·论　著·

时间序列 ARIMA 模型在艾滋病疫情预测中的应用*

罗　静[1]，杨　书[2△]，张　强[1]，王　璐[1]

（1.四川大学公共卫生学院卫生统计教研室，成都 610041；2.成都医学院公共卫生系，成都 610083）

摘　要：目的　探讨应用时间序列求和自回归移动平均（ARIMA）模型预测艾滋病发病率的可行性。**方法**　利用重庆市疾病控制部门提供的 1993～2009 年艾滋病发病情况数据建立 ARIMA 预测模型。**结果**　ARIMA$(1,1,1)\times(0,1,0)_{12}$很好地拟合了艾滋病发病率，2009 年 7～12 月的预测值符合实际发病率变动趋势。**结论**　ARIMA 模型很好地模拟艾滋病发病率在时间序列上的变动趋势，可以为疫情防控提供借鉴。

关键词：预测；获得性免疫缺陷综合征；时间序列；ARIMA 模型

doi：10.3969/j.issn.1671-8348.2012.13.003　　**文献标识码：**A　　**文章编号：**1671-8348(2012)13-1255-02

ARIMA model of time series for forecasting epidemic situation of AIDS*

Luo Jing[1]，*Yang Shu*[2△]，*Zhang Qiang*[1]，*Wang Lu*[1]

（1. *Department of Health Statistics*，*School of Public Health*，*Sichuan University*，*Chengdu*，*Sichuan* 610041，*China*；2. *Department of Public Health*，*Chengdu Medical College*，*Chengdu*，*Sichuan* 610083，*China*）

Abstract：Objective　To explore the feasibility of auto regressive integrated moving average（ARIMA）model of time series to predict the incidence of AIDS. **Methods**　The ARIMA model was established basing on the data of AIDS incidences in Chongqing during 1993－2009. **Results**　The model of ARIMA$(1,1,1)\times(0,1,0)_{12}$ exactly fitted the incidence of AIDS. The predicting values of incidence in July to December 2009 were consistent with the actual change trend of incidence. **Conclusion**　The ARIMA model can be used to exactly simulate the change trend of the incidence of AIDS in time series，which can provide reference for prevention and control of AIDS.

Key words：forecasting；acquired immunode ficiency syndrome；；time series；ARIMA model

艾滋病，即获得性免疫缺陷综合征（acquired immune deficiency syndrome，AIDS），是由艾滋病病毒（HIV）破坏人体免疫系统，使其丧失抵抗各种疫病能力的一种严重危害人类生命安全的疾病。2000 年以后，特别是 2005 年以来，中国的艾滋病感染人数迅速上涨。在艾滋病的防控工作中，如果能在局部范围内对未来感染人数做一定程度预判，为“三间分布”提供信息，对制定正确的防控政策和卫生资源配置提供依据，具有一定的指导意义。本文以重庆市疾病控制部门提供的艾滋病疫情发展为例，采用求和自回归移动平均（auto regressive integrated moving average，ARIMA）时间序列模型拟合预测发病率，探讨模型的可行性，对相关问题进行探索性研究。

1　资料与方法

1.1　一般资料　相关数据由重庆市疾病控制部门提供，包括 1993～2009 年重庆市辖区月度新发艾滋病感染人数，以及该市 2010 年卫生统计年鉴。

1.2　模型建立　ARIMA 模型是以序列不同时期内的相关度量为基础进行的一种精确度较高的短期预测分析方法。该法由美国学者 Box 和英国统计学者 Jenkins 于 1976 年提出，故又称为 Box-Jenkins 模型[1]。在 ARIMA 模型中，变量的未来取值可以表达为过去若干个取值和随机误差的线性函数。

$$\begin{cases}\Phi(B)\nabla_d^{\nabla}X_t=\Theta(B)\varepsilon_t,\\ E(\varepsilon_t)=0,Var(\varepsilon_t)=\sigma_\varepsilon^{\,2},E(\varepsilon_t\varepsilon_s)=0,s\neq t,\\ E(x_s\varepsilon_t)=0,\forall\, s<t.\end{cases}$$

式中：

$$\nabla^d=(1-B)^d$$

$$\Phi(B)=1-\varphi_1B-\cdots-\varphi_pB^p$$

$$\Theta(B)=1-\theta_1B-\cdots-\theta_qB^q$$

其中 B 是后移算子，ε_t 为各期的随机扰动或随机误差，d 为差分阶数，p 和 q 分别表示自回归阶数和移动平均阶数，X_t 为各期的观察值（$t=1,2,\cdots,k$）[2-3]。

建立 ARIMA 时间序列模型可归纳为 3 个阶段，即序列的平稳化、模型识别以及参数估计和模型诊断，通过这 3 个阶段处理的反复进行，最终确定一个用于预报的“最优”模型[4]。

1.2.1　序列的平稳化　序列的平稳性是 ARIMA 模型分析的前提条件，即要求均数不随时间变化；方差不随时间变化；自相关系数只与时间间隔有关，而与所处的时间无关[5]。对于非平稳的序列，可以通过差分和 Box-Cox 变换使均数和方差平稳化。

1.2.2　模型识别　通过观察序列自相关（auto correction fuction，ACF）和偏自相关（partial auto correction fuction，PACF）的截尾、拖尾性初步为序列定阶，提供几个粗模型以便进一步分析完善[6-7]。

1.2.3　参数估计和模型诊断　根据模型阶数，运用最大似然法估计或最小二乘法估计，计算出求和自回归移动平均过程的各项系数，并做假设检验。在模型的拟合中，应满足模型的残差序列是白噪声序列，即 Box-Ljung Q 统计量相比较差异无统计学意义（$P>0.05$）。若几个模型均满足参数相比较差异有统计学意义，残差序列为白噪声序列的要求，则使拟合优度统

*　**基金项目：**国家自然科学基金青年基金资助项目（81001295）。　△　通讯作者，Tel：13730665374；E-mail：lttys-1983@163.com。

表 1 备选模型的参数估计

参数	ARIMA$(1,1,1)\times(0,1,0)_{12}$			ARIMA$(1,1,1)\times(0,1,1)_{12}$			ARIMA$(1,1,0)\times(0,1,0)_{12}$		
	B	t	P	B	t	P	B	t	P
自回归系数	−0.545	−3.351	0.003	−0.493	−3.111	0.005	−0.752	−5.914	0.000
平均移动系数	0.928	2.341	0.028	0.979	0.667	0.511	—	—	—
季节移动平均系数	—	—	—	0.520	1.580	0.127	—	—	—
常数	0.000	0.038	0.970	0.002	0.176	0.862	−0.024	−0.214	0.832

—:表示无数据。

计量赤池信息准则(akaike' s information cnitenion,AIC)和贝叶斯算法(selective bayes classifiers,SBC)均达到最小的模型为最优模型。反之,模型参数间比较差异无统计学意义,或残差序列不是白噪声序列,都需要返回识别阶段,重新调整各个阶数的值,再进行参数估计和模型诊断。

1.3 统计学处理 应用 SPSS 13.0 统计软件建立 ARIMA 时间序列模型并进行数据处理和分析[8-9]。

2 结 果

2.1 数据处理 对 1993～2009 年重庆市疾病控制部门提供的艾滋病月发病率作序列图,发现数据总体呈上升趋势。其中,1993～2003 年月发病率较低,其大多数月份为 0,最大值为 0.073 9(1/10 万);2005 年 1 月和 3 月呈现 2 个高峰,其后数据波动幅度增大,序列的方差在前后差别明显。因此,以 2005 年 1 月为切点,将数据分为两个部分。以 2005 年 1 月至 2009 年 6 月发病率作建模数据,2009 年 7～12 月的数据作验证数据,对序列进行自然对数变换,差分和季节差分后,序列平稳。

2.2 模型识别 观察处理后序列的 ACF 和 PACF(图 1、2),发现自相关函数和偏自相关函数呈现递减且拖尾。可初步判断模型为模型一 ARIMA$(1,1,1)\times(0,1,0)_{12}$、模型二 ARIMA$(1,1,1)\times(0,1,1)_{12}$或模型三 ARIMA$(1,1,0)\times(0,1,0)_{12}$。

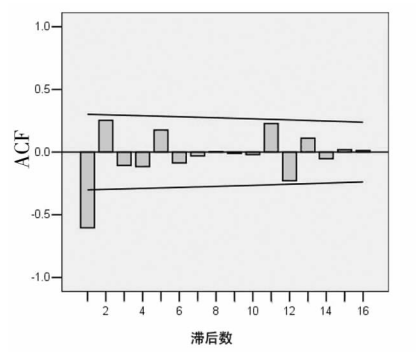
2.3 参数估计及检验 模型一和模型三的参数间比较差异有统计学意义,模型二中 MA1 和 SMA 比较差异无有统计学意义。见表 1。

2.4 模型诊断 表 2 所示,在备选模型中,模型一拟合优度较小,且参数间无明显相关性($r=0.267$)。此外,观察其残差的自相关图,结果显示该模型的 Box-Ljung Q 统计量间比较差异均无统计学意义($P>0.05$),可以认为残差序列为白噪声[10]。综上分析,模型一为最优模型。可以确定重庆市艾滋病发病率的预测模型为 ARIMA$(1,1,1)\times(0,1,0)_{12}$,其表达式为:$(1+0.545B)\nabla_{12}\nabla\ \ln X_t=(1-0.928B)\varepsilon_t$。

表 2 备选模型拟合优度统计量

项目	ARIMA $(1,1,1)\times(0,1,0)_{12}$	ARIMA $(1,1,1)\times(0,1,1)_{12}$	ARIMA $(1,1,0)\times(0,1,0)_{12}$
标准误差	0.752	0.665	1.022
对数似然	−31.713	−30.225	−39.737
AIC	69.425	68.449	83.475
BIC	73.422	73.778	86.139

2.5 模型预测 用 ARIMA$(1,1,1)\times(0,1,0)_{12}$模型预测重庆市 2009 年 7～12 月艾滋病发病率,结果如表 3 所示。可以看出模型预测值的动态趋势与实际情况基本一致,模型对未来的情况进行了很好的跟踪和预测。2009 年 7～12 月的实际发病率虽然与预测值不完全一样,但是各月实际值都落入了预测值 95%的可信区间范围。

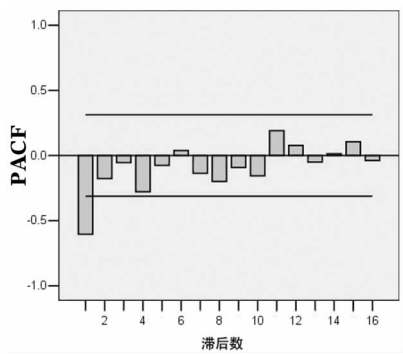


□:系数;—:可信限。

图 1 原序列经过对数转换和两次差分后的 ACF 图

表 3 2009 年 7～12 月重庆市实际发病率与预测发病率(1/10 万)

发病率	7 月	8 月	9 月	10 月	11 月	12 月
实际发病率	0.711 3	0.573 9	0.656 4	0.488 5	0.494 6	0.317 5
预测发病率	0.657 1	0.446 2	0.668 5	0.800 1	0.585 4	0.738 4



□:系数;—:可信限。

图 2 原序列经过对数转换和两次差分后的 PACF 图

3 讨 论

3.1 艾滋病发病率预测的意义 根据模型预测并结合实际情况,重庆市艾滋病感染速度呈上升趋势。相关部门可以有针对性地采取预防控制措施。如整合艾滋病医疗资源,大力提升其诊治能力;建立"重庆市艾滋病关爱之家"[11],动员全社会参与艾滋病防治,消除对艾滋病患者的恐惧和歧视;组建艾滋病职业暴露药品库,降低全市艾滋病职业暴露人员感染 HIV 的危险性等[12]。由于近几年重庆市艾滋病感染者基数较大且不断增加,致使发病率仍然不断上升,所以,还应加大其预防控制工作强度并且在预防控制手段上有所创新,加大对高危人群及高发地区的监测和行为干预[13]。 (下转第 1259 页)

疗的疗效分析[J]. 重庆医学,2011,40(14):1419-1421.
[2] 郑捍东,吴文友. 基底节区高血压脑出血开瓣手术与双管钻孔引流术的对比研究[J]. 重庆医学,2009,38(10):1200-1201.
[3] 朱成明,姚文华,王贵富,等. 颅骨钻孔尿激酶溶解引流术与小骨窗开颅血肿清除术治疗高血压脑出血的比较分析[J]. 重庆医学,2011,40(13):1318-1320.
[4] 陶子荣. 中国脑卒中患者临床神经功能缺损评分标准信度、效度及敏感度的评价[J]. 第二军医大学学报,2009,30(3):283-285.
[5] 王新,王拥军,颜振瀛. 脑卒中患者临床神经功能缺损评分标准的信度和效度研究[J]. 卒中与神经疾病,1999,6(3):22-24.
[6] 方波,袁鹏,朱政鸣,等. 高血压脑出血微创血肿清除治疗[J]. 重庆医学,2005,34(11):1605-607.
[7] Nguyen JP, Decq P, Brugieres P, et al. A technique for stereotactic of deep intracerebral hematomas under computed tomographic contol using a new device[J]. Neurosurgery,1992,32(2):330-334.
[8] 辛东,李希福. 高血压脑出血外科手术探讨[J]. 中国实用医药,2009,4(14):100-101.
[9] Zia E, Hedblad B, Pessah-Rasmussen H, et al. Blood pressure in relation to the incidence of cerebral infarction and intracerebral hemorrhage. Hypertensive hemorrhage: debated nomenclature is still relevant[J]. Stroke, 2007, 38(10):2681-2685.
[10] Murakami M, Fujioka S, Oyama T, et al. Serial changes in the regional cerebral blood flow of patients with hypertensive intracerebral hemorrhage-long-term follow-up SPECT study[J]. J Neurosurg Sci, 2005, 49(3):117-114.
[11] 潘进钱,叶盛,张宇,等. 基底节出血抽吸术靶点设定与术中及术后再出血的关系[J]. 中华神经外科杂志,2004,20(3):225-227.
[12] 赵春平,秦家振,魏群,等. 采用微创手术治疗高血压脑出血[J]. 中华神经外科疾病研究杂志,2003,2(1):80-81.
[13] 王多姿,郭富强. 脑出血血肿周围继发性损伤[J]. 国际脑血管病杂志,2007,15(11):838-841.
[14] Hirohata M, Yoshita M, Ishida C, et al. Clinical features of non-hypertensive lobar intracerebral hemorrhage related to cerebral amyloid angiopathy[J]. Eur J Neurol, 2010,17(6):823-829.
[15] Zhao X, Wang Y, Wang C, et al. Quantitative evaluation for secondary injury to perihematoma of hypertensive cerebral hemorrhage by functional MR and correlation analysis with ischemic factors[J]. Neurol Res, 2006, 28(1):66-70.

(收稿日期:2011-12-09 修回日期:2012-01-29)

(上接第 1256 页)

3.2 ARIMA 模型的应用 时间序列分析是在不需要考虑预测变量的相关因素及其关系的情况下,利用事物发展的延续性,建立时间序列模型来预测未来的变化[14]。而传统的时间序列模型要求序列具有平稳的线性趋势,但实际上疾病的发病情况一般有着明显的周期变化,如果不考虑这些因素的影响,做出的预测往往不准确。本研究采用的 ARIMA 模型,综合考虑了序列的趋势变化、周期变化及随机干扰等因素的影响,对艾滋病发病拟合度较好[15]。由于疫情波动受到诸多未知随机因素的影响,所建立的模型不是一成不变的,它较适合进行短期的预测,同时需要不断加入新的实际数据,以不断新拟合更能反映实际情况的预测模型,并提高预测的敏感性。

参考文献:

[1] Geoge EP, Gwilym M. 时间序列分析预测与控制[M]. 北京:中国统计出版社,1997.
[2] 肖枝洪,郭明月. 时间序列分析与 SAS 应用[M]. 武昌:武汉大学出版社,2009.
[3] 何书元. 应用时间序列分析[M]. 北京:北京大学出版社,2003.
[4] 孙振球,徐勇勇. 医学统计学[M]. 北京:人民卫生出版社,2002.
[5] 张文增,冀国强,史继新,等. ARIMA 模型在细菌性痢疾预测预警中的应用[J]. 中国卫生统计,2009,26(6):636-639.
[6] 吴家兵,叶临湘,尤尔科. 时间序列模型在传染病发病率预测中的应用[J]. 中国卫生统计,2006,23(3):276.
[7] 刘晓宏,金丕焕,陈启明. ARIMA 模型中时间序列平稳性的统计检验方法及应用[J]. 中国卫生统计,1998,15(3):12-14.
[8] 张文彤. SPSS11 统计分析教程高级篇[M]. 北京:北京希望电子出版社,2002.
[9] 薛薇. SPSS 统计分析方法及应用[M]. 2 版. 北京:电子工业出版社,2009.
[10] 孟蕾,王玉明. ARIMA 模型在肺结核发病预测中的应用[J]. 中国卫生统计,2010,27(5):507-509.
[11] 王治伦,晏治碧,陈思源,等. 建立重庆市艾滋病关爱之家体会[J]. 中国感染控制杂志,2004,3(3):275-276.
[12] 李颖,汪洋,刘琴,等. 重庆市高危人群中艾滋病防治的定性研究[J]. 中国卫生事业管理,2005(2):96-97.
[13] 丁贤彬,邝富国,凌华,等. 重庆市艾滋病流行现状及防治策略[J]. 疾病控制杂志,2005,9(4):340-341.
[14] 邓丹,王润华,周燕荣. 时间序列分析及其在卫生事业中的应用[J]. 数理医学杂志,2002,15(5):455-457.
[15] 冯超,白彬. 时间序列模型拟合艾滋病发病趋势预测[J]. 中国公共卫生,2005,21(7):893.

(收稿日期:2011-10-09 修回日期:2012-01-22)