**39. 时间序列分析Ⅱ——ARIMA模型**

随着对时间序列分析方法的深入研究，人们发现非平稳序列的确定性因素分解方法（如季节模型、趋势模型、移动平均、指数平滑等）只能提取显著的确定性信息，对随机性信息浪费严重，同时也无法对确定性因素之间的关系进行分析。

而非平稳序列随机分析的发展就是为了弥补确定性因素分解方法的不足。时间序列数据分析的第一步都是要通过有效手段提取序列中所蕴藏的确定性信息。Box和Jenkins使用大量的案例分析证明差分方法是一种非常简便有效的确定性信息的提取方法。而Gramer分解定理则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。

**（一）ARMA模型**

即自回归移动平均模型，是最常用的拟合平稳时间序列的模型，分为三类：AR模型、MA模型和ARMA模型。

**一、AR(***p***)模型——p阶自回归模型**

1. 模型：



其中，，随机干扰序列ε*t*为0均值、方差的白噪声序列（, t≠s），且当期的干扰与过去的序列值无关，即E(*xtεt*)=0.

由于是平稳序列，可推得均值. 若，称为中心化的AR**(***p***)**模型，对于非中心化的平稳时间序列，可以令，转化为中心化。

记B为延迟算子，称为*p*阶自回归多项式，则AR**(***p***)**模型可表示为：.

2. 格林函数

用来描述系统记忆扰动程度的函数，反映了影响效应衰减的快慢程度（回到平衡位置的速度），Gj表示扰动ε*t-j*对系统现在行为影响的权数。

例如，AR(1)模型（一阶非齐次差分方程），

模型解为.

3. 模型的方差

对于AR(1)模型，.

4. 模型的自协方差

对中心化的平稳模型，可推得自协方差函数的递推公式：



用格林函数显示表示：



对于AR(1)模型，



5. 模型的自相关函数

递推公式：



对于AR(1)模型，.

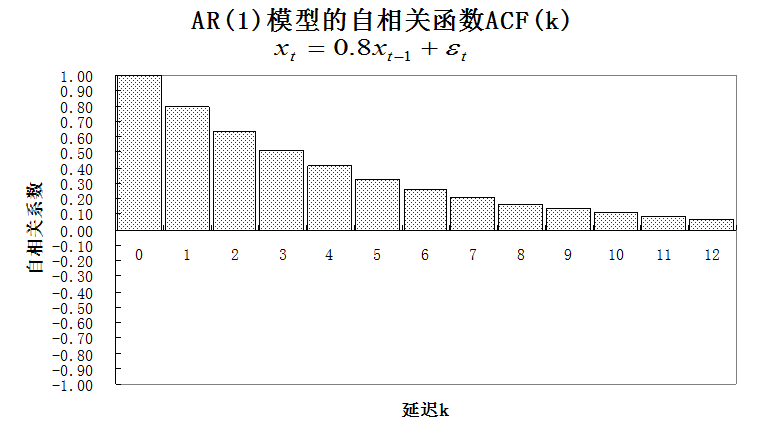
平稳AR(*p*)模型的自相关函数有两个显著的性质：

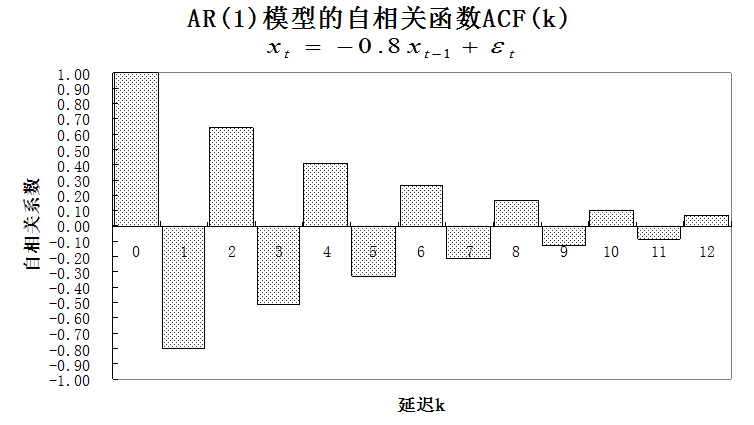
（1）拖尾性

指自相关函数ρ(k)始终有非零取值，不会在*k*大于某个常数之后就恒等于零；

（2）负指数衰减

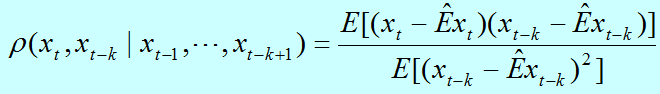
随着时间的推移，自相关函数ρ(k)会迅速衰减，且以负指数（其中为自相关函数差分方程的特征根）的速度在减小。





6. 模型的偏自相关函数

自相关函数ρ(k)实际上并不只是*xt*与*xt-k*之间的相关关系，它还会受到中间k-1个随机变量*xt-*1, …, *xt-k+*1的影响。为了能剔除了中间k-1个随机变量的干扰，单纯测度*xt*与*xt-k*之间的相关关系，引入了滞后k偏自相关函数（PACF），计算公式为：



其中，



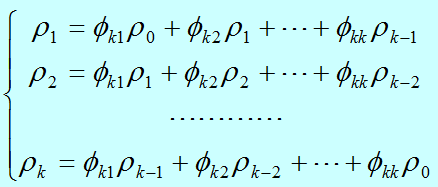
滞后k偏自相关函数实际上等于k阶自回归模型第k个回归系数：



两边同乘以*xt-k*，求期望再除以得到

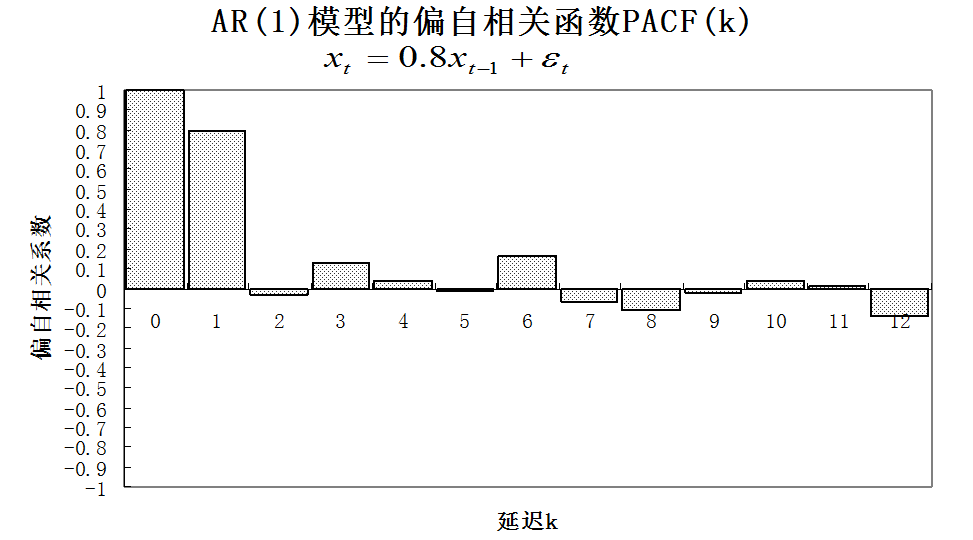


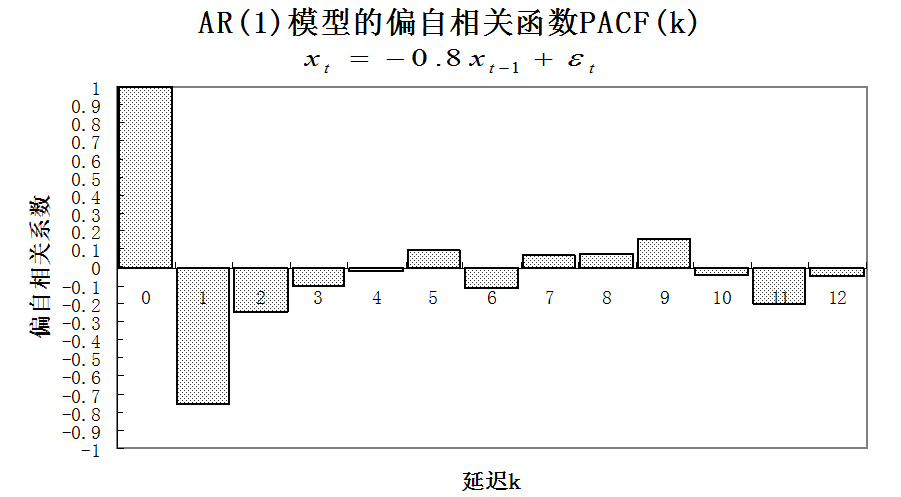
取前k个方程构成的方程组：



称为Yule-Walker方程，可以解出.

可以证明平稳AR(p)模型，当k>p时，. 即平稳AR(p)模型的偏自相关函数具有p步截尾性。





注：实际上样本的随机性使得偏自相关函数不是严格截尾，例如上面两图都1阶显著不为0，1阶之后都近似为0.

**二、MA(q)模型——q阶移动平均模型**

1. 模型：

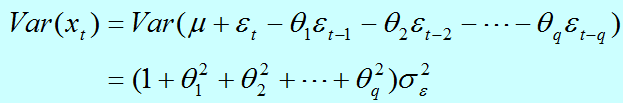


其中，，随机干扰序列为0均值、方差的白噪声序列（, t≠s）。

若μ=0，称为中心化的MA(q)模型，非中心化的MA(q)模型可以通过转化为中心化。

记B为延迟算子，称为q阶自移动平均系数多项式，则中心化MA(q)模型可以表示为.

2. 模型的方差

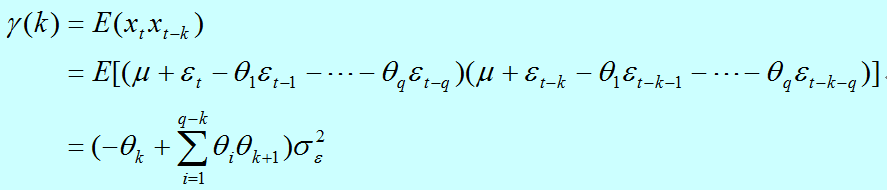


3. 模型的自协方差

只与滞后阶数k相关，且q阶截尾。当k=0时，

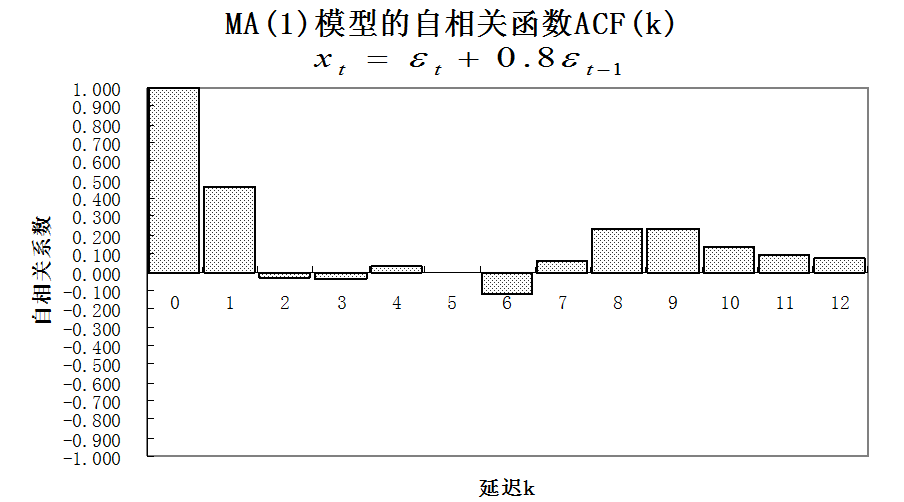


当1≤k≤q时，

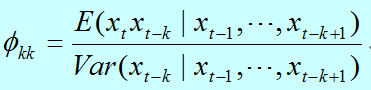


当k>q时，.

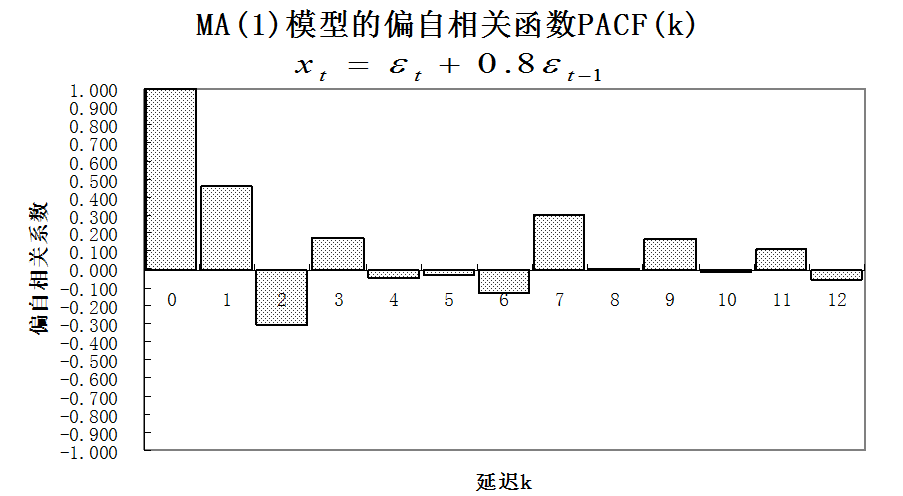
4. 模型的自相关函数：（q阶截尾性）



5. 模型的滞后k阶偏自相关函数（中心化）



可以证明滞后k阶偏自相关函数具有拖尾性。



6. 模型的可逆性

以MR(1)为例，

模型Ⅰ：

模型Ⅱ：

它们的自相关函数相同（即相同的自相关函数对应不同的回归模型），为了保证对应的唯一性，需要增加约束条件，即MR(q)模型的可逆性条件。

观察两个模型的第二种表示：当时，模型Ⅰ收敛、模型Ⅱ不收敛；当时，模型Ⅰ不收敛、模型Ⅱ收敛。

表示成收敛形式的MR(q)模型称为可逆MR(q)模型。一个自相关函数只对应唯一一个可逆MR(q)模型。

**三、ARMA(p, q)模型——自回归移动平均模型**

1. 模型



其中，，，随机干扰序列ε*t*为0均值、方差的白噪声序列（, t≠s），且当期的干扰与过去的序列值无关，即E(*xtεt*)=0.

若，则称为中心化的ARMA(p,q)模型。引入延迟算子，中心化的ARMA(p,q)模型可表示为：.

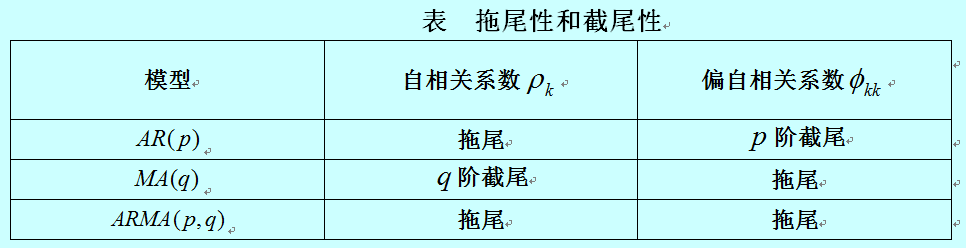
显然，AR(p)和MA(q)模型是ARMA(p,q)模型的特例。

2. 数字特征

（1）均值：;

（2）自协方差函数：，其中*Gi*为格林函数；

（3）自相关函数：



3. 模型的初步定阶

对于平稳非白噪声序列，计算出样本自相关系数（ACF）和偏自相关系数（PACF），根据其性质估计自相关阶数和移动平均阶数，称为ARMA(p,q)模型的定阶。

可以推导出：样本自相关函数和偏自相关函数都近似服从正态分布.

取显著水平α=0.05，若样本自相关系数和样本偏自相关系数在最初的k阶明显大于2倍标准差，而后几乎95%的系数都落在2倍标准差的范围内，且非零系数衰减为小值波动的过程非常突然，通常视为k阶截尾；若有超过5%的样本相关系数大于2倍标准差，或者非零系数衰减为小值波动的过程比较缓慢或连续，通常视为拖尾。

4. 参数估计

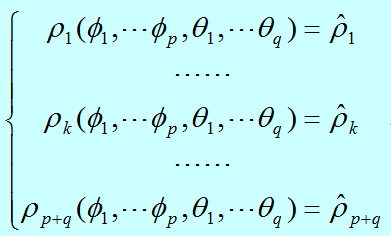
对非中心化的ARMA(p,q)模型

.

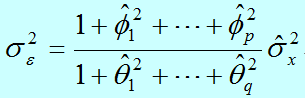
参数μ可用样本均值来估计总体均值（矩估计法），初步定阶估计出自相关阶数和移动平均阶数后，模型共有p+q+1个未知参数：.

（1）参数的矩估计

用时间序列样本数据计算出延迟1阶到p+q阶的样本自相关函数，延迟k阶的总体自相关函数为. 用计算出的样本自相关函数来估计总体自相函数，得到p+q个联立方程组：



从中解出的值作为未知参数估计值. ARMA(p,q)模型的两边同时求方差，并把前面的参数的估计值代入，可得白噪声序列的方差估计为：



（2）参数的极大似然估计

当总体分布类型已知时，极大似然估计是常用的估计方法。其基本思想是，认为样本来自使该样本出现概率最大的总体。

因此，未知参数的极大似然估计，就是使得似然函数（即联合密度函数）达到最大值的参数值：



在时间序列分析中，序列的总体分布通常是未知的。为了便于分析和计算，通常假设序列服从多元正态分布，它的联合密度函数是可导的。在求极大似然估计时，为了求导方便，常对似然函数取对数，然后对对数似然函数中的未知参数求偏导数，得到似然方程组。理论上，只要求解似然方程组即可得到未知参数的极大似然估计。但在实际上是使用计算机经过复杂的迭代算法求出未知参数的极大似然估计。

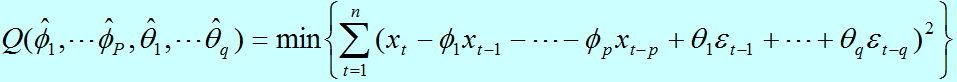
两种估计的比较：

矩估计的优点是不要求知道总体的分布，计算量小，估计思想简单直观。但缺点是只用到了样本自相关系数的信息，序列中的其他信息被忽略了，这导致估计精度一般较差。因此，它常被作为极大似然估计和最小二乘估计的迭代计算的初始值。

极大似然估计的优点是充分应用了每一个观察值所提供的信息，因而它的估计精度高，同时，还具有估计的一致性、渐近正态性和渐近有效性等优良统计性质，是一种非常优良的参数估计方法。

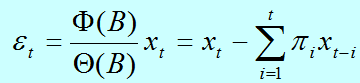
（3）参数的最小二乘估计

使ARMA(p,q)模型的残差平方和达到最小的那组参数值：

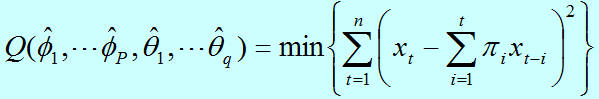


通过计算机借助迭代方法求出。由于充分利用了序列的信息，该方法估计精度最高。

在实际运用中，最常用的是条件最小二乘估计，假定时间序列过去未观察到序列值等于序列均值，可得到残差的有限项表达式：

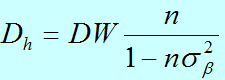


于是残差平方和达到最小的那组参数值为：



5. 模型和参数的显著性检验

ARMA(p,q)模型中，使用QLB统计量检验残差序列的自相关性，为了克服DW检验的有偏性，Durbin在1970年提出了修正的Durbin h统计量：



其中，*n*为观察值序列的长度，为延迟因变量系数的最小二乘估计的方差。

参数的显著性检验是要检验每一个模型参数是否显著非零。若某个参数为零，模型中包含这个参数的乘积项就为零，可以简化模型。因此，该检验的是为了精简模型。

原假设H0：某未知参数βj=0；H1：βj≠0. 可以构造出检验未知参数显著性的t(n-m)检验统计量，其中m为参数的个数。

6. 模型优化

当一个拟合模型在置信水平α下通过了检验，说明了在该置信水平下该拟合模型能有效地拟合时间序列观察值的波动。但是这种有效的拟合模型并不是惟一的。

如果同一个时间序列可以构造两个拟合模型，且两个模型都显著有效，那么应该选择哪个拟合模型用于统计推断呢？通常采用AIC和SBC信息准则来进行模型优化。

（1）AIC准则——最小信息量准则

由日本统计学家赤池弘次（Akaike）于1973年提出，是一种考评综合最优配置的指标，它是拟合精度和参数未知个数的加权函数：

AIC=－2ln(模型中极大似然函数值)+2(模型中未知参数个数)

使其达到最小值的模型被认为是最优模型。

（2）BIC/SBC准则

AIC准则的不足：若时间序列很长，相关信息就越分散，需要多自变量复杂拟合模型才能使拟合精度比较高。在AIC准则中拟合误差等于，即随样本容量*n*增大，但模型参数个数的惩罚因子（始终=2）却与*n*无关。因此在样本容量*n*趋于无穷大时，由AIC准则选择的拟合模型不收敛于真实模型，它通常比真实模型所含的未知参数个数要多。

为了弥补AIC准则的不足，Akaike于1976年提出BIC准则。而Schwartz在1978年根据贝叶斯理论也得出同样的判别准则，称为SBC准则。SBC准则定义为：

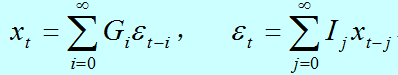
SBC=－2ln(模型中极大似然函数值)+ln(*n*)(模型中未知参数个数)

即将未知参数个数的惩罚权重由常数2变成了ln(*n*)。在所有通过检验的模型中使得AIC或SBC函数达到最小的模型为相对最优模型（因为不可能比较所有模型）。

7. 模型预测

即利用时间序列已观察到的样本值对时间序列在未来某个时刻的取值进行估计。常用的预测方法是线性最小方差预测。

根据ARMA(p,q)模型的平稳性和可逆性，可以用格林函数的传递形式和逆转函数的逆转形式等价描述该序列：

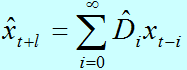


右式代入左式得：



可见，*xt*是历史数据*xt-*1, *xt-*2, …的线性函数。

对于任意一个将来时刻*t*+*l*，也可以用上式预测，但*xt+l-*1, …, *xt+*1未知。根据线性函数的可加性，所有未知信息都可以用已知信息的线性函数表示出来，并用该线性函数进行估计：



用来衡量预测误差，最常用的预测原则是预测误差的方差最小法：



在线性预测方差最小法下得到的估计值是在序列*xt*, *xt-*1, …已知的情况下得到的条件无偏最小方差估计值。且预测方差只与预测步长*l*有关，而与预测起始点*t*无关。

预测步长*l*越大预测值的方差越大，因此只适合于短期预测。在正态假定下，估计值的1-α的置信区间为：



**（二）ARIMA模型——混和自回归移动平均模型**

**一、原理**

也称Box-Jenkins模型，用来处理单变量同方差的非平稳时间序列，通过差分法或适当的变换转化为平稳序列，再使用ARMA模型。

注：残差的条件方差是异方差的时间序列，适合用GARCH模型。

ARIMA(p,d,q)模型的形式如下：



其中，为d阶差分，

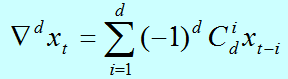




为平稳可逆ARMA(p,q)模型的自回归和移动平均系数多项式。

可见，ARIMA模型的实质就是差分运算与ARMA模型的组合。任何非平稳序列只要通过适当阶数的差分实现平稳，就可以对差分后序列进行ARMA模型的拟合了。

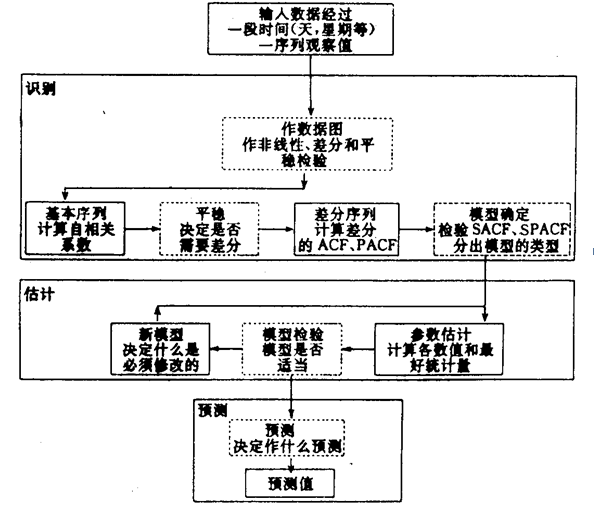
d阶差分后的序列可表示为：



其中，为组合数，即d阶差分后序列等于原来序列的若干序列值的某种加权和。

**二、建模步骤**

分为三个阶段：识别阶段、估计阶段和预测阶段。



1. 识别阶段

使用identify语句来指定响应变量序列并且识别候选ARIMA模型。一般先对序列进行非线性、差分和平稳性检验，可能对序列进行差分，然后计算自相关系数ACF、逆自相关系数IACF、偏自相关系数PACF和互相关系数。此阶段的输出通常会建议一个或多个可拟合的ARIMA模型。如果模型确定，还可以检验样本自相关系数SACF和样本偏自相关系数SPACF，以分出模型的类型。

2. 估计阶段

使用estimate语句来指定ARIMA模型去拟合在前面identify语句中指定的响应变量，并且估计该模型的参数。estimate语句也生成诊断统计量从而帮助判断该模型的适用性。

关于参数估计值的显著性检验可以指出模型里的一些项是否不需要：拟合优度统计量R2可帮助比较该模型和其他模型的优劣；白噪声残差检验可指明残差序列是否包含可被其他更复杂模型采用的额外信息，如果诊断检验表明模型不适用，则可尝试另一个模型然后重复估计和诊断。

3. 预测阶段

使用forecast语句来预测时间序列的未来值，并对这些来自前面estimate语句生成的ARIMA模型的预测值产生置信区间。

**（三）PROC ARIMA过程**

ARIMA过程采用Box-Jenkins方法建立模型，是集一元时间序列模型判定、参数估计和预测为一体的多功能综合工具。当ARIMA模型包括其他时间序列作为输入变量时，有时也被称为ARIMAX模型。ARIMA模型还支持干预或中断时间序列模型、误差的多元回归分析、任意复杂程度的有理转移函数模型。

基本语法：

proc arima data=数据集 out=输出数据集;

where 条件表达式;

identify var=变量(…) <选项列表> ;

estimate <选项列表>;

forecast <选项列表>;

说明：

（1）where语句

指定用于分析的时间间隔，通常条件表达式是有关日期变量的条件表达式，例如：’31dec98’d < 日期变量 < ’31dec99’d

（2）identify语句

主要完成时间序列的差分计算，样本ACF、IACF和PACF函数的计算、卡方检验统计量和白噪声自相关检验的p值的计算。主要选项：

 var=变量(d1,…,dk)——是必选项，指定要分析的时间序列变量，按括号内列出的差分周期列表来计算时间序列的滞后差分。例如：

var=X(1) 为对滞后1项的序列差分，即Xt-Xt-1;

var=X(2) 为对滞后2项的序列差分，即Xt-Xt-2;

var=X(1,1) 为X进行二阶差分，即(Xt-Xt-1)-(Xt-1-Xt-2);

 nlag=数字——指定计算自相关的滞后数，其值应大于p+d+q，小于观测数，默认值为24；

 crosscorr=(干预变量(d1))——列出有var=指定的响应序列的交叉相关变量。干预变量在交叉相关变量中。交叉相关变量的差分由圆括号内的差分滞后数确定。

（3）estimate语句

对已执行的identify语句中的响应变量规定一个模型，主要选项：

 p=(p1,p2,…)…(p1,p2,…)——定义一个在p中指定的滞后处具有自回归参数的模型，p的默认值为0;

 q=(q1,q2,…)…(q1,q2,…)——定义一个在q中指定的滞后处具有滑动平均参数的模型，q的默认值为0。如果p=和q=都没有指定，则拟合随机模型；

 noconstant——在模型中舍弃常数项μ;

 noint——在该模型中不拟合截距参数；

 method=ml | uls | cls——指定估计时使用的方法，分别为极大似然方法、无条件最小二乘法、有条件最小二乘法，默认为cls;

 outest=数据集——将参数估计值输出到指定的数据集；

 outmode=数据集——将模型和参数估计值输出到指定的数据集；

 outstat=数据集——将模型诊断统计量输出到指定的数据集；

 plot——可以绘制残差自回归函数等；

（4）forecast语句

利用estimate语句所产生的参数估计生成时间序列的预测值，主要选项：

 alpha=α——设置预测置信限的大小，上下置信限的置信水平为1-α，默认值为0.05;

 lead=n——指定要计算的多步向前预测值的次数，默认为24；

 back=n——指定在数据末尾前n个观测值开始进行分步预测，默认为0;

 interval=时间间隔——指定观测之间的时间间隔，常用的时间间隔为year、qtr、month、week、weekday、day、hour、minute、second;

 id=变量名——指明输入数据集中一个变量，用于识别与观测有关的时间周期；

 out=数据集——将预测值和其他值输出到一个指定的数据集中。

**例1**有一组1949年至1961年国际航线旅客月度人数的记录：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **YEAR** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
| 1949 | 112 | 118 | 132 | 129 | 121 | 135 | 148 | 148 | 136 | 119 | 104 | 118 |
| 1950 | 115 | 126 | 141 | 135 | 125 | 149 | 170 | 170 | 158 | 133 | 114 | 140 |
| 1951 | 145 | 150 | 178 | 163 | 172 | 178 | 199 | 199 | 184 | 162 | 146 | 166 |
| 1952 | 171 | 180 | 193 | 181 | 183 | 218 | 230 | 242 | 209 | 191 | 172 | 194 |
| 1953 | 196 | 196 | 236 | 235 | 229 | 243 | 264 | 272 | 237 | 211 | 180 | 201 |
| 1954 | 204 | 188 | 235 | 227 | 234 | 264 | 302 | 293 | 259 | 229 | 203 | 229 |
| 1955 | 242 | 233 | 267 | 269 | 270 | 315 | 364 | 347 | 312 | 274 | 237 | 278 |
| 1956 | 284 | 277 | 317 | 313 | 318 | 374 | 413 | 405 | 355 | 306 | 271 | 306 |
| 1957 | 315 | 301 | 356 | 348 | 355 | 422 | 465 | 467 | 404 | 347 | 305 | 336 |
| 1958 | 340 | 318 | 362 | 348 | 363 | 435 | 491 | 505 | 404 | 359 | 310 | 337 |
| 1959 | 360 | 342 | 406 | 396 | 420 | 472 | 548 | 559 | 463 | 407 | 362 | 405 |
| 1960 | 417 | 391 | 419 | 461 | 472 | 535 | 622 | 606 | 408 | 461 | 390 | 432 |

使用ARIMA过程进行建模和预测。

（一）**读入数据、绘制时间序列图、判断平稳性，**代码：

**data** arimad01;

date=intnx('month',**'31dec1948'd**,\_n\_);

input x @@;

format date monyy5.;

datalines;

112 118 132 129 121 135 148 148 136 119 104 118

115 126 141 135 125 149 170 170 158 133 114 140

145 150 178 163 172 178 199 199 184 162 146 166

171 180 193 181 183 218 230 242 209 191 172 194

196 196 236 235 229 243 264 272 237 211 180 201

204 188 235 227 234 264 302 293 259 229 203 229

242 233 267 269 270 315 364 347 312 274 237 278

284 277 317 313 318 374 413 405 355 306 271 306

315 301 356 348 355 422 465 467 404 347 305 336

340 318 362 348 363 435 491 505 404 359 310 337

360 342 406 396 420 472 548 559 463 407 362 405

417 391 419 461 472 535 622 606 408 461 390 432

;

**run**;

**proc** **sgplot** data=arimad01;

series x=date y=x / markers;

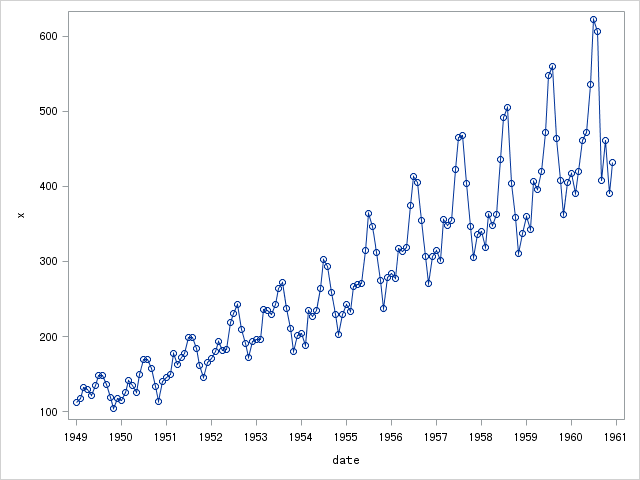
**run**;

**proc** **arima** data=arimad01;

identify var=x;

**run**;

运行结果及说明：



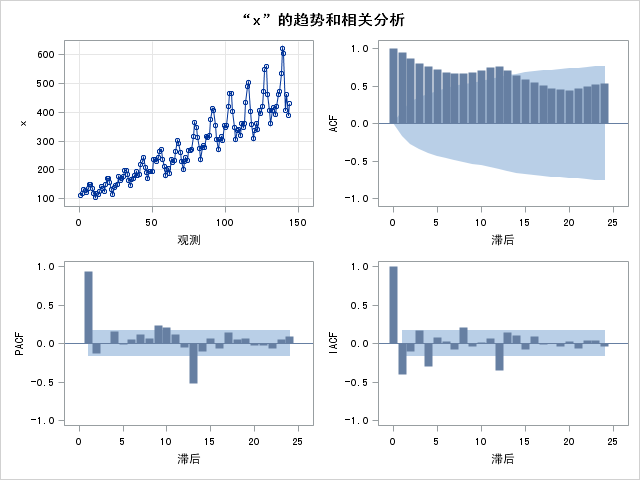
为了识别时间序列的数据是否平稳和是否明显有季节性，绘制时间序列图，可见不平稳。更精确的判断，通常利用proc arima过程的identify语句来检测序列是否平稳：

ARIMA 过程

| **变量名 = x** | |
| --- | --- |
| **工作序列的均值** | 279.6042 |
| **标准差** | 118.5103 |
| **观测数** | 144 |

| **白噪声的自相关检查** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **至滞后** | **卡方** | **自由度** | **Pr > 卡方** | **自相关** | | | | | |
| **6** | 572.74 | 6 | <.0001 | 0.940 | 0.869 | 0.801 | 0.756 | 0.717 | 0.687 |
| **12** | 1041.31 | 12 | <.0001 | 0.669 | 0.660 | 0.676 | 0.711 | 0.750 | 0.765 |
| **18** | 1372.45 | 18 | <.0001 | 0.711 | 0.644 | 0.587 | 0.540 | 0.503 | 0.471 |
| **24** | 1616.11 | 24 | <.0001 | 0.454 | 0.445 | 0.462 | 0.489 | 0.522 | 0.535 |

P值<0.0001<α=0.05，拒绝序列为白噪声的原假设，即序列是非平稳的。



自相关系数ACF，发现自相关系数长期位于零轴正的一边，这是具有单调递增趋势序列的典型特征，同时自相关图呈现出明显的正弦波动规律，这是具有周期变化规律的非平稳序列的典型特征。自相关图显示出来的这两个性质和该序列时序图显示出的带长期递增趋势的周期性质是非常吻合的。

对于不同的时间序列类型，它们相应ACF的特性有：

 ①若时间序列是随机无趋势的，所有时滞的自相关系数都等于0;

 ②若时间序列是上升或下降趋势的，那么对于短时滞来说，自相关系数大且为正，而且随着时滞lag的增加而缓慢地下降。ACF本身似乎随着时滞长度的增加而呈下降趋势;

 ③若时间序列无趋势但具有季节性，那么对于按月(L=12)采集的数据来说，时滞12，24，…的自相关系数达到最大（即nL处），而随着时滞长度增加时，变得较小;

 ④若时间序列有趋势且具有季节性，那么自相关系数特性类同于有趋势序列，但是它们是摆动的，对于按月的数据，在时滞12，24…等处具有峰态;

 ⑤若时间序列是平稳的，它具有正自相关的误差项，那么低阶的自相关系数较大，但是随着时滞长度的增加而迅速地衰减;

另外，在ACF图中，通过自相关系数值是否在零点附近的两条对称的虚线之内，完成不同时滞的总体自相关系数为0的检验。

**（二）变换不平稳序列为平稳序列**

最常用的变换方法有：

 ①若时间序列呈线性趋势，均值不是常数，利用一阶差分将产生一个平稳序列；

 ②若时间序列呈二次趋势，均值不是常数，利用二阶差分将产生一个平稳序列；

 ③若时间序列呈现出随时间的上升或下降而偏差，方差不是常数，通常可利用取自然对数转化为平稳序列；

 ④若时间序列呈现指数趋势，均值和方差都不是常数，通常也可利用取自然对数转化为平稳序列；

 ⑤若时间序列呈现“相对环”趋势，通常将数据除以同时发生的时间序列的相应值转化为平稳序列。

**一、取对数消除振幅变大趋势**

从原始序列的时序图或ACF图发现，该序列具有季节性变化同时有增大的趋势，而且季节变化的振幅越来越大。因此，先对数据取对数来先消除增幅越来越大的现象。

代码：

**data** arimad02;

set arimad01 ;

xlog=log(x);

**run**;

**proc** **print** data = arimad02;

**run**;

goptions reset=global gunit=pct cback=white border

htitle=**6** htext=**3** ftext=swissb colors=(black);

**proc** **gplot** data=arimad02;

plot xlog\*date / vaxis=axis1 haxis=axis2

href=**'31dec1949'd** to **'1jan61'd** by year;

plot2 x\*date /vaxis=axis3 vref=**100**;

symbol1 i=join v=c h=**3** l=**1** r=**1** font=swissb c=green;

symbol2 i=join v=c h=**3** l=**1** r=**1** font=swissb c=blue;

axis1 label=('Log') order=(**4.5** to **6.5** by **0.5**) offset=(**0**,**45**);

axis2 label=('12 Month') order=(**'1jan49'd** to **'1jan61'd** by year);

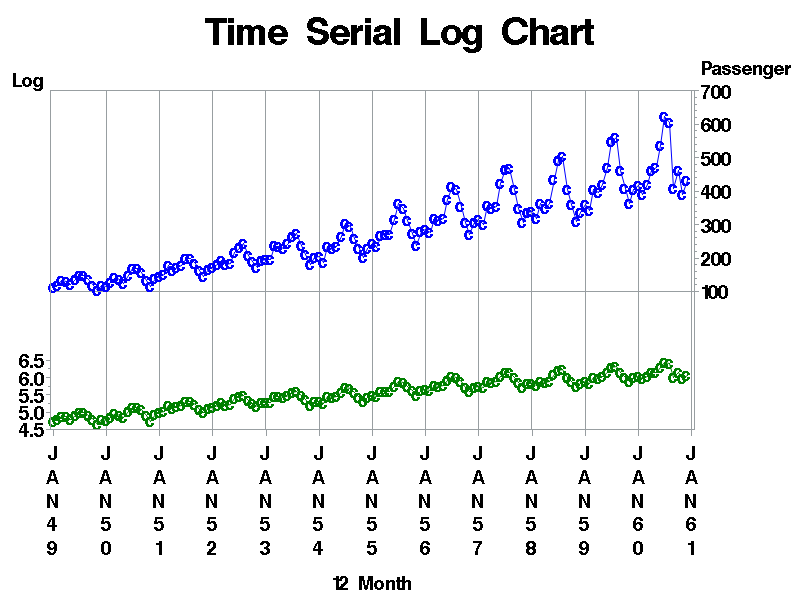
axis3 label=('Passenger') order=(**100** to **700** by **100**) offset=(**23**,**0**);

format date monyy. ;

title1 'Time Serial Log Chart';

**run**;

运行结果（**为便于对比，按12个时间单位间隔画在同一图上**）：



**二、取两次差分消除季节增长趋势**

对取对数后的序列，明显呈现季节性的增长的趋势，仔细分析每12个单位的周期还有增长趋势。所以需要对该序列再进行滞后一次和滞后12次共两次差分最终转换为平稳序列。

代码：

**data** arimad03;

set arimad02;

dif12=dif1(xlog)-(lag1(xlog)-lag12(xlog));

**run**;

**proc** **gplot** data=arimad03 ;

plot xlog\*date /vaxis=axis1 haxis=axis2

href=**'31dec1949'd** to **'1jan61'd** by year;

plot2 dif12\*date /vaxis=axis3 vref=-**1**;

symbol1 i=join v=c h=**3** l=**1** r=**1** font=swissb c=green;

symbol2 i=join v=c h=**3** l=**1** r=**1** font=swissb c=blue;

axis1 label=('Log') order=(**4.5** to **6.5** by **0.5**) offset=(**0**,**45**);

axis2 label=('12 Month') order=(**'1jan49'd** to **'1jan61'd** by year);

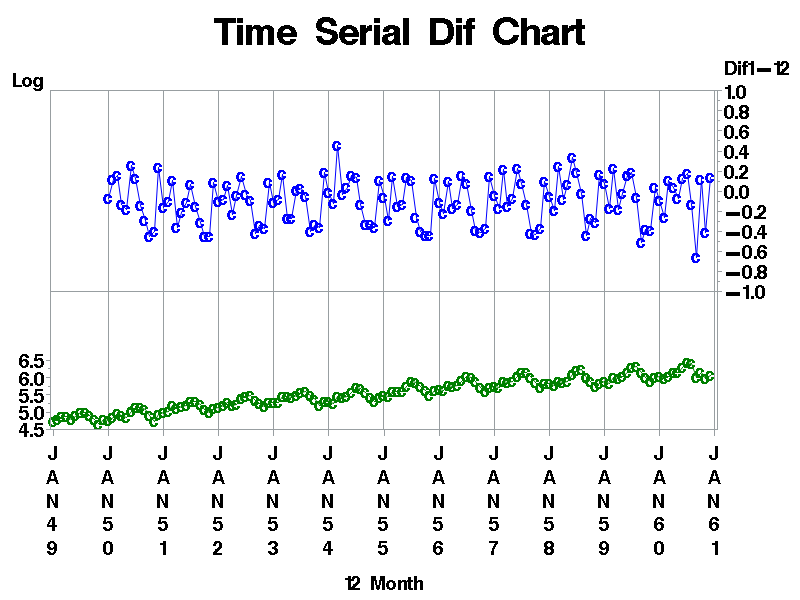
axis3 label=('Dif1-12') order=(-**1** to **1** by **0.2**) offset=(**23**,**0**);

format date monyy. ;

title1 'Time Serial Dif Chart';

**run**;

运行结果：



**（三）检验待选的时间序列模型的自相关函数**

对需要转换为平稳时间序列的数据，若最终是要用差分的方法来转换，通常可直接调用PROC ARIMA过程的identify语句来实现对指定变量xlog所选差分的时滞数（如1和12）的检验。

代码：

**proc** **arima** data=arimad02;

identify var=xlog(**1**,**12**) nlag=**15**;

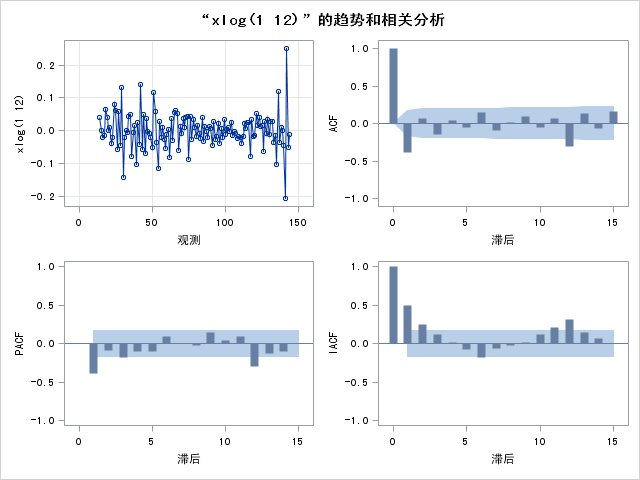
**run**;

运行结果：

ARIMA 过程

| **变量名 = xlog** | |
| --- | --- |
| **差分期间** | 1,12 |
| **工作序列的均值** | 0.000291 |
| **标准差** | 0.053721 |
| **观测数** | 131 |
| **差分剔除的观测** | 13 |

| **白噪声的自相关检查** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **至滞后** | **卡方** | **自由度** | **Pr > 卡方** | **自相关** | | | | | |
| **6** | 27.44 | 6 | 0.0001 | -0.388 | 0.072 | -0.144 | 0.041 | -0.050 | 0.151 |
| **12** | 44.65 | 12 | <.0001 | -0.092 | 0.013 | 0.089 | -0.056 | 0.070 | -0.306 |



ACF图中，自相关系数在延迟1阶后都落入2倍标准差内，然后在延迟12阶处突然有一个较大的自相关系数，紧接着又落入2倍标准差内，很象在1,12处截尾；又根据PACF图中，偏自相关系数很象在1,12处拖尾，在3和9处有超过2倍标准差，因此可选择MA的阶数为1,12。从IACF及PACF图看可选择AR的阶数为1,12。

自相关系数的白噪声检验，明显看出延迟6期和12期的QLB统计量分别为27.44和44.65，p值<α=0.05，拒绝序列为白噪声的原假设，即拒绝延迟期数小于等于6和12的序列值之间相互独立，说明经过滞后1次和滞后12次共两次差分转换后序列是平稳非白噪声序列，还蕴藏着相关信息需要提取出来，因此该序列蕴含ARIMA模型。

**（四）估计备选时间序列模型的参数**

代码：

**proc** **arima** data=arimad03;

identify var=xlog(**1**,**12**) nlag=**15**;

estimate q=(**1**)(**12**) p=(**1**)(**12**) noconstant outmodel=xmode;

**run**;

运行结果及说明：

| **条件最小二乘估计** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **数** | **估计值** | **标准误差** | **t 值** | **近似Pr > |t|** | **滞后** |
| **MA1,1** | 0.55123 | 0.14697 | 3.75 | 0.0003 | 1 |
| **MA2,1** | 0.42423 | 0.19218 | 2.21 | 0.0291 | 12 |
| **AR1,1** | 0.06398 | 0.17671 | 0.36 | 0.7179 | 1 |
| **AR2,1** | -0.15038 | 0.20519 | -0.73 | 0.4650 | 12 |

|  |  |
| --- | --- |
| **方差估计** | 0.002005 |
| **标准误差估计** | 0.044778 |
| **AIC** | -438.081 |
| **SBC** | -426.581 |
| **残差数** | 131 |

\* AIC 和 SBC 不包括对数行列式。

| **参数估计相关性** | | | | |
| --- | --- | --- | --- | --- |
| **参数** | **MA1,1** | **MA2,1** | **AR1,1** | **AR2,1** |
| **MA1,1** | 1.000 | 0.066 | 0.862 | 0.065 |
| **MA2,1** | 0.066 | 1.000 | 0.127 | 0.849 |
| **AR1,1** | 0.862 | 0.127 | 1.000 | 0.112 |
| **AR2,1** | 0.065 | 0.849 | 0.112 | 1.000 |

| **残差的自相关检查** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **至滞后** | **卡方** | **自由度** | **Pr > 卡方** | **自相关** | | | | | |
| **6** | 5.51 | 2 | 0.0637 | -0.002 | 0.029 | -0.043 | -0.064 | 0.011 | 0.181 |
| **12** | 6.50 | 8 | 0.5918 | -0.032 | -0.016 | 0.057 | -0.036 | 0.033 | 0.002 |
| **18** | 10.39 | 14 | 0.7332 | -0.019 | 0.011 | 0.081 | -0.132 | 0.031 | -0.020 |
| **24** | 14.45 | 20 | 0.8067 | -0.067 | -0.070 | 0.036 | -0.018 | 0.121 | 0.005 |

| **变量“xlog”的模型** | |
| --- | --- |
| **差分期间** | 1,12 |

在该模型中没有均值项。

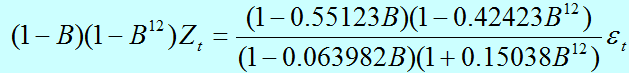
| **自回归因子** | |
| --- | --- |
| **因子 1:** | 1 - 0.06398 B\*\*(1) |
| **因子 2:** | 1 + 0.15038 B\*\*(12) |

| **移动平均因子** | |
| --- | --- |
| **因子 1:** | 1 - 0.55123 B\*\*(1) |
| **因子 2:** | 1 - 0.42423 B\*\*(12) |

不带均值的条件最小二乘估计输出：该参数的估计值、标准误差和t值，以及该参数在模型中滞后数。

进一步诊断检验模型，若拟合的模型是可行的，则所有参数估计应该具有显著的t值，残差近似地为正态，残差序列具有随机无趋势序列的ACF与PACF，且不含有过度的p、d、q。

t统计量（其精确性依赖于模型的有效性和观察序列的长度等因素）用来检验该项系数是否为0. 拟合的ARIMA(1,1,1)×(1,1,1)12模型的具体形式为：



由于AR1,1及AR2,1的t值分别为0.36和-0.73（很小），所以该项系数为0的假设检验并不显著p值<α=0.05，故可以丢弃这两项。重新估计：

**（五）估计模型参数、用模型进行预测**

代码：

**proc** **arima** data=arimad03;

identify var=xlog(**1**,**12**) nlag=**15**;

estimate q=(**1**)(**12**) noconstant outmodel=xmode;

forecast lead=**12** interval=month id=date out=forxlog;

**run**;

**proc** **print** data=forxlog;

**run**;

运行结果及说明：

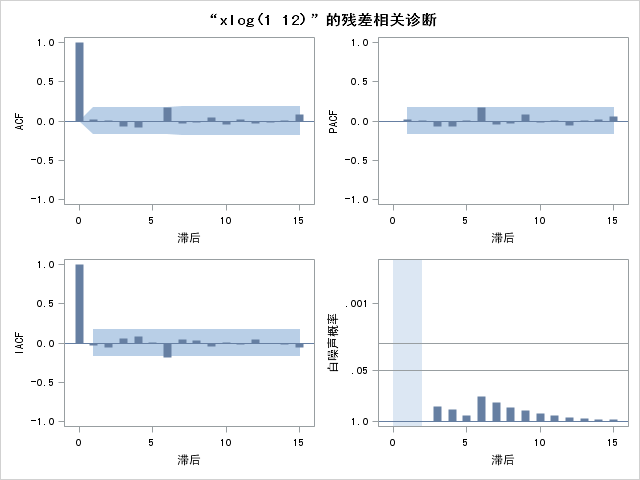
| **条件最小二乘估计** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **参数** | **估计值** | **标准误差** | **t 值** | **近似Pr > |t|** | **滞后** |
| **MA1,1** | 0.50275 | 0.07635 | 6.58 | <.0001 | 1 |
| **MA2,1** | 0.52993 | 0.09564 | 5.54 | <.0001 | 12 |

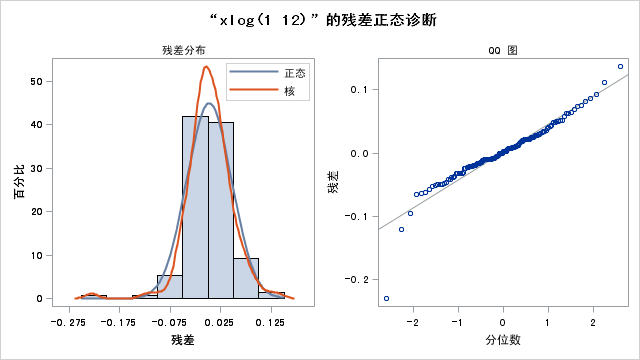
|  |  |
| --- | --- |
| **方差估计** | 0.001986 |
| **标准误差估计** | 0.044559 |
| **AIC** | -441.319 |
| **SBC** | -435.569 |
| **残差数** | 131 |

\* AIC 和 SBC 不包括对数行列式。

| **参数估计相关性** | | |
| --- | --- | --- |
| **参数** | **MA1,1** | **MA2,1** |
| **MA1,1** | 1.000 | -0.064 |
| **MA2,1** | -0.064 | 1.000 |

| **残差的自相关检查** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **至滞后** | **卡方** | **自由度** | **Pr > 卡方** | **自相关** | | | | | |
| **6** | 5.60 | 4 | 0.2312 | 0.018 | 0.007 | -0.068 | -0.074 | 0.003 | 0.173 |
| **12** | 6.54 | 10 | 0.7684 | -0.032 | -0.018 | 0.051 | -0.036 | 0.019 | -0.031 |
| **18** | 10.12 | 16 | 0.8604 | -0.014 | 0.013 | 0.082 | -0.126 | 0.023 | -0.018 |
| **24** | 15.41 | 22 | 0.8440 | -0.064 | -0.082 | 0.029 | -0.001 | 0.137 | 0.052 |





| **变量“xlog”的模型** | |
| --- | --- |
| **差分期间** | 1,12 |

在该模型中没有均值项。

| **移动平均因子** | |
| --- | --- |
| **因子 1:** | 1 - 0.50275 B\*\*(1) |
| **因子 2:** | 1 - 0.52993 B\*\*(12) |

参数估计的t值为6.58和5.54（较大），故要保留。拟合优度统计量表中给出了残差序列的方差、标准误差，以及按AIC和SBC标准计算的统计量，这两个值较小，表明对模型拟合得较好。

另外从-441.3195<-438.08143，-435.56911<-426.58065可看出本次拟合的ARIMA(0,1,1)×(0,1,1)12模型比上次拟合的模型要好。

估计值之间的相关系数为-0.064，这是一个较小的相关系数，如果这个相关系数较大时，就需要考虑是否删除其中一个参数。

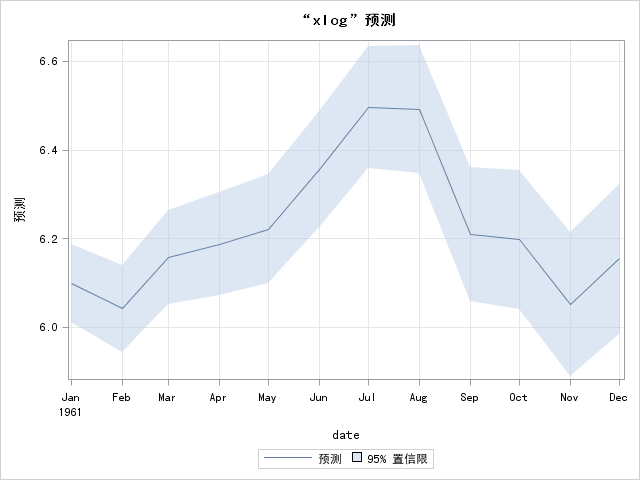
对模型拟合后残差为白噪声原假设进行检验，延迟6期和12期的QLB统计量分别为5.60和6.54，相应p值=0.2312和0.7684>α=0.05，故不能拒绝拟合模型的残差为白噪声，说明这个拟合模型，延迟数小于等于6期和12期的所有残差自相关系数为零，即残差中蕴涵信息已经完全被提取出来了。

最后将表中移动平均MA的两个因子（1-0.50275B）和（1-0.52993B12）代入ARIMA(0,1,1)×(0,1,1)12模型得到：



其中，Z=log(*xt*).

| **以下变量的预测:xlog** | | | | |
| --- | --- | --- | --- | --- |
| **观测** | **预测** | **标准误差** | **95% 置信限** | |
| **145** | 6.0987 | 0.0446 | 6.0114 | 6.1861 |
| **146** | 6.0415 | 0.0498 | 5.9439 | 6.1390 |
| **147** | 6.1573 | 0.0545 | 6.0505 | 6.2640 |
| **148** | 6.1881 | 0.0588 | 6.0728 | 6.3033 |
| **149** | 6.2216 | 0.0628 | 6.0984 | 6.3448 |
| **150** | 6.3570 | 0.0666 | 6.2264 | 6.4876 |
| **151** | 6.4967 | 0.0702 | 6.3590 | 6.6343 |
| **152** | 6.4917 | 0.0736 | 6.3474 | 6.6360 |
| **153** | 6.2095 | 0.0769 | 6.0588 | 6.3602 |
| **154** | 6.1979 | 0.0800 | 6.0410 | 6.3547 |
| **155** | 6.0515 | 0.0830 | 5.8887 | 6.2142 |
| **156** | 6.1556 | 0.0859 | 5.9872 | 6.3241 |



预测今后一年（12期）国际航线各个月度的旅客人数，结果存在数据集forxlog中：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **144** | DEC60 | 6.06843 | 6.05720 | 0.044559 | 5.96987 | 6.14453 | 0.01123 |
| **145** | JAN61 | . | 6.09873 | 0.044559 | 6.01139 | 6.18606 | . |
| **146** | FEB61 | . | 6.04148 | 0.049764 | 5.94395 | 6.13902 | . |
| **147** | MAR61 | . | 6.15726 | 0.054474 | 6.05049 | 6.26403 | . |
| **148** | APR61 | . | 6.18807 | 0.058808 | 6.07281 | 6.30333 | . |
| **149** | MAY61 | . | 6.22160 | 0.062843 | 6.09843 | 6.34477 | . |
| **150** | JUN61 | . | 6.35699 | 0.066635 | 6.22638 | 6.48759 | . |
| **151** | JUL61 | . | 6.49668 | 0.070222 | 6.35904 | 6.63431 | . |
| **152** | AUG61 | . | 6.49172 | 0.073635 | 6.34740 | 6.63604 | . |
| **153** | SEP61 | . | 6.20952 | 0.076896 | 6.05880 | 6.36023 | . |
| **154** | OCT61 | . | 6.19788 | 0.080025 | 6.04104 | 6.35473 | . |
| **155** | NOV61 | . | 6.05147 | 0.083035 | 5.88873 | 6.21422 | . |
| **156** | DEC61 | . | 6.15565 | 0.085941 | 5.98720 | 6.32409 | . |

最后，将xlog预测值转换为原来的*x*值，并绘制带上下95%置信度线的原始数据线和预测数据线的时间序列图。

代码：

**proc** **gplot** data=arimad04;

where date>=**'1jan57'd**;

plot x\*date forecast\*date l95\*date u95\*date /overlay vaxis=axis1 haxis=axis2 href=**'31dec60'd** ;

symbol1 i=join v=C h=**2.5** l=**1** font=swissb c=red;

symbol2 i=join v=F h=**3** l=**1** font=swissb c=blue;

symbol3 i=join l=**1** font=swissb c=green;

symbol4 i=join l=**1** font=swissb c=green;

axis1 label=('Passenger') order=(**250** to **800** by **50**);

axis2 label=('Month') order=(**'1jan57'd** to **'1jan62'd** by year);

format date monyy. ;

title1 'Forecast Chart';

title2 'C--x';

title3 'F--forecast';

title4 'None--u95 and l95';

**run**;

运行结果：

