时间序列分析

董雪梅

Outline

01 什么是时间序列

02 时间序列的预处理

03 平稳时间序列分析：ARMA模型

04 非平稳时间序列分析：ARIMA模型

05 案例分析

01 什么是时间序列

- 最早的时间序列分析可以追溯到7000年前的古埃及。
  - 古埃及人把尼罗河涨落的情况逐天记录下来，就构成所谓的时间序列。对这个时间序列长期的观察使他们发现尼罗河的涨落非常有规律。由于掌握了尼罗河泛滥的规律，使得古埃及的农业迅速发展，从而创建了埃及灿烂的史前文明。
- 按照时间的顺序把随机事件变化发展的过程记录下来就构成了一个时间序列。**对时间序列进行观察、研究，寻找它变化发展的规律，预测它将来的走势就是时间序列分析。**

时间序列的数学定义

- 随机序列:按时间顺序排列的一组随机变量

$$\cdots, X_1, X_2, \cdots, X_t, \cdots$$

- 观察值序列:随机序列的 n个有序观察值，称之为序列长度为n 的观察值序列

$$x_1, x_2, \cdots, x_t$$

- 随机序列和观察值序列的关系
  - 观察值序列是随机序列的一个实现
  - 我们研究的目的是想揭示随机时序的性质
  - 实现的手段都是通过观察值序列的性质进行推断

时间序列分析方法

- 描述性时序分析

- 统计时序分析

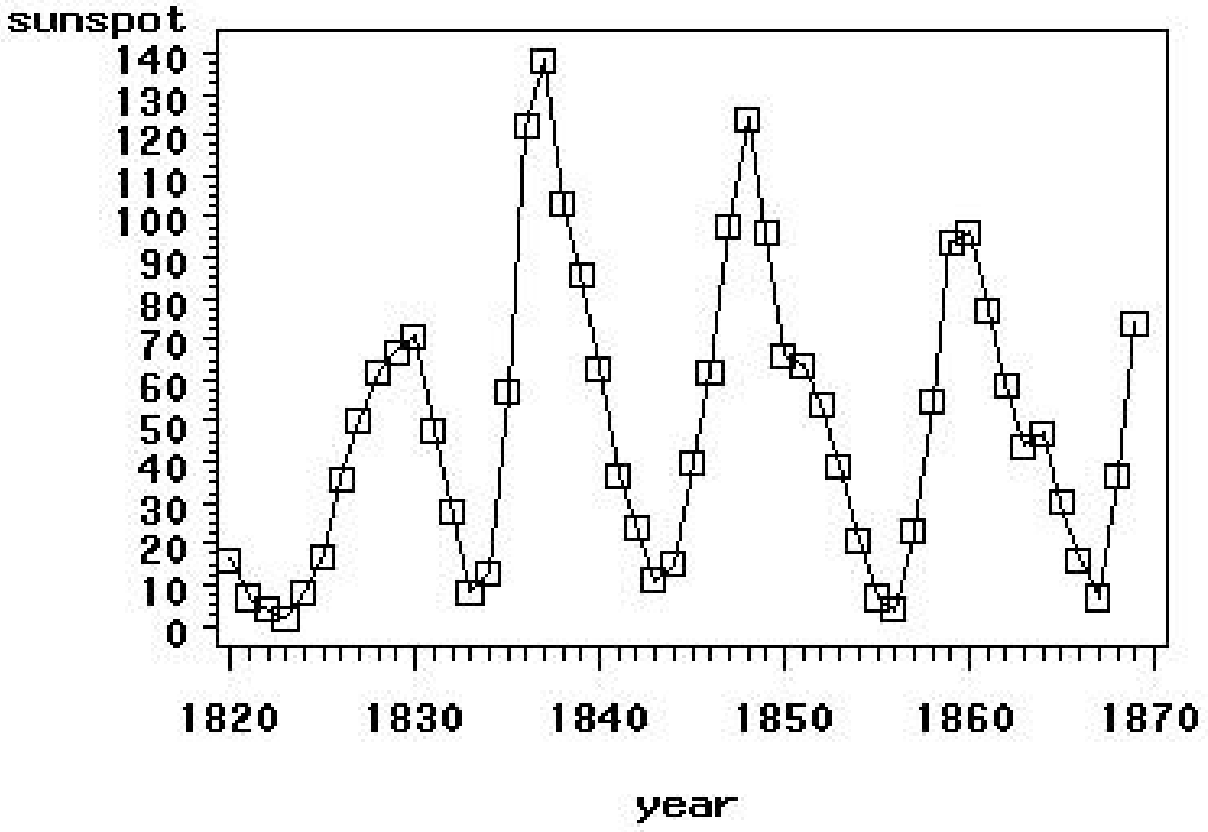
描述性时序分析

- 通过直观的数据比较或绘图观测，寻找序列中蕴含的发展规律，这种分析方法就称为描述性时序分析
- 描述性时序分析方法具有操作简单、直观有效的特点，它通常是人们进行统计时序分析的第一步。

----探索性分析

描述性时序分析案例

- 德国业余天文学家施瓦尔发现太阳黑子的活动具有11年左右的周期。



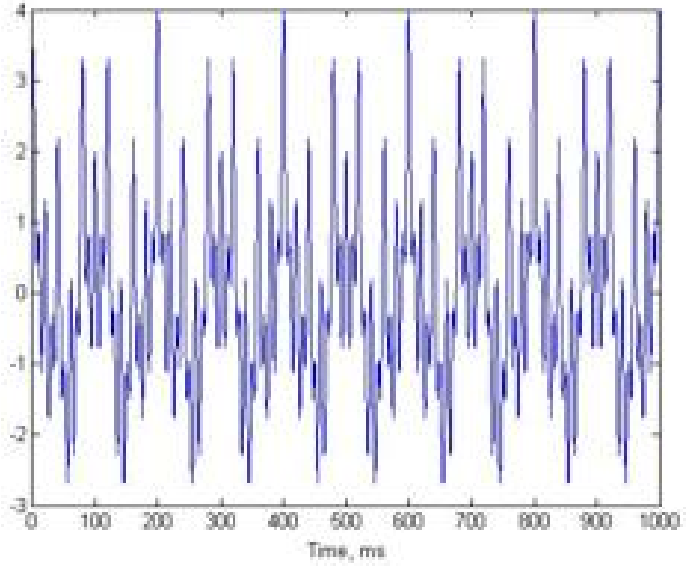
统计时序分析

- 频域分析方法

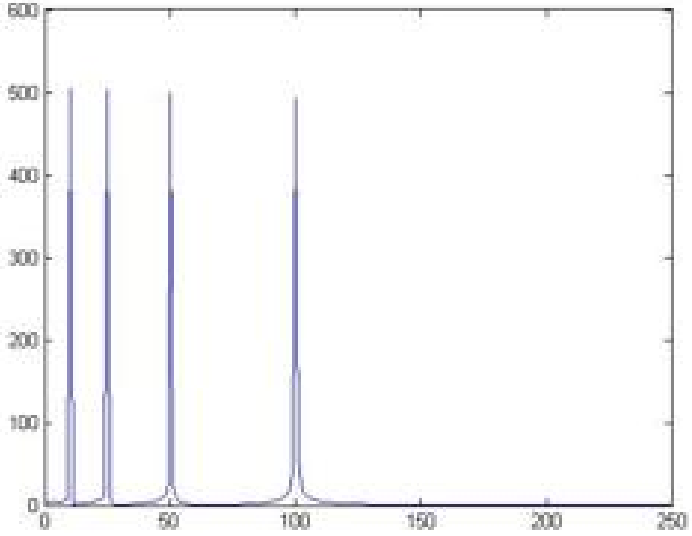
- 时域分析方法

频域分析方法

- 原理
  - 假设任何一种无趋势的时间序列都可以分解成若干不同频率的周期波动。
- 发展过程
  - 早期的频域分析方法借助**傅里叶分析**从频率的角度揭示时间序列的规律。
  - 后来借助了傅里叶变换，用正弦、余弦项之和来逼近某个函数。
  - 20世纪60年代，引入最大熵谱估计理论，进入现代谱分析阶段。
- 特点
  - 非常有用的动态数据分析方法，但是由于分析方法复杂，结果抽象，有一定的使用局限性。



FFT



$$x(t) = \cos(2\pi * 10t) + \cos(2\pi * 25t) + \cos(2\pi * 50t) + \cos(2\pi * 100t)$$

10, 25, 50, 100Hz

时域分析方法

- 原理
  - 事件的发展通常都具有一定的惯性，这种惯性用统计的语言来描述就是序列值之间存在着一定的相关关系，这种相关关系通常具有某种统计规律。
- 目的
  - 寻找出序列值之间相关关系的统计规律，并拟合出适当的数学模型来描述这种规律，进而利用这个拟合模型预测序列未来的走势。
- 特点
  - 理论基础扎实，操作步骤规范，分析结果易于解释，是时间序列分析的主流方法。

时域分析方法的发展过程

- 基础阶段
- 核心阶段
- 完善阶段

基础阶段

- G.U.Yule
  - 1927年，AR模型
- G.T.Walker
  - 1931年，MA模型，ARMA模型

核心阶段

- G.E.P.Box和 G.M.Jenkins
  - 1970年，出版《Time Series Analysis Forecasting and Control》
  - 提出ARIMA模型（Box—Jenkins 模型）
  - Box—Jenkins模型实际上是主要运用于单变量、同方差场合的线性模型

完善阶段

- 异方差场合
  - Robert F.Engle，1982年，ARCH模型
  - Bollerslov，1985年GARCH模型
- 多变量场合
  - C.Granger，1987年，提出了协整（co-integration）理论
- 非线性场合
  - 汤家豪等，1980年，门限自回归模型
  - 2015年，谷歌LSTM(Long Short-Term Memory)
  - 近几年，机器学习方法

时间序列分析软件

- 常用软件
  - S-plus，Matlab，Gauss，SAS ，python，DataRobot...
- 案例分析中用的matlab

时域分析方法的一般步骤(Figure 1)

- 观测数据（均值，周期等）
- 数据预处理
- 平稳化 - 去趋势与去周期，剩余随机项
- 用类似于回归/滑动平均的思想来拟合随机项

1. 判断去趋势去周期后的数据是否平稳
2. 计算数据的自相关函数和偏相关函数
3. 根据自相关/偏相关函数性质决定选用什么模型来拟合随机项
4. 模型定阶和拟合参数的求解
5. 模型检验

02 时间序列的预处理

时间序列的数字特征

- 均值 $\mu_t = Ex_t$
- 方差 $\sigma_t^2 = E(x_t - \mu_t)^2$
- 自协方差 $\gamma(t，k) = Cov(x_t，x_k) = E(x_t - \mu_t)(x_k - \mu_k)$
- 自相关系数 $\rho(t，k) = \dfrac{\gamma(t，k)}{\sqrt{\sigma_t^2 \times \sigma_k^2}} = \dfrac{\gamma(t，k)}{\sigma_t \times \sigma_k}$

时间序列数据的平稳性

定义：假定某个时间序列是由某一随机过程（stochastic process）生成的，如果满足下列条件：

1）均值$E(X_t)=\mu$是与时间t无关的常数；

2）方差$Var(X_t)=\sigma^2$是与时间t无关的常数；

3）协方差$Cov(X_t, X_{t+k})=\gamma_k$是只与时期间隔k有关，与时间t无关的常数；

则称该随机时间序列是平稳的（stationary)，而该随机过程是一平稳随机过程（stationary stochastic process）。

例1．一个最简单的随机时间序列是一具有零均值同方差的独立分布序列：

$$X_t=\mu_t \quad , \quad \mu_t \sim N(0,\sigma^2)$$

该序列常被称为是一个白噪声（white noise）。

由于X_t具有相同的均值与方差，且协方差为零，由定义，一个白噪声序列是平稳的。

例2．另一个简单的随机时间列序被称为随机游走（random walk），该序列由如下随机过程生成：

$$X_t=X_{t-1}+\mu_t$$

这里，μ_t是一个白噪声。

容易知道该序列有相同的均值：$E(X_t)=E(X_{t-1})$

为了检验该序列是否具有相同的方差，可假设X_t的初值为X_0，则易知：

$X_1=X_0+\mu_1$

$X_2=X_1+\mu_2=X_0+\mu_1+\mu_2$

… …

$X_t=X_0+\mu_1+\mu_2+...+\mu_t$

由于X_0为常数，μ_t是一个白噪声，因此：$Var(X_t)=t\sigma^2$，即X_t的方差与时间t有关而非常数，它是一非平稳序列。

然而，对X取一阶差分（first difference）：

$$\Delta X_t = X_t - X_{t-1} = \mu_t$$

由于μ_t是一个白噪声，则序列$\{\Delta X_t\}$是平稳的。

后面将会看到：如果一个时间序列是非平稳的，它常常可通过取差分的方法而形成平稳序列。

- 事实上，随机游走过程是下面我们称之为1阶自回归AR(1)过程的特例：　$X_t=\phi X_{t-1}+\mu_t$

  不难验证：

1)$|\phi|>1$时，该随机过程生成的时间序列是发散的，表现为持续上升($\phi>1$)或持续下降($\phi<-1$)，因此是非平稳的；

2)$\phi=1$时，是一个随机游走过程，也是非平稳的。

事实上可以证明:只有当$-1<\phi<1$时，该随机过程才是平稳的

平稳性检验的图示判断

- 给出一个随机时间序列，首先可通过该序列的时间路径图来粗略地判断它是否是平稳的。
- 一个平稳的时间序列在图形上往往表现出一种围绕其均值不断波动的过程。
- 而非平稳序列则往往表现出在不同的时间段具有不同的均值（如持续上升或持续下降）。

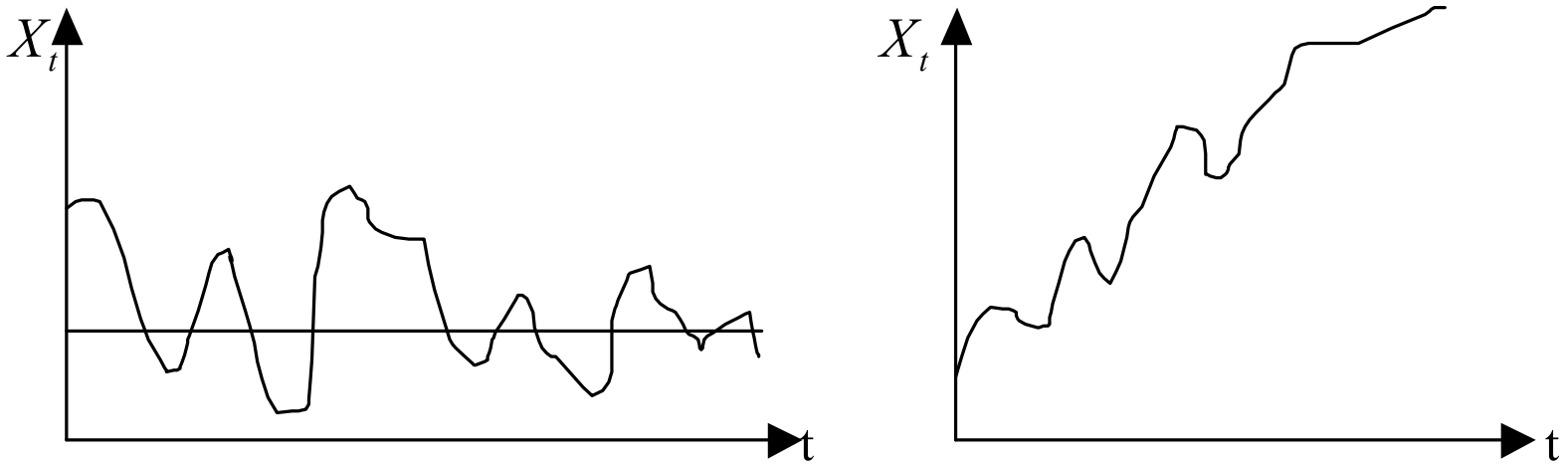


图 9.1　平稳时间序列与非平稳时间序列图

- 进一步的判断:检验样本自相关函数及其图形

定义随机时间序列的自相关函数（**autocorrelation function, ACF**）如下：

$$\rho_k=\gamma_k/\gamma_0$$

自相关函数是关于滞后期k的递减函数。

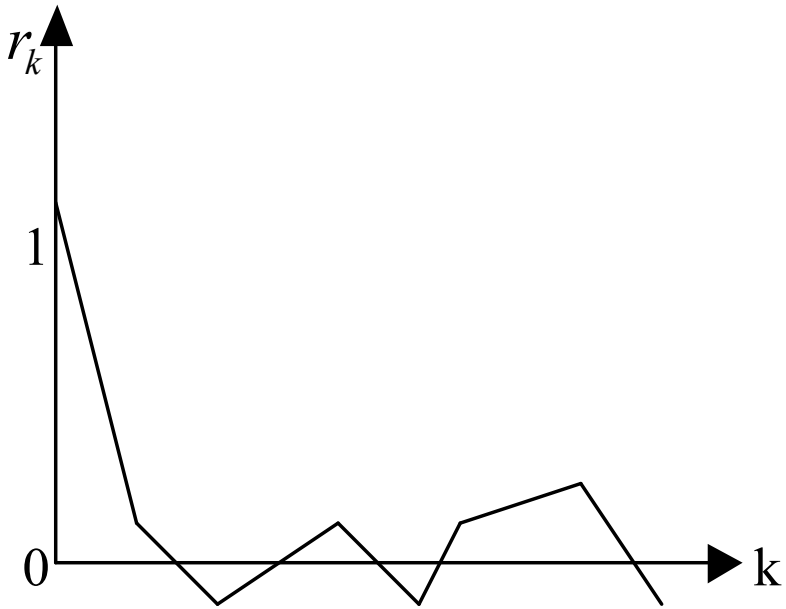
实际上,对一个随机过程只有一个实现（样本），因此，只能计算样本自相关函数（Sample autocorrelation function）。

Matlab命令：[ACF, lags, bounds] = autocorr( )

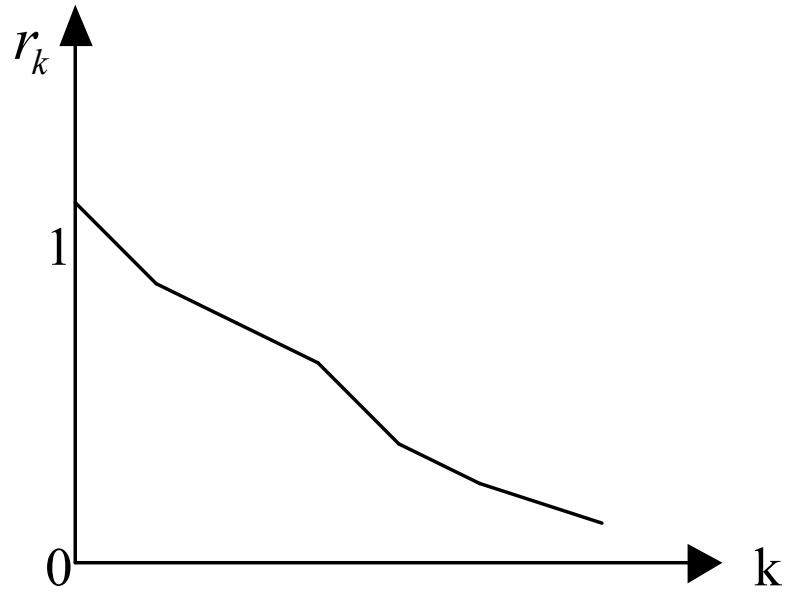
- 一个时间序列的样本自相关函数定义为：

$$r_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} \qquad k = 1,2,3,\cdots$$

随着k的增加，样本自相关函数下降且趋于零。但从下降速度来看，平稳序列要比非平稳序列快得多。



(a)



(b)

平稳时间序列与非平稳时间序列样本相关图

可检验对所有k>0，自相关系数都为0的联合假设，这可通过如下Q_{LB}统计量进行：

$$Q_{LB} = n(n+2)\sum_{k=1}^{m}\left(\frac{r_k^2}{n-k}\right)$$

该统计量近似地服从自由度为m的χ^2分布（m为滞后长度）。

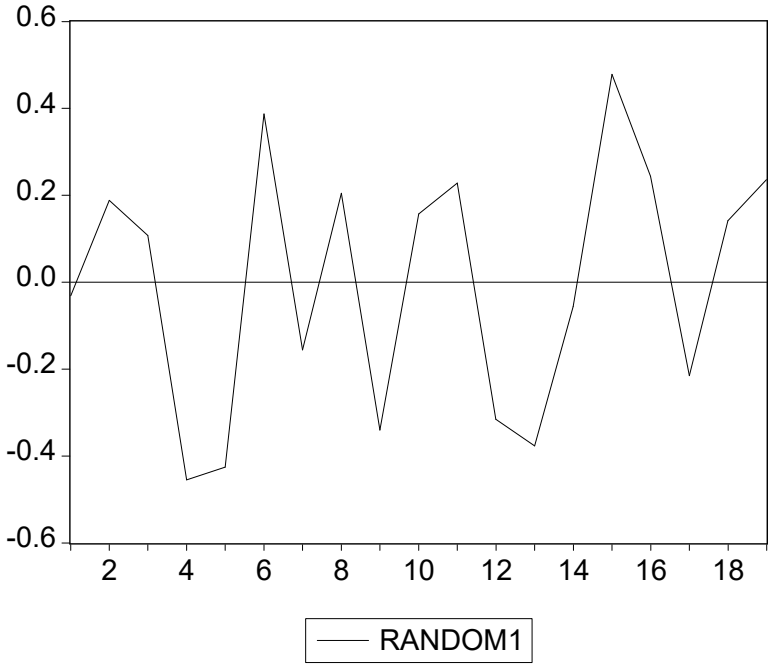
因此：如果计算的Q值大于显著性水平为α的临界值，则有1-α的把握拒绝所有ρ_k(k>0)同时为0的假设。

一个纯随机序列与随机游走序列的检验

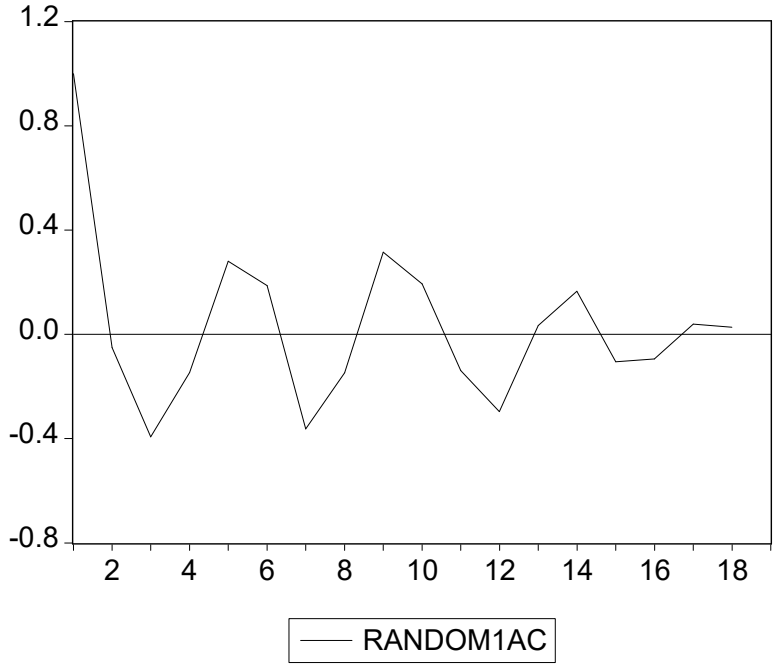
序号	Random1	自相关系数 r_k (k=0,1,...17)		Q_{LB}	Random2	自相关系数 r_k (k=0,1,...17)		Q_{LB}
1	-0.031	K=0,	1.000		-0.031		1.000	
2	0.188	K=1,	-0.051	0.059	0.157		0.480	5.116
3	0.108	K=2,	-0.393	3.679	0.264		0.018	5.123
4	-0.455	K=3,	-0.147	4.216	-0.191		-0.069	5.241
5	-0.426	K=4,	0.280	6.300	-0.616		0.028	5.261
6	0.387	K=5,	0.187	7.297	-0.229		-0.016	5.269
7	-0.156	K=6,	-0.363	11.332	-0.385		-0.219	6.745
8	0.204	K=7,	-0.148	12.058	-0.181		-0.063	6.876
9	-0.340	K=8,	0.315	15.646	-0.521		0.126	7.454
10	0.157	K=9,	0.194	17.153	-0.364		0.024	7.477
11	0.228	K=10,	-0.139	18.010	-0.136		-0.249	10.229
12	-0.315	K=11,	-0.297	22.414	-0.451		-0.404	18.389
13	-0.377	K=12,	0.034	22.481	-0.828		-0.284	22.994
14	-0.056	K=13,	0.165	24.288	-0.884		-0.088	23.514
15	0.478	K=14,	-0.105	25.162	-0.406		-0.066	23.866
16	0.244	K=15,	-0.094	26.036	-0.162		0.037	24.004
17	-0.215	K=16,	0.039	26.240	-0.377		0.105	25.483
18	0.141	K=17,	0.027	26.381	-0.236		0.093	27.198
19	0.236				0.000			

- 容易验证：该样本序列的均值为0，方差为0.0789。

- 从图形看：它在其样本均值0附近上下波动，且样本自相关系数迅速下降到0，随后在0附近波动且逐渐收敛于0。



（a）



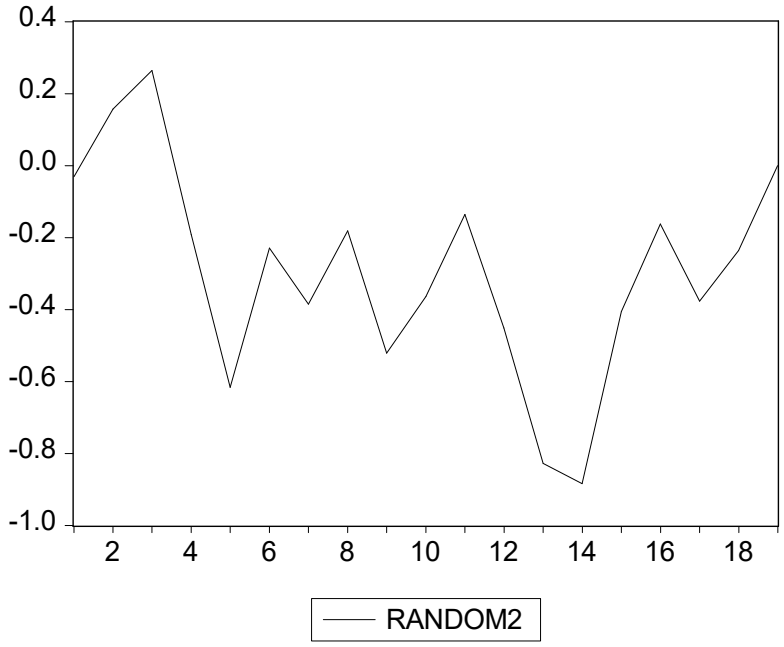
（b）

- 从Q_{LB}统计量的计算值看，滞后17期的计算值为26.38，未超过5%显著性水平的临界值27.58，因此,可以接受所有的自相关系数$\rho_k(k>0)$都为0的假设。
- 因此，Random1是一个平稳过程。

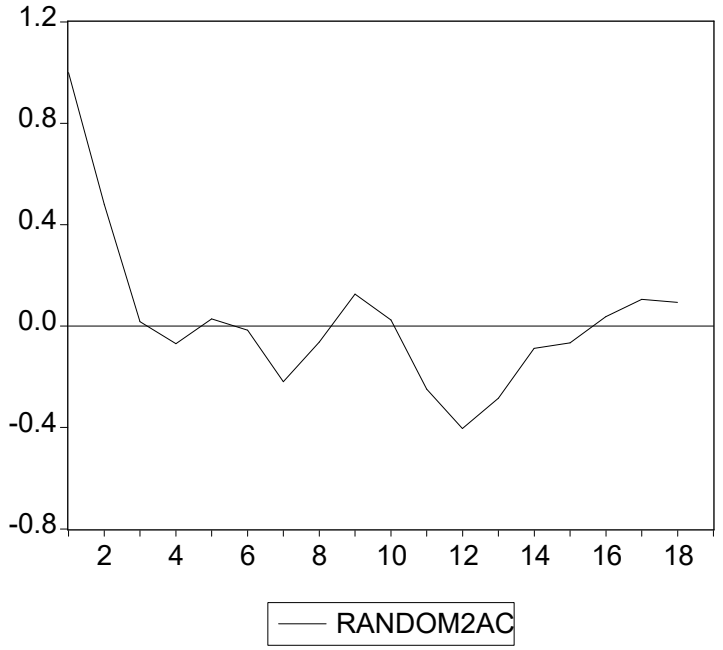
- 序列Random2是由一随机游走过程

$$X_t=X_{t-1}+\mu_t$$

生成的一随机游走时间序列样本。其中，μ_t是由Random1表示的白噪声。



（a）



（b）

图形表示出：该序列具有相同的均值，但从样本自相关图看，虽然自相关系数迅速下降到0，但随着时间的推移，则在0附近波动且呈发散趋势。

从Q_{LB}统计量的计算值看，滞后1期的计算值为5.116，超过5%显著性水平的临界值3.84，因此,拒绝自相关系数$\rho_k(k>0)$都为0的假设。

该随机游走序列是非平稳的。

平稳性的单位根检验（unit root test）

1. DF检验

2. ADF检验

（adftest.m）

03 ARMA模型

方法性工具

- 差分运算
- 延迟算子
- 线性差分方程

差分运算

- 一阶差分

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

- p 阶差分

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

- k 步差分

$$\nabla_k = x_t - x_{t-k}$$

延迟算子

- 延迟算子类似于一个时间指针，当前序列值乘以一个延迟算子，就相当于把当前序列值的时间向过去拨了一个时刻
- 记B为延迟算子，有

$$x_{t-p} = B^p x_t, \forall p \geq 1$$

延迟算子的性质

- $B^0 = 1$
- $B(c \cdot x_t) = c \cdot B(x_t) = c \cdot x_{t-1}$，$c$为任意常数
- $B(x_t \pm y_t) = x_{t-1} \pm y_{t-1}$
- $B^n x_t = x_{t-n}$
- $(1-B)^n = \sum_{i=0}^{n} (-1)^n C_n^i B^i$ ， 其中 $C_n^i = \dfrac{n!}{i!(n-i)!}$

用延迟算子表示差分运算

- p 阶差分

$$\nabla^p x_t = (1-B)^p x_t = \sum_{i=0}^{p} (-1)^p C_p^i x_{t-i}$$

- k步差分

$$\nabla_k = x_t - x_{t-k} = (1-B^k)x_t$$

线性差分方程

- 线性差分方程

$$z_t + a_1 z_{t-1} + a_2 z_{t-2} + \cdots + a_p z_{t-p} = h(t)$$

- 齐次线性差分方程

$$z_t + a_1 z_{t-1} + a_2 z_{t-2} + \cdots + a_p z_{t-p} = 0$$

齐次线性差分方程的解

- 特征方程

$$\lambda^p + a_1\lambda^{p-1} + a_2\lambda^{p-2} + \cdots + a_p = 0$$

- 特征方程的根称为特征根，记作 $\lambda_1, \lambda_2, \cdots, \lambda_p$
- 齐次线性差分方程的通解
  - 不相等实数根场合

$$z_t = c_1\lambda_1^t + c_2\lambda_2^t + \cdots + c_p\lambda_p^t$$

  - 有相等实根场合

$$z_t = (c_1 + c_2 t + \cdots + c_d t^{d-1})\lambda_1^t + c_{d+1}\lambda_{d+1}^t + \cdots + c_p\lambda_p^t$$

  - 复根场合

$$z_t = r^t(c_1 e^{it\varpi} + c_2 e^{-it\varpi}) + c_3\lambda_3^t + \cdots + c_p\lambda_p^t$$

非齐次线性差分方程的解

- 非齐次线性差分方程的特解
  - 使得非齐次线性差分方程成立的任意一个解z_t''

$$z_t'' + a_1 z_{t-1}'' + a_2 z_{t-2}'' + \cdots + a_p z_{t-p}'' = h(t)$$

- 非齐次线性差分方程的通解
  - 齐次线性差分方程的通解和非齐次线性差分方程的特解之和 z_t

$$z_t = z_t' + z_t''$$

ARMA模型的形式

自回归移动平均（autoregressive moving average model，简称ARMA）模型是平稳时间序列分析的经典方法。

ARMA模型的典型形式

ARMA(*p*，*q*)模型

$$x_t = c + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q}$$

AR(*p*)模型

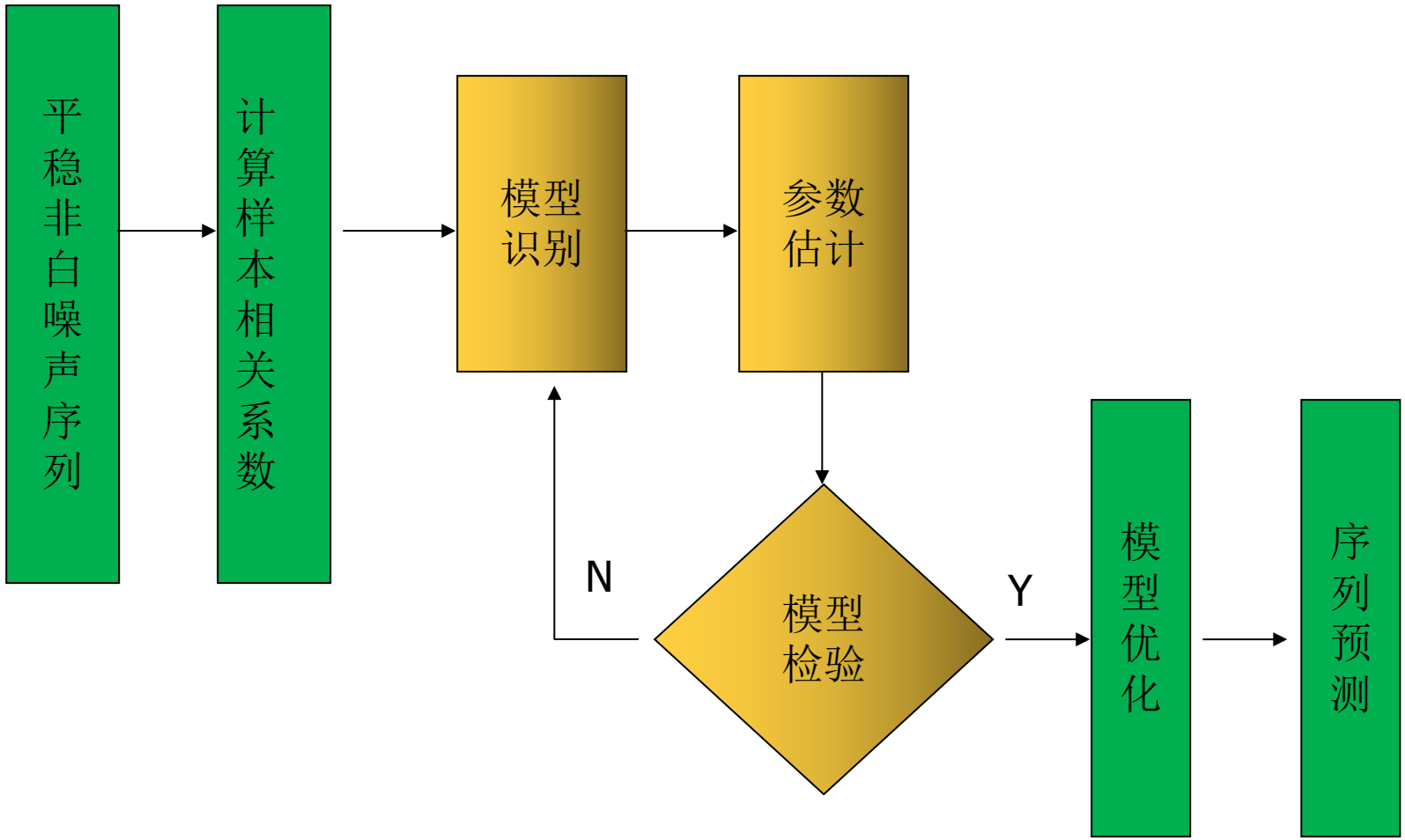
$$x_t = c + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \varepsilon_t$$

MA(*q*)模型

$$x_t = c + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_q \varepsilon_{t-q}$$

Box - Jenkins建模步骤

Box和Jenkins为**平稳时间序列**建模提供了一套标准的策略.



平稳性检验

在上述建模步骤之前，首先要对时间序列作图以观察其是否具有趋势特征，并对序列的平稳性做出初步的判断，然后需要对序列进行正规的**单位根检验**以判断其平稳性。

- **平稳时间序列的图形特征：**

（1）平稳时间序列的**时序图**应表现为**在特定水平值（即均值）附近的有界波动**；

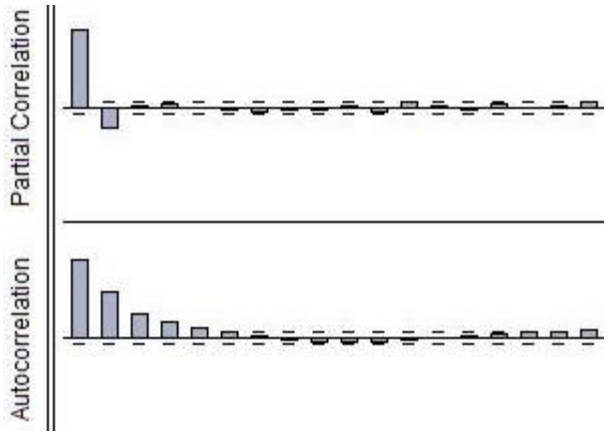
（2）平稳时间序列的**自相关系数图**应表现为**很快衰减向零**（AR和ARMA过程表现为以指数速度向零衰减的拖尾特性，MA过程表现为截尾特性）。

（1） 模型识别

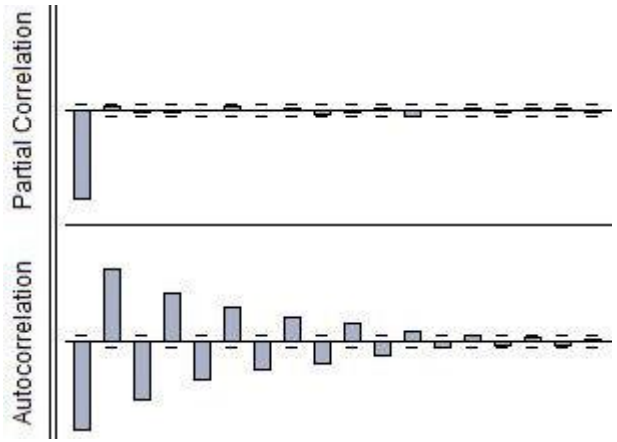
模型	自相关函数（ACF）	偏自相关函数（PACF）
AR(p)	拖尾	p阶截尾
MA(q)	q阶截尾	拖尾
ARMA(p，q)	拖尾	拖尾

AR模型的样本自相关图和偏自相关图示例

$$x_t = 0.8x_{t-1} - 0.15x_{t-2} + \varepsilon_t$$

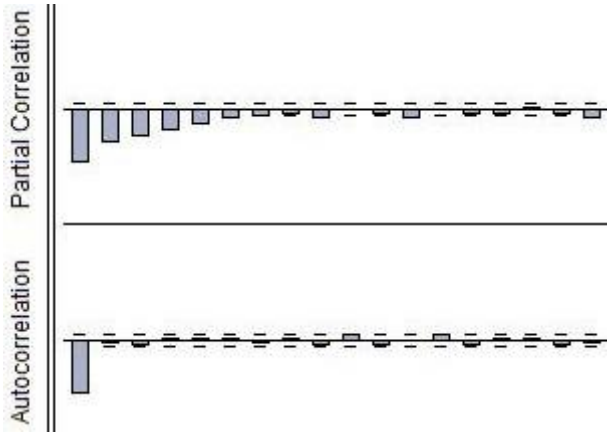


$$x_t = -0.8x_{t-1} + \varepsilon_t$$

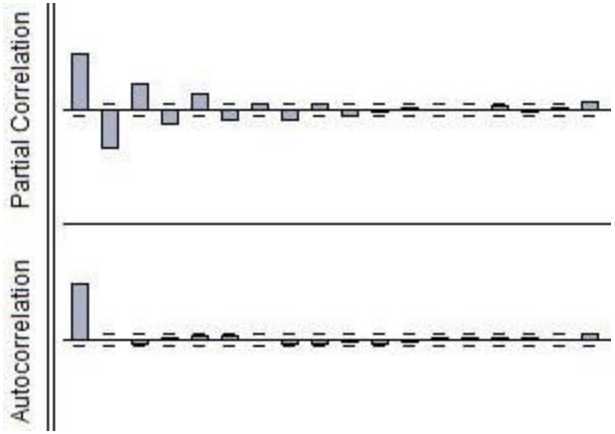


MA模型的样本自相关图和偏自相关图示例

$$x_t = \varepsilon_t - 0.8\varepsilon_{t-1}$$

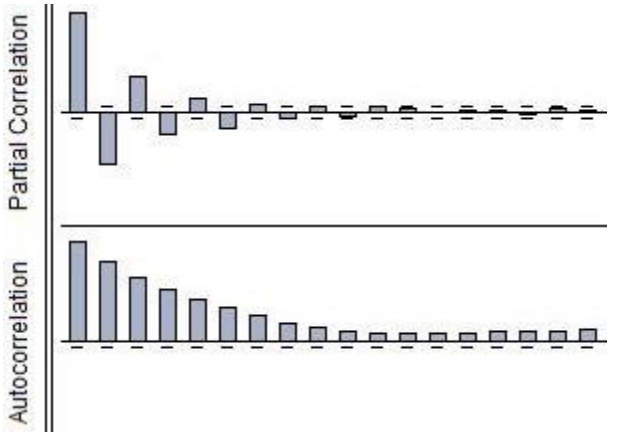


$$x_t = \varepsilon_t + 0.8\varepsilon_{t-1}$$

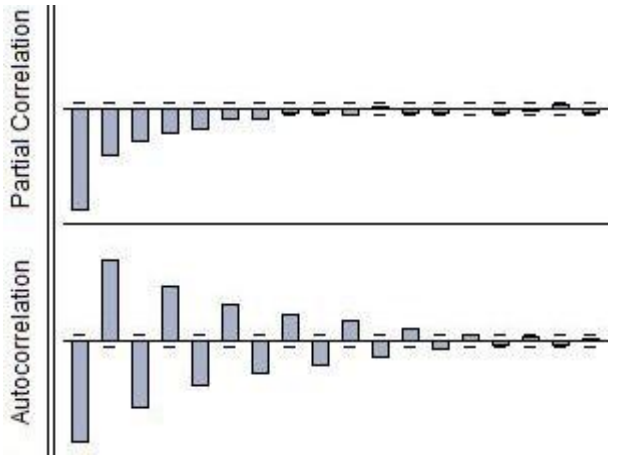


ARMA模型的样本自相关图和偏自相关图示例

$$x_t = 0.8x_{t-1} + \varepsilon_t + 0.8\varepsilon_{t-1}$$



$$x_t = -0.8x_{t-1} + \varepsilon_t - 0.8\varepsilon_{t-1}$$



特别地，如果序列具有纯随机性（例如白噪声过程），此时该序列的水平值并没有相关性规律可供建模，因此无需对该序列构建ARMA模型。

（2）参数估计

经过模型识别环节确定好合适的模型形式，即可通过各种统计软件通过极大似然估计或非线性最小二乘估计等方法得出模型的参数估计结果。

（3）诊断检验

ARMA模型的诊断检验主要包括三项内容：

- 参数显著性
- 平稳可逆性
- 残差的纯随机性

Q统计量

原假设：$\hat{\rho}_1 = \hat{\rho}_2 = \cdots = \hat{\rho}_s = 0$

Q统计量：$Q = T\sum_{k=1}^{s} \hat{\rho}_k^2$

修正的Q统计量：$Q = T(T+2)\sum_{k=1}^{s} \hat{\rho}_k^2 / (T-k)$

原假设成立条件下，Q统计量近似服从自由度为s的χ^2分布 。如果Q统计量是用于检验ARMA(p，q)模型残差的纯随机性，则上述两种Q统计量服从χ^2分布的自由度为s-p-q（若模型中还包含常数项，则自由度为s-p-q-1），其自由度不再是s

模型选择

应用上述Box-Jenkins建模方法，有可能有多个模型都能够通过诊断检验。

为了选出一个更加合适的模型，可以考虑采用AIC（Akaike Information Criterion）或SBC（Schwarz Bayesian Criterion）等信息判断准则，来筛选出一个更优的模型。

根据AIC和SBC信息判断准则的统计原理，应选择AIC和SBC统计量数值最小的模型作为更优模型 。

ARMA模型的预测

- 最小均方误差预测
- 条件期望
- 预测误差
- AR(p)过程的预测
- MA(q)过程的预测
- ARMA(p，q)过程的预测

最小均方误差预测

所谓时间序列的**预测**，就是根据所有已知的历史信息，对时间序列未来某个时期的发展水平进行预估。

常用**均方误差**（mean square error，MSE）来评价预测效果。

$$MSE(\hat{y}_{t+k|t}) \equiv E(y_{t+k} - \hat{y}_{t+k|t})^2$$

AR(p)过程的预测

考虑AR(p)过程 $$x_t = c + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots \alpha_p x_{t-p} + \varepsilon_t$$

1. 条件期望

$$\begin{aligned} E_t x_{t+k} &= E\left(c + \alpha_1 x_{t+k-1} + \alpha_2 x_{t+k-2} + \cdots + \alpha_p x_{t+k-p} + \varepsilon_{t+k} \middle| x_t,\ x_{t-1}, \cdots\right) \\ &= c + \alpha_1 E_t x_{t+k-1} + \alpha_2 E_t x_{t+k-2} + \cdots + \alpha_p E_t x_{t+k-p} + E_t \varepsilon_{t+k} \\ &= c + \alpha_1 E_t x_{t+k-1} + \alpha_2 E_t x_{t+k-2} + \cdots + \alpha_p E_t x_{t+k-p} \end{aligned}$$

2. 预测误差的方差

$$Var\left[e_t(k)\right] = \left[1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2\right]\sigma_\varepsilon^2$$

权系数ψ_j可以根据AR (p)过程的格林函数递推公式计算得到

MA(*q*)过程的预测

考虑MA(*q*)过程 $$x_t = c + \varepsilon_t - \beta_1\varepsilon_{t-1} - \beta_2\varepsilon_{t-2} - \cdots - \beta_q\varepsilon_{t-q}$$

1. 条件期望

$$\begin{aligned} E_t x_{t+k} &= E\left(c + \varepsilon_{t+k} - \beta_1\varepsilon_{t+k-1} - \beta_2\varepsilon_{t+k-2} - \cdots - \beta_q\varepsilon_{t+k-q} \middle| y_t,\ y_{t-1}, \cdots\right) \\ &= c + E_t\varepsilon_{t+k} - \beta_1 E_t\varepsilon_{t+k-1} - \beta_2 E_t\varepsilon_{t+k-2} - \cdots - \beta_q E_t\varepsilon_{t+k-q} \\ &= \begin{cases} c - \beta_k\varepsilon_t - \cdots - \beta_q\varepsilon_{t+k-q}, & k \le q \\ c, & k > q \end{cases} \end{aligned}$$

2. 预测误差的方差

$$Var\left[e_t(k)\right] = \left[1 + \beta_1^2 + \beta_2^2 + \cdots + \beta_{k-1}^2\right]\sigma_\varepsilon^2$$

ARMA(p，q)过程的预测

考虑ARMA(p，q)过程

$$x_t = c + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots \alpha_p x_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_q \varepsilon_{t-q}$$

1. 条件期望

$$\begin{aligned}
&E_t x_{t+k} \\
&= E\left(c + \alpha_1 x_{t+k-1} + \cdots + \alpha_p x_{t+k-p} + \varepsilon_{t+k} - \beta_1 \varepsilon_{t+k-1} - \cdots - \beta_q \varepsilon_{t+k-q} \middle| x_t,\ x_{t-1}, \cdots\right) \\
&= c + \alpha_1 E_t x_{t+k-1} \cdots + \alpha_p E_t x_{t+k-p} + E_t \varepsilon_{t+k} - \beta_1 E_t \varepsilon_{t+k-1} - \cdots - \beta_q E_t \varepsilon_{t+k-q} \\
&= \begin{cases} c + \alpha_1 E_t x_{t+k-1} + \cdots + \alpha_p E_t x_{t+k-p} - \beta_k \varepsilon_t - \cdots - \beta_q \varepsilon_{t+k-q}, & k \le q \\ c + \alpha_1 E_t x_{t+k-1} + \cdots + \alpha_p E_t x_{t+k-p}, & k > q \end{cases}
\end{aligned}$$

2. 预测误差的方差

$$Var\left[e_t(k)\right] = \left[1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{k-1}^2\right]\sigma_\varepsilon^2$$

04 ARIMA模型

确定性因素分解

- 传统的因素分解
  - 长期趋势
  - 循环波动
  - 季节性变化
  - 随机波动
- 现在的因素分解
  - 长期趋势波动
  - 季节性变化
  - 随机波动

一　趋势分析

- 目的
  - 有些时间序列具有非常显著的趋势，我们分析的目的就是要找到序列中的这种趋势，并利用这种趋势对序列的发展作出合理的预测
- 常用方法
  - 趋势拟合法
  - 平滑法

（1）趋势拟合法

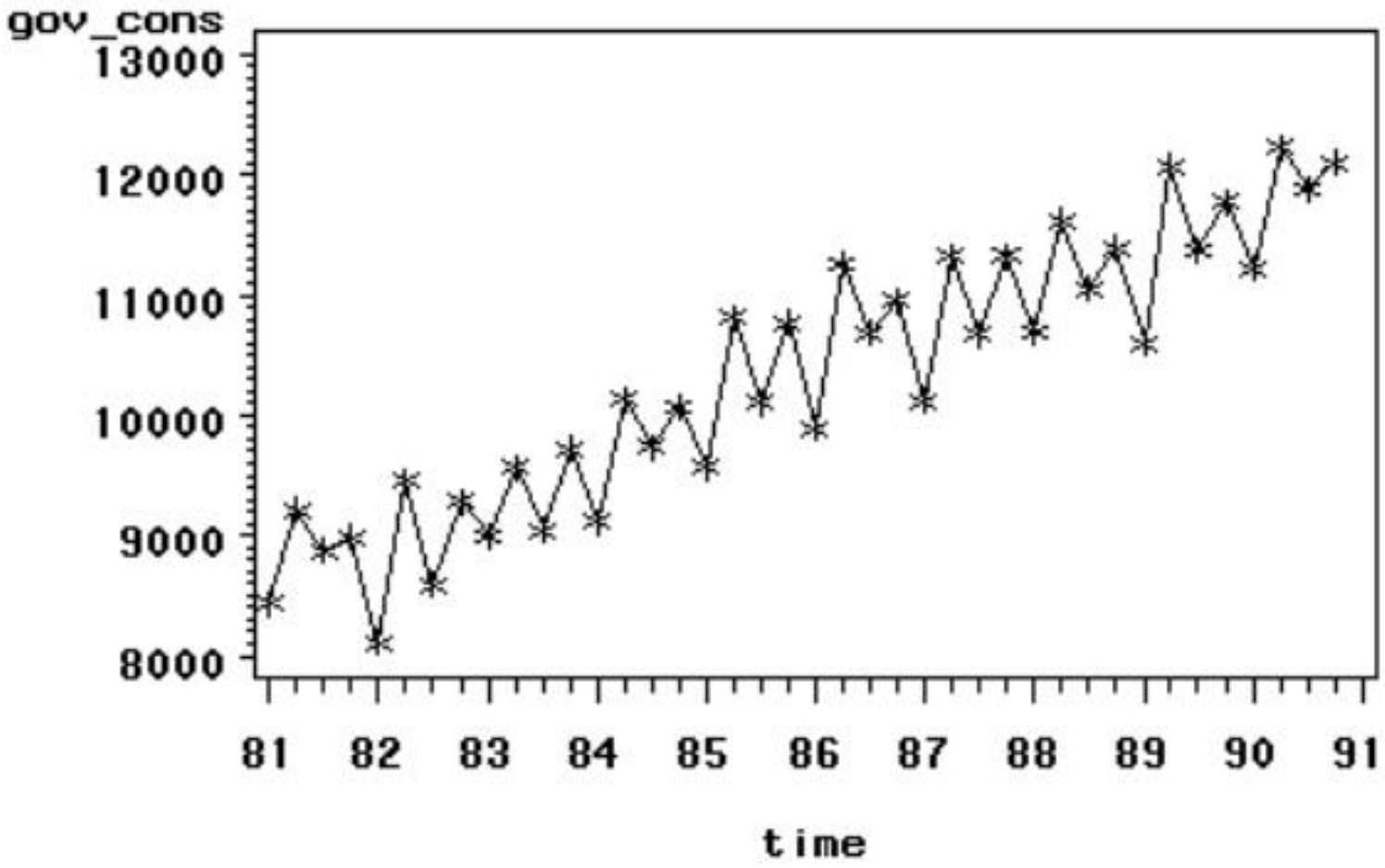
- 趋势拟合法就是把时间作为自变量，相应的序列观察值作为因变量，建立序列值随时间变化的回归模型的方法
- 分类
  - 线性拟合
  - 非线性拟合

线性拟合

- 使用场合
  - 长期趋势呈现出线形特征
- 模型结构

$$\begin{cases} x_t = a + bt + I_t \\ E(I_t) = 0, Var(I_t) \end{cases}$$

例4.1:拟合澳大利亚政府1981——1990年每季度的消费支出序列



线性拟合

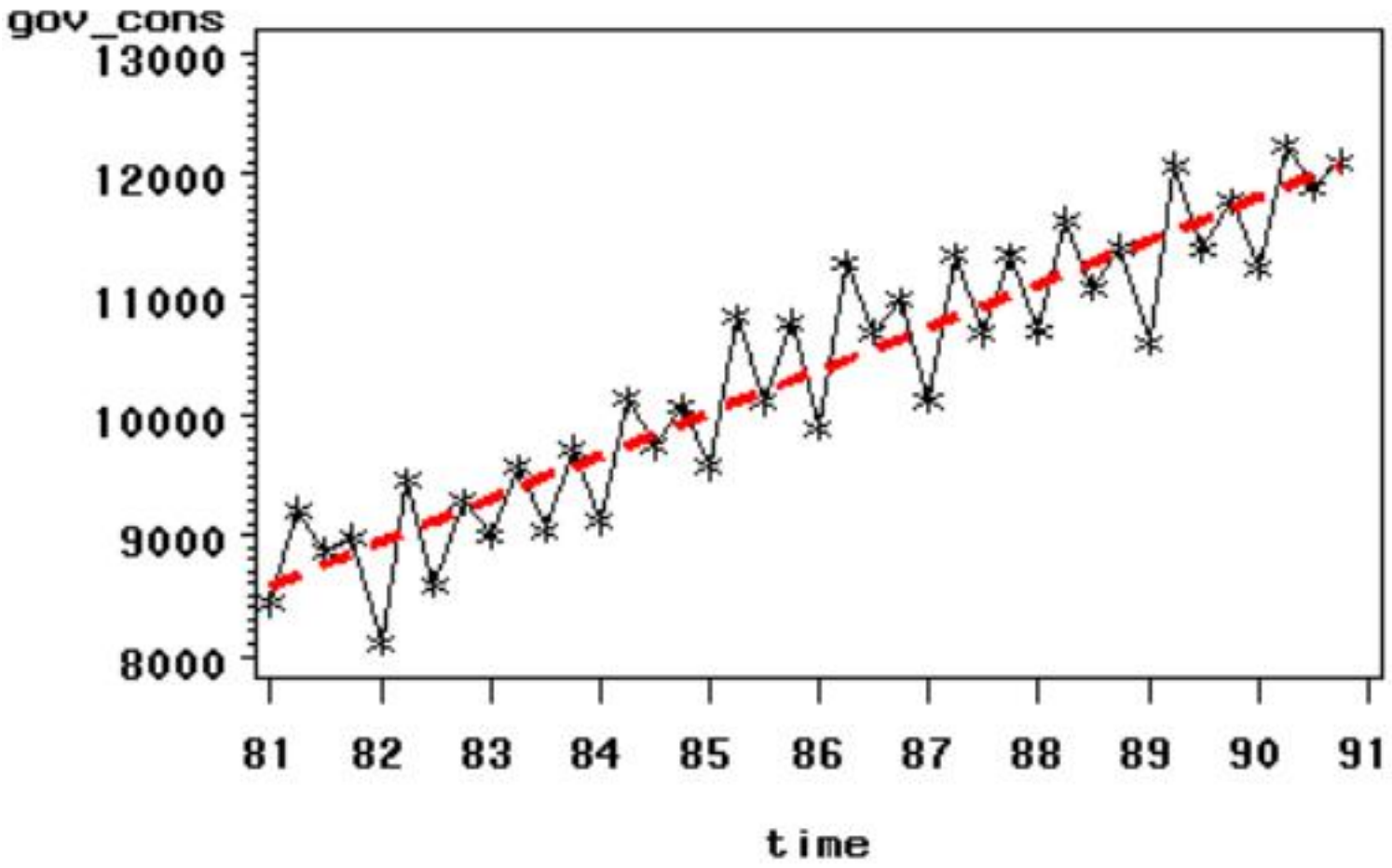
- 模型

$$\begin{cases} x_t = a + bt + I_t & ,t = 1,2\cdots,40 \\ E(I_t) = 0, Var(I_t) = \sigma^2 \end{cases}$$

- 参数估计方法
  - 最小二乘估计
- 参数估计值

$$\hat{a} = 8498.69 \quad , \quad \hat{b} = 89.12$$

拟合效果图


gov_cons
13000
12000
11000
10000
9000
8000
81
82
83
84
85
86
87
88
89
90
91
time

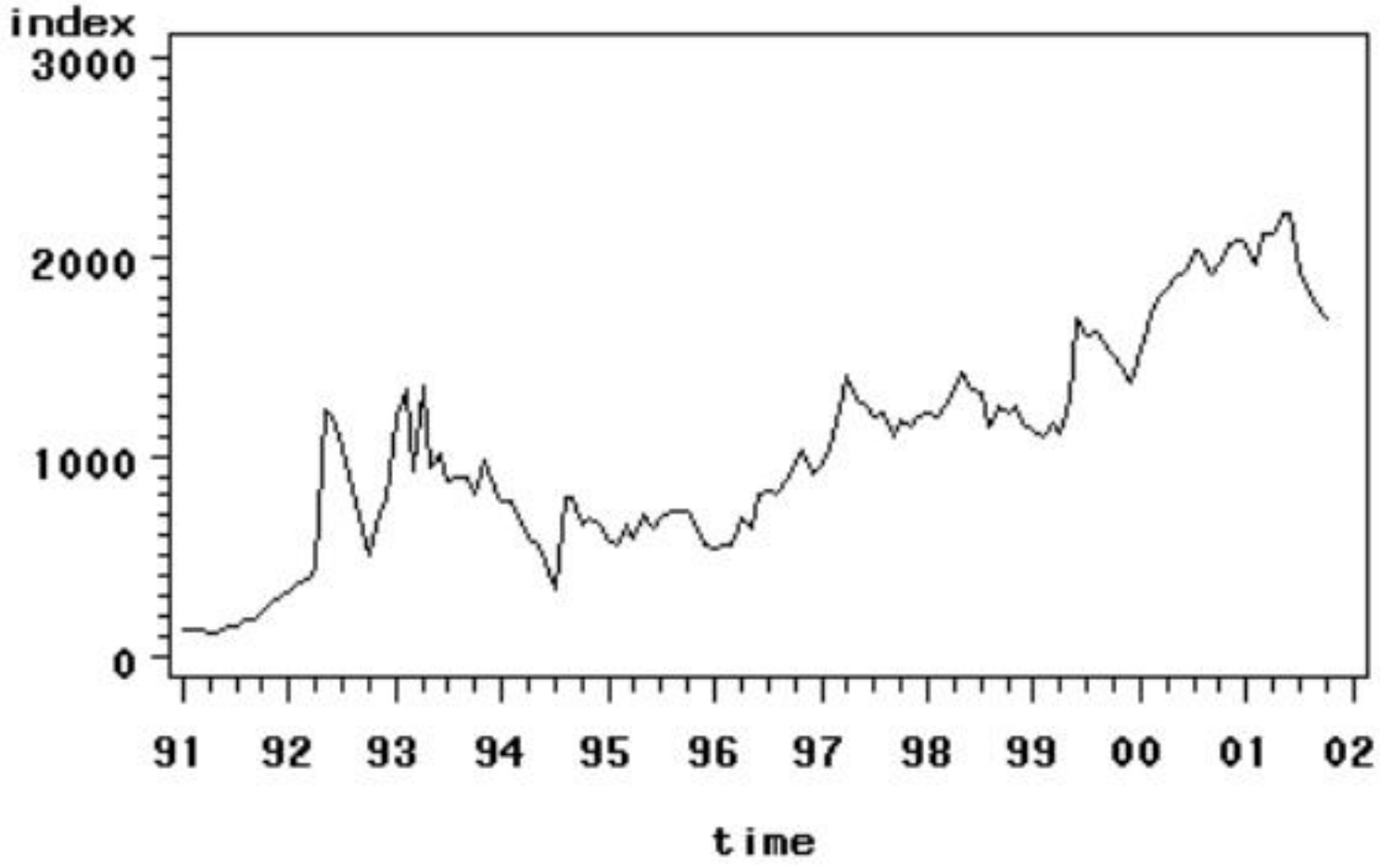
非线性拟合

- 使用场合
  - 长期趋势呈现出非线形特征
- 参数估计指导思想
  - 能转换成线性模型的都转换成线性模型，用线性最小二乘法进行参数估计
  - 实在不能转换成线性的，就用迭代法进行参数估计

常用非线性模型

模型	变换	变换后模型	参数估计方法
$T_t = a + bt + ct^2$	$t_2 = t^2$	$T_t = a + bt + ct_2$	线性最小二乘估计
$T_t = ab^t$	$T_t' = \ln T_t$ $a' = \ln a$ $b' = \ln b$	$T_t' = a' + b't$	线性最小二乘估计
$T_t = a + bc^t$	—	—	迭代法
$T_t = e^{a+bc^t}$	—	—	迭代法
$T_t = \dfrac{1}{a + bc^t}$	—	—	迭代法

例4.2：对上海证券交易所每月末上证指数序列进行模型拟合



非线性拟合

- 模型

$$T_t = a + bt + ct^2$$

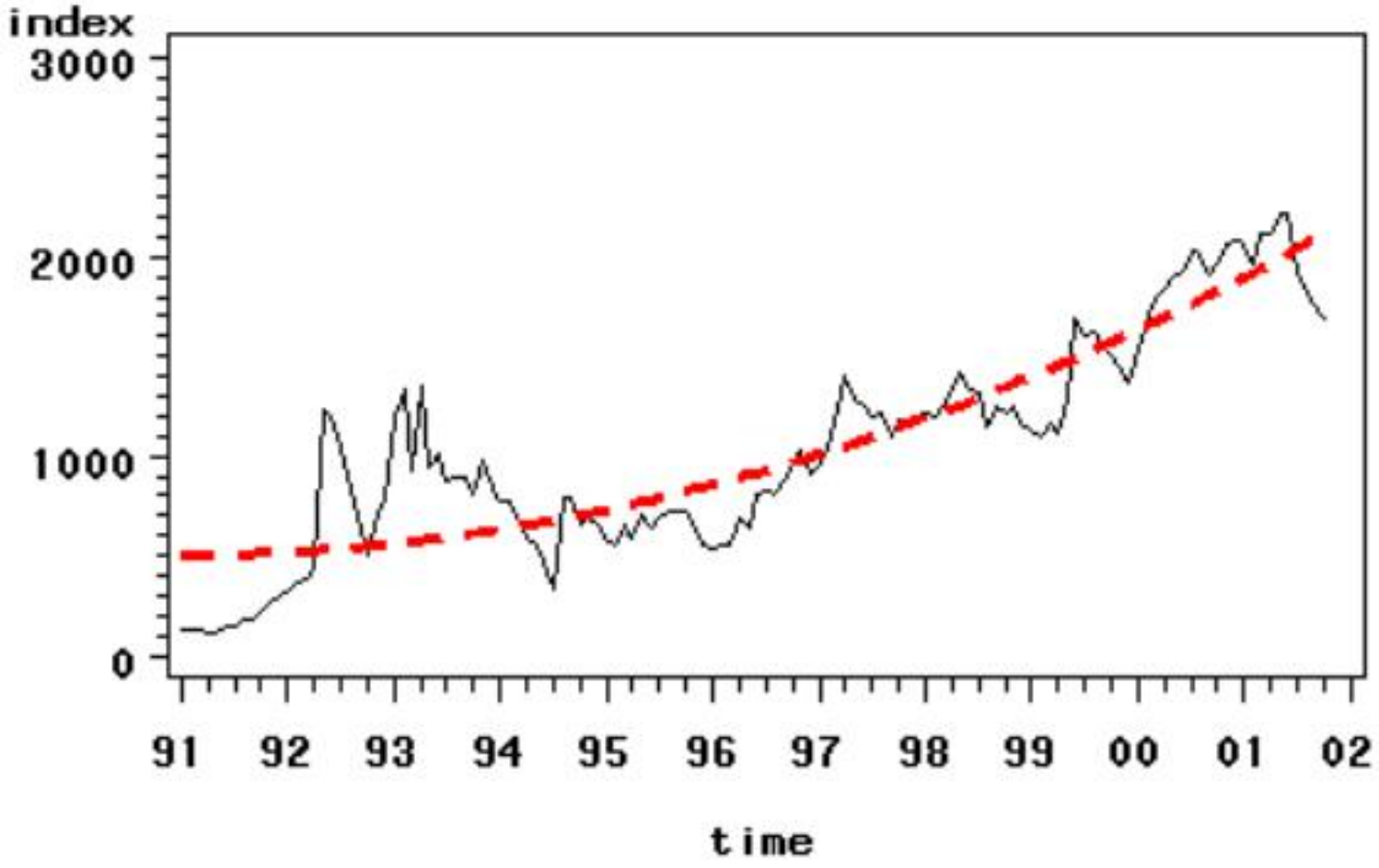
- 变换

$$t_2 = t^2$$

- 参数估计方法
  - 线性最小二乘估计
- 拟合模型口径

$$T_t = 502.2517 + 0.0952t^2$$

拟合效果图



（2）平滑法

- 平滑法是进行趋势分析和预测时常用的一种方法。它是利用修匀技术，削弱短期随机波动对序列的影响，使序列平滑化，从而显示出长期趋势变化的规律
- 常用平滑方法
  - 移动平均法
  - 指数平滑法

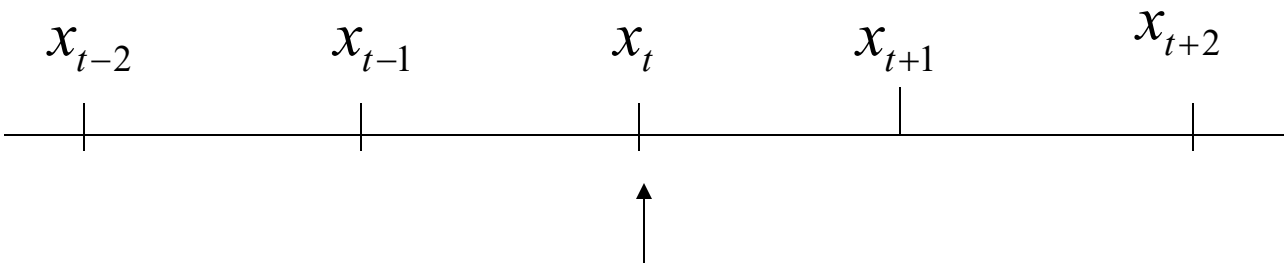
移动平均法

- 基本思想
  - 假定在一个比较短的时间间隔里，序列值之间的差异主要是由随机波动造成的。根据这种假定，我们可以用一定时间间隔内的平均值作为某一期的估计值
- 分类
  - n期中心移动平均
  - n期移动平均

n期中心移动平均

$$\tilde{x}_t = \begin{cases} \frac{1}{n}\left(x_{t-\frac{n-1}{2}} + x_{t-\frac{n-1}{2}+1} + \cdots + x_t + \cdots + x_{t+\frac{n-1}{2}-1} + x_{t+\frac{n-1}{2}}\right), & n\text{为奇数} \\ \frac{1}{n}\left(\frac{1}{2}x_{t-\frac{n}{2}} + x_{t-\frac{n}{2}+1} + \cdots + x_t + \cdots + x_{t+\frac{n}{2}-1} + \frac{1}{2}x_{t+\frac{n}{2}}\right), & n\text{为偶数} \end{cases}$$

5期中心移动平均

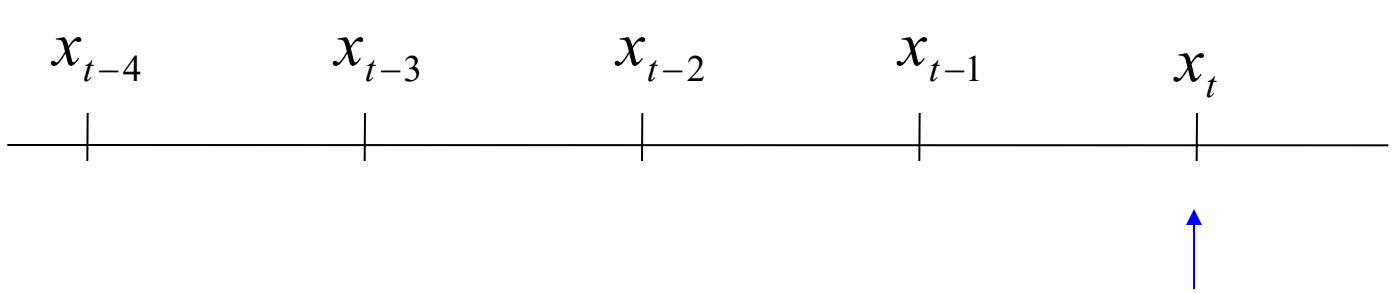


$$\tilde{x}_t = \frac{x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2}}{5}$$

n期移动平均

$$\tilde{x}_t = \frac{1}{n}(x_t + x_{t-1} + \cdots + x_{t-n+1})$$

5期移动平均



$$\tilde{x}_t = \frac{x_{t-4} + x_{t-3} + x_{t-2} + x_{t-1} + x_t}{5}$$

移动平均期数确定的原则

- 事件的发展有无周期性
  - 以周期长度作为移动平均的间隔长度，以消除周期效应的影响
- 对趋势平滑的要求
  - 移动平均的期数越多，拟合趋势越平滑
- 对趋势反映近期变化敏感程度的要求
  - 移动平均的期数越少，拟合趋势越敏感

移动平均预测

$$\hat{x}_{T+l} = \frac{1}{n}(x'_{T+l-1} + x'_{T+l-2} + \cdots + x'_{T+l-n})$$

$$x'_{T+l-i} = \begin{cases} \hat{x}_{T+l-i} & ,l > i \\ x_{T+l-i} & ,l \le i \end{cases}$$

例4.3

- 某一观察值序列最后4期的观察值为：

  5，5.5，5.8，6.2

（1）使用4期移动平均法预测 $\hat{x}_{T+2}$.

（2）求在二期预测值 $\hat{x}_{T+2}$中 x_T前面的系数等于多少？

解

（1） $\hat{x}_{T+1}=\frac{1}{4}\left(x_T+x_{T-1}+x_{T-2}+x_{T-3}\right)=\frac{5+5.4+5.8+6.2}{4}=5.6$

$\hat{x}_{T+2}=\frac{1}{4}\left(\hat{x}_{T+1}+x_T+x_{T-1}+x_{T-2}\right)=\frac{5.6+5+5.4+5.8}{4}=5.45$

（2） $\hat{x}_{T+2}=\frac{1}{4}\left(\hat{x}_{T+1}+x_T+x_{T-1}+x_{T-2}\right)$

$=\frac{1}{4}\left[\frac{1}{4}\left(x_T+x_{T-1}+x_{T-2}+x_{T-2}\right)+x_T+x_{T-1}+x_{T-2}\right]$

$=\frac{5}{16}\left(x_T+x_{T-1}+x_{T-2}\right)+\frac{1}{16}x_{T-3}$

在二期预测值中 x_T 前面的系数等于 $\frac{5}{16}$

指数平滑法

- 指数平滑方法的基本思想
  - 在实际生活中，我们会发现对大多数随机事件而言，一般都是近期的结果对现在的影响会大些，远期的结果对现在的影响会小些。为了更好地反映这种影响作用，我们将考虑到时间间隔对事件发展的影响，各期权重随时间间隔的增大而呈指数衰减。这就是指数平滑法的基本思想
- 分类
  - 简单指数平滑
  - Holt两参数指数平滑

简单指数平滑

- 基本公式

$$\tilde{x}_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \cdots$$

- 等价公式

$$\tilde{x}_t = \alpha x_t + (1-\alpha)\tilde{x}_{t-1}$$

经验确定

- 初始值的确定

$$\tilde{x}_0 = x_1$$

- 平滑系数的确定
  - 一般对于变化缓慢的序列，α 常取较小的值
  - 对于变化迅速的序列，α常取较大的值
  - 经验表明 α的值介于0.05至0.3之间，修匀效果比较好。

简单指数平滑预测

- 一期预测值

$$\hat{x}_{T+1} = \tilde{x}_T$$
$$= \alpha x_T + \alpha(1-\alpha)x_{T-1} + \alpha(1-\alpha)^2 x_{T-2} + \cdots$$

- 二期预测值

$$\hat{x}_{T+2} = \alpha\hat{x}_{T+1} + \alpha(1-\alpha)x_T + \alpha(1-\alpha)^2 x_{T-1} + \cdots$$
$$= \alpha\hat{x}_{T+1} + (1-\alpha)\hat{x}_{T+1} = \hat{x}_{T+1}$$

- l 期预测值

$$\hat{x}_{T+l} = \hat{x}_{T+1} \quad , \quad l \geq 2$$

例4.4

- 对某一观察值序列$\{x_t\}$使用指数平滑法。已知 $x_T=10$，$\tilde{x}_{T-1}=10.5$，平滑系数 $\alpha=0.25$

  (1) 求二期预测值 $\hat{x}_{T+2}$。

  (2)求在二期预测值 $\hat{x}_{T+2}$中x_T前面的系数等于多少？

解

（1）$\hat{x}_{T+1}=\tilde{x}_T=0.25x_T+0.75\tilde{x}_{T-1}=10.3$

$\hat{x}_{T+2}=\hat{x}_{T+1}=10.3$

（2）$\hat{x}_{T+2}=\hat{x}_{T+1}=\alpha x_T+\alpha(1-\alpha)x_{T-1}+\cdots$

所以使用简单指数平滑法二期预测值中x_T前面的系数就等于平滑系数 $\alpha=0.25$

Holt两参数指数平滑

- 使用场合
  - 适用于对含有线性趋势的序列进行修匀
- 构造思想
  - 假定序列有一个比较固定的线性趋势

$$\hat{x}_t = x_{t-1} + r$$

  - 两参数修匀

$$\begin{cases} \tilde{x}_t = \alpha x_t + (1-\alpha)(\tilde{x}_{t-1} + r_{t-1}) \\ r_t = \gamma(\tilde{x}_t - \tilde{x}_{t-1}) + (1-\gamma) r_{t-1} \end{cases}$$

初始值的确定

- 平滑序列的初始值

$$\tilde{x}_0 = x_1$$

- 趋势序列的初始值

$$r_0 = \frac{x_{n+1} - x_1}{n}$$

Holt两参数指数平滑预测

- l 期预测值

$$\hat{x}_{T+l} = \tilde{x}_T + l \cdot r_T$$

例

- 对北京市1978——2000年报纸发行量序列进行Holt两参数指数平滑。指定

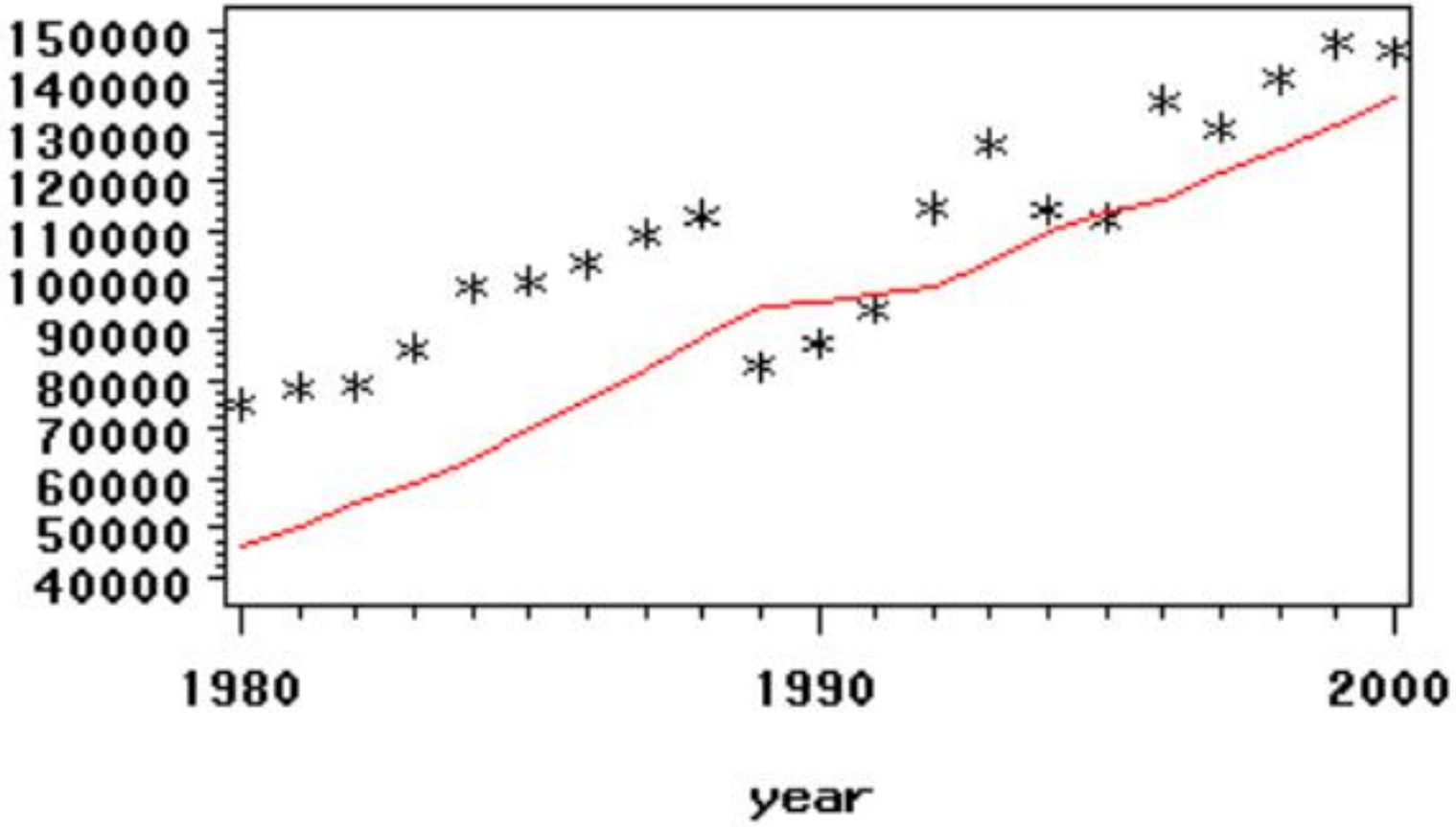
$$\tilde{x}_0 = x_1 = 51259$$

$$r_0 = \frac{x_{23} - x_1}{23} = 4325$$

$$\alpha = 0.15$$

$$\gamma = 0.1$$

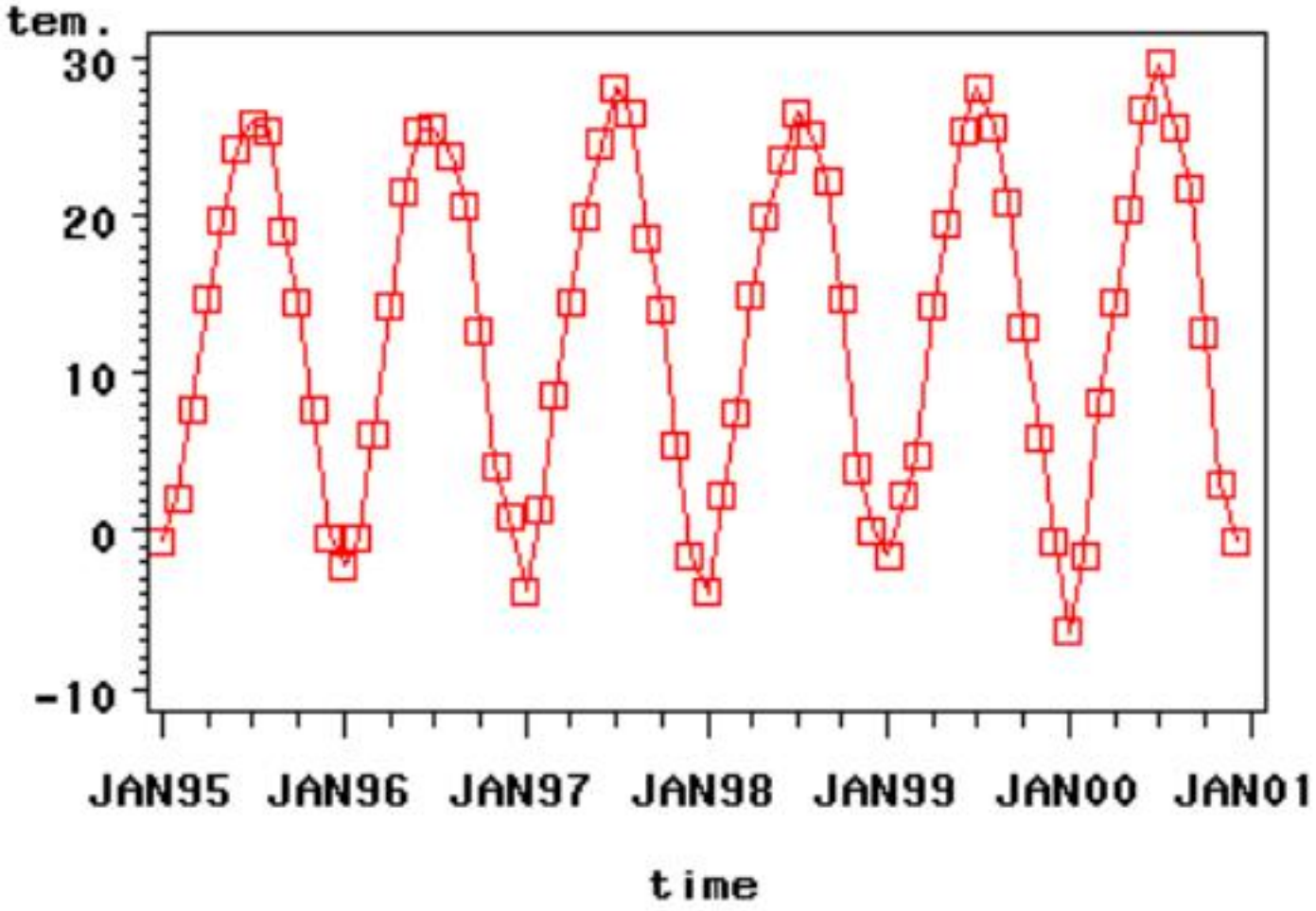
平滑效果图


150000
140000
130000
120000
110000
100000
90000
80000
70000
60000
50000
40000
1980
1990
2000
year

二　季节效应分析

【例】以北京市1995年——2000年月平均气温序列为例，介绍季节效应分析的基本思想和具体操作步骤。

时序图


tem.
30
20
10
0
-10
JAN95
JAN96
JAN97
JAN98
JAN99
JAN00
JAN01
time

季节指数

- 季节指数的概念
  - 所谓季节指数就是用简单平均法计算的周期内各时期季节性影响的相对数
- 季节模型

$$x_{ij} = \bar{x} \cdot S_j + I_{ij}$$

季节指数的计算

- 计算**周期**内各期平均数

$$\bar{x}_k = \frac{\sum_{i=1}^{n} x_{ik}}{n} \quad ,k = 1,2,\cdots,m$$

- 计算总平均数

$$\bar{x} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} x_{ik}}{nm}$$

- 计算季节指数

$$S_k = \frac{\bar{x}_k}{\bar{x}} \quad ,k = 1,2,\cdots,m$$

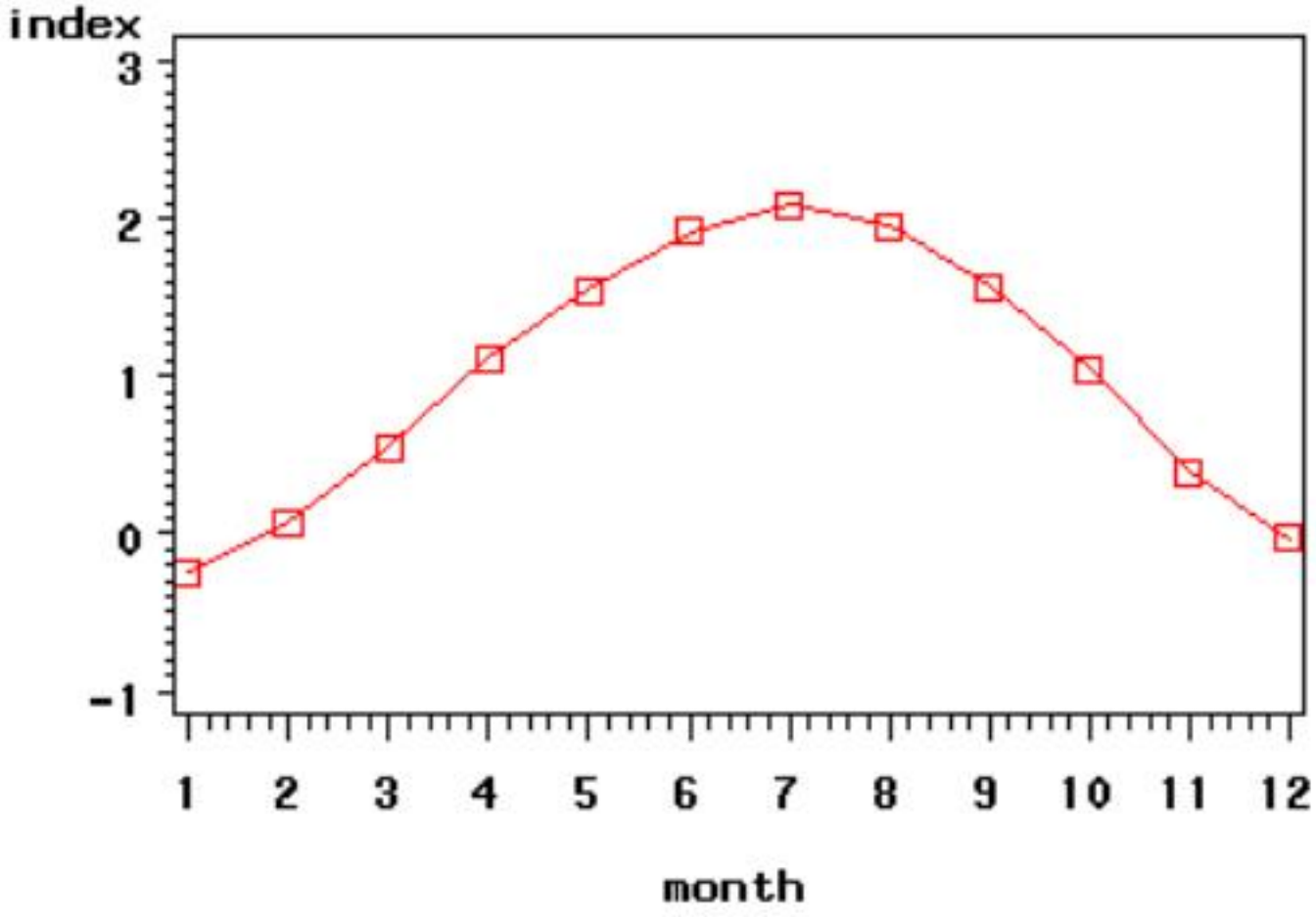
季节指数的理解

- 季节指数反映了该季度与总平均值之间的一种比较稳定的关系
- 如果这个比值大于1，就说明该季度的值常常会高于总平均值
- 如果这个比值小于1，就说明该季度的值常常低于总平均值
- 如果序列的季节指数都近似等于1，那就说明该序列没有明显的季节效应

例4.5：季节指数的计算

年 / 月	平均气温（℃）						月平均 $\bar{x}_{i.}$	季节指数 S_i
	1995	1996	1997	1998	1999	2000		
1	-0.7	-2.2	-3.8	-3.9	-1.6	-6.4	-3.10	-0.238
2	2.1	-0.4	1.3	2.4	2.2	-1.5	1.02	0.078
3	7.7	6.2	8.7	7.6	4.8	8.1	7.18	0.551
4	14.7	14.3	14.5	15.0	14.4	14.6	14.58	1.119
5	19.8	21.6	20.0	19.9	19.5	20.4	20.20	1.550
6	24.3	25.4	24.6	23.6	25.4	26.7	25.00	1.919
7	25.9	25.5	28.2	26.5	28.1	29.6	27.30	2.095
8	25.4	23.9	26.6	25.1	25.6	25.7	25.38	1.948
9	19.0	20.7	18.6	22.2	20.9	21.8	20.53	1.576
10	14.5	12.8	14.0	14.8	13.0	12.6	13.62	1.045
11	7.7	4.2	5.4	4.0	5.9	3.0	5.03	0.386
12	-0.4	0.9	-1.5	0.1	-0.6	-0.6	-0.35	-0.027
总平均 $\bar{x}$ =13.03								

例4.5季节指数图



综合分析

- 常用综合分析模型
  - 加法模型

$$x_t = T_t + S_t + I_t$$

  - 乘法模型

$$x_t = T_t \cdot S_t \cdot I_t$$

  - 混合模型

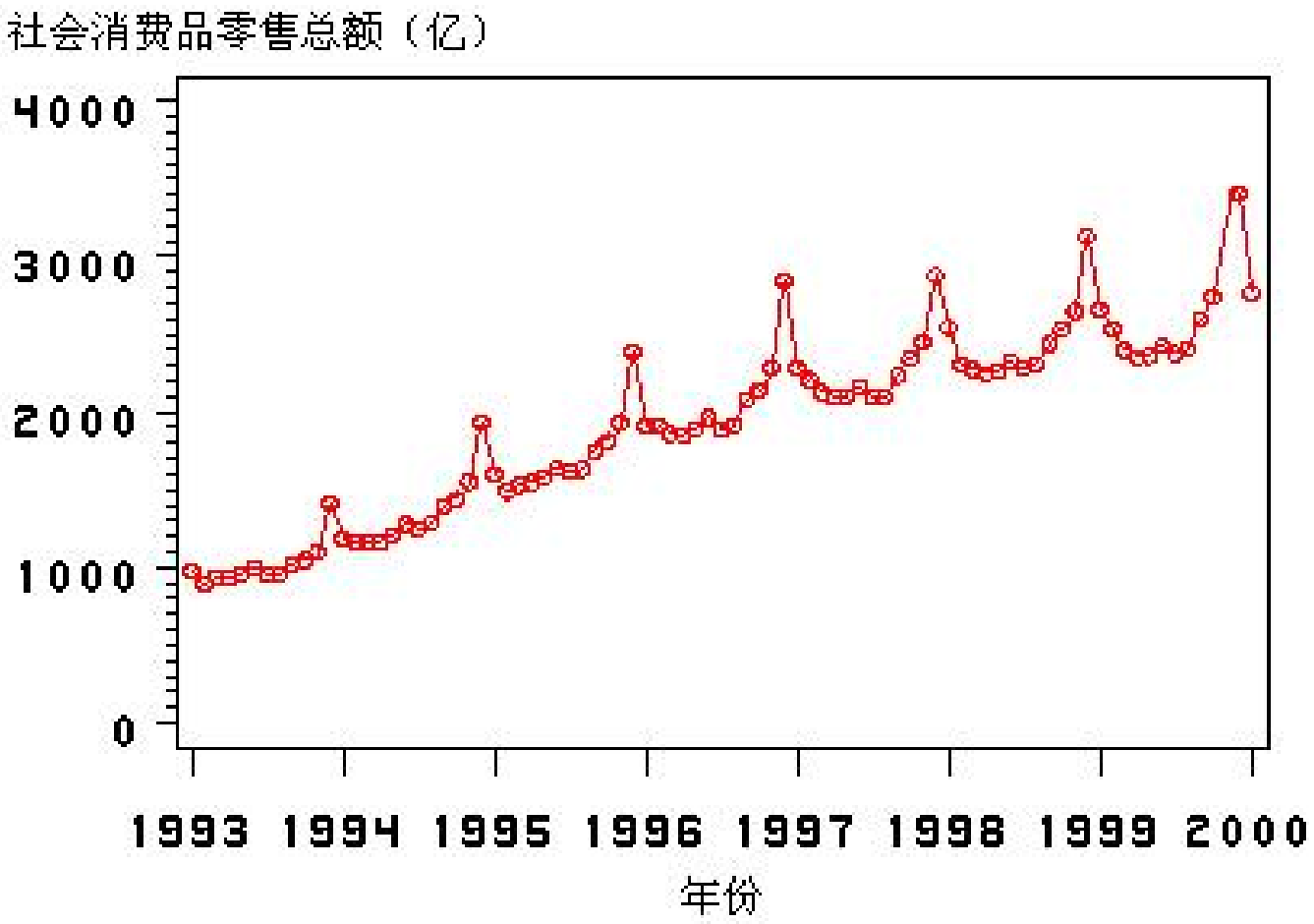
$$a) \quad x_t = S_t \cdot T_t + I_t$$

$$b) \quad x_t = S_t \cdot (T_t + I_t)$$

例4.6

- 对1993年——2000年中国社会消费品零售总额序列（数据见附录1.11）进行确定性时序分析。

(1)绘制时序图


社会消费品零售总额（亿）
4000
3000
2000
1000
0
1993
1994
1995
1996
1997
1998
1999
2000
年份

(2)选择拟合模型

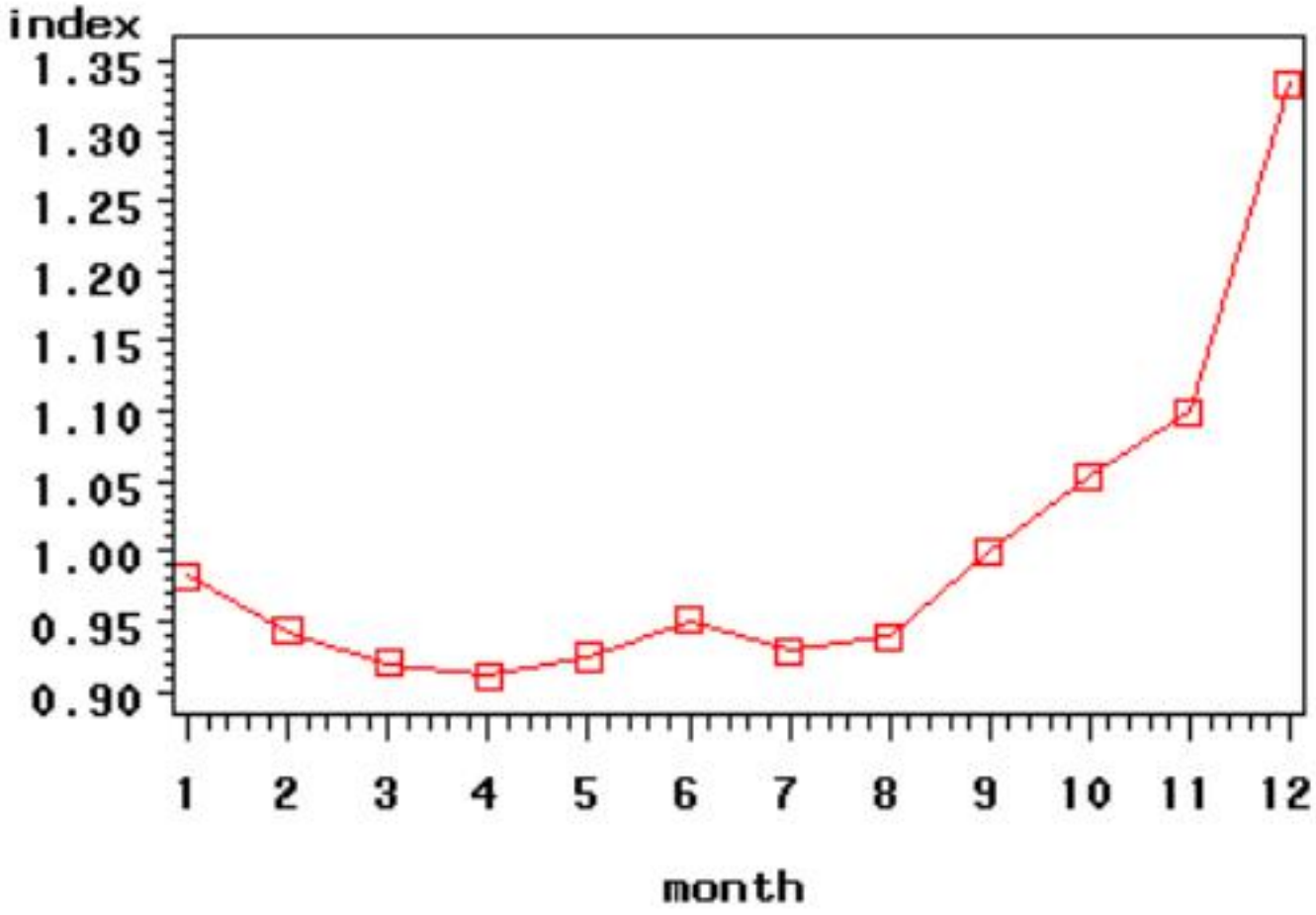
- 长期递增趋势和以年为固定周期的季节波动同时作用于该序列，因而尝试使用混合模型（b）拟合该序列的发展

$$x_t = S_t \cdot (T_t + I_t)$$

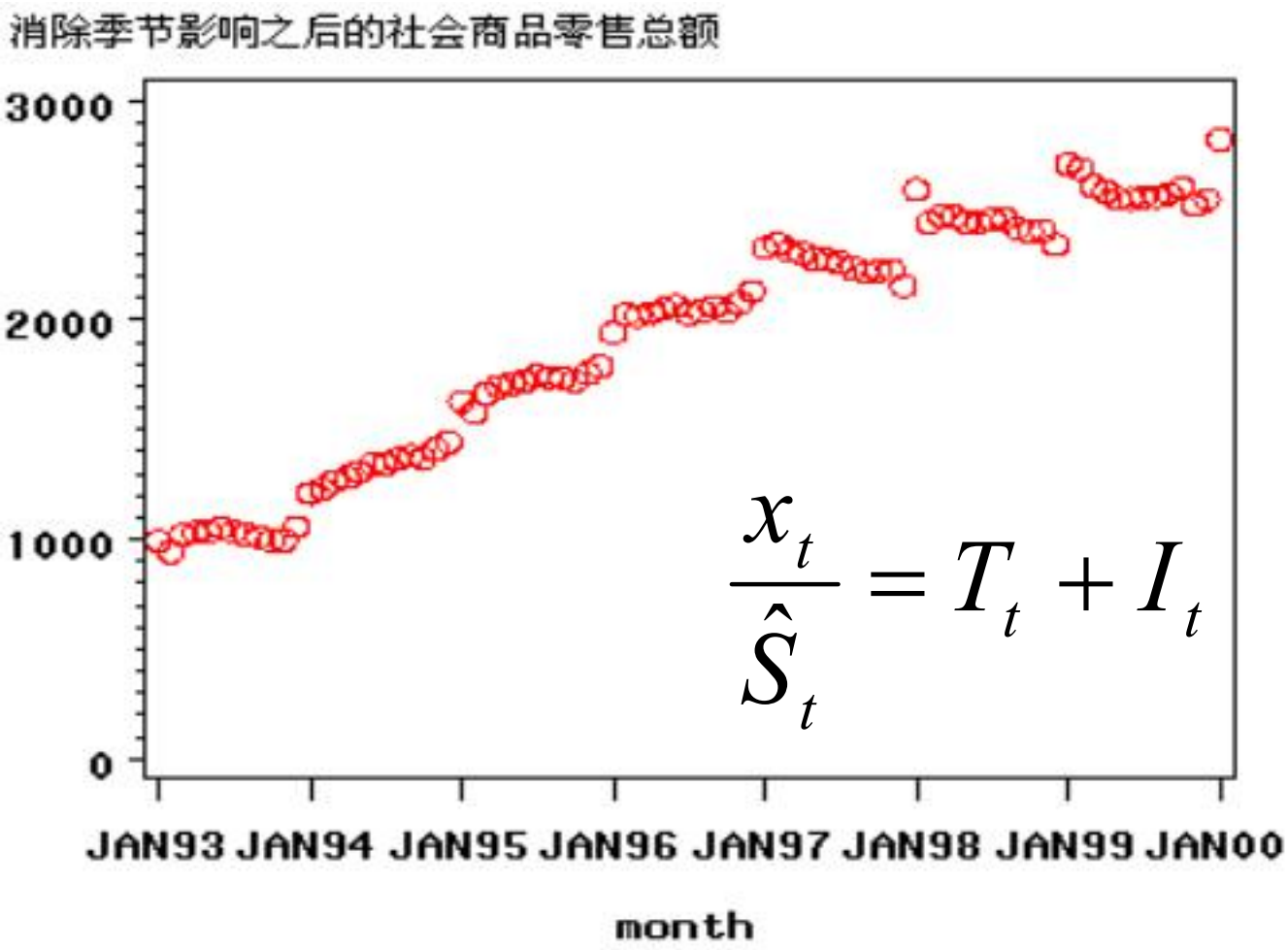
(3)计算季节指数

月份	季节指数	月份	季节指数
1	0.982	7	0.929
2	0.943	8	0.940
3	0.920	9	1.001
4	0.911	10	1.054
5	0.925	11	1.100
6	0.951	12	1.335

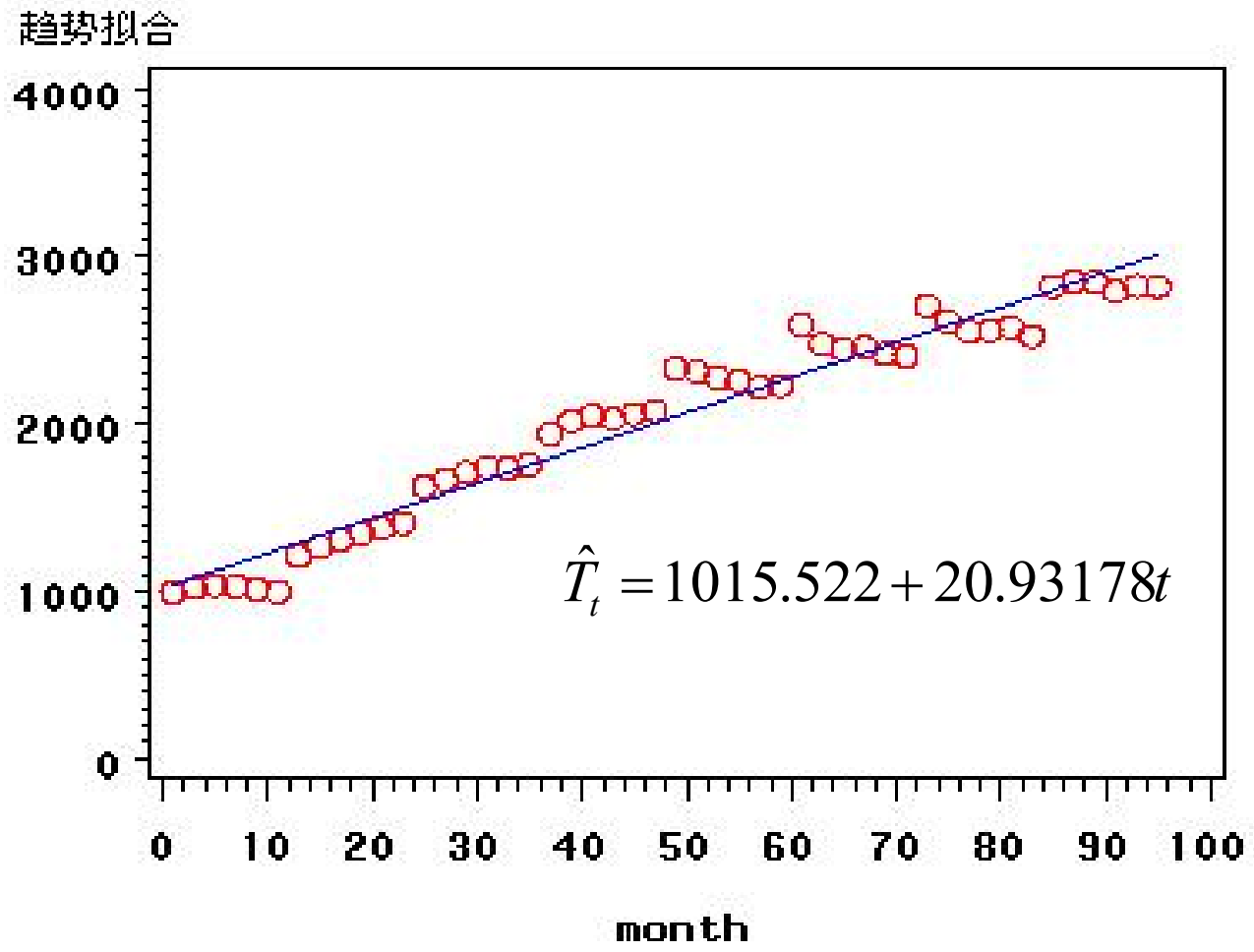
季节指数图


index
1.35
1.30
1.25
1.20
1.15
1.10
1.05
1.00
0.95
0.90
1
2
3
4
5
6
7
8
9
10
11
12
month

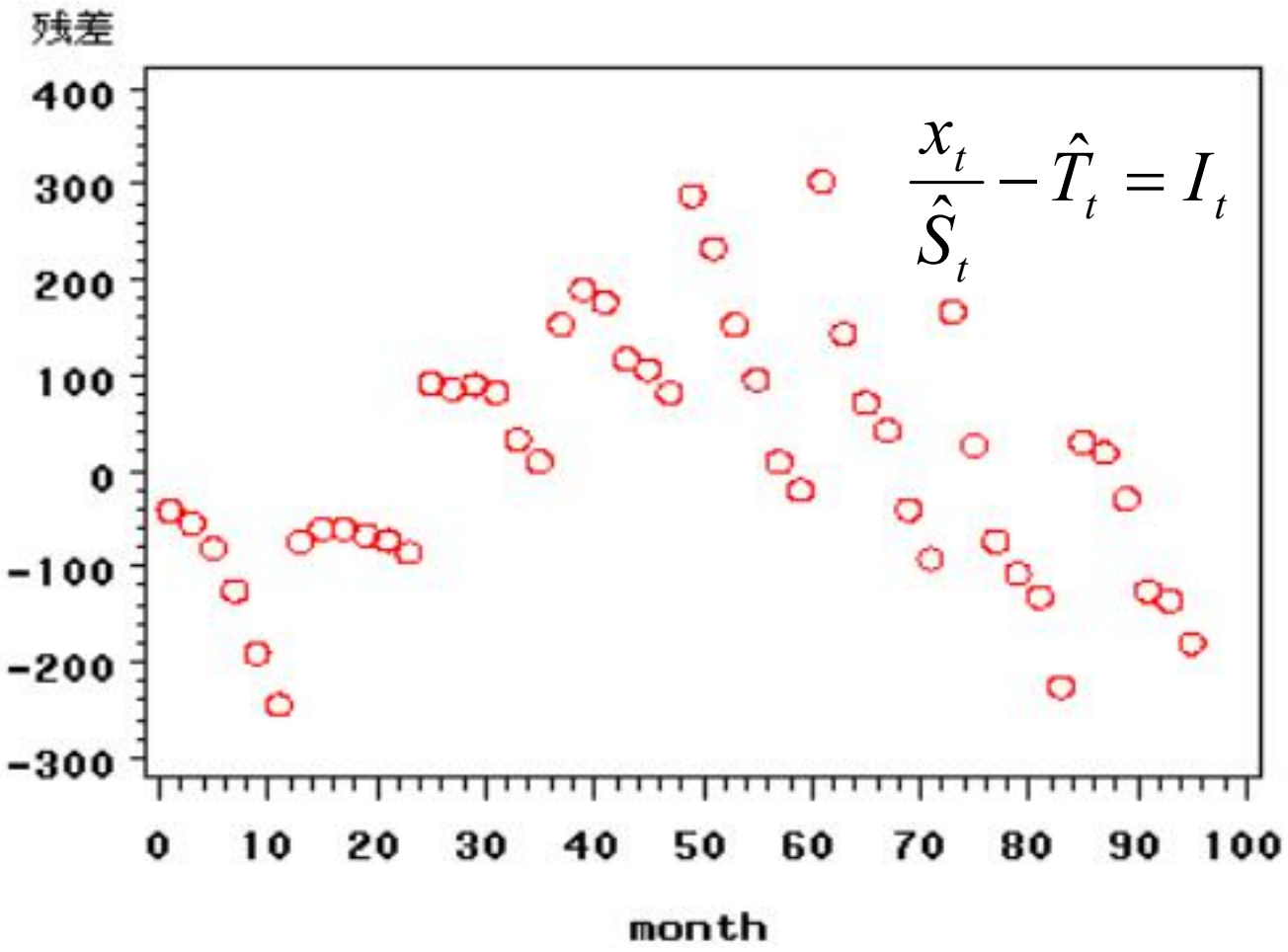
季节调整后的序列图


消除季节影响之后的社会商品零售总额
3000
2000
1000
0
JAN93 JAN94 JAN95 JAN96 JAN97 JAN98 JAN99 JAN00
month
$\frac{x_t}{\hat{S}_t} = T_t + I_t$

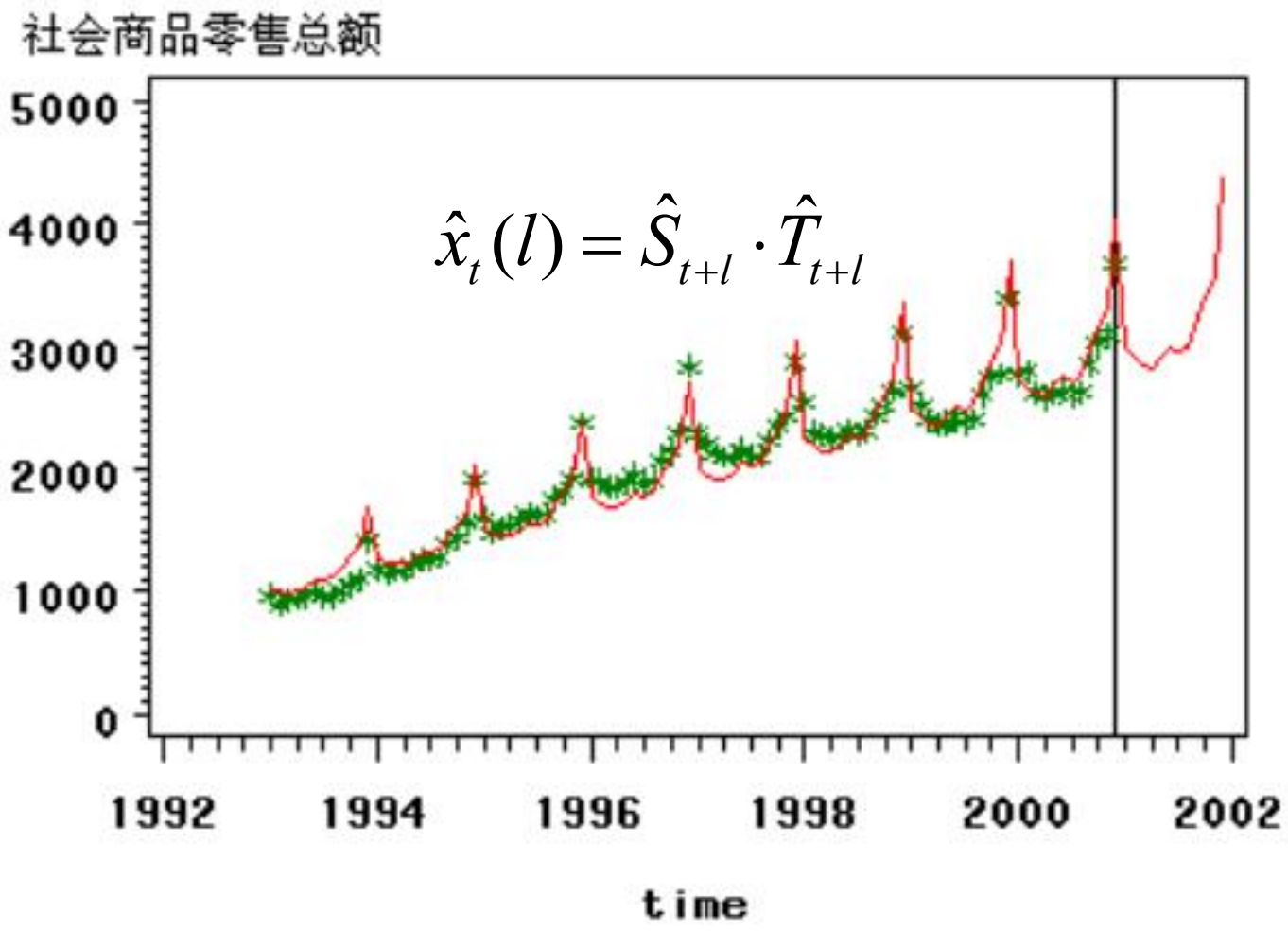
(4)拟合长期趋势


趋势拟合
4000
3000
2000
1000
0
0
10
20
30
40
50
60
70
80
90
100
month
$\hat{T}_t = 1015.522 + 20.93178t$

(5)残差检验


残差
400
300
200
100
0
-100
-200
-300
0
10
20
30
40
50
60
70
80
90
100
month
$\frac{x_t}{\hat{S}_t} - \hat{T}_t = I_t$

(6)短期预测



$$\hat{x}_t(l) = \hat{S}_{t+l} \cdot \hat{T}_{t+l}$$

X-11过程

- 简介
  - X-11过程是美国国情调查局编制的时间序列季节调整过程。它的基本原理就是时间序列的确定性因素分解方法
- 因素分解
  - 长期趋势起伏
  - 季节波动
  - 不规则波动
  - 交易日影响
- 模型
  - 加法模型
  - 乘法模型

方法特色

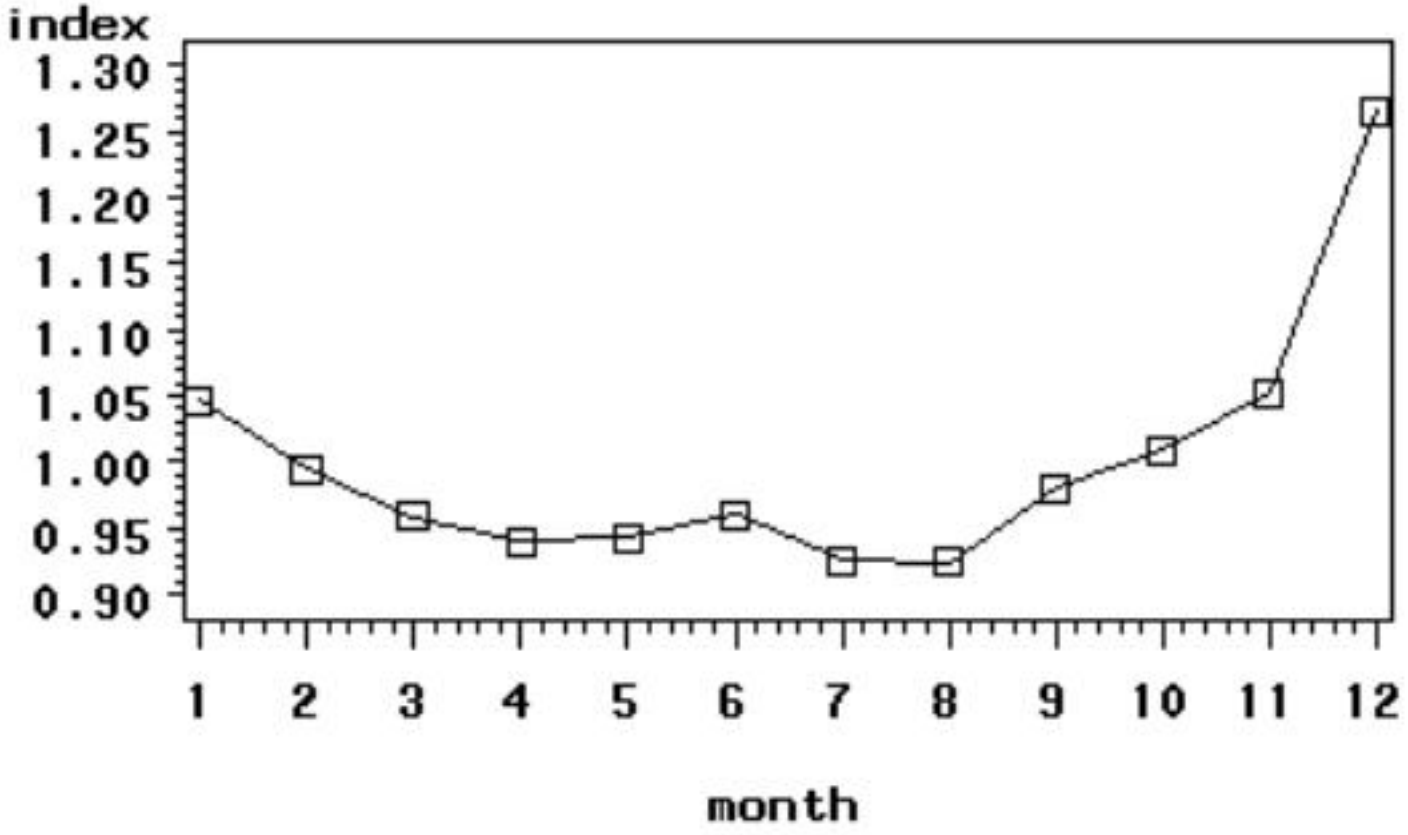
- 普遍采用移动平均的方法
  - 用多次短期中心移动平均消除随机波动
  - 用周期移动平均消除趋势
  - 用交易周期移动平均消除交易日影响

例4.6续

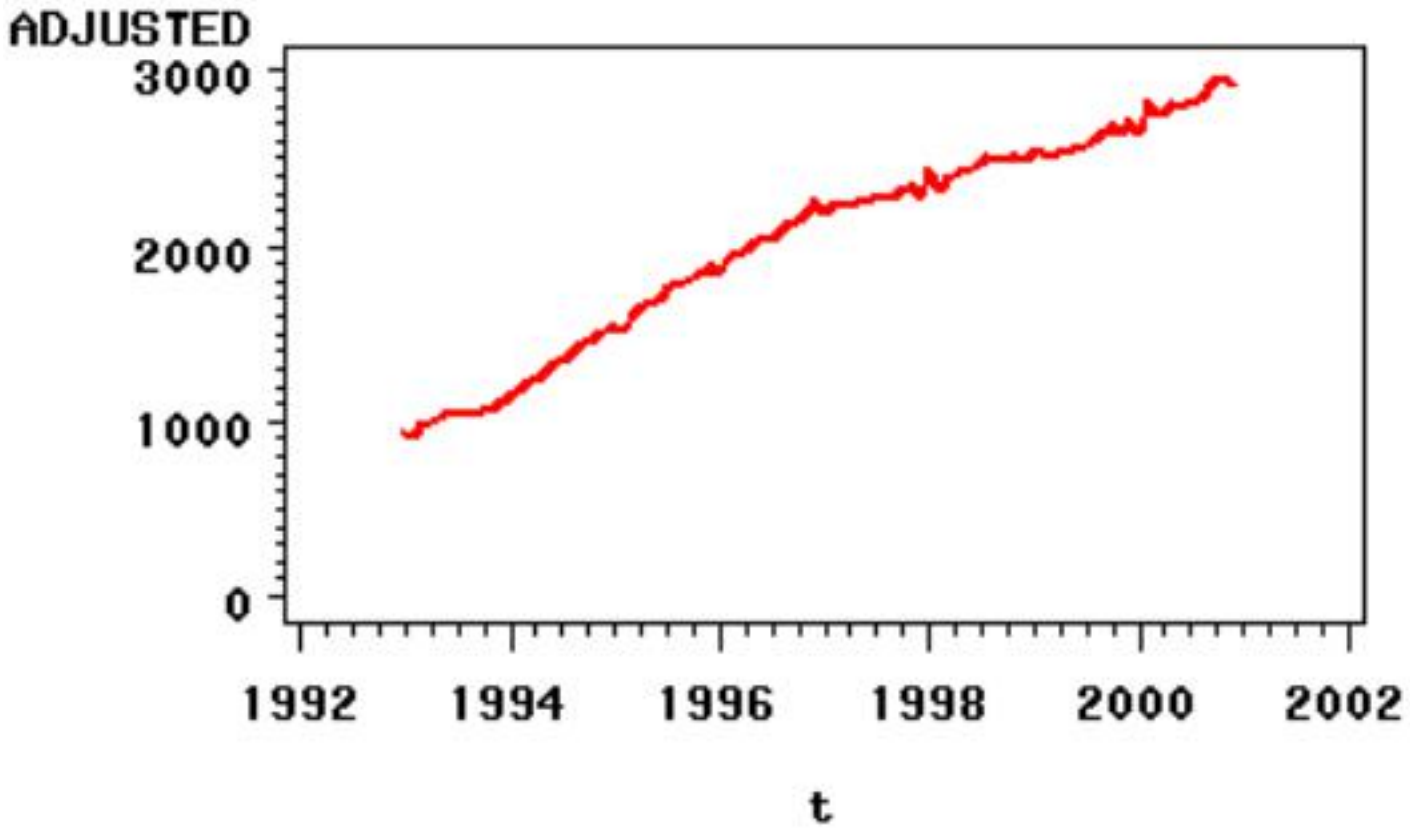
- 对1993年——2000年中国社会消费品零售总额序列使用X-11过程进行季节调整
- 选择模型（无交易日影响）

$$x_t = T_t S_t I_t$$

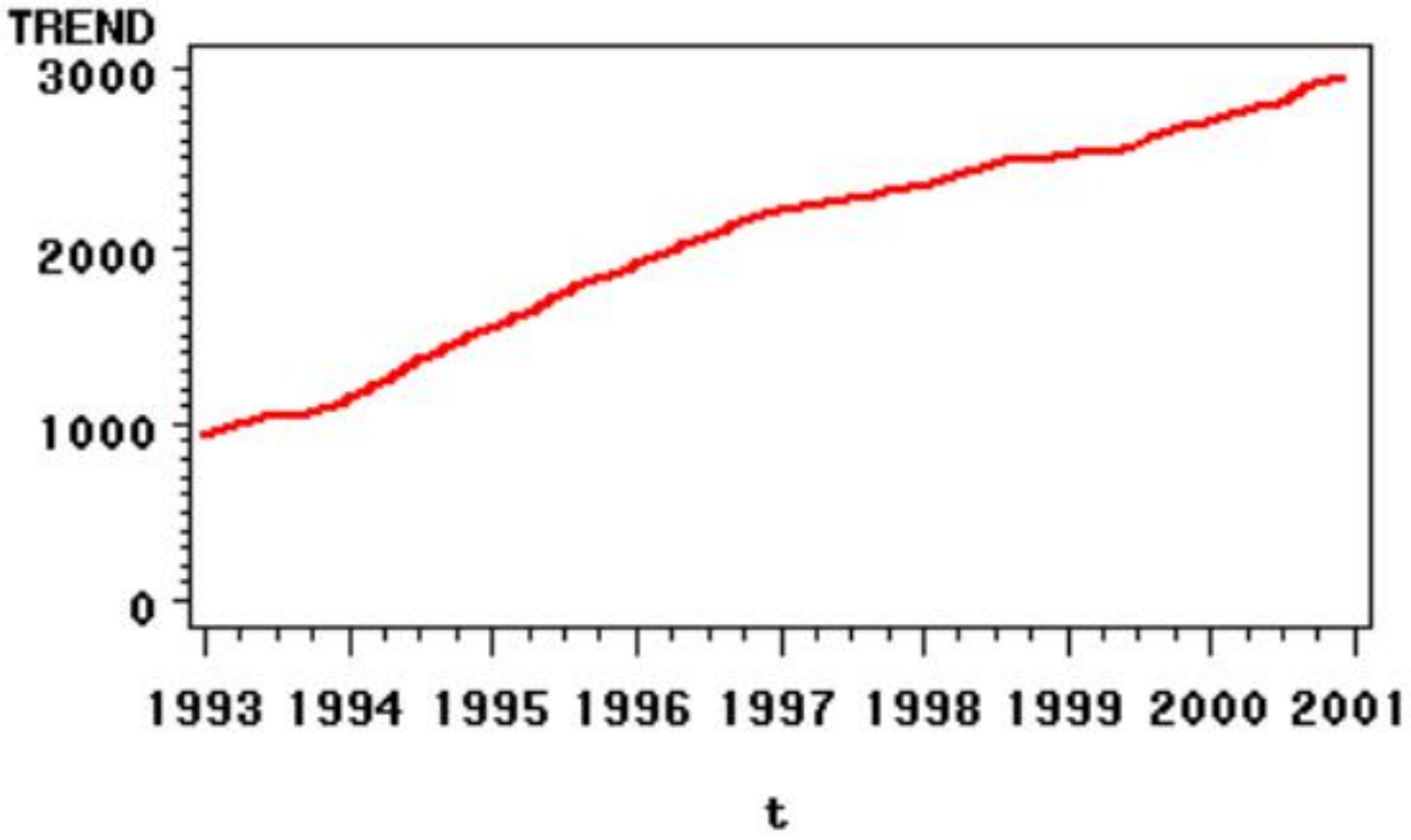
X11过程获得的季节指数图



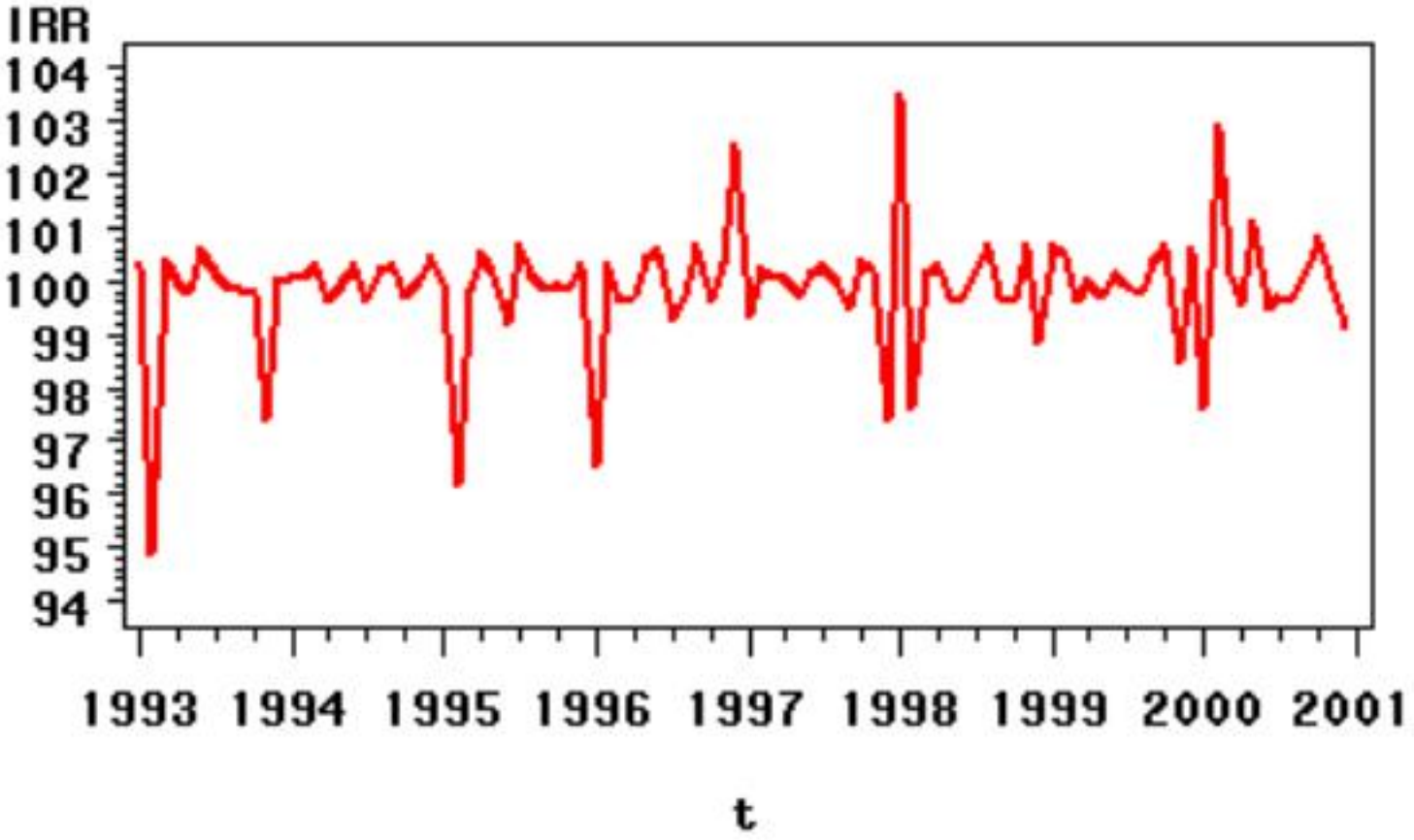
季节调整后的序列图



趋势拟合图


TREND
3000
2000
1000
0
1993 1994 1995 1996 1997 1998 1999 2000 2001
t

随机波动序列图


IRR
104
103
102
101
100
99
98
97
96
95
94
1993
1994
1995
1996
1997
1998
1999
2000
2001
t

三　差分运算

- 差分运算的实质
- 差分方式的选择
- 过差分

差分运算的实质

- 差分方法是一种非常简便、有效的确定性信息提取方法
- Cramer分解定理在理论上保证了适当阶数的差分一定可以充分提取确定性信息
- 差分运算的实质是使用自回归的方式提取确定性信息

$$\nabla^d x_t = (1-B)^d x_t = \sum_{i=0}^{d} (-1)^i C_d^i x_{t-i}$$

差分方式的选择

- 序列蕴含着显著的线性趋势，一阶差分就可以实现趋势平稳
- 序列蕴含着曲线趋势，通常低阶（二阶或三阶）差分就可以提取出曲线趋势的影响
- 对于蕴含着固定周期的序列进行步长为周期长度的差分运算，通常可以较好地提取周期信息

例4.7

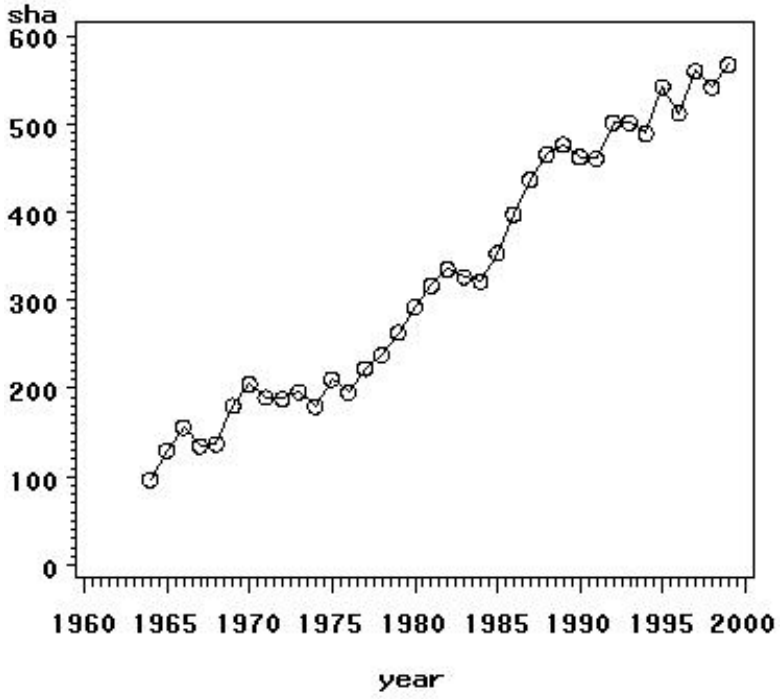
【例】1964年——1999年中国纱年产量序列蕴含着一个近似线性的递增趋势。对该序列进行一阶差分运算

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

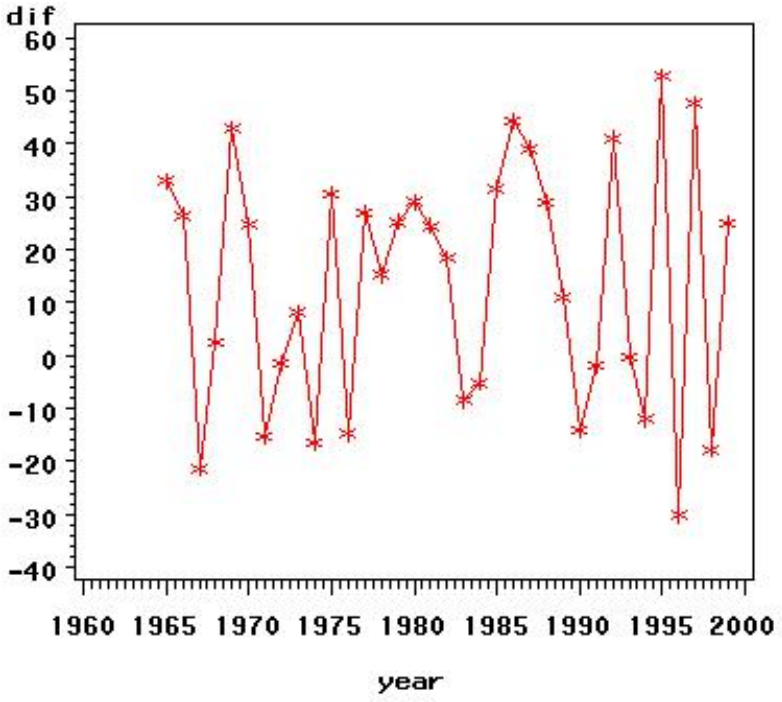
考察差分运算对该序列线性趋势信息的提取作用

差分前后时序图

原序列时序图

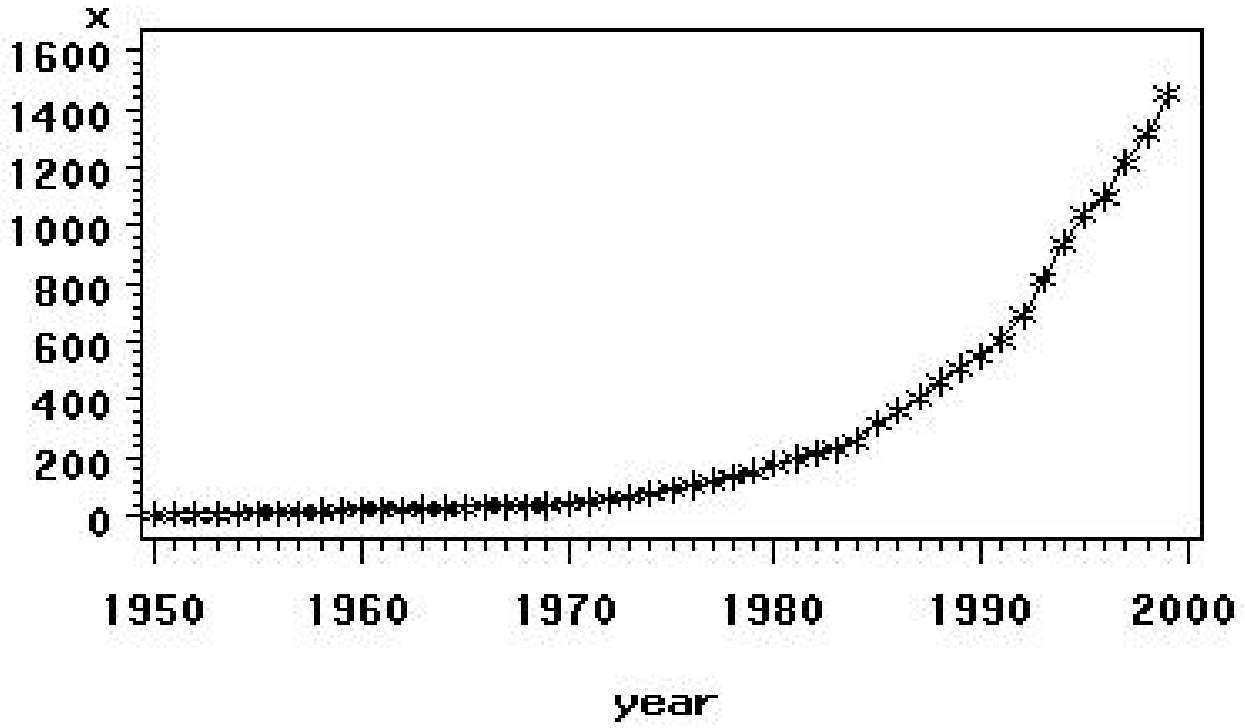


差分后序列时序图



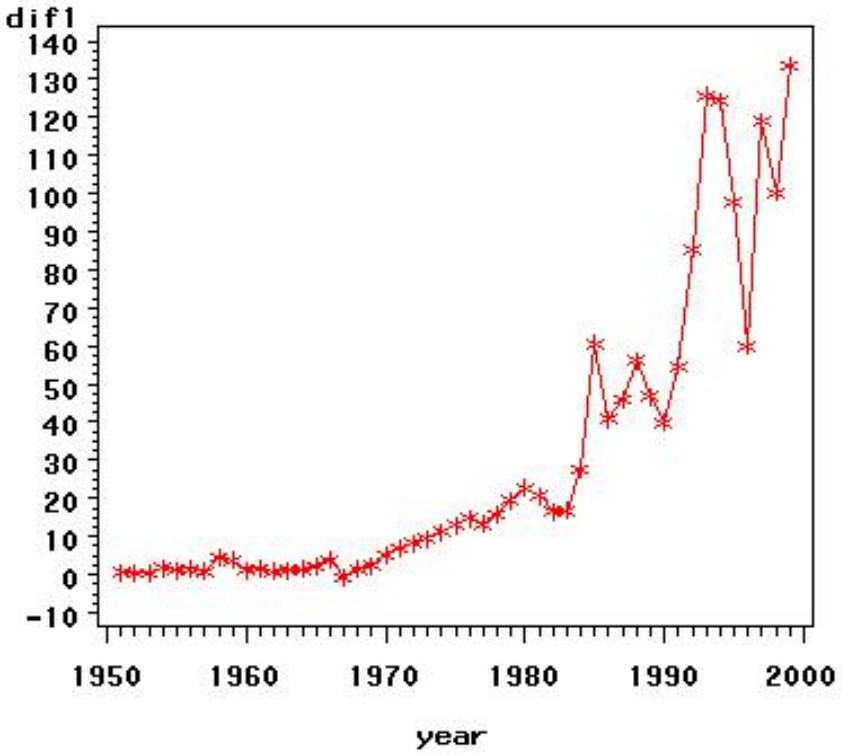
例4.8

- 尝试提取1950年——1999年北京市民用车辆拥有量序列的确定性信息

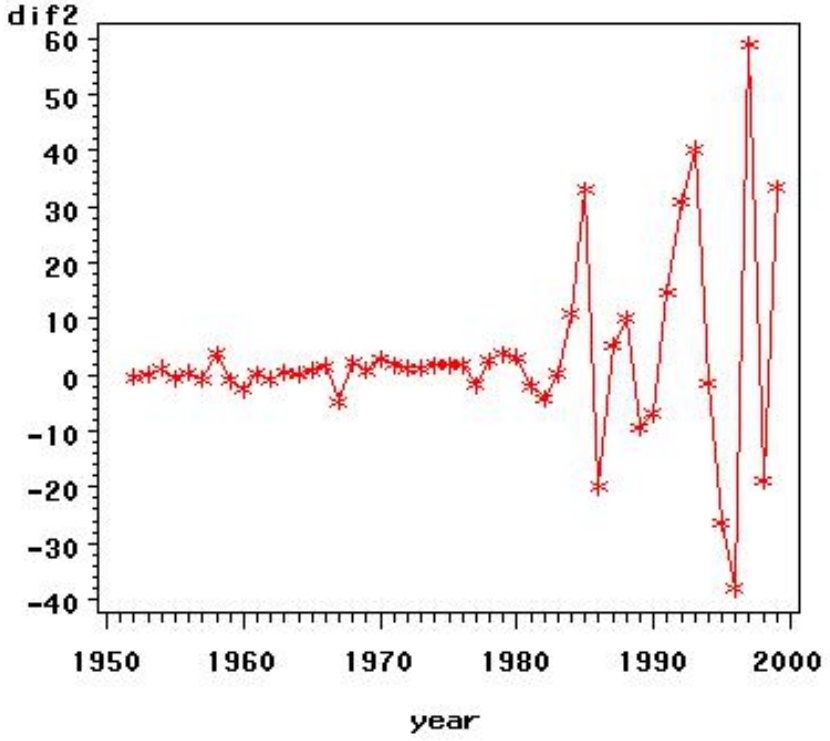


差分后序列时序图

一阶差分

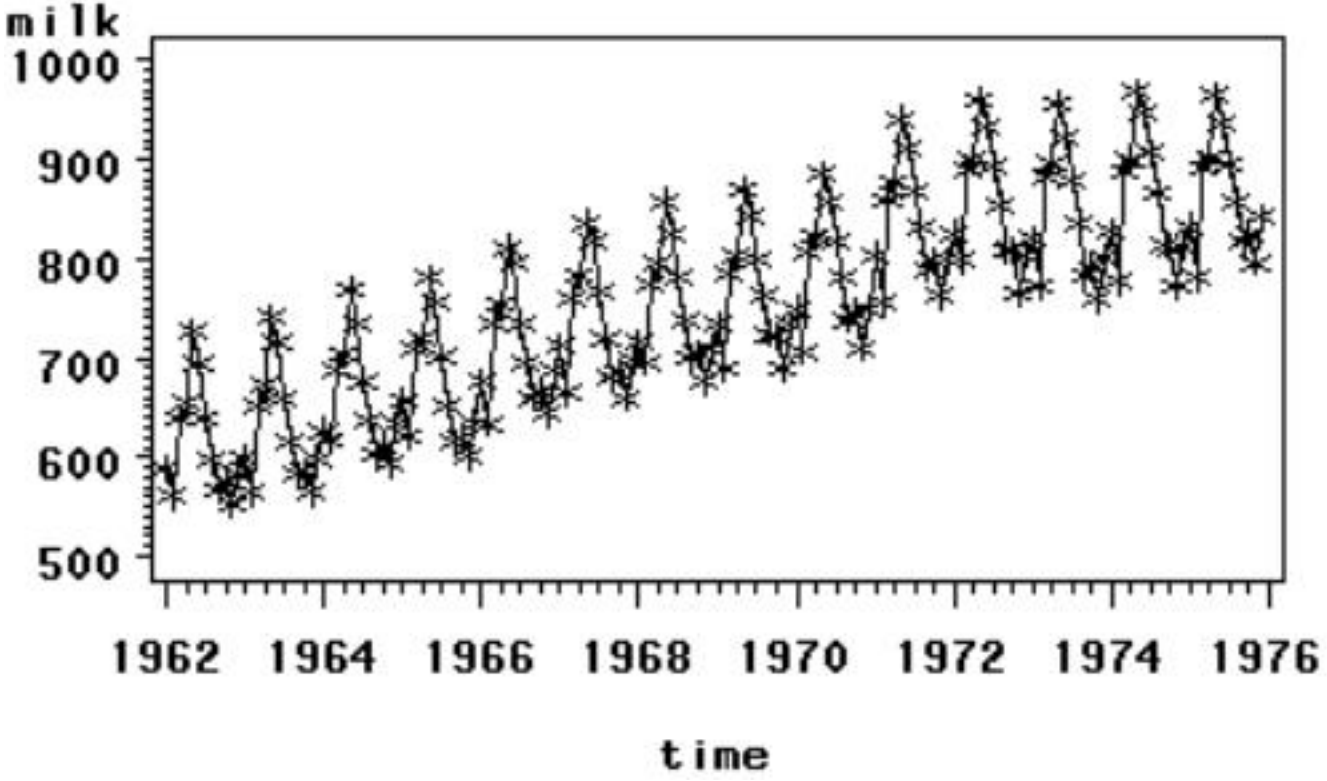


二阶差分



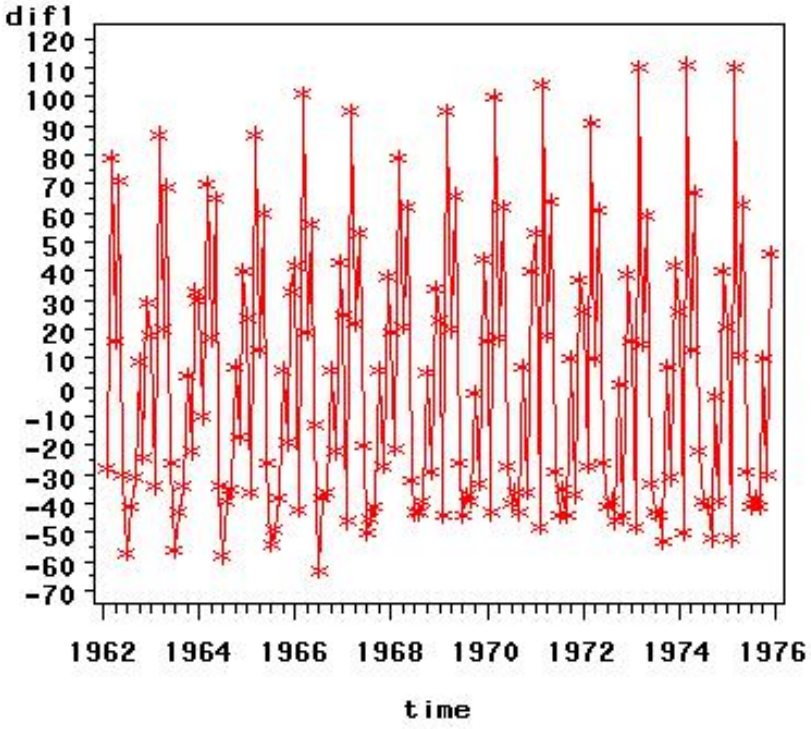
例4.9

- 差分运算提取1962年1月——1975年12月平均每头奶牛的月产奶量序列中的确定性信息

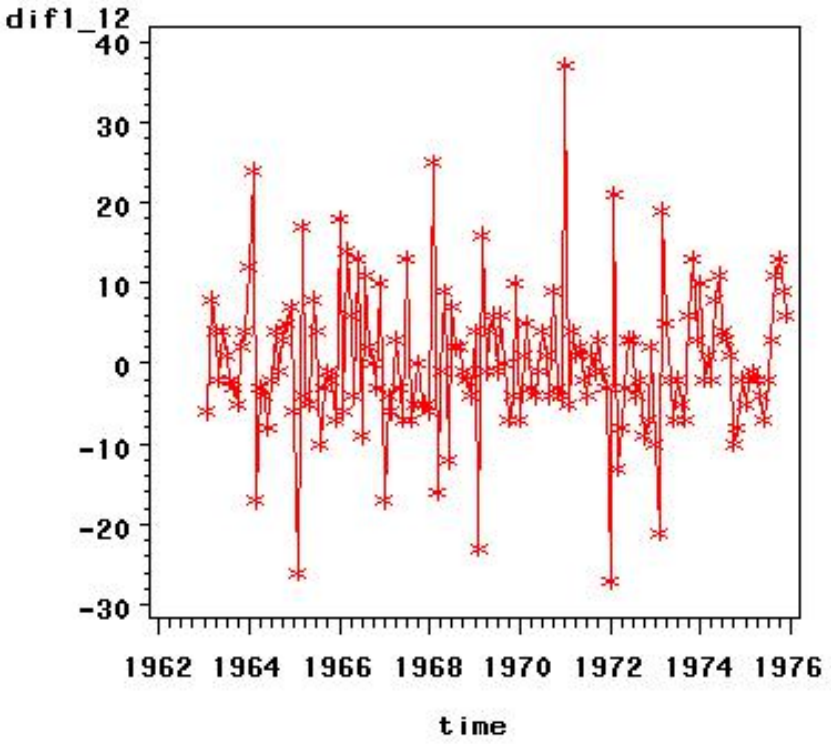


差分后序列时序图

一阶差分



1阶－12步差分



过差分

- 足够多次的差分运算可以充分地提取原序列中的非平稳确定性信息
- 但过度的差分会造成有用信息的浪费

四　ARIMA模型结构

- 使用场合
  - 差分平稳序列拟合
- 模型结构

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, \quad Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ Ex_s \varepsilon_t = 0, \forall s < t \end{cases}$$

ARIMA 模型族

- d=0

  ARIMA(p,d,q)=ARMA(p,q)

- P=0

  ARIMA(P,d,q)=IMA(d,q)

- q=0

  ARIMA(P,d,q)=ARI(p,d)

- d=1,P=q=0

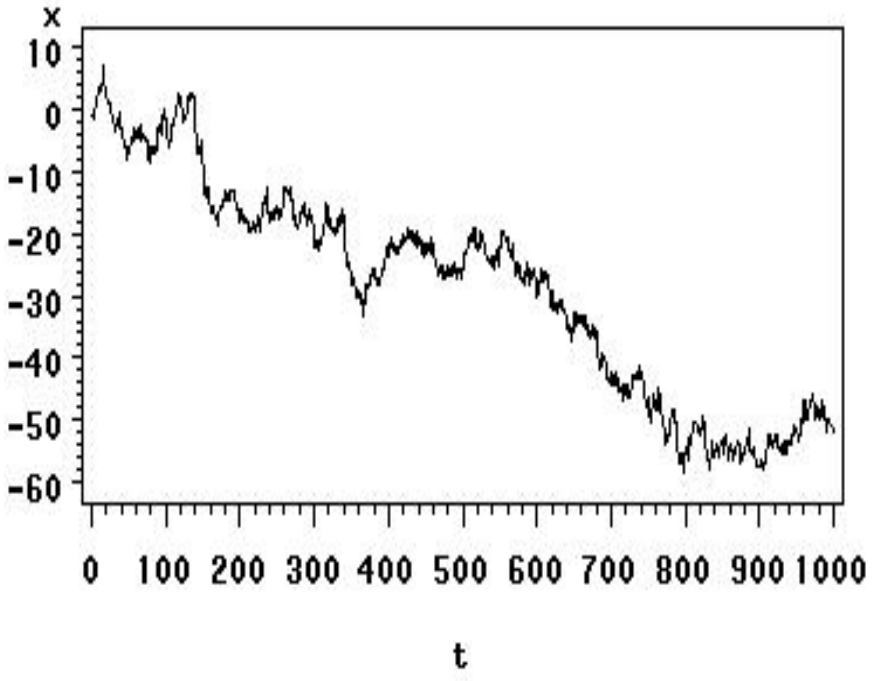
  ARIMA(P,d,q)=random walk model

ARIMA模型的平稳性

- ARIMA(p,d,q)模型共有p+d个特征根，其中p个在单位圆内，d个在单位圆上。所以当$d \neq 0$时ARIMA(p,d,q)模型非平稳。

例

ARIMA(0,1,0)时序图



ARIMA模型的方差齐性

- $d \neq 0$ 时，原序列方差非齐性

$ARIMA(0,1,0)$模型

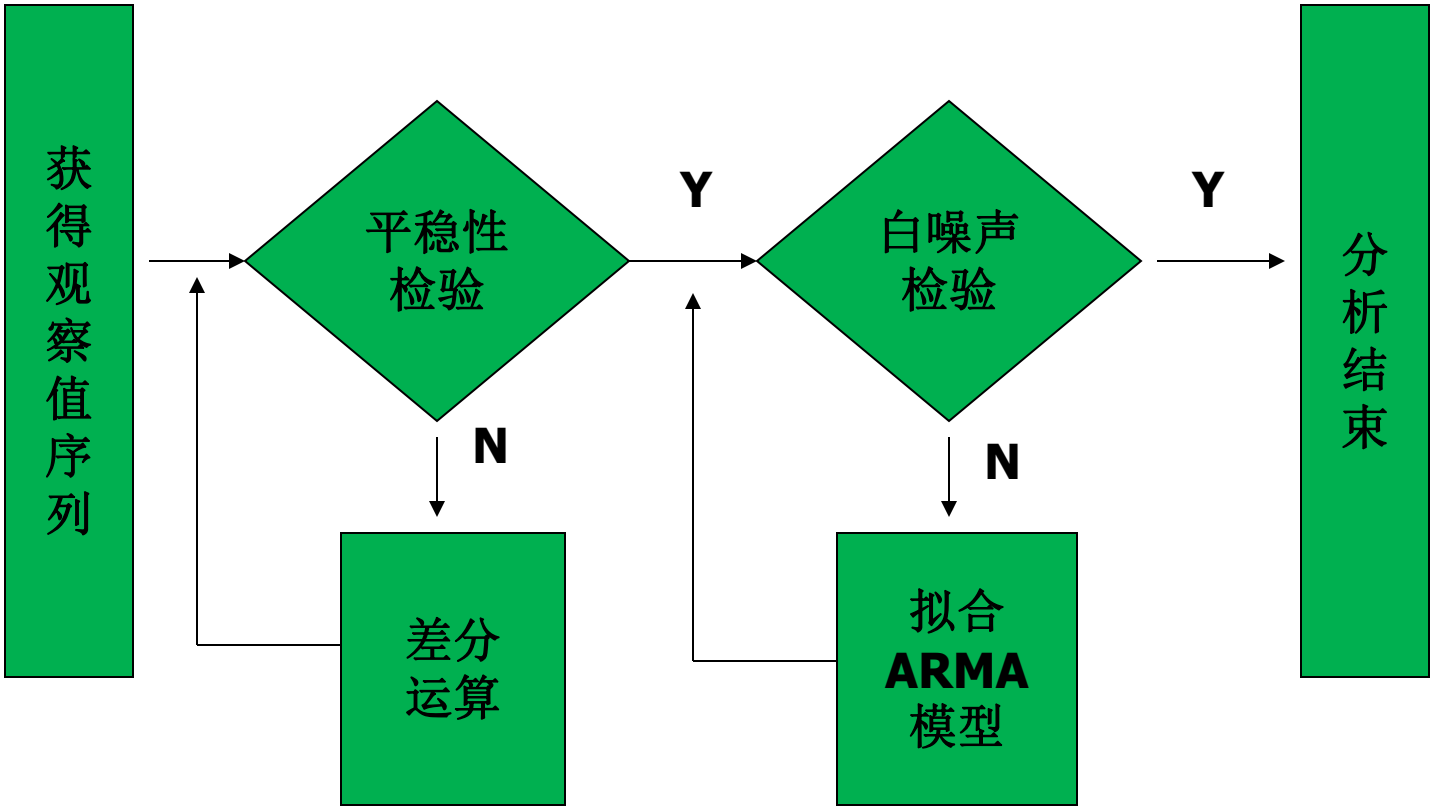
$$Var(x_t) = Var(x_0 + \varepsilon_t + \varepsilon_{t-1} + \cdots \varepsilon_1) = t\sigma_\varepsilon^2$$

- d阶差分后，差分后序列方差齐性

$ARIMA(0,1,0)$模型

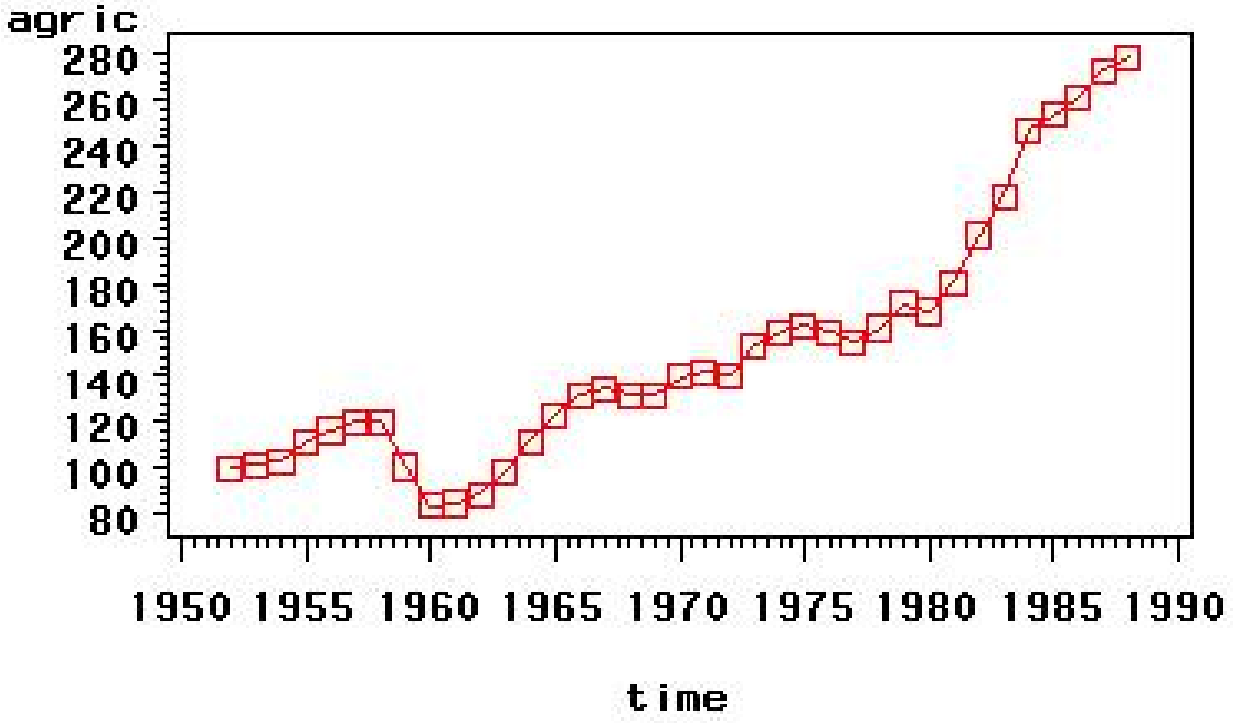
$$Var(\nabla x_t) = Var(\varepsilon_t) = \sigma_\varepsilon^2$$

ARIMA模型建模步骤

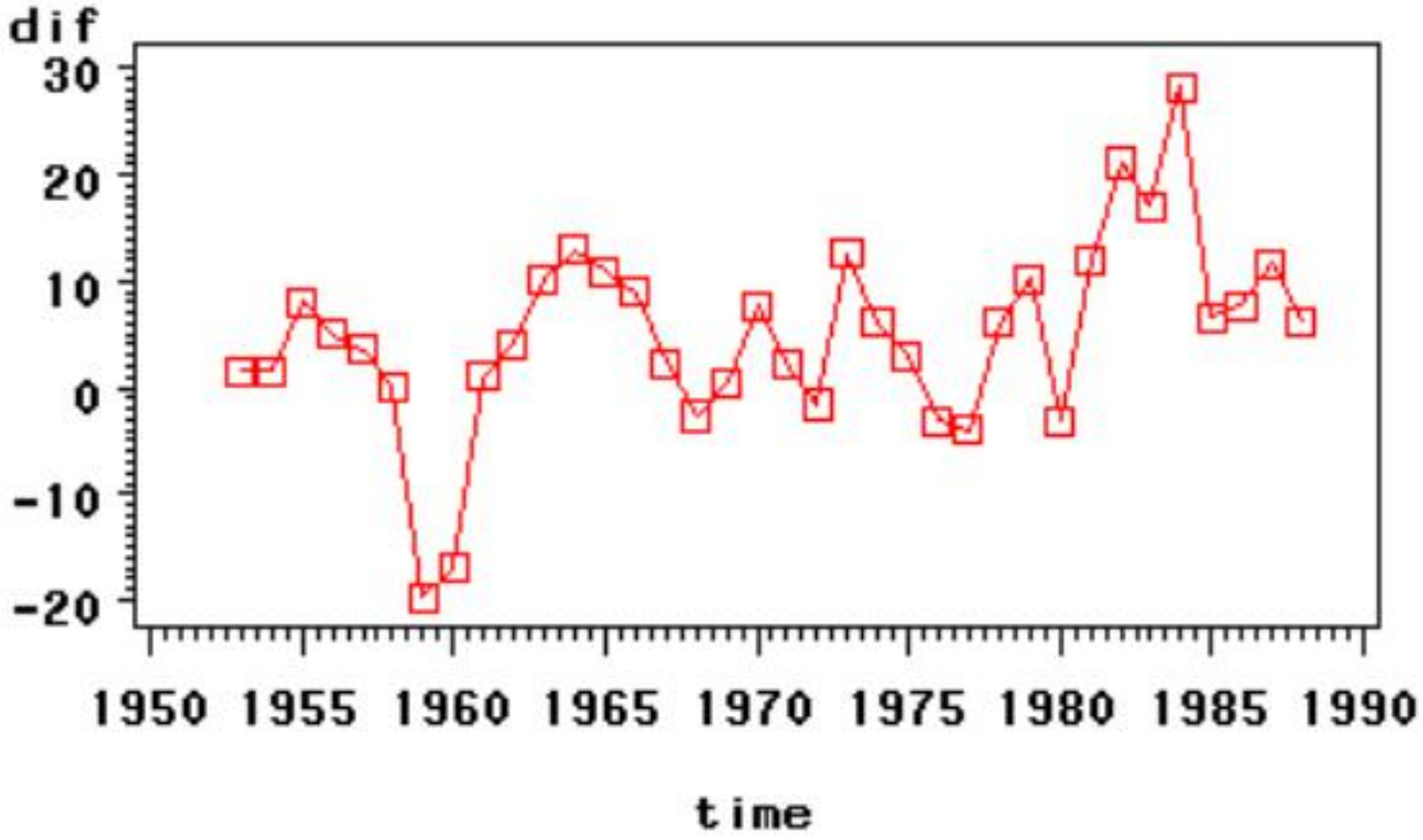


例4.10

- 对1952年——1988年中国农业实际国民收入指数序列建模



一阶差分序列时序图


dif
30
20
10
0
-10
-20
1950
1955
1960
1965
1970
1975
1980
1985
1990
time

一阶差分序列自相关图

Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	78.239167	1.00000	|********************|
1	42.075733	0.53778	|********** . |
2	16.246605	0.20765	|**** . |
3	7.058588	0.09022	|** . |
4	-11.132207	-.14228	|*** . |
5	-7.917076	-.10119	|** . |
6	-9.245185	-.11817	|** . |
7	-11.564313	-.14781	|*** . |
8	7.108735	0.09086	|** . |
9	12.965116	0.16571	|*** . |
10	-0.909105	-.01162	| . |
11	-2.455085	-.03138	| . |
12	-3.501852	-.04476	| . |
13	-6.583063	-.08414	|** . |
14	-7.883765	-.10076	|** . |
15	-4.783310	-.06114	| . |
16	2.087515	0.02668	|* . |
17	12.894776	0.16481	|*** . |
18	15.631250	0.19979	|**** . |

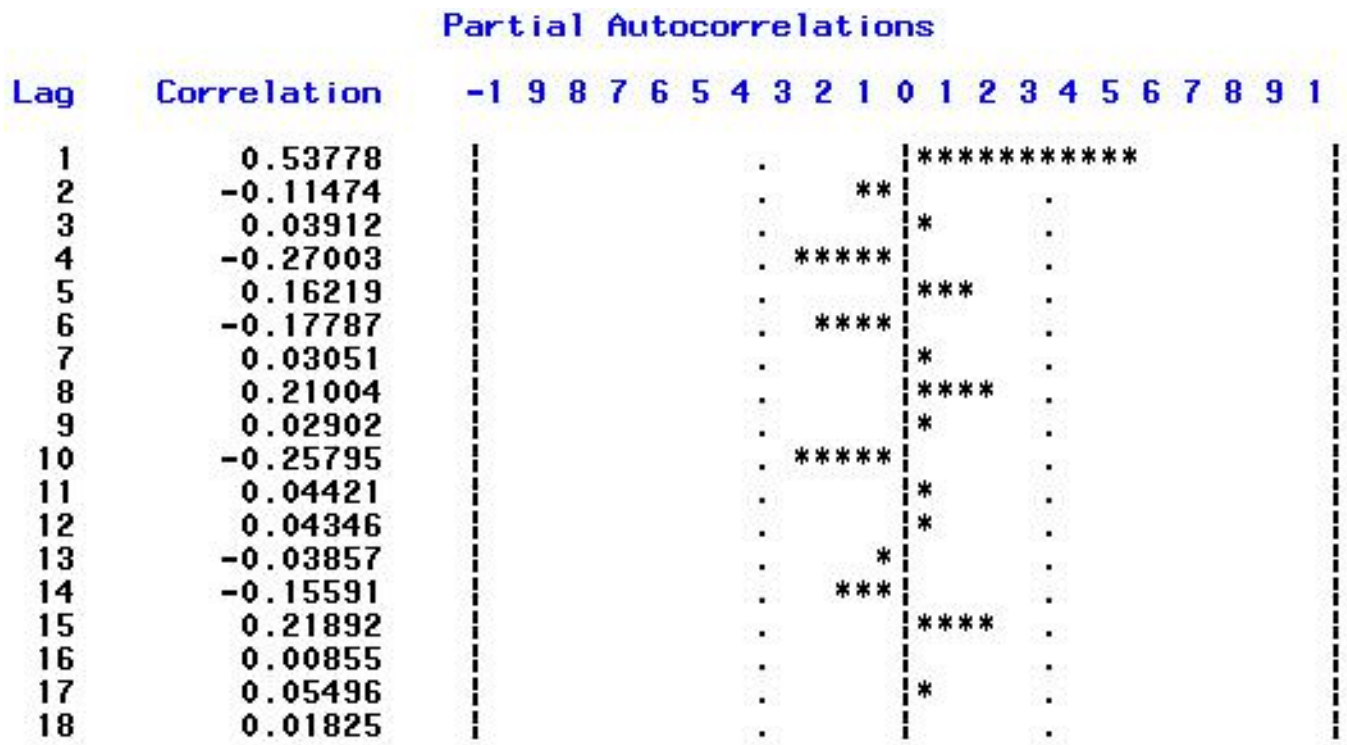
"." marks two standard errors

一阶差分后序列白噪声检验

延迟阶数	χ^2 统计量	P值
6	15.33	0.0178
12	18.33	0.1060
18	24.66	0.1344

拟合ARMA模型

- 偏自相关图



Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.53778	|. |********** |
2	-0.11474	|. ** | . |
3	0.03912	|. | * | . |
4	-0.27003	|. ***** | . |
5	0.16219	|. | *** | . |
6	-0.17787	|. **** | . |
7	0.03051	|. | * | . |
8	0.21004	|. | **** | . |
9	0.02902	|. | * | . |
10	-0.25795	|. ***** | . |
11	0.04421	|. | * | . |
12	0.04346	|. | * | . |
13	-0.03857	|. * | . |
14	-0.15591	|. *** | . |
15	0.21892	|. | **** | . |
16	0.00855	|. | . |
17	0.05496	|. | * | . |
18	0.01825	|. | . |

建模

- 定阶
  - ARIMA(0,1,1)
- 参数估计

$$(1-B)x_t = 4.99661 + (1+0.70766B)\varepsilon_t$$

$$Var(\varepsilon_t) = 56.48763$$

- 模型检验
  - 模型显著
  - 参数显著

ARIMA模型预测

- 原则
  - 最小均方误差预测原理
- Green函数递推公式

$$\begin{cases} \psi_1 = \phi_1 - \theta_1 \\ \psi_2 = \phi_1\psi_1 + \phi_2 - \theta_2 \\ \vdots \\ \psi_j = \phi_1\psi_{j-1} + \cdots + \phi_{p+d}\psi_{j-p-d} - \theta_j \end{cases}$$

预测值

$$x_{t+l} = (\varepsilon_{t+l} + \psi_1 \varepsilon_{t+l-1} + \cdots + \psi_{l-1} \varepsilon_{t+1}) + (\psi_l \varepsilon_t + \psi_{l+1} \varepsilon_{t-1} + \cdots)$$

$e_t(l)$

$\hat{x}_t(l)$

$$E[e_t(l)] = 0$$

$$Var[e_t(l)] = (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_\varepsilon^2$$

例4.11

- 已知ARIMA(1,1,1)模型为

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

且

$$x_{t-1}=4.5 \qquad x_t=5.3 \qquad \varepsilon_t=0.8 \quad \sigma_\varepsilon^2=1$$

- 求 x_{t+3} 的95％的置信区间

预测值

- 等价形式

$$(1-1.8B+0.8B^2)x_t=(1-0.6B)\varepsilon_t$$

- 计算预测值 $x_t=1.8x_{t-1}-0.8x_{t-2}+\varepsilon_t-0.6\varepsilon_{t-1}$

$$\hat{x}_t(1)=1.8x_t-0.8x_{t-1}-0.6\varepsilon_t=5.46$$
$$\hat{x}_t(2)=1.8\hat{x}_t(1)-0.8x_t=5.59$$
$$\hat{x}_t(3)=1.8\hat{x}_t(2)-0.8\hat{x}_t(1)=5.69$$

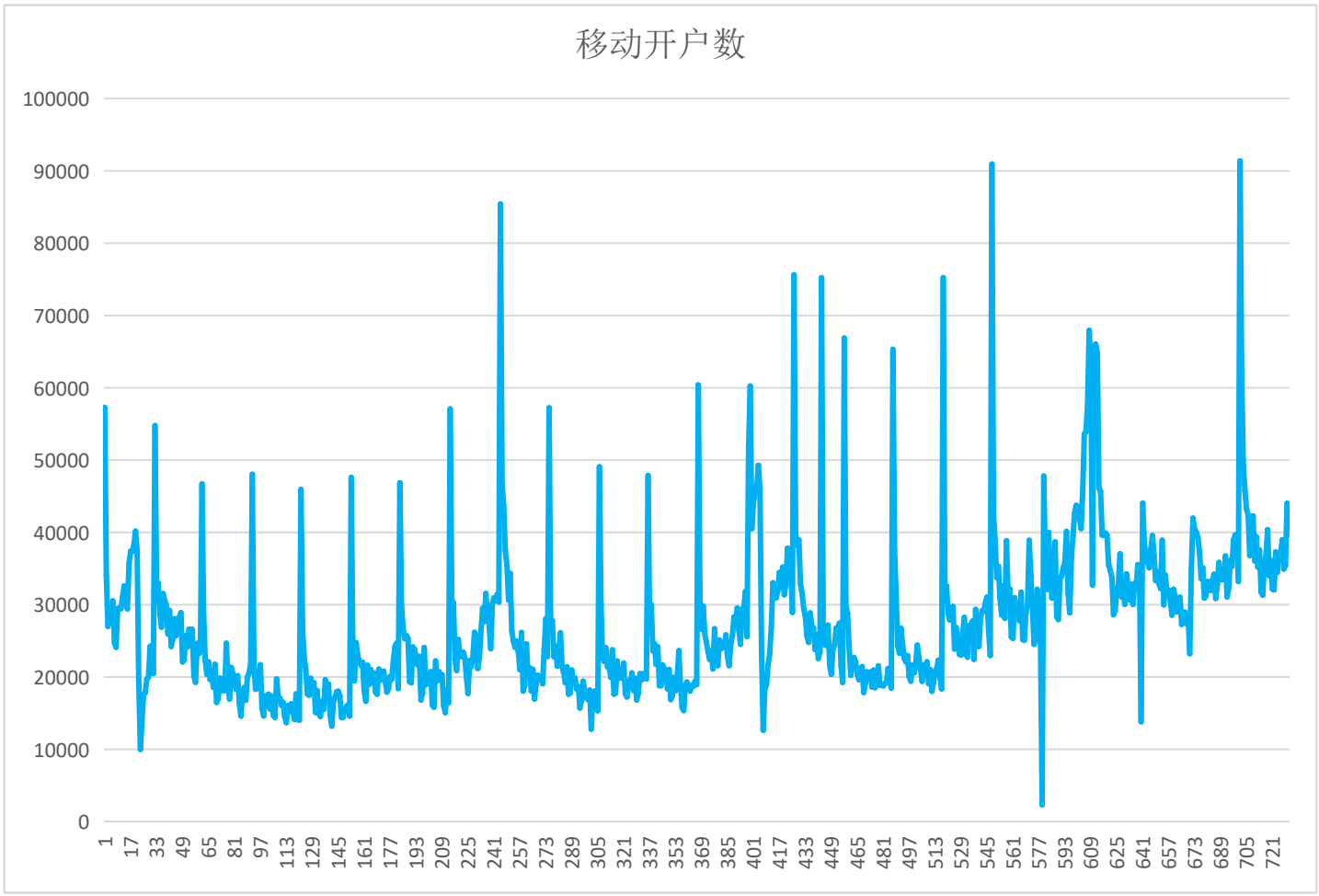
05 案例分析

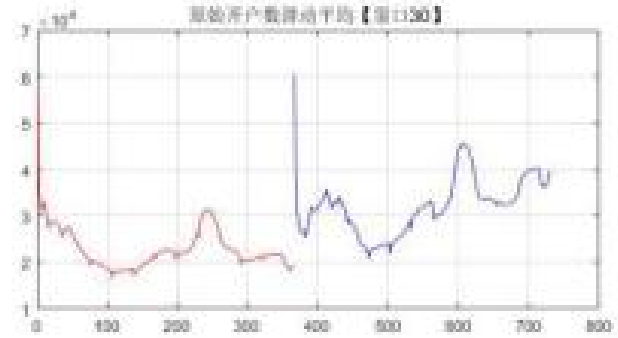
参考博客：
https://blog.csdn.net/qq_40527086/article/details/84033957#_376.

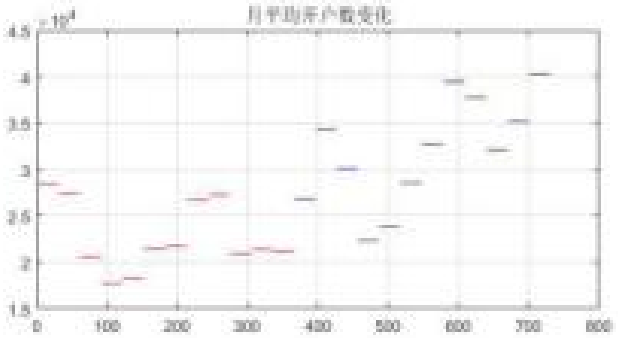
分析步骤

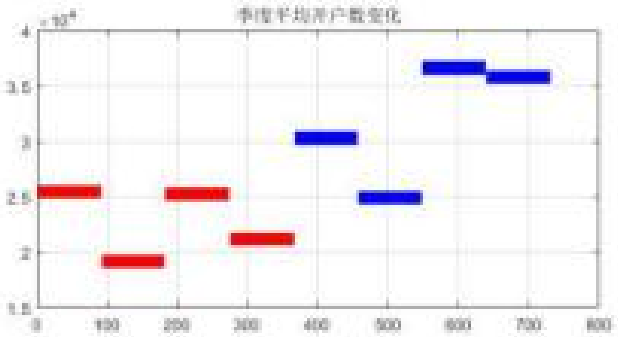
- 观测数据（均值，周期等）
- 数据预处理
- 平稳化 - 去趋势与去周期，剩余随机项
- 用类似于回归/滑动平均的思想来拟合随机项

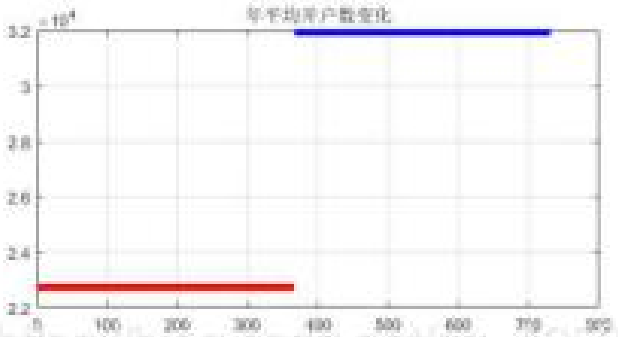
1. 判断去趋势去周期后的数据是否平稳
2. 计算数据的自相关函数和偏相关函数
3. 根据自相关/偏相关函数性质决定选用什么模型来拟合随机项
4. 模型定阶和拟合参数的求解
5. 模型检验


移动开户数
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
1
17
33
49
65
81
97
113
129
145
161
177
193
209
225
241
257
273
289
305
321
337
353
369
385
401
417
433
449
465
481
497
513
529
545
561
577
593
609
625
641
657
673
689
705
721









开户数据趋势变化周期的观察

综上观察，我们可以得出以下结论：

- 一个月应该是一个最小周期，两个季度是一个中周期，一年为一个确定的大周期。
- 开户人数有明显的上升趋势。
- 由观测值我们可以认为存在异常值。

数据预处理

- 缺失值处理

数据中不存在缺失值

- 异常值处理

拟采用三倍标准差即拉以达法则筛选异常值，考虑到时间序列的短期影响性，用其周围值平均代替。

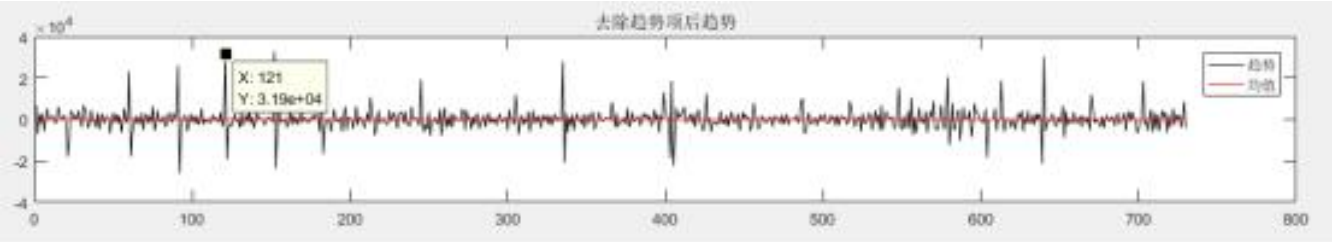
去趋势与去周期

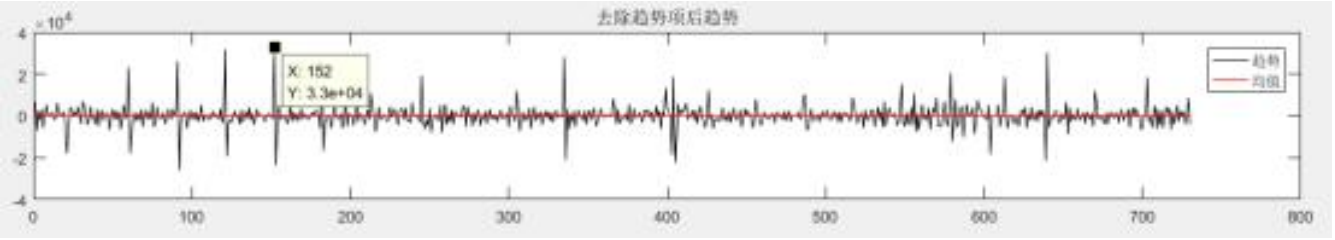
进行了简单的观察和数据预处理后，我们开始正式进行时间序列平稳化的操作。考虑到有两年，我们从单独考虑2012，2013，然后集中考虑2012和2013来进行处理和检验。

- 去趋势

由于数据变化有一定的集中性，且有滑动平均看来近似可以用一次函数拟合。所以采用一次差分的方法去除数据的趋势项。这里采用diff函数实现一次差分。

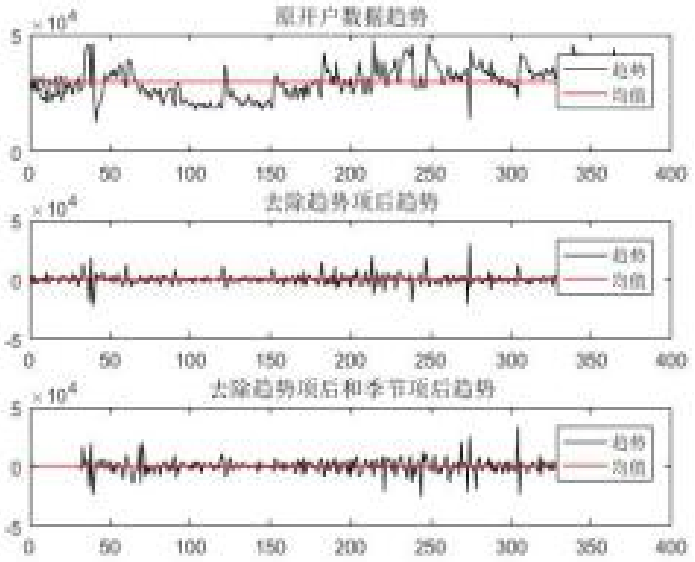
去趋势后，我们进行了如下的观察：

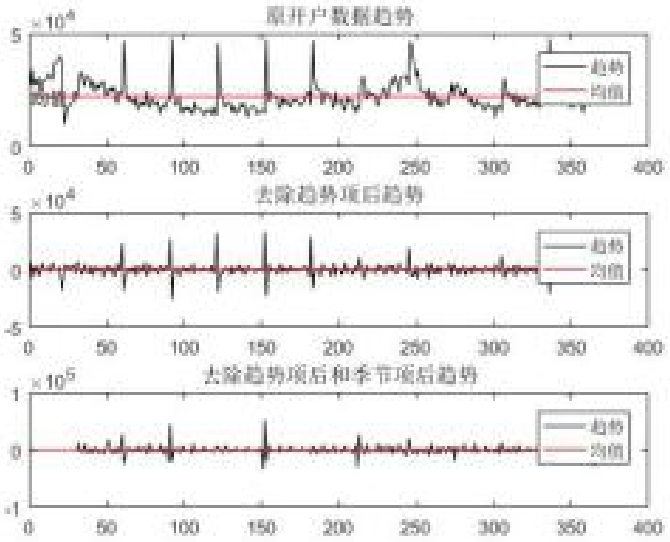


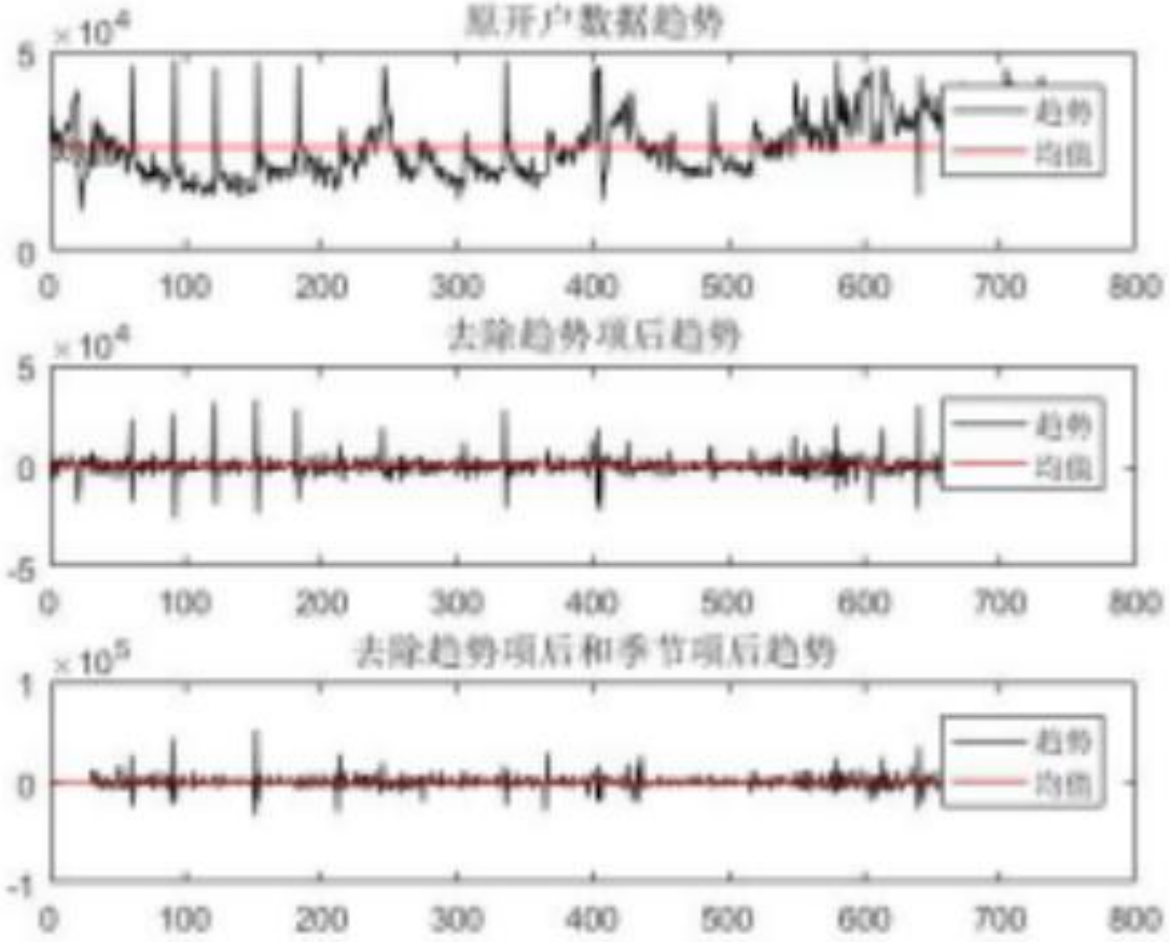


- 去周期

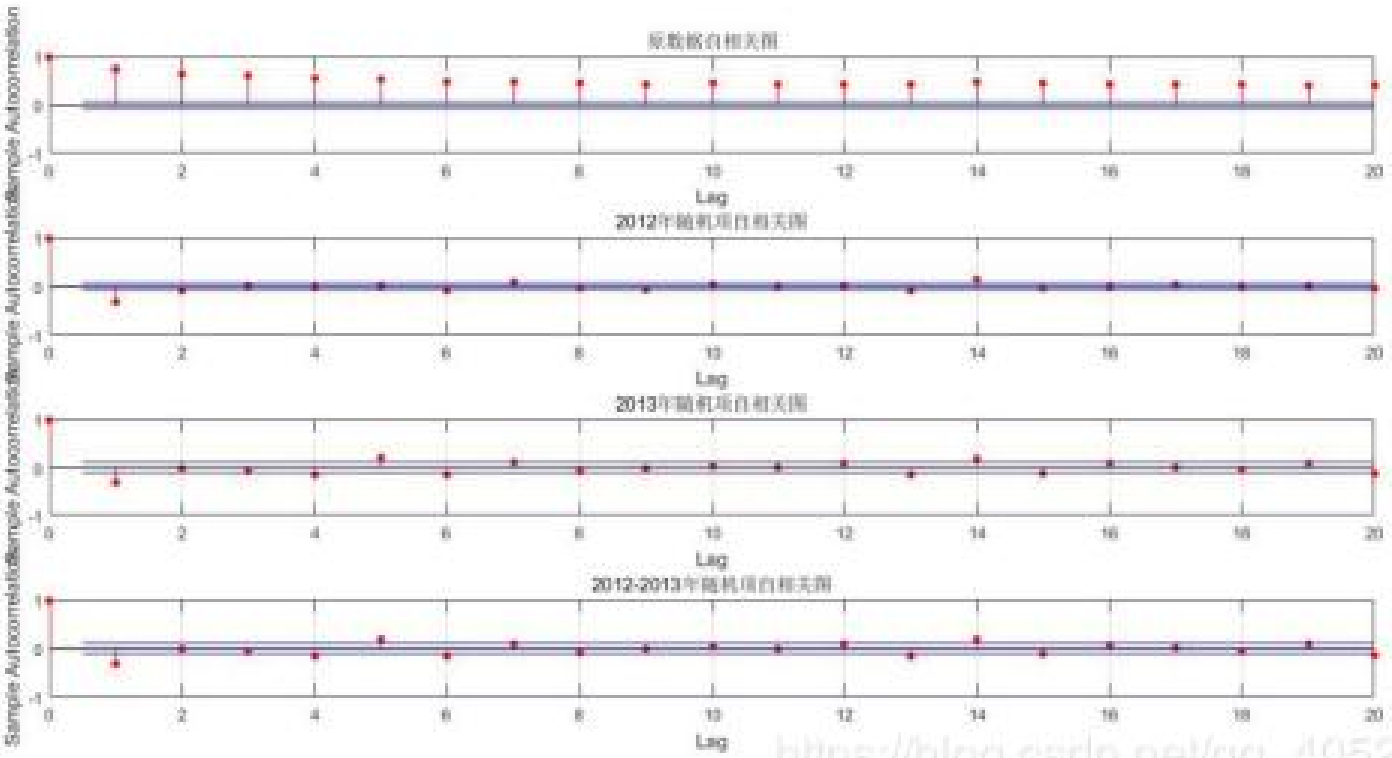
由以上的观察数据和去趋势后数据综合考虑，采用一个月作为周期是个不错的选择。这里自定义d步差分函数对去趋势后的数据进行去周期的处理。







2012-2013年去趋势和去周期比较图



自相关图检验平稳性

%% 计算月平均，季度平均，年平均意图找寻周期规律，输入开户数据，输出各有用的统计量

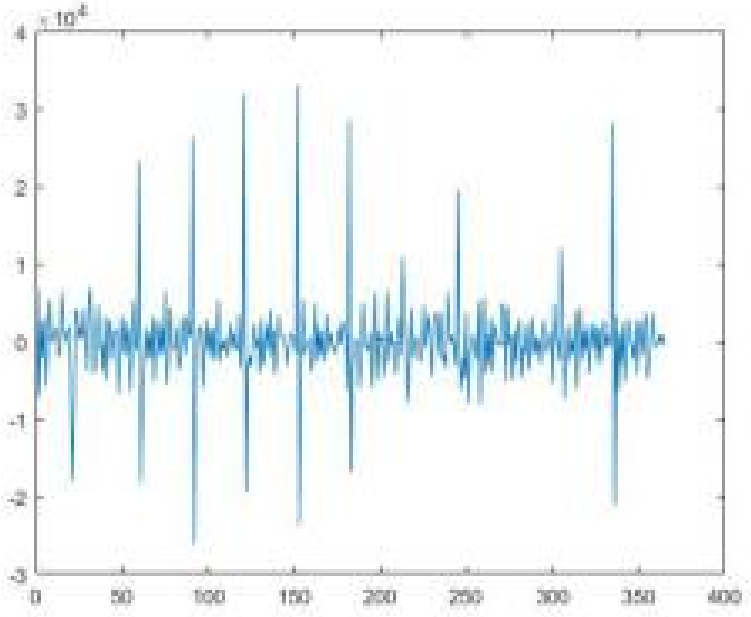
```
function
[months_mean,seasons_mean,years_mean,month_starts,month_ends,season_starts,season_ends] =
Calu_mean(open_nums)
t = 1:length(open_nums);
nums = open_nums;
month_days = [0,31,29,31,30,31,30,31,31,30,31,30,31,31,28,31,30,31,30,31,31,30,31,30,31];
season_days =
[0,sum(month_days(2:4)),sum(month_days(5:7)),sum(month_days(8:10)),sum(month_days(11:13))...
    ,sum(month_days(14:16)),sum(month_days(17:19)),sum(month_days(20:22)),sum(month_days(23:
25))];
% calu the month_mean nums
months_mean = zeros(1,24);
month_starts = zeros(1,24);
month_ends = zeros(1,24);
for ii = 1:length(month_days)-1
    start_day = sum(month_days(1:ii))+1;
    end_day = sum(month_days(1:ii+1));
    month_starts(ii) = start_day;
    month_ends(ii) = end_day;
    months_mean(ii) = mean(nums(start_day:end_day));
end
```

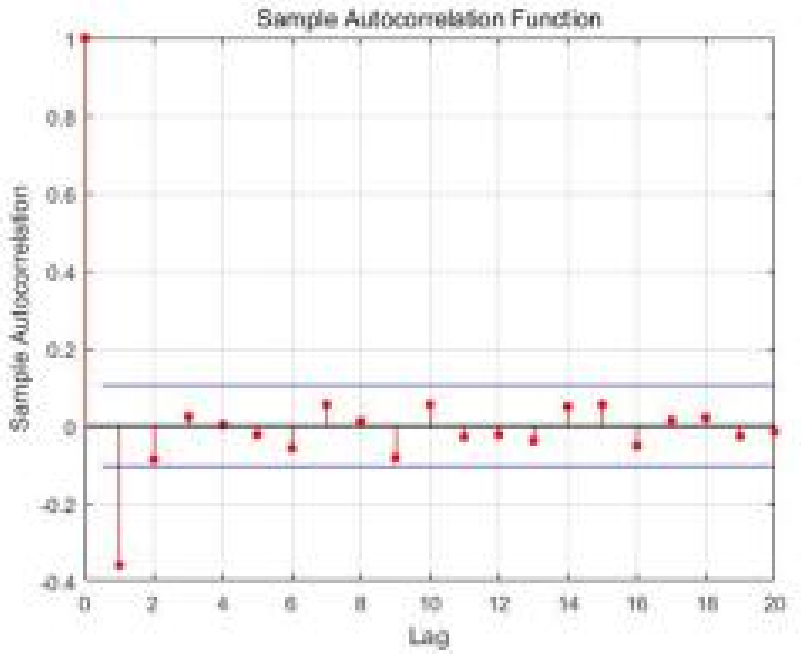
```
%% 去趋势和周期函数
% 输入预处理后的时序数据向量vector和观察数据变化人为认定的周期T
% 输出去趋势后的向量数据detrend_data和去趋势去周期后的向量数据
detrend_deT_data
function [detrend_data,detrend_deT_data] = Detrend_plot(vector,T)
    subplot(3,1,1)
    plot(1:length(vector),vector,'k')
    hold on
    plot(0:length(vector),mean(vector)*ones(1,length(0:length(vector))),'r')
    text(1,mean(vector),'均值')
    legend('趋势','均值')
    title('原开户数据趋势')
    % 去趋势项
    detrend_data = diff(vector,1);
    subplot(3,1,2)
    plot(1:length(detrend_data),detrend_data,'k')
    hold on

plot(0:length(detrend_data),mean(detrend_data)*ones(1,length(0:length(detrend_
data))),'r')
    legend('趋势','均值')
    title('去除趋势项后趋势')
```

```
    % 去季节项
    detrend_deT_data = zeros(length(detrend_data),1);
    for i=1:length(detrend_data)
        if(i<=T)
            detrend_deT_data(i) = [];
        else
            detrend_deT_data(i) = detrend_data(i)-detrend_data(i-T);
        end
    end
    subplot(3,1,3)
    plot(1:length(detrend_deT_data),detrend_deT_data,'k')
    hold on
plot(0:length(detrend_deT_data),mean(detrend_deT_data)*ones(1,length(0:lengt
h(detrend_deT_data))),'r')
    legend('趋势','均值')
    title('去除趋势项后和季节项后趋势')
end
```

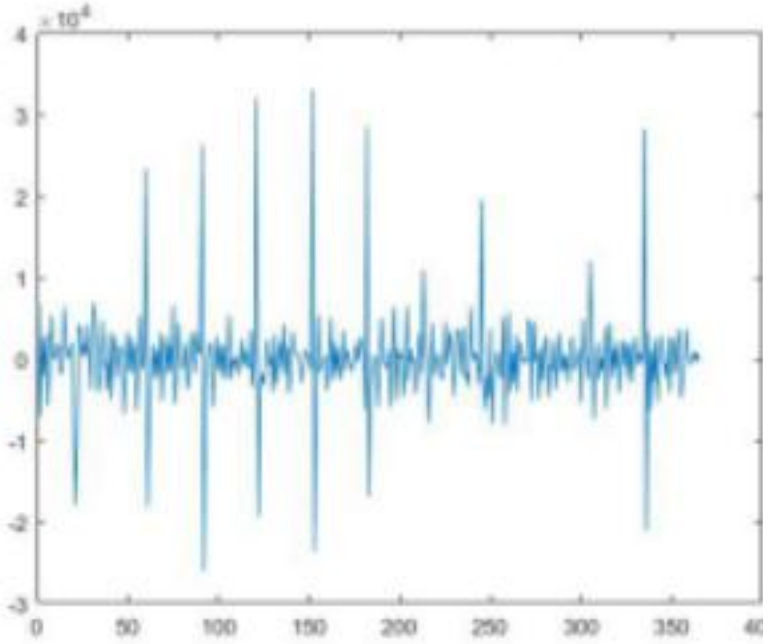
```
%% 主函数脚本，运行该脚本将输出以上所有结果
clearvars
%% 读取数据并预处理
open_nums = xlsread('移动通知户开户数.xlsx',1,'B2:B732');
%% 计算月平均，季度平均，年平均并绘图找周期规律,并处理异常值
figure(1),[months_mean,seasons_mean,years_mean,month_starts,month_ends,season_starts,s
eason_ends] = Calu_mean(open_nums);
[m,n] = find(abs(open_nums-mean(open_nums))>2*std(open_nums));
open_nums(m) = mean(open_nums);
%% 尝试进行去趋势和去周期,考虑到存在多重周期，将年份分开按照月为周期去除周
期（其中每月为多少天根据两年分别不同）
data_2012 = open_nums(1:366);
data_2013 = open_nums(367:end);
figure(2),[detrend_data2012,detrend_deT_data2012] = Detrend_plot(data_2012,30);% 2012
开户数变化趋势
figure(3),[detrend_data2013,detrend_deT_data2013] = Detrend_plot(data_2013,31);% 2013
开户数变化趋势
figure(4),[detrend_data,detrend_deT_data] = Detrend_plot(open_nums,30);% 2012-2013开
户数变化趋势
figure(5),subplot(411),autocorr(open_nums),title('原数据自相关图')% 自相关图分析
subplot(412),autocorr(detrend_data),title('2012年随机项自相关图')
subplot(413),autocorr(detrend_deT_data2013),title('2013年随机项自相关图')
subplot(414),autocorr(detrend_deT_data2013),title('2012-2013年随机项自相关图')
clearvars month_starts month_ends season_starts season_ends data_2012 data2013
```

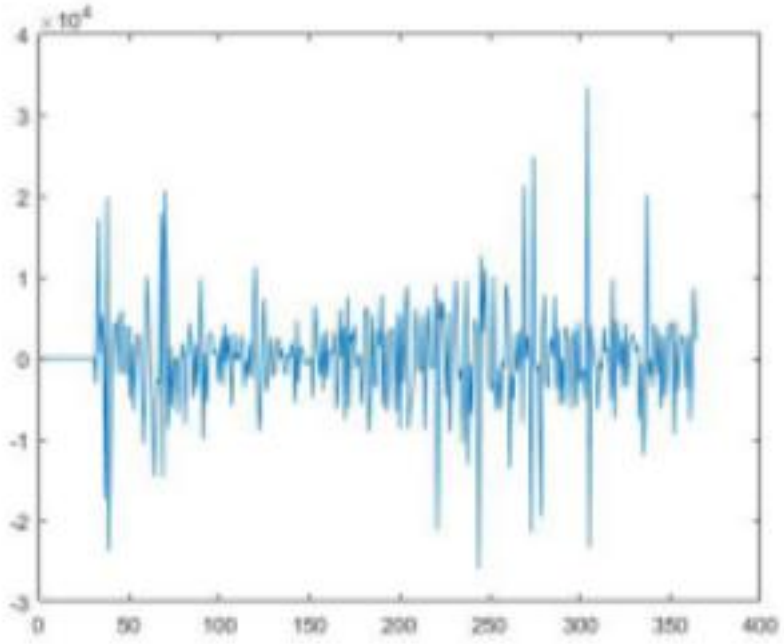
- 判断去趋势和去周期后数据的平稳性





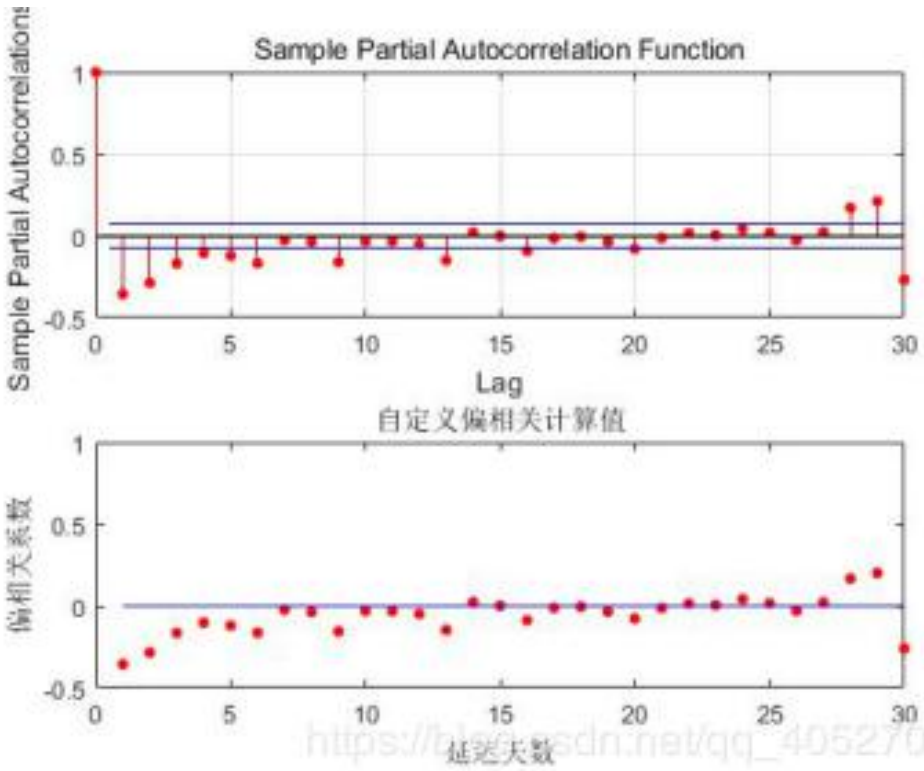
- 转换为零均值平稳序列





通过上面的分析，自相关系数和自偏相关系数均呈现拖尾性，故使用ARMA模型进行进一步分析。

- 自定义偏相关系数计算值与parcorr前30期计算值比较



通过上面的分析，自相关系数和自偏相关系数均呈现拖尾性，故使用ARMA模型进行进一步分析。

- **ARMA模型的参数估计和定阶**

拟函数法进行求解，求解步骤如下：

1. 初定阶数p和q,运用Yule-Walker求解逆函数I
2. 求解MA滑动平均系数
3. 求解AR自回归系数
4. 在求出所有参数后，利用极大似然法残差平方和求方差，算出AIC/BIC指标
5. 重复步骤1-4，在一定大范围内找出极小BIC值对应的阶数p，q

- 参数检验和定阶的进一步修正

通过模型的显著性检验和参数的显著性检验，在参数估计和定阶最小BIC情况下对应的结束p,q附近进行，最终找到最佳的ARMA阶数为ARMA（1，1），参数求解为

$$X_t - 0.1686X_{t-1} = \varepsilon_t - 0.7095\varepsilon_{t-1}$$

- **定阶的进一步修正-模型的检验**

确定好最小BIC准则下的参数估计值和模型阶数p和q后，我们知道——以上各个问题是相互关联的，需整体进行系统化的模型优化. 而根据已有的一组样本数据建立模型，对模型阶数和参数作出判断和估计，是重要的两部分工作.所以最后，我们需要在已经初定的阶数周围小范围内进行搜索和检验，通过模型的有效性检验的参数的显著性检验进一步优化模型，寻找最优的阶数和参数估计值。统计模型只是对生成观测数据真实过程的近似，在模型拟合后，还需要进行模型的有效性检验，及检验拟合模型对序列中信息的提取是否充分。模型的有效性检验即对残差的白噪声检验。

可以发现序列间没有相关性，不具有拖尾和截尾的性质，可以初步判断该模型是可靠的。为了进一步验证残差序列是否为白噪声序列，下面进行LB统计量的检验结果如下：

延迟期数	LB统计量	卡方统计量
6	2.64	1.63
12	4.67	5.23
18	7.67	9.93

课后延伸阅读及实验：

- Matlab时间序列分析：

https://blog.csdn.net/qq_40527086/article/details/84033957#_376

- Kaggle网站流量预测任务第一名解决方案：从模型到代码详解时序预测

https://blog.csdn.net/uwr44uouqcnsuqb60zk2/article/details/78794503

Questions ?