文章编号: 1007-4619 (2003) 04-0245-06

SARS 疫情预测预报中的分段非线性回归方法

崔恒建[1], 李仲来[1], 杨 华[2], 李小文[2,3]
(1. 北京师范大学 数学系, 数据统计与分析中心, 北京 100875; 2. 北京师范大学 遥感与 GIS 研究中心, 北京 100875;
3. 中国科学院 遥感应用研究所, 北京 100101)

摘 要: 介绍了几种对累计 SARS 疫情预测预报中的非线性增长曲线模型, 说明了 Richards 增长曲线在这次 SARS 疫情预测预报中合理性和可行性, 由此建立了累计 SARS 疫情预测预报中的非线性回归点模型。并具体对北京 SARS 疫情进行了跟踪预测预报, 包括整体和分时间段的预测预报, 获得了北京 SARS 疫情随时间的预测预报结果, 说明了北京 4 月底的一系列控制措施对北京 SARS 疫情所带来的影响, 为进一步的后续研究打下了良好基础。

关键词: SARS; 增长曲线; 非线性回归点模型; Richards 曲线; 分段拟合
中图分类号: R181.8/O21 文献标识码: A

1 引 言

SARS(Severe Acute Respiratory Syndrome, 简称非典) 是一种通过近距离飞沫、粪便、接触病人分泌物等途径传染, 具有很强传染性的严重急性呼吸道疾病, 它给人民的身体健康, 生命安全带来严重威胁。这是一种新的突发传染病, 人类对它的防治还处于初步摸索阶段。为了进一步认识这一疾病, 弄清 SARS 疫情目前在中国大陆范围内的形势, 特别是首都北京的 SARS 疫情的趋势, 克服人们的恐惧心理, 采取客观冷静的态度, 科学地进行防治, 我们对北京 SARS 疫情数据进行了统计分析, 建立了非线性回归模型的整体和分时间段的预测预报, 获得良好效果 (所用数据来自中国卫生部网站[1] 和世界卫生组织网站(WHO)[2])。

2 SARS 预测预报非线性回归模型及其机理

自 SARS 从中国广东省、香港向北京及周边地区蔓延时, 我们对世界和中国 SARS 疫情的数据进行统计分析, 并通过分别观察加拿大、新加坡、中国香港、广东的累计病例的走势, 首次提出了用累积增长的"S"形曲线去拟合和预测 SARS 累计病例, 新增病例的走势, 由于人们已认识 SARS 是一种传染性很强的传染病, 且具有聚集性传染的特点, 随着人们的日常交往的频繁, 假设 $A=A(t_0)$ 为从时刻 t_0 开始以后的最终感染总人数, $I(t_0)=I_0$ 和 $I(t)$ 分别为 t_0 和 $t(t\geqslant t_0)$ 时刻的累积病例数。令 $y=[I(t)-I(t_0)]/A$, 则 y 满足如下增长率方程:

$$\frac{\mathrm{dlog}(y)}{\mathrm{d}t}=f(y,\ t-t_0) \tag{1}$$

其中 $f(y, t)$ 是一形式已知的关于 $0<y<1$, t 的连续函数。常见的 $f(y, t)$ 有如下两类形式。第一类为 Turner 型:

$$f(y,\ t;p,\ m,\ \lambda)=\lambda[1-y^m]^{1-p}[y^{-m}-1]^p$$

其中 λ, $m>0$(可参见文献[3], [4])。当在 Turner 型中取 $p=0$, $m=1$, 则有 $\mathrm{dlog}(y)/\mathrm{d}t=\lambda[1-y]$,

$$I(t)\approx I(t_0)+A/[1+\exp\{-a(t-t_0)+b\}] \tag{2}$$

即为人们熟知的 Logistic 增长模型。当 $p=1$, $m>0$ 时, 则有 $\mathrm{dlog}(y)/\mathrm{d}t=\lambda[y^{-m}-1]$, 即为 Richards 增长方程, 它有如下形式的解:

$$I(t)=I(t_0)+A[1-\exp\{-K(t-t_0)\}]^B。 \tag{3}$$

第二类为 GAMMA-型:

收稿日期: 2003-06-05; 修订日期: 2003-06-07
基金项目: 国家自然科学基金主任基金项目"SARS 传播时空模型研究"(40341002) 和 863 计划课题"SARS 流行病学资料的实时收集、分析和趋势预测"(2003AA208401) 资助。

作者简介: 崔恒建(1963—), 男, 北京师范大学数学系教授, 博士导师, 现任中国概率统计学会常务理事, 从事统计学理论和应用方面研究, 发表论文 50 余篇。

$$f(y,\ t;\ \alpha,\ \beta,\ \lambda)=\frac{\lambda\beta^{\alpha}y^{-1}}{\Gamma(\alpha)}t^{\alpha-1}\exp\{-\beta t\},$$

对应于(1), 它有解:

$$I(t)=I(t_0)+A\int_0^{t-t_0}\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}\exp\{-\beta x\}\,dx. \tag{4}$$

由于 Richards 增长曲线(3)为一显式解, 且具有精确的初始条件: $t=t_0$ 时, $I(t)=I(t_0)$, 具有整体以及分段拟合与预测的灵活性, 参数估计可借助线性回归方法获得(参见第 3 节)。不仅如此, 它还满足"负反馈"微分方程

$$\frac{dI}{dt}=aI^{-m+1}-bI \tag{5}$$

其中 $a>b>0$, 且满足 $A=(a/b)^{1/m}$, $B=1/m$, $K=bm$。由于SARS 在开始时, 具有聚集性传染, 来势凶猛, 人们对它认识不清楚, 与其它通常传染病有所不同的特点, 因而认为它的累计病例 I 是以指数速度自然增长的, 即 aI^{-m+1}(m 为幂指数), 随着疫情的发展, 人们对它有所认识, 病死、病愈、抗体人群的产生以及采取的控制措施, 作用速度认为是 $-b$, 使得累计病例有所减缓, 因而它的累计病例数应近似遵循速率方程(5)。鉴于(3)式的上述优点, 我们建议利用(5)式导出的非线性增长曲线(3)式作为准均匀介质下的非线性回归方程来拟合SARS 的累积病例数据。

3 非线性回归模型参数的估计方法

令 K 为迭代参数, 对(3)式做如下变换:

$$I'=\log(I-I_0),\ A'=\log(A),$$
$$t'=\log(1-\exp\{-K(t-t_0)\}),$$

则(3)式化为: $I'=A'+Bt'$。当我们获得 n 个数据时, 其非线性回归模型如下:

$$I'_i=A'+Bt_i'+\varepsilon_i \tag{6}$$

其中,

$$I_i'=\log(I_i-I_0),$$
$$t_i'=\log(1-\exp\{-K(t_i-t_0)\}),$$

ε_i 为误差($1\leqslant i\leqslant n$)。这样, 我们可利用线性模型回归中的最小二乘法, 获得参数 A', B 的最小二乘估计, 然后将 $A=\exp\{A'\}$, B, K 的最小二乘估计视为A, B, K 的初值, 带回到方程(3), 使得非线性回归残差平方和

$$\sum_{i=1}^{n}(I_i-I_0-A[1-\exp\{-K(t-t_0)\}]^B)^2=\min$$

而最后获得参数估计 A, B, K, 这里 t_0, I_0 给定。

由于从(5)导出的非线性 Richards 增长曲线(3), 具有模型简单、机理清楚、算法简明快速, 并已对中国香港、新加坡、加拿大获得了成功应用, 我们继续利用此模型对北京以及周边地区(如山西、内蒙古自治区)、广州等地进行预测预报, 包括动态的即时预报(一周以内)以及中长期预报, 给出新发病人数、累积病例、疑似病例等的预报以及预报误差范围。

4 北京 SARS 疫情数据的整体和分时间段的预测预报

4.1 整体 SARS 疫情预测

依据北京的SARS 疫情数据, 可以计算模型(3)中的参数 A, B, K。北京的模型参数与预测见表 1。可以看出参数 A, B, K 随时间的变化逐渐趋于稳定。特别是 5 月 1 日前后是北京SARS 的暴发期, 这期间我们预测北京 SARS 累计病例的上限是 3300 左右, 对消除当时人们的恐惧心理为政府部门制定进一步的相关决策是至关重要的, 即使从现在来看, 当时的这一预测也是合理可行的。

4.2 北京疫情的分时间段拟合与预测

在(3)式中, A 的物理意义是最明确的, 为疫情稳定后(除 I_0 外)总发病人数。K 基本上标志控制措施生效的强度, K 越大, 疫情越快稳定下来。由于控制措施在疫情发展过程中是变数, 用疫情发展过程中所有数据来拟合(3)式, 发现预测结果总体略偏向于保守, 我们的创新点在于, 根据采取控制措施的时间, 分段拟合出不同的参数, 用于评估措施的有效性, 并预测最终的累积发病人数。如北京在 4 月 20 日左右, 采取了一些有力措施, 拟合了 3 月 5 日到 4 月 20 日, 21 日, 22 日, 23 日的数据, 发现参数基本上是稳定的。4 月 24 日以后, 参数有明显变化。这说明 4 月 20 日左右的措施, 滞后了大约 4 天, 开始影响疫情, 此前我们利用 3 月 5 日至 4 月 23 日数据进行拟合, 结果及参数见表 2 和图 1。换言之, 没有 4 月 20 日及以后的措施, 北京累计将有 4134 例确诊非典患者。由于 4 月下旬的一系列措施, 5 月 5 日长假以后模型三参数逐渐稳定, 这期间 A 的估计区间为2517-2533, 即 4 月 20 日及以后的措施明显地减少了可能的累计发病人数(净减约 1600 人), 加速了疫情的控制。对 5 月 5 日至 6 月 5 日数据进行拟合, 结果及参数见表 2 和图 2。但由于这一系列措施时间上比较集中, 发生效果的时延不一, 不易从目前的数据中分析出来, 我们可用"系统动力学模型"

表 1　北京整体 SARS 疫情拟合预测结果及参数变化

Table 1　The forecasting results of Beijing SARS situation based on consecutive time

日期	*A*	置信区间(1σ)	*B*	*K*	次日新增病例预测(实测)	10d 新增病例预测(实测)	标准差
2003-05-01	3345.9	(3329.9, 3361.9)	109.42	0.086179	102(113)	69(38)	16.02
2003-05-02	3260.3	(3244.9, 3275.7)	117.08	0.087889	98(83)	63(39)	15.65
2003-05-03	3278.3	(3264.4, 3292.3)	115.29	0.087504	95(105)	60(43)	15.23
2003-05-04	3109.9	(3093.8, 3125.9)	135.48	0.091485	87(62)	51(27)	15.73
2003-05-05	3068.5	(3053.2, 3083.8)	141.82	0.092598	82(94)	46(18)	15.45
2003-05-06	2990.3	(2975.1, 3005.5)	156.16	0.094912	76(63)	40(17)	15.64
2003-05-07	2986.4	(2972.4, 3000.4)	156.98	0.095036	71(89)	37(15)	15.30
2003-05-08	3018.2	(3004, 3032.3)	150.17	0.09399	67(87)	35(14)	15.17
2003-05-09	2986.4	(2972.4, 3000.4)	157.33	0.09508	62(41)	31(3)	15.13
2003-05-10	2950.3	(2936.2, 2964.4)	166.64	0.096413	57(50)	27(8)	15.34
2003-05-11	2906.5	(2891.8, 2921.2)	179.87	0.098168	51(38)	24(0)	16.10
2003-05-12	2868.4	(2852.9, 2883.9)	193.49	0.099829	46(39)	21(12)	16.91
2003-05-13	2842.6	(2826.4, 2858.8)	204.2	0.10105	41(43)	18(11)	17.30
2003-05-14	2812.9	(2796.1, 2829.7)	218.35	0.10255	37(27)	16(25)	18.22
2003-05-15	2782.2	(2764.4, 2799.9)	235.43	0.10422	33(18)	14(9)	19.60
2003-05-16	2753.7	(2733.6, 2773.8)	253.92	0.10589	29(17)	12(5)	21.11
2003-05-17	2728.3	(2705.6, 2751)	273.16	0.10749	25(15)	10(8)	22.61
2003-05-18	2706.3	(2681.3, 2731.4)	292.38	0.10897	22(14)	8(3)	23.96
2003-05-19	2684.3	(2657.1, 2711.6)	314.53	0.11055	19(3)	7(3)	25.69
2003-05-20	2664.8	(2635.6, 2694.1)	337.09	0.11205	16(8)	6(1)	27.31
2003-05-21	2646.1	(2614.9, 2677.4)	361.81	0.11357	14(0)	5(1)	29.12
2003-05-22	2631.5	(2599, 2664)	383.72	0.11482	12(12)	4(1)	30.32
2003-05-23	2619.5	(2586.2, 2652.9)	403.7	0.1159	11(11)	4(0)	31.20
2003-05-24	2612.7	(2580, 2645.4)	416.12	0.11654	10(25)	3(0)	31.30
2003-05-25	2607.1	(2575, 2639.2)	426.95	0.11709	8(9)	3(0)	31.31
2003-05-26	2602.1	(2570.7, 2633.5)	437.38	0.1176	7(5)	2(0)	31.30
2003-05-27	2598	(2567.5, 2628.6)	446.25	0.11802	7(8)	2(0)	31.22
2003-05-28	2594.1	(2564.4, 2623.8)	455.33	0.11844	5(2)	2	31.17
2003-05-29	2590.5	(2561.5, 2619.4)	464.1	0.11884	5(3)	1	31.12
2003-05-30	2587.2	(2559, 2615.3)	472.42	0.11921	4(3)	1	31.05
2003-05-31	2584	(2556.7, 2611.4)	480.74	0.11957	3(1)	1	31.01
2003-06-01	2581.1	(2554.4, 2607.7)	488.93	0.11992	3(1)	1	30.98
2003-06-02	2578.2	(2552.2, 2604.2)	497.16	0.12027	3(0)	1	30.97
2003-06-03	2575.5	(2550.1, 2600.9)	505.33	0.12061	2(0)	0	30.97
2003-06-04	2572.9	(2548.1, 2597.7)	513.36	0.12093	2(0)	0	30.98
2003-06-05	2570.5	(2546.2, 2594.7)	521.2	0.12125	2(0)	0	30.99

表 2　数据分段拟合结果及参数

Table 2　Piecewise time fitting results and model parameters

时间段	*A*	*B*	*K*	*b*	标准差
3月 5 日至 4 月 23 日	4134	108.12	0.08367	9.05	11.0
5月 5 日至 6月 5 日	2524	5623.5	0.1608	904.3	10.3

解决。参数 $K \cdot B = b$ 是表征控制力度的一个量。可以概略理解为每天有多少病源被切断了进一步传染的通径(死亡、隔离、治愈)。表 3 则是利用 5 月 5 日至 6 月 5 日的数据进行拟合的结果, 其预测预报效果比整体预测预报误差小。

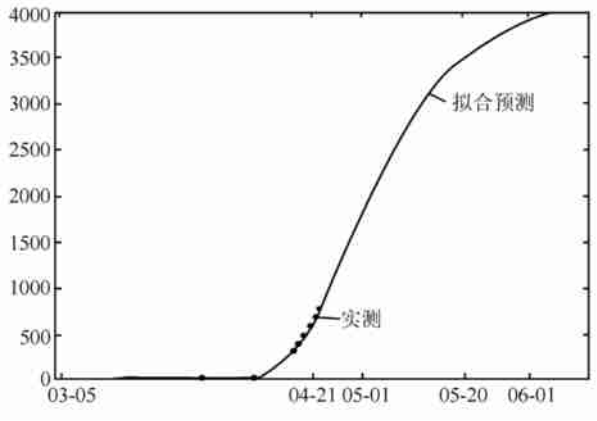

图 1 北京 SARS 病例累计拟合(2003-03-05—2003-06-05)

Fig. 1 Piecewise time fitting and forecasting of Beijing cumulative SARS patients during Mar. 5—Apr. 23, 2003

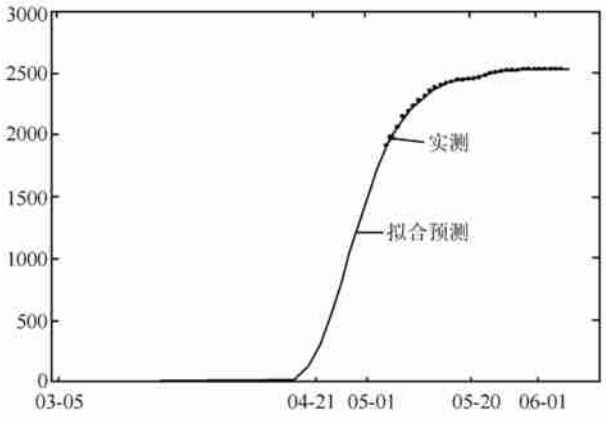

图 2 北京 SARS 分段拟合预测图(2003-03-05—2003-06-05)

Fig. 2 Piecewise time fitting and forecasting of Beijing cumulative SARS patients during Mar. 5—Jun. 5, 2003

表 3 北京 **2003-05-05** 后 **SARS** 疫情分段拟合预测结果及参数变化

Tbale 3 The forecasting results of Beijing SARS situation based on piecewise time fitting and model parameters after May, 5, 2003

日期	*A*	置信区间	*B*	*K*	次日新增病例预测(实测)	10d 新增病例预测(实测)	标准差
2003-05-10	2690.2	(2682, 2698.3)	922.09	0.128	50(50)	17(8)	9.65
2003-05-11	2635.7	(2628, 2643.3)	1264.8	0.13421	42(38)	13(0)	10.18
2003-05-12	2593.2	(2585.7, 2600.7)	1995	0.14244	36(39)	11(12)	9.74
2003-05-13	2616.8	(2609.5, 2624.1)	1452.6	0.13683	33(43)	11(11)	9.62
2003-05-14	2583.3	(2576.7, 2589.9)	2298.9	0.14495	27(27)	8(25)	8.88
2003-05-15	2555.8	(2549.4, 2562.1)	3727.1	0.15337	22(18)	6(9)	8.49
2003-05-16	2540.3	(2533.9, 2546.7)	5195.9	0.15904	18(17)	5(5)	8.21
2003-05-17	2522.2	(2516.3, 2528.1)	8399.7	0.16716	15(15)	4(8)	7.65
2003-05-18	2525.2	(2519.2, 2531.1)	7083.2	0.16439	13(14)	3(2)	7.80
2003-05-19	2512.4	(2506.9, 2518)	10150	0.1705	10(3)	2(3)	7.69
2003-05-20	2506.1	(2500.3, 2511.9)	12905	0.17451	8(8)	2(3)	7.69
2003-05-21	2500.3	(2493.4, 2507.3)	13994	0.17602	6(0)	2(1)	8.30
2003-05-22	2490.5	(2484, 2497.1)	23054	0.18418	5(12)	1(1)	7.56
2003-05-23	2491.4	(2485.2, 2497.6)	22844	0.18398	4(11)	1(0)	7.38
2003-05-24	2494.1	(2486.7, 2501.5)	21517	0.18294	4(25)	1(0)	8.28
2003-05-25	2499.9	(2488.9, 2510.9)	16752	0.17879	3(9)	1(0)	9.21
2003-05-26	2504.7	(2490.9, 2518.6)	13582	0.17532	3(5)	1(0)	9.82
2003-05-27	2509.6	(2494.7, 2524.5)	11008	0.17184	3(8)	1	10.46
2003-05-28	2513.3	(2498.9, 2527.6)	9391.6	0.16922	2(2)	1	10.78
2003-05-29	2516.3	(2502.9, 2529.7)	8237.1	0.16705	2(3)	0	10.96
2003-05-30	2518.8	(2506.5, 2531.2)	7362.4	0.1652	2(3)	0	11.07
2003-05-31	2520.8	(2509.4, 2532.1)	6761.7	0.1638	1(1)	0	11.07
2003-06-01	2522.2	(2511.8, 2532.7)	6331.3	0.16272	1(1)	0	11.01
2003-06-02	2523.2	(2513.5, 2532.9)	6047.9	0.16196	1(0)	0	10.90
2003-06-03	2523.9	(2514.8, 2533)	5861.7	0.16145	1(0)	0	10.76
2003-06-04	2524.3	(2515.8, 2532.9)	5741.4	0.16111	0(0)	0	10.61
2003-06-05	2524.6	(2516.6, 2532.7)	5666.5	0.16089	0(0)	0	10.46

5　5 月 19 日前广东、山西、内蒙古自治区 SARS 疫情的分段拟合

广东疫情据报导在 2 月中旬已趋平稳, 但随着春假结束, 民工返回, 学校开学, 广交会如期举办等等, 疫情在 3、4 月明显反弹。但此期间无可靠数据。从 4—5 月数据看, 到 4 月底本应再次趋于平稳, 但从 5 月 1 日开始, 再次反弹, 5 月 7 日后重新得到有力控制, 见图 3。山西 SARS 疫情预测见图 4, 内蒙古 SARS 疫情预测见图 5。从结果看, 广东在 5 月底进入了 SARS 的尾声, 而北京及其周边地区将在 6 月同步进入 SARS 的尾声。但需提高警惕, 以防疫情反弹。

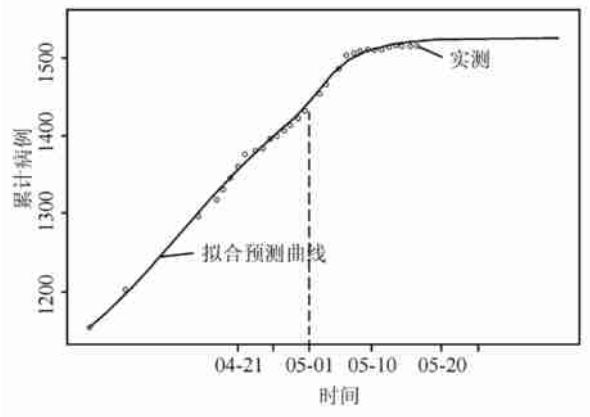

图 3　广东 SARS 预测曲线图(2003-05-19)

Fig. 3　Piecewise time fitting and forecasting for Guangdong SARS situation (up to May 19, 2003)

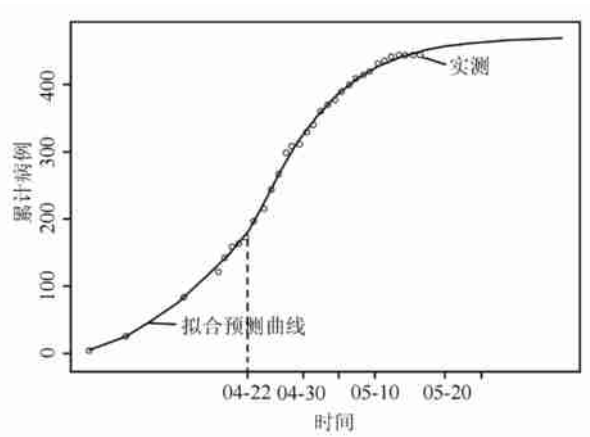

图 4　山西 SARS 预测曲线图(2003-05-19)

Fig. 4　Piecewise time fitting and forecasting for Shanxi SARS situation (up to May 19, 2003)

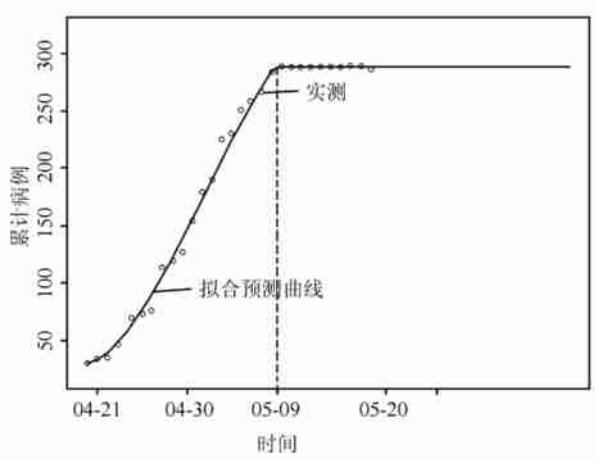

图 5　内蒙古 SARS 预测曲线图(2003-05-19)

Fig. 5　Piecewise time fitting and forecasting for Neimenggu SARS situation (up to May 19, 2003)

6　结论与讨论

以 Richards 增长曲线作为 SARS 累计病例的非线性回归曲线所建立的模型, 具有模型简单, 机理清楚, 算法简明快速等特点, 从北京 SARS 的整体预测预报上看效果较好, 模型三参数 A, B, K 的总体拟合效果随时间的变化逐渐趋于稳定, 其预测预报合理可行。由于控制等因素, 使得 SARS 累计病例的趋势有所变化, 因而对模型分时间段来拟合预测更为合理, 模型参数的变化(特别是 A, b) 说明控制措施对 SARS 疫情的影响, 并且预测预报效果要比整体随时间变化的预测预报效果要好, 误差要小。另外, 模型整体上预测预报结果偏向于保守, 对分时间段预测预报的时间段把握标准, 模型的异方差以及与其它相关模型比较等还需进一步从理论和实际数据中加以探讨和研究, 参见文献[5]。本文的结果为下一步的研究工作打下了良好基础。

参 考 文 献 (References)

[1] Webpage of Ministry of Health PRC(中国国家卫生部网站): http://www.moh.gov.cn/zhgl/yqfb/.

[2] Webpage of World Health Organization (WHO 网站): http://www.who.int/csr/sars/country/en/.

[3] France J, Dijkstra J, Thornley J H M, *et al*. A Simple but Flexible Growth Function[J]. *Growth Development & Aging*, 1996, **60**: 71—83.

[4] Zeger S L, Harlow S D. Mathematical Models from Laws of Growth to Tools for Biological Analysis: Fifty Years of Growth[J]. Growth, 1987, **51**: 1—21.

[5] Webpage of Cui Hengjian, http://math.bnu.edn.cn/~chj/(Fighting SARS).

Nonlinear Regression in SARS Forecasting

CUI Heng-jian[1], LI Zhong-lai[1], YANG Hua[2], LI Xiao-wen[2,3]

(1. *Department of Mathematics, Statistical Data Analysis Center of Beijing Normal University, Beijing* 100875, *China*;
2. *Remote Sensing and GIS Research Center of Beijing Normal University, Beijing* 100875, *China*;
3. *Institute of Remote Sensing Appliations, Chinese Academy of Sciences, Beijing* 100101, *China*)

Abstract: This paper introduces some kinds of nonlinear growth curve for forecasting cumulative SARS patients, it is shown that the Richards curve is reasonable and flexible in this SARS forecasting. The nonlinear growth curve regression model is established for forecasting cumulative SARS patients. Specifically, the SARS situation forecasting in Beijing is made well which includes forecasting based on comsecutive and piecewise time fitting. It means some control policies in Beijing at the end of this April play important role for anti-spread of SARS, and also provides a good basis for future works.

Key words: SARS; growth curve; nonlinear regression model; Richards growth curve; piecewise time fitting