浙 江 林 学 院 学 报 1997, **14**(4): 382~ 387
Journal of Zhejiang Forestry College

自回归模型的预测方差估计*

葛宏立 项小强 何时珍 顾金荣
(林业部华东调查规划设计院, 金华 321001)

摘 要 讨论了自回归模型的预测误差的方差估计问题, 给出了详细的递推计算公式, 并给出了平稳序列与非平稳序列的例子各 1 个。

关键词 自回归; 预测; 方差
中图分类号 **S758**

1 关于模型

自回归的预测误差的方差估计问题, 文献 [1] 与 [2] 都作过一些介绍, 但除了 [2] 给出了一阶自回归的具体计算式子外, 没有给出任意阶的可供计算的具体式子及具体步骤。本文给出多步预测的方差递推计算式子。本文只讨论标量模型。设带输入项的自回归模型为:

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + b_0 x_t + b_1 x_{t-1} + \cdots + b_q x_{t-q} + X \qquad (1)$$

序列 $\{y_t\}$ 可以是平稳的, 也可以是非平稳的, $\{x_t\}$ 为确定性的输入变量; X为白噪声, $E(X^2) = e$(常数), $E(X \cdot X) = 0 (i \neq j)$; $a_i (i = 0, 1, \cdots, p)$ 和 $b_i (i = 0, 1, \cdots, q)$ 为已知常数; p为模型的阶。在以下的讨论中, 假定参数是真值而非估计值, 即假定它们的方差与协方差均为 0。

这里不限制序列 $\{y_t\}$ 为平稳的, 是因为在实际应用中, 对于具有趋势性的非平稳序列, 直接用自回归模型拟合, 建立预测模型是可行的。其优点是简便实用。它保留了原始观察数据所包含的信息, 如果进行差分, 会带来一定的信息损失。但是由于模型是非平稳的, 所得到的预测曲线是发散的, 所以只适宜进行短期预测, 用以进行长期预测则效果不佳[1, 3]。

为简单起见, 令 $X_t = \sum_{j=0}^{q} b_j x_{t-j}$, 则 (1) 可改写为

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + X_t + X \qquad (2)$$

2 计算式子推导

设目前时间为 $t-1$, 即目前有观测值 $y_{t-1}, y_{t-2}, \cdots, y_{t-p}, y_t, y_{t+1} \cdots$, 未知。设 $r+1$为向前

收稿日期: 1997-02-21
* 联合国发展规划署资助项目
第 1作者简介: 葛宏立, 男, 1960年生, 工程师, 硕士。

预测的步数,即从目前时间 $t-1$到 $t+r$, $r=0,1,2,\cdots$。用 $y_{t-1},y_{t-2},\cdots,y_{t-p}$预测 y_{t+r},其误差是 $X_t,X_{t+1},\cdots,X_{t+r}$的某种积累,显然它是 $X_t,X_{t+1},\cdots,X_{t+r}$的线性组合。设

$$y_{t+r}=c_{0r}+c_{1r}y_{t-1}+c_{2r}y_{t-2}+\cdots+c_{pr}y_{t-p}+U_r+(d_0,d_1,d_2,\cdots,d_r)(X_{t+r},X_{t+r},\cdots,X_{t+1},X_t)^T \tag{3}$$

$c_{0r},c_{1r},\cdots,c_{pr}$为可通过原模型的参数计算得到的参数, U_r为输入变量 $X_t,X_{t+1},\cdots,X_{t+r}$的组合, $d_0,d_1,\cdots,d_r$也为可计算得到的参数 据(3)有

$$\begin{cases} y_{t+r-m}=c_{0(r-m)}+c_{1(r-m)}y_{t-1}+c_{2(r-m)}y_{t-2}+\cdots+c_{p(r-m)}y_{t-p}+U_{r-m}+(d_0,d_1,\cdots,d_{r-m})(X_{t+r-m},X_{t+r-m-1},\cdots,X_{t+1},X_t)^T \\ m=1,2,\cdots,r \end{cases} \tag{4}$$

而据(2)有

$$y_{t+r}=a_0+a_1y_{t+r-1}+a_2y_{t+r-2}\cdots+\cdots+a_py_{t+r-p}+X_{t+r}+X_{t+r} \tag{5}$$

将(4)代入(5),经整理,可得下列递推式子:

$$\begin{cases} c_{0r}=a_0+a_1c_{0(r-1)}+a_2c_{0(r-2)}+\cdots+a_pc_{0(r-p)} \\ c_{0(-1)}=c_{0(-2)}=\cdots=c_{0(-p)}=0 \end{cases} \tag{6}$$

$$\begin{cases} a_{1r}=a_1c_{1(r-1)}+a_2c_{1(r-2)}+\cdots+a_pc_{1(r-p)} \\ a_{2r}=a_1c_{2(r-1)}+a_2c_{2(r-2)}+\cdots+a_pc_{2(r-p)} \\ \cdots \\ c_{pr}=a_1c_{p(r-1)}+a_2c_{p(r-2)}+\cdots+a_pc_{p(r-p)} \\ a_{(-j)}=\begin{cases} 1 & \text{当 } i=j,i,j=1,2,\cdots,p \\ 0 & \text{其他} \end{cases} \end{cases} \tag{7}$$

$$\begin{cases} d_r=a_1d_{r-1}+a_2d_{r-2}+\cdots+a_pd_{r-p} \\ d_0=1,d_{-1}=d_{-2}=\cdots=d_{-p}=0 \end{cases} \tag{8}$$

$$\begin{cases} U_r=a_1U_{r-1}+a_2U_{r-2}+\cdots+a_pU_{r-p}+X_{t+r} \\ U_{-1}=U_{-2}=\cdots+U_{-p}=0 \end{cases} \tag{9}$$

(9)也可写成

$$U_r=(d_0,d_1,d_2,\cdots,d_r)(X_{t+r},X_{t+r-1},\cdots,X_t)^T \tag{9}$$

其中 $d_i(i=0,1,\cdots,r)$由(8)式确定。

令

$$e_r=(d_0,d_1,\cdots,d_r)(X_{t+r},X_{t+r-1},\cdots,X_{t+1},X_t)^T$$

则(3)可写成

$$\hat{y}_{t+r}=c_{0r}+c_{1r}y_{t-1}+c_{2r}y_{t-2}+\cdots+c_{pr}y_{t-p}+U_r+e_r \tag{10}$$

y_{t+r}的估计值 y_{t+r}为

$$\hat{y}_{t+r}=c_{0r}+c_{1r}y_{t-1}+c_{2r}y_{t-2}+\cdots+c_{pr}y_{t-p}+U_r \tag{11}$$

$\hat{y}_{t+r}$作为 y_{t+r}的估计值,其方差为 $D(\hat{y}_{t+r}-y_{t+r})$,并根据参数真值已知的假定,有

$$D(\hat{y}_{t+r}-y_{t+r})=D(e_r)=(d_0^2+d_1^2+\cdots+d_r^2)\sigma^2 \tag{12}$$

(12)式即为向前预测 $r+1$步时的误差的方差的估计式,它与(8)式一起组成方差的递推计算公式 要指出的是,在实际中,模型参数一般是估计值,带有误差,尤其在小样本情况下,所以实

际预测误差的方差要比用 (12) 式算得的数值大[2]。y_{n+r}的估计区间为

$$\hat{y}_{n+r} \pm U_\alpha s_r \tag{13}$$

其中 s_r 为标准差

$$s_r = (d_0^2 + d_1^2 + \cdots + d_r^2)^{1/2} e \tag{14}$$

u_α 为正态分布的双侧分位数。

3 例子

原始数据引自文献 [4], 是一株紫果云杉解析木的树高 (H) 生长数据。原始数据见表 1中的第 2、3列。例子分为平稳序列与非平稳序列。

表 1 例 1拟合数据范围内的计算结果

Table 1 Result within the fitting data range for example 1

t	年龄	H_t	$\hat{H}_{n+0}$	X	r	$\hat{H}_{n+r}$	e_r	s_r
1	10	1.41						
2	20	2.49						
3	30	3.50	3.506 3	-0.006 3	0	3.506 3	-0.006 3	1.134 4
4	40	4.50	4.444 6	0.055 4	1	4.456 9	0.043 1	2.483 6
5	50	5.57	5.430 0	0.140 0	2	5.340 1	0.229 9	4.068 4
6	60	6.90	6.561 4	0.338 6	3	6.154 6	0.745 4	5.826 6
7	70	10.30	8.132 4	2.167 6	4	6.899 7	3.400 3	7.713 1
8	80	15.30	13.487 2	1.812 8	5	7.574 9	7.725 1	9.692 6
9	90	19.70	19.986 3	-0.286 3	6	8.180 1	11.519 9	11.736 1
10	100	23.30	23.792 7	-0.492 7	7	8.715 5	14.584 5	13.819 2
11	110	27.96	26.612 5	1.347 5	8	9.181 9	18.778 1	15.921 2
12	120	31.30	32.258 9	-0.958 9	9	9.580 1	21.719 9	18.023 8
13	130	34.30	34.324 7	-0.024 7	10	9.911 4	24.388 6	20.111 1
14	140	36.50	36.985 8	-0.485 8	11	10.177 3	26.322 7	22.169 1
15	150	38.18	38.412 7	-0.232 7	12	10.379 5	27.800 5	24.185 8
16	160	41.39	39.588 9	1.801 1	13	10.520 1	30.869 9	26.150 4
17	170	41.85	44.240 3	-2.390 3	14	10.601 3	31.248 7	28.053 7
18	180	42.31	42.078 3	0.231 7	15	10.625 4	31.684 6	29.887 8
19	190	42.77	42.536 0	0.234 0	16	40.595 1	32.174 9	31.646 1
20	200	43.22	42.993 7	0.226 3	17	10.513 1	32.706 9	33.323 1

3.1 非平稳序列的例子 (例 1)

树高值明显随年龄的增大而增大,所以树高序列是一典型的非平稳序列。设其模型为

$$H_t = a_1 H_{t-1} + a_2 H_{t-2} + X \tag{15}$$

即 $p = 2$, 不带输入项, 不带常数项。拟合结果为 $a_1 = 1.947\ 598, a_2 = -0.952\ 614$, $e = 1.134\ 410$。其他计算结果见表 1和表 2。表 1是为了与实测值比较,在拟合数据范围内进行计算。$\hat{H}_{n+0}$一列是令 $r = 0$进行预测,即每次预测时,都用实测值作为自变量,而不用预测值作为自变量,其值与实测值拟合得很好,X为其相应的误差。$\hat{H}_{n+r}$则是令 r从 0一直到 17,只根据 $H_1 = 1.41$和 $H_2 = 2.49$这 2个实测值一直预测到 $\hat{H}_{20}$, e_r为其相应的误差, s_r为其相应的标准差。很明显, e_r与 s_r随 r的增大而迅速增大。对于 (15) 式而言,最初几个 d值为

$$\begin{cases} d_0 = 1 \\ d_1 = a_1 \\ d_2 = a_1^2 - a_2 \\ d_3 = a_1^3 + 2a_1a_2 \\ d_4 = a_1^4 + 3a_1^2a_2 + a_2^2 \end{cases}$$

为了观察拟合数据范围之外,其预测值的变化情况,令 r 从 0到 17,根据最后 2个实测值 $H_{19} = 42.77$和 $H_{20} = 43.22$,一直预测到 $\hat{H}_{38}(\hat{H}_{21+17})$,结果见表 2 误差 e_r 无法计算,标准差 s_r同表 1,因为 s_r 只与 r 有关,而与起点无关。从 $\hat{H}_{21+r}$可看出,预测值愈来愈偏离客观实际,因为预测值越来越小。这显然与树高随年龄增大而增大的规律相违背。但如取 $u_{\alpha} = 2$,根据 (13)式算得的估计区间,根据树高的生长规律,实际值还不会跑出这个区间。由此可看出,这样计算的标准差还是符合实际的。

表 2　例 1拟合数据范围外的计算结果

Table 2　Result without the fitting data range for example 1

r	0	1	2	3	4	5	6	7	8
$21+r$	21	22	23	24	25	26	27	28	29
$\hat{H}_{21+r}$	43.431 9	43.415 9	43.182 9	42.744 2	42.112 0	41.298 5	40.316 4	39.178 7	37.898 3
r	9	10	11	12	13	14	15	16	17
$21+r$	30	31	32	33	34	35	36	37	38
$\hat{H}_{21+r}$	36.488 4	34.962 4	33.333 3	31.614 1	29.817 8	27.957 2	26.044 5	24.091 8	22.110 7

3.2　平稳序列的例子 (例 **2**)

据 [5],将树高值减去其趋势项,得平稳序列,即

$$y_t = H_t - H(t) \tag{16}$$

这里取 $H(t)$ 为

$$H(t) = c_0(1 - e^{-c_1(t-10)c_2}) \tag{17}$$

$c_0 = 47.779\,80, c_1 = 0.021\,661\,1, c_2 = 5.663\,95, t$· 10即为年龄。新的序列 y_t 显然是平稳的,每处的期望值均为 0,与例 1一样,分拟合数据范围内与拟合数据范围外 2种情况进行计算。设其模型同 (15),即

$$y_t = a_1y_{t-1} + a_2y_{t-2} + X \tag{18}$$

$a_1 = 1.016\,905, a_2 = -0.414\,086, e = 0.700\,84$ 拟合范围内的计算结果见表 3。

$\hat{y}_{t+0}$ 与表 1中的 $\hat{H}_{t+0}$ 的算法相同,$\hat{y}_{3+r}$ 与 $\hat{H}_{3+r}$ 的算法相同,$\hat{y}_{3+r}$ 的绝对值呈下降的趋势,s_r 呈上升趋势。拟合数据范围外的计算结果见表 4 $\hat{y}_{21+r}$ 的绝对值也呈下降趋势。当序列平稳时,当 r 趋向无穷时,预测值趋向于序列的期望值 [2]。本例 y_t 的期望值为 0,所以预测值趋向于 0。

表 3 例 2拟合数据范围内的计算结果

Table 3 Result within the fitting data range for example 2

t	y_t	$\hat{y}_{t-1}(1)$	X_t	r	$\hat{y}_{3+r}$	e_r	s_r
1	1.405 5						
2	2.361 8						
3	2.770 8	1.819 7	0.951 1	0	1.819 7	0.951 1	0.700 8
4	2.325 2	1.839 6	0.485 5	1	0.872 5	1.452 7	0.999 6
5	0.972 6	1.217 1	−0.244 5	2	0.133 7	0.838 9	1.089 9
6	−0.975 1	0.026 2	−1.001 4	3	−0.225 3	−0.749 8	1.099 8
7	−1.437 2	−1.394 4	−0.042 8	4	−0.284 5	−1.152 7	1.100 2
8	−0.576 4	−1.057 7	0.481 3	5	−0.196 0	−0.380 4	1.104 0
9	−0.323 0	0.009 0	−0.332 0	6	−0.081 5	−0.241 5	1.106 9
10	−0.676 3	−0.089 8	−0.586 5	7	−0.001 7	−0.674 6	1.107 9
11	0.352 0	−0.554 0	0.906 0	8	0.032 0	0.320 0	1.107 9
12	0.448 8	0.638 0	−0.189 2	9	0.033 3	0.415 5	1.107 9
13	0.614 8	0.310 6	0.304 2	10	0.020 6	0.594 3	1.108 0
14	0.380 0	0.439 4	−0.059 4	11	0.007 1	0.372 8	1.108 0
15	−0.004 3	0.131 8	−0.136 2	12	−0.001 3	−0.003 1	1.108 0
16	1.473 6	−0.161 8	1.635 4	13	−0.004 2	1.477 8	1.108 0
17	0.492 3	1.500 3	−1.008 0	14	−0.003 8	0.496 1	1.108 0
18	−0.239 2	−0.109 6	−0.129 6	15	−0.002 1	−0.237 1	1.108 0
19	−0.759 0	−0.447 1	−0.312 0	16	−0.000 6	−0.758 4	1.108 0
20	−1.111 5	−0.672 8	−0.438 7	17	0.000 3	−1.111 8	1.108 0

表 4 例 2拟合数据范围外的计算结果

Table 4 Result without the fitting data range for example 2

r	0	1	2	3	4	5	6	7	8
$21+r$	21	22	23	24	25	26	27	28	29
$\hat{y}_{21+r}$	−0.816 0	−0.369 5	−0.037 9	0.114 5	0.132 1	0.086 9	0.033 7	−0.001 7	−0.015 7
r	9	10	11	12	13	14	15	16	17
$21+r$	30	31	32	33	34	35	36	37	38
$\hat{y}_{21+r}$	−0.015 3	−0.009 0	−0.002 8	0.000 8	0.002 0	0.001 7	0.000 9	0.000 2	−0.000 2

从这 2个例子可以进一步看出,当预测时期较长时,对于平稳序列来说,预测值趋向于期望值,太长期的预测将失去意义,但预测的结果还不至于出现太荒唐的情况,而对于非平稳序列来说,长期预测时,还可能出现荒唐的结果。无论哪种情况,方差总随预测步数的增大而单调增大,至于是否有界,有待进一步研究。总之,对于自回归模型来说,不能作太长期的预测,尤其是非平稳序列的自回归模型。可以根据计算的标准差 s_r 计算预测精度,当预测精度小到一定程度时,就该停止预测。另外还要指出,对于平稳序列来说,若其期望 $E(y_t) = \mu \neq 0$,则模型中不应省略常数项,若省略了常数项,则应限制 $a_1 + a_2 + \cdots + a_p = 1$,因为

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + X_t$$

两边取期望,有

$$\_ = a_0 + a_1\_ + \cdots + a_p\_$$

即

$$- = \frac{a_0}{1 - (a_1 + \cdots + a_p)}$$

若省去常数项 a_0,相当于 $a_0 = 0$,而又不限制 $a_1 + \cdots + a_p = 1$,这时有 $\_ = 0$,这与序列的期望不为 0的事实矛盾。

参 考 文 献

1 陈玉祥, 张汉亚. 预测技术与应用. 北京: 机械工业出版社, 1985. 124~ 129
2 李卓立. 实用经济计量学. 北京: 清华大学出版社, 1987. 158~ 160
3 复旦大学. 概率论: 第 3册—— 随机过程. 北京: 人民教育出版社, 1985. 340~ 343
4 北京林业大学, 测树学. 北京: 中国林业出版社, 1987. 90
5 朱明德, 佘光辉. 统计预测与控制. 北京: 中国林业出版社, 1993

Ge Hongli (East China Forest Inventory Institute, jinhua 321001, Zhejiang, PRC), Xiang Xiaoqiang, He Shizhen and Gu Jinrong. **Variances Estimation of A Sequence Forecasted by An Autoregression Model**. *J Zhejiang For Coll*, 1997, **14** (4): 382~ 387

Abstract This paper introduces a recurrence formula for estimating the variances of errors of a sequence forecasted by an autoregression model and gives two examples under both smooth and nonsmooth conditions.

Key words autoregression; forecasting; variance