots, p_l$ 为非零的自回归系数;如果只是移动平均系数中有部分缺省，那么该疏系数模型简记为ARIMA(p,d,($q_1, q_2,\cdots, q_m$))，

求和自回归移动平均模型的建模

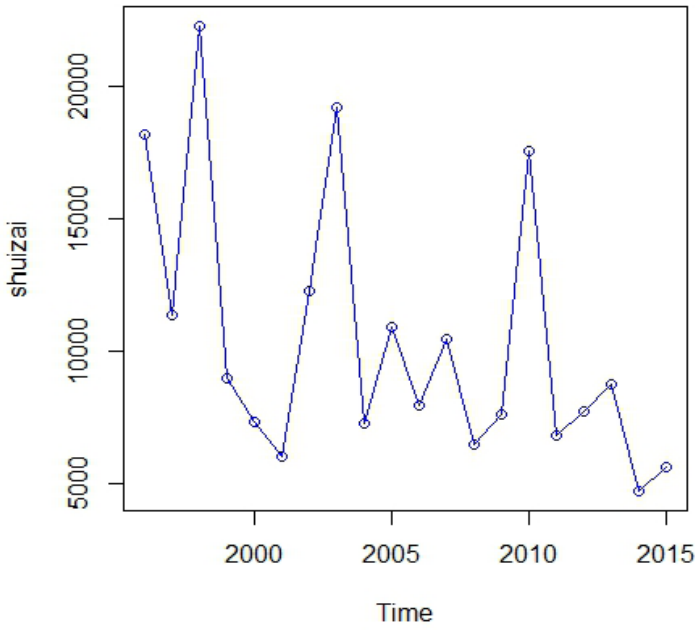
- 其中 $q_1, q_2, \cdots, q_m$ 为非零的移动平均系数; 如果自回归系数中和移动平均系数中都有部分缺省, 那么该疏系数模型简记为ARIMA($(p_1, p_2, \cdots, p_l)$,d,$(q_1, q_2, \cdots, q_m)$).
- 在 R 语言中, 使用函数 arima( ) 来拟合 ARIMA 疏系数模型. 函数 arima( ) 拟合疏系数模型的命令格式如下:
- arima(x, order=, method=, transform.pars=, fixed=)
- 该函数的参数说明 (仅说明后两个参数的使用):
- - transform.pars: 指定参数估计是否由系统自动完成. transform.pars=T 表示系统根据order 选项设置的模型阶数自动完成参数估计. 这是系统默认设置. transform.pars=F 表示需要拟合疏系数模型.
- - fixed: 对疏系数模型指定疏系数的位置.

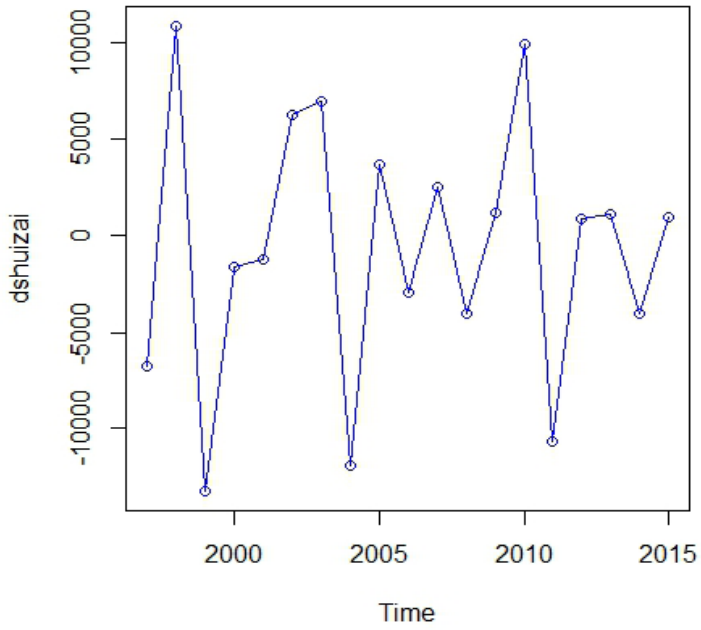
求和自回归移动平均模型的建模

- **例 5.6** 对 1996 年至 2015 年, 我国农业受水灾面积序列进行分析, 建立 ARIMA 模型, 并预测未来 3 年的受水灾面积大小 (单位: 千公顷).
- **解:** 读取数据, 并作出序列时序图. 具体命令如下, 运行结果见下图.
- > x <- read.csv("E:/DATA/CHAP5/e5.3.csv",header=T)
- > shuizai <- ts(x$SY,start=1996)
- > par(mfrow=c(1,2))
- > plot(shuizai,type="o",col="blue")
- > dshuizai <- diff(shuizai); plot(dshuizai,type="o",col="blue")

求和自回归移动平均模型的建模

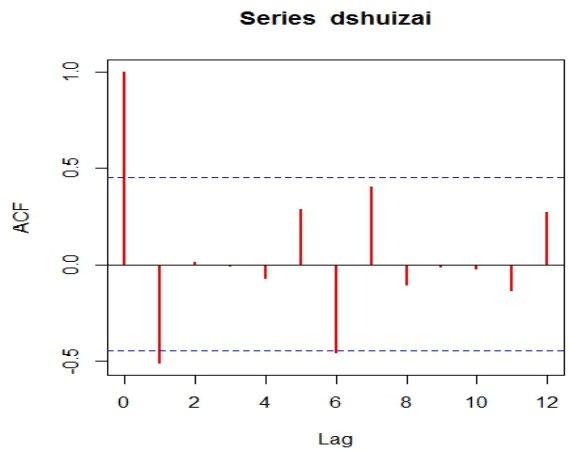
- **解:** 读取数据，并作出序列时序图. 具体命令如下，运行结果见下图.

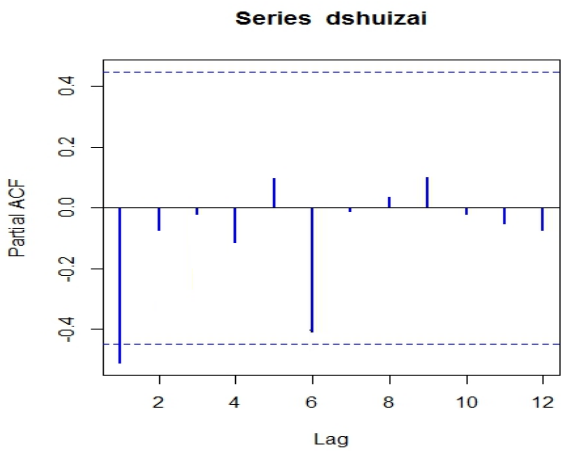




求和自回归移动平均模型的建模

- 可见我国近 20 年来农业水灾面积呈现明显下滑趋势. 所以, 做 1 阶差分提取线性趋势, 并绘制差分序列时序图. 右图表明, 1 阶差分序列呈现平稳趋势. 接下来, 绘制自相关图和偏自相关图. 具体命令如下, 运行结果见图 .
- > acf(dshuizai,lwd=2,col="red")
- > pacf(dshuizai,lwd=2,col="blue")





求和自回归移动平均模型的建模

- 从上知，可对该序列尝试进行疏系数模型拟合. 再考虑到自相关函数在 6 阶之后拖尾，而偏自相关函数在 6 阶之后截尾，所以选用模型 ARIMA((1,6),1,0) 进行数据拟合,并且对拟合结果的残差进行白噪声检验. 具体命令及运行结果如下:
- > Nihe <- Arima(shuizai,order=c(6,1,0),transform.pars = F,fixed=
- + c(NA,0,0,0,0,NA))
- > Nihe
- Series: shuizai
- ARIMA(6,1,0)

求和自回归移动平均模型的建模

- Coefficients:
- ar1 ar2 ar3 ar4 ar5 ar6
- -0.2529 0 0 0 0 -0.6172
- s.e. 0.1686 0 0 0 0 0.1869
- sigma^2 estimated as 29948231: log likelihood=-188.63
- AIC=383.26 AICc=384.86 BIC=386.1
- > for(i in 1:2)print(Box.test(Nihe$residuals,lag=6*i))
- Box-Pierce test
- data: Nihe$residuals
- X-squared = 5.4781, df = 6, p-value = 0.4841

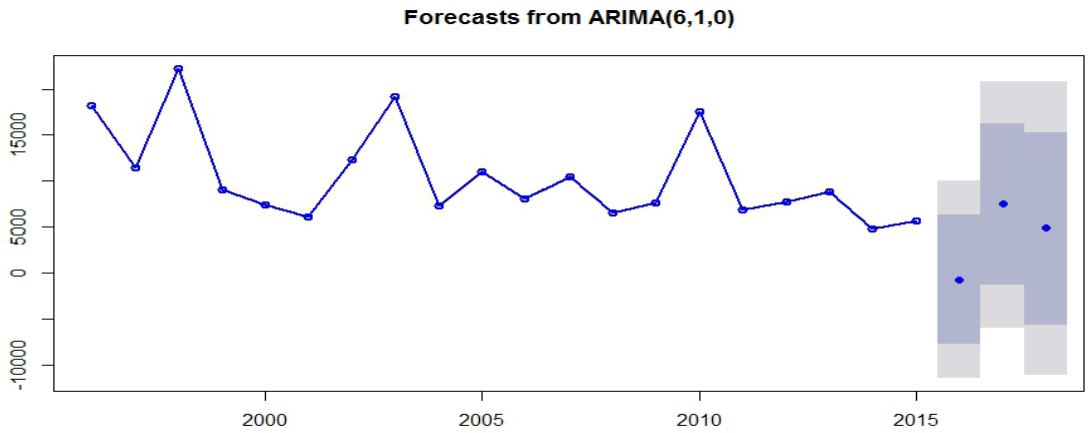
求和自回归移动平均模型的建模

- Box-Pierce test
- data: Nihe$residuals
- X-squared = 6.5059, df = 12, p-value = 0.8885
- 分别做延迟 6 阶和 12 阶的白噪声检验. 检验结果表明, 残差序列显著为白噪声, 因此数据基本符合拟合模型 ARIMA((1,6),1,0). 最后, 预测未来 3 年水灾面积情况. 具体命令及运行结果如下, 预测图如图所示.
- > sz.fore <- forecast(Nihe,h=3)
- > plot(sz.fore,col="blue",type="o",lwd=2)
- > sz.fore

求和自回归移动平均模型的建模

- > sz.fore

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2016	-725.4174	-7738.705	6287.871	-11451.316	10000.48
2017	7459.2924	-1295.300	16213.885	-5929.702	20848.29
2018	4854.6304	-5585.674	15294.935	-11112.438	20821.70



求和自回归移动平均模型的预测理论

- 设$\{x_t\}$服从ARIMA(p,d,q)模型，则x_t的d阶差分序列$\{\nabla^d x_t\}$服从平稳可逆的ARMA(p,q)模型，即

$$\Phi(B)(1-B)^d = \Theta(B)\varepsilon_t \qquad (5.10)$$

因此，类似于ARMA(p,q)模型的传递形式，(5.10)式可以写成

$$x_t = \sum_{i=0}^{\infty} G_i^* \varepsilon_{t-i} = G^*(B)\varepsilon_t \qquad (5.11)$$

其中，$G_0^* = 1$，$G^*(B)$满足

$$\Phi(B)(1-B)^d G^*(B) = \Theta(B)\varepsilon_t$$

令广义自回归系数多项式$\varphi(B)$为

$$\varphi(B) = 1 - \phi_1^* B - \cdots - \phi_{p+d}^* B^{p+d}$$

求和自回归移动平均模型的预测理论

- 设$\{x_t\}$服从ARIMA(p,d,q)模型，则x_t的d阶差分序列$\{\nabla^d x_t\}$服从平稳可逆的ARMA(p,q)模型，即

$$\Phi(B)(1-B)^d = \Theta(B)\varepsilon_t \qquad (5.10)$$

因此，类似于ARMA(p,q)模型的传递形式，(5.10)式可以写成

$$x_t = \sum_{i=0}^{\infty} G_i^* \varepsilon_{t-i} = G^*(B)\varepsilon_t \qquad (5.11)$$

其中，$G_0^* = 1$，$G^*(B)$满足

$$\Phi(B)(1-B)^d G^*(B) = \Theta(B)\varepsilon_t$$

令广义自回归系数多项式$\varphi(B)$为

$$\varphi(B) = 1 - \phi_1^* B - \cdots - \phi_{p+d}^* B^{p+d}$$

求和自回归移动平均模型的预测理论

- 则由待定系数法得

$$\begin{cases} G_1^* = \phi_1^* - \theta_{1;} \\ G_2^* = \phi_1^* G_1^* + \phi_2^* - \theta_{2;} \\ \quad \vdots \\ G_k^* = \phi_1^* G_{k-1}^* + \cdots + \phi_{p+d}^* G_{k-p-d}^* - \theta_k \end{cases} \tag{5.12}$$

其中，当 $k<0$ 时，$G_k^*=0$；当 $k=0$ 时，$G_k^*=1$；当 $k>q$ 时，$\theta_k=0$.

根据(5.11)式知，x_{t+l} 真实值为

$$x_{t+l} = \left(\varepsilon_{t+l} + G_1^* \varepsilon_{t+l-1} + \cdots + G_{l-1}^* \varepsilon_{t+1}\right) + \left(G_l^* \varepsilon_t + G_{l+1}^* \varepsilon_{t-1} + \cdots\right).$$

因为 $\varepsilon_{t+l}, \varepsilon_{t+l-1}, \cdots, \varepsilon_{t+1}$ 不可预测，所以 x_{t+l} 只能用 $\varepsilon_t, \varepsilon_{t+1}, \cdots$ 来估计:

求和自回归移动平均模型的预测理论

$$\hat{x}_{t+l} = \tilde{G}_0\varepsilon_t + \tilde{G}_1\varepsilon_{t-1} + \tilde{G}_2\varepsilon_{t-2} + \cdots.$$

于是，真实值与预测值之间的均方误差为

$$E\left(x_{t+l} - \hat{x}_{t+l}\right)^2 = \left(1 + G_1^{*2} + G_2^{*2} + \cdots + G_{l-1}^{*2}\right)\sigma_\varepsilon^2 + \sum_{i=0}^{\infty}\left(G_{l+i}^{*} - \tilde{G}_i\right)^2 \sigma_\varepsilon^2.$$

为使均方误差最小，当且仅当$G_{l+i}^{*} = \tilde{G}_i$. 因此，在均方误差最小原则下，$l$ 期预测值为

$$\hat{x}_{t+l} = G_l^{*}\varepsilon_t + G_{l+1}^{*}\varepsilon_{t-1} + G_{l+2}^{*}\varepsilon_{t-2} + \cdots, \qquad (5.13)$$

l 期的预测误差的方差为

$$Var\left[e_t(l)\right] = \left(1 + G_1^{*2} + G_2^{*2} + \cdots + G_{l-1}^{*2}\right)\sigma_\varepsilon^2. \qquad (5.14)$$

求和自回归移动平均模型的预测理论

在实际的预测中，一般不会使用预测公式(5.13)，而是根据模型递推. 但是预测误差却可由公式(5.14) 给出.

例 5.7 已知某序列服从ARIMA(1,1,1) 模型:

$(1+0.5B)(1-B)x_t=(1-0.8B)\varepsilon_t$, 且 $x_1=3.9, x_2=4.8, \varepsilon_2=0.6, \sigma_\varepsilon^2=1$,

求: x_5 的 95% 的置信区间.

解 将原模型写成: $x_t=0.5x_{t-1}+0.5x_{t-2}+\varepsilon_t-0.8\varepsilon_{t-1},$

得到预测递推公式 $\hat{x}_3=0.5x_2+0.5x_1+\varepsilon_2=4.95,$

$\hat{x}_4=0.5\hat{x}_3+0.5x_2=4.875,$

$\hat{x}_5=0.5\hat{x}_4+0.5\hat{x}_3=4.9125,$

求和自回归移动平均模型的预测理论

由模型的广义自回归系数多项式 $\varphi(B)=1-0.5B-0.5B^2$，得 $\phi_1^*=0.5,\phi_2^*=0.5$. 根据 (5.12) 式得 $G_1^*=-0.3,G_2^*=-1.34$. 第 5 期预测误差的方差为 $\mathrm{Var}\left[e_2(3)\right]=\left(1+G_1^{*2}+G_2^{*2}\right)\sigma_\varepsilon^2=2.8856.$ x_5 的 95% 的置信区间为

$$\left(\hat{x}_5-1.96\sqrt{\mathrm{Var}\left[e_2(3)\right]},\ \hat{x}_5+1.96\sqrt{\mathrm{Var}\left[e_2(3)\right]}\right),$$ 即

$(1.583,8.242)$.

本章结构

1. 非平稳序列的概念
2. 趋势的消除
3. 求和自回归移动平均模型
4. 残差自回归模型

残差自回归模型的概念

我们知道, 对序列做 1 阶差分可以消除线性趋势; 做 2 阶、3 阶等低阶向后差分可以消除曲线趋势, 但是我们也同时分析了这样做会有过差分的风险, 即人为造成的非平稳或“不好”的信息. 因此, 在进行 ARIMA 建模时, 一方面要看到差分运算能够充分提取确定性信息, 另一方面也要看到差分运算解释性不强, 同时有过差分风险.

当序列呈现出很强烈的趋势时, 传统的消除确定性趋势的方法显示出一定的优越性, 但是又担心可能浪费残差信息. 不过当残差中含有自相关关系时, 可继续对残差序列建立自回归模型,这样就自然地提出了残差自回归模型.

残差自回归模型的概念

5.4.1 残差自回归模型的概念

根据数据分解的形式，当序列的长期趋势非常显著时，我们可以将序列分解为 $x_t = T_t + S_t + \varepsilon_t,$ (5.15)

其中，T_t 为长期的递增或递减趋势，S_t 为季节变化.

一般来讲，(5.15) 式中的趋势项 T_t 和 S_t 不一定能够把数据中的确定性信息充分提取. 当残差中含有显著的自相关关系时，进一步对残差序列进行自回归拟合，从而再次提取相关信息.于是得到如下残差自回归模型的概念.我们称具有下列结构的模型为残差自回归模型:

残差自回归模型的概念

$$\begin{cases} x_t = T_t + S_t + \varepsilon_t, \\ \varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + \omega_t; \\ \mathrm{E}(\omega_t) = 0,\ \mathrm{Var}(\omega_t) = \sigma^2, \mathrm{Cov}(\omega_t, \omega_{t-k}) = 0,\ k \geq 1. \end{cases} \tag{5.16}$$

在建模中, 对趋势项 T t 的拟合有两种常用形式:

$T_t = a_0 + a_1 t + \cdots + a_k t^k$ 和 $T_t = a_0 + a_1 x_{t-1} + \cdots + a_k x_{t-k}$

对季节变化项 S_t 的拟合也有两种常用形式:

$S_t = S_t'$ 和 $S_t = a_0 + a_1 x_{t-m} + \cdots + a_k x_{t-km}$，其中 S_t' 为季节指数; m 为季节变化的周期.

残差的自相关检验

在进行残差自回归模型建模时，首先拟合趋势项和季节变化项，然后进行残差检验. 当残差序列自相关性不显著时，则建模结束; 当残差序列自相关性显著时，再对残差进行建模. 残差的建模步骤与 ARIMA 模型建模步骤一致.

- 残差的自相关检验

1. Durbin-Watson 检验

我们可以使用由 J.Durbin 和 G. S. Watson 于 1950 年提出的所谓 Durbin-Watson 检验(简称 DW 检验) 来检验序列残差的自相关性. 下面以 1 阶自相关性检验为例介绍 DW 检验原理.

原假设 H_0 : 残差序列不存在 1 阶自相关性:

残差的自相关检验

$E(\varepsilon_t \varepsilon_{t-1})=0$, 即 $\rho(1)=0$;

备择假设 H_1 : 残差序列存在 1 阶自相关性:

$E(\varepsilon_t \varepsilon_{t-1}) \neq 0$,即 $\rho(1) \neq 0$.

构造 DW 检验统计量:

$$DW = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n}\varepsilon_t^2} \qquad (5.17)$$

当观测样本 n 很大时, 有

$$\sum_{t=2}^{n}\varepsilon_t^2 \approx \sum_{t=2}^{n}\varepsilon_{t-1}^2 \approx \sum_{t=1}^{n}\varepsilon_t^2 . \qquad (5.18)$$

残差的自相关检验

将 (5.18) 式代入 (5.17) 式得

$$DW \approx 2\left(1-\frac{\sum_{t=2}^{n}\varepsilon_t-\varepsilon_{t-1}}{\sum_{t=1}^{n}\varepsilon_t^2}\right)$$

回顾自相关函数的定义, 得

$$\rho(1)=\frac{\sum_{t=2}^{n}\varepsilon_t-\varepsilon_{t-1}}{\sum_{t=1}^{n}\varepsilon_t^2}$$

残差的自相关检验

于是, 有

$$DW \approx 2\left[1-\rho(1)\right]. \qquad (5.19)$$

由于自相关函数的范围为 $[-1,1]$, 所以 DW 的范围也大约介于 $[0,4]$.

(1) 当 $0<\rho(1)\leq 1$ 时, 序列正相关.

当 $\rho(1)\to 1$ 时, $DW\to 0$; 当 $\rho(1)\to 0$ 时, $DW\to 2$. 由此可确定两个临界值 $0\leq m_L, m_U \leq 2$. 当 $DW<m_L$ 时, 序列显著正相关; 当 $DW>m_L$ 时, 序列显著不相关; 当 $m_L \leq DW \leq m_U$ 时, 无法断定序列的相关性.

(2) 当 $-1<\rho(1)\leq 0$ 时, 序列负相关.

残差的自相关检验

当 $\rho(1)\to -1$ 时, $DW\to 4$; 当 $\rho(1)\to 0$ 时, $DW\to 2$. 同样由此可确定两个临界值 $2\le m_L^*, m_U^*\le 4$. 当 $DW>m_U^*$ 时, 序列显著负相关; 当 $DW<m_L^*$ 时, 序列显著不相关; 当 $m_L^*\le DW\le m_U^*$ 时, 无法断定序列的相关性.

根据 $\rho(1)$ 的对称性, 可令

$$\begin{cases} 2-m_U=m_L^*-2; \\ 2-m_L=m_U^*-2, \end{cases}$$

则

$$\begin{cases} m_L^*=4-m_U; \\ m_U^*=4-m_L. \end{cases}$$

残差的自相关检验

由此得到相关性判断表 (见表 5.1).

表 5.1 相关性判断表

DW 的取值区间	$(0,m_L)$	$[m_L,m_U]$	$[m_U,4-m_U]$	$[4-m_U,4-m_L]$	$[4-m_L,4]$
序列相关性	正相关	相关性待定	不相关	相关性待定	负相关

2. Durbin h 检验

DW 统计量是为回归模型残差自相关性的检验而提出, 它要求自变量“独立”. 在自回归情况下, 即当回归因子包含延迟因变量时, 有 $x_t = a_0 + a_1 x_{t-1} + \cdots + a_k x_{t-k} + \varepsilon_t$.

此时, 残差序列$\{\varepsilon_t\}$的 DW 统计量是个有偏的统计量. 因而, 当$\rho(1)$ 趋于零时, DW≠2.在这种情况下, 使用 DW 统计量会导致残差序列自相关性不显著的误判.

残差自回归模型建模

为了克服 DW 检验的有偏性, Durbin 提出了 DW 统计量的修正统计量:

$$D_h = D_w \frac{n}{1 - n\sigma_a^2},$$

式中, n 为观察值序列的长度; $n\sigma_a^2$ 为延迟因变量系数的最小二乘估计的方差. 这大大提高了检验的精度.

- **残差自回归模型建模**

本小节, 举例说明残差自回归模型建模.

例 5.8 分析 1882 年至 1936 年苏格兰离婚数序列, 并建立残差自回归模型.

残差自回归模型建模

解 读取数据, 并绘制时序图. 从时序图分析, 该序列有显著的线性递增趋势, 但没有季节效应. 因此, 考虑建立如下结构的残差自回归模型:

$$\begin{cases} x_t = T_t + \varepsilon_t; \\ \varepsilon_t = \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p} + \omega_t; \\ E(\omega_t) = 0, \mathrm{Var}(\omega_t) = \sigma^2, \mathrm{Cov} = (\omega_t, \omega_{t-k}) = 0, k \geq 1. \end{cases}$$

对 T_t 分别尝试构造如下两个确定性趋势:

(1) 变量为时间 t 的线性函数

$$T_t = a_0 + a_1 .t, t = 1, 2, \cdots;$$

残差自回归模型建模

(2) 变量为 1 阶延迟序列值 x_{t-1}

$$T_t = a_0 + a_1.x_{t-1}$$

具体命令及运行结果如下, 趋势拟合图如图 5.17 所示.

```
> x <- read.csv("E:/DATA/CHAP5/e5.8.csv",header=T)
> Divorces <- ts(x$Divorces,start=1882)
> plot(Divorces,type="p",col="blue",pch=20) #绘制时序图
> t <- 1:55
> x.fix1 <- lm(Divorces~t) #线性拟合
> summary(x.fix1)
```

残差自回归模型建模

Call:

lm(formula = Divorces ~ t)

Residuals:

Min 1Q Median 3Q Max

-88.329 -43.110 -4.268 33.263 125.234

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.293 14.421 -0.228 0.82

t 9.812 0.448 21.900 <2e-16 ***

残差自回归模型建模

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.75 on 53 degrees of freedom

Multiple R-squared: 0.9005, Adjusted R-squared: 0.8986

F-statistic: 479.6 on 1 and 53 DF, p-value: < 2.2e-16

> LI <- ts(x.fix1$fitted.values,start=1882)

> lines(LI,col="red",lwd=2) #绘制线性拟合图

> X1 <- Divorces[2:55]

> X2 <- Divorces[1:54]

残差自回归模型建模

```
> x.fit2 <- lm(X2~X1) #自回归拟合
> summary(x.fit2)


Call:
lm(formula = X2 ~ X1)
Residuals:
Min 1Q Median 3Q Max
-160.384 -21.735 -1.087 13.487 137.268
Coefficients:
```

残差自回归模型建模

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.6956 13.1215 0.891 0.377

X1 0.9189 0.0410 22.412 <2e-16 ***

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.21 on 52 degrees of freedom

Multiple R-squared: 0.9062, Adjusted R-squared: 0.9044

F-statistic: 502.3 on 1 and 52 DF, p-value: < 2.2e-16

> AR <- ts(x.fit2$fitted.values,start=1882)

> lines(AR,type="o") #绘制自回归拟合图

残差自回归模型建模

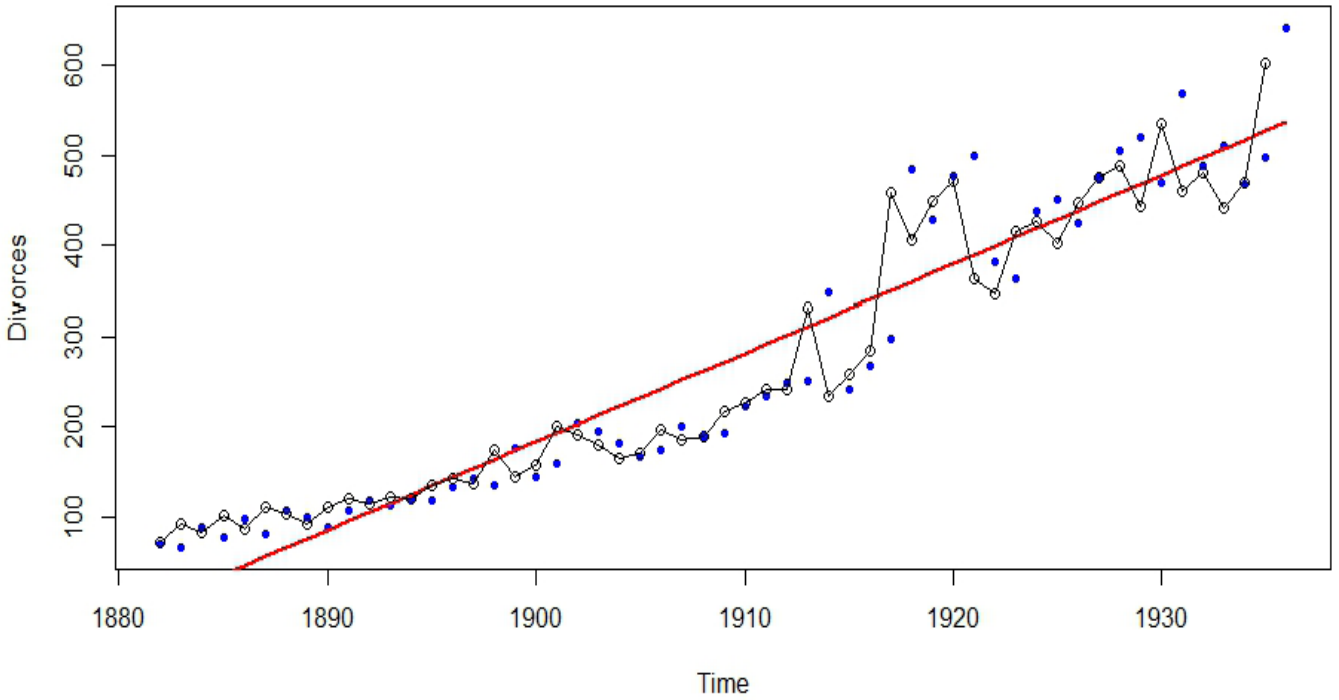


图 5.17 趋势拟合图

残差自回归模型建模

在图 5.17 中, 点状图为序列时序图; 直线为关于时间 t 的线性拟合图; 曲线为关于延迟变量的自回归拟合图.

根据输出结果, 得到如下两个确定性趋势拟合模型:

$$(1)\ x_t = -3.293 + 9.812t + \varepsilon_t,\ \ \varepsilon_t \sim N\left(0, 52.75^2\right)$$

$$(2)\ x_t = 11.6956 + 0.9189x_{t-1},\ \ \varepsilon_t \sim N\left(0, 49.21^2\right)$$

在 R 语言中, 使用函数 dwtest() 对残差作 DW 检验. 检验之前, 首先安装并下载程序包lmtest. 函数 dwtest() 的命令格式如下:

```
dwtest(x.fix,order.by=)
```

该函数的参数说明:

- x.fix: 为拟合结果变量名.

残差自回归模型建模

- order.by: order.by 的取值为延迟因变量, 默认值为无延迟情形.

对残差做 DW 检验的命令及运行结果如下:

```
> dwtest(x.fix1)

Durbin-Watson test

data: x.fix1

DW = 0.91823, p-value = 2.818e-06

alternative hypothesis: true autocorrelation is greater than 0

> dwtest(x.fit2,order.by=X1)

Durbin-Watson test

data: x.fit2
```

残差自回归模型建模

DW = 1.8349, p-value = 0.2253

alternative hypothesis: true autocorrelation is greater than 0

对拟合 (1) 的残差检验表明, DW 统计量的值小于 1, 且 p 很小. 这说明残差序列高度正相关, 有必要对残差序列继续提取信息. 对拟合 (2) 的残差检验表明, DW 统计量的值接近 2, 且 p大于 0.05, 说明残差序列不存在显著的相关性, 不需要再进行拟合.

下面对 (1) 的残差序列拟合自相关模型. 首先对残差序列作自相关函数图和偏自相关图.具体命令如下, 运行结果如图 5.18 和图 5.19 所示.

残差自回归模型建模

```
> par(mfrow=c(1,2))
> acf(x.fix1$residuals)
>pacf(x.fix1$residuals)
```

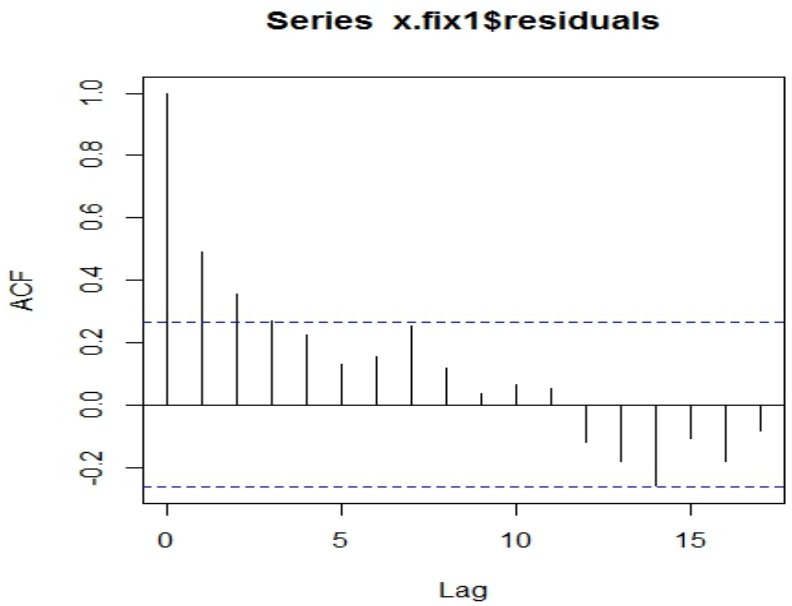


图5.18 残差序列的自相关图

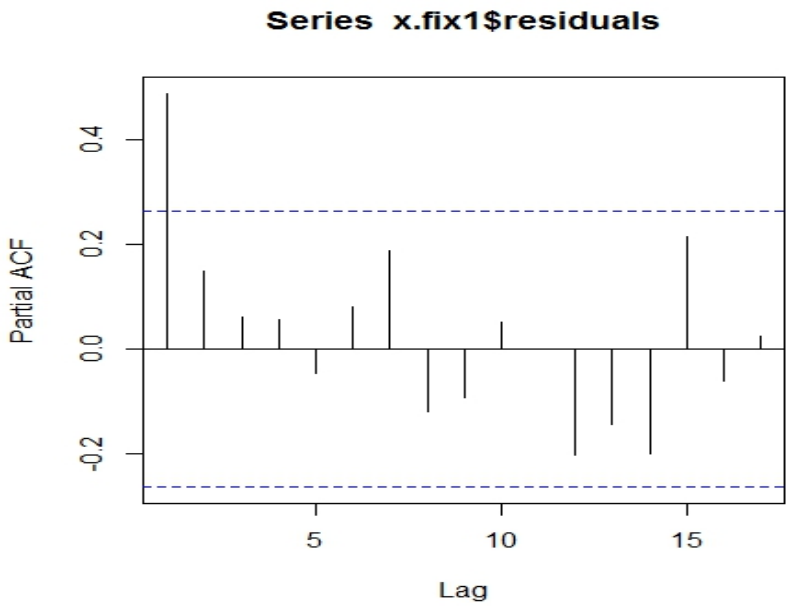


图 5.19 残差序列的偏自相关图

残差自回归模型建模

由图 5.18 的拖尾性和图 5.19 的 1 阶截尾性知, 可以用 AR(1) 建模. 具体的建模命令及运行结果如下:

```
> r.fit <- Arima(x.fix1$residuals,order=c(1,0,0),include.mean=F)

> r.fit

Series: x.fix1$residuals

ARIMA(1,0,0) with zero mean

Coefficients:

ar1

0.5344

s.e. 0.1196

sigma^2 estimated as 2001: log likelihood=-286.75
```

残差自回归模型建模

AIC=577.5 AICc=577.73 BIC=581.51

> for(i in 1:2)print(Box.test(r.fit$residuals,lag=6*i))

Box-Pierce test

data: r.fit$residuals

X-squared = 1.4962, df = 6, p-value = 0.9597

Box-Pierce test

data: r.fit$residuals

X-squared = 6.2098, df = 12, p-value = 0.9051

拟合结果为

$$\varepsilon_t = 0.5344\varepsilon_{t-1} + \omega_t, \quad \omega_t \sim N(0, 2001).$$

习题 5

模型显著性检验显示该拟合模型显著成立.

综合前面的分析, 对 1882 年至 1936 年苏格兰离婚数序列, 我们建立如下残差自回归模型:

$$\begin{cases} x_t = -3.293 + 9.812t + \varepsilon_t \\ \varepsilon_t = 0.5344\varepsilon_{t-1} + \omega_t, \ \omega_t \sim N(0, 2001). \end{cases}$$

- 习题 5

1. 简述差分运算的本质和趋势信息提取的关系.

2. 举例说明过差分现象产生的本质以及如何最大程度地避免过差分现象发生.

3. 举例说明 ARIMA 模型的建模过程和预测理论.

习题 5

4. 举例说明为何要建立残差自回归模型?

5. 将下列模型识别成特定的 ARIMA 模型, 请写出 p,d,q 的值和各项系数的值.

(1) $x_t = x_{t-1} - 0.25x_{t-2} + \varepsilon_t - 0.1\varepsilon_{t-1;}$

(2) $x_t = 0.5x_{t-1} - 0.5x_{t-2} + \varepsilon_t - 0.5\varepsilon_{t-1} + 0.25\varepsilon_{t-2}.$

6. 获得 100 个 ARIMA$(0,1,1)$ 模型的序列观察值$x_{1,}x_{2,}\cdots,x_{100.}$

(1) 已知$\theta_1 = 0.3, x_{100} = 50, \hat{x}_{101} = 51$,求 $\hat{x}_{102}$的值.

(2) 假定新获得 $x_{100} = 52$, 求 $\hat{x}_{102}$ 的值.

7. 已知下列 ARIMA 模型, 试求 $\mathrm{E}\nabla x_t$ 与 $\mathrm{Var}(\nabla x_t)$.

习题 5

(1) $x_t = 3 + x_{t-1} + \varepsilon_t - 0.75\varepsilon_{t-1}$;

(2) $x_t = 10 + 1.25x_{t-1} - 0.25x_{t-2} + \varepsilon_t - 0.1\varepsilon_{t-1}$

(3) $x_t = 5 + 2x_{t-1} - 1.7x_{t-2} + 0.7x_{t-3} + \varepsilon_t - 0.5\varepsilon_{t-1} + 0.25\varepsilon_{t-2}$.

8. 已知一个序列 y_t 由 $y_t = a_0 + a_{1t} + x_t$ 给出, 其中 $\{x_t\}$ 是一个随机游走序列, 并假设 a_0 与 a_1 是常数.

(1) $\{y_t\}$ 是否是平稳的?

(2) $\{\nabla y_t\}$ 是否是平稳的?

9. 已知 ARIMA(1,1,1) 模型为

$$(1-0.8B)(1-B)x_t = (1-0.6B)\varepsilon_t,$$

且 $x_{t-1} = 4.5, x_t = 5.3, \varepsilon_t = 0.8, \sigma^2 = 1$, 求 x_{t+3} 的 95% 的置信区间.