状态空间模型及其在传染病发病率预测中的应用

陈友春*,朱文婕

(蚌埠医学院计算机教研室,安徽 蚌埠 233000)

[摘 要] **目的**:研究状态空间模型的建立及其在传染病发病预测中的应用,并探讨提高模型准确性和实用性的途径。**方法**:以2005 年 1 月~2010 年 12 月我国肺结核发病资料建立模型,以 2011 年的发病资料作为模型预测效果的考核样本。首先采用移动平均比率法和 HP 滤波对资料进行初步分析,然后根据分析的结果进行定阶、初始化并估计参数,建立状态空间模型,最后对预测结果进行检验和分析。**结果**:状态空间模型可以将发病率变化过程中的各种特征成分分解出来,其年内逐月发病率的预测精度在 90%以上。**结论**:状态空间模型对我国肺结核发病情况的拟合度较高,预测效果良好。

[关键词] 状态空间模型;卡尔曼滤波;传染病;预测
[中图分类号] O212 [文献标志码] A [文章编号] 1007-4368(2015)02-275-04
doi:10.7655/NYDXBNS20150233

State space model and its application on forecasting in incidence of infectious disease

Chen Youchun*, Zhu Wenjie

(*Computer T&R Section, Bengbu Medical College, Bengbu 233000, China*)

[Abstract] **Objective**: To research the application of State Space Model forecasting in infectious disease incidence, and discuss the method to improve its veracity and practicability. **Methods**: A model was fitted by the historical data of the incidence of tuberculosis in China. Firstly, used the ratio to moving average method and HP filter to make preanalysis. Secondly, parameters of model is estimated and a State Space Model was set up by decison of the rank of it. Finally, the paper tests the result of forecast and analysis it. **Results**: State Space Model can decompose some characteristic components from the changing process of the incidence. Forecast accuracy of the Monthly incidence in a year is above 90%. **Conclusion**: The fit values of incidence are consistent with the actual data of incidence and the forecasting effect is good.

[Key words] state space model; kalman filter; infectious disease; forecasting

[Acta Univ Med Nanjing, 2015, 35(02): 275-278]

对传染病流行趋势进行预测,以便采取控制措施,是疾病预防与控制的一项重要工作。但是传染病的发病受到许多因素的影响,包括各种自然因素和社会因素,这些影响因素之间存在着错综复杂的联系,很难运用静态的因果模型加以解释。同时,传染病的发生往往具有长期趋势、季节性、周期性、短期波动和不规则变动等特征,因此对这类资料进行预测分析时需要同时考虑这些特征[1]。

状态空间模型作为一种广泛应用的建模工具,被用来估计不可观测的时间变量,将不可观测的变量(状态变量)并入可观测模型,通过强有力的迭代算法—Kalman 滤波来估计不同的状态成分以达到分析和预测的目的[2-3]。本文采用状态空间模型对我国肺结核逐月发病率进行分析,获得其变化过程中的各种特征成分,在此基础上对逐月发病率进行预测与检验,为我国肺结核的监测和防治提供科学依据。

1 状态空间模型

1.1 状态空间模型构成

本文所研究线性高斯状态空间模型主要由以下方程构成:

$$y_t=Z_t\alpha_t+\varepsilon_t,\ \varepsilon_t\sim N(0,H_t) \quad (1)$$

$$\alpha_t=T_t\alpha_{t-1}+\eta_t,\ \eta_t\sim N(0,Q_t) \quad (2)$$

$$\alpha_0\sim N(\alpha_0,P_0),\ t=1,2,\cdots,n \quad (3)$$

式(1)为量测方程,式(2)为状态方程,式(3)为

[基金项目] 蚌埠医学院科研课题计划(BYKY1301)
*通信作者(Corresponding author),E-mail:chenyouchun1@sina.com

初始状态向量的分布。模型中 y_t 是包含 k 个变量的可观测向量(单变量模型时 k=1),这些向量与 m×1 维向量 α_t 有关,α_t 被称为状态向量。T_t 为 m×m 阶状态转移矩阵,Z_t 为 k×m 阶量测矩阵。ε_t 和 η_t 分别为量测方程和状态方程的随机扰动项,并且满足:

$$\begin{pmatrix}\varepsilon_t\\ \eta_t\end{pmatrix}\sim N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}H_t & 0\\0 & Q_t\end{bmatrix}\right) \tag{4}$$

其中系统矩阵 Z_t、H_t、T_t 和 Q_t 在模型中至少有部分元素是未知的,通过 Ψ 向量表示这些未知参数的集合,为区别于状态向量中的未知元素,称之为超参数(hyperparameters)。

1.2 Kalman 滤波和预测

Kalman 滤波及预测对状态空间模型的估计给出了线性无偏最小方差的递推算法,假定已建立了上述状态空间模型,我们可以用 Kalman 滤波对状态向量及其方差进行递推计算。

记 $\alpha_{t|s}=E(\alpha_t|y_1,\cdots,y_s)$,$P_{t,U|s}=E[(\alpha_t-\alpha_{t|s})(\alpha_U-\alpha_{U|s})|y_1,\cdots,y_s]$,适当选取如式(3)中状态向量的初值后,Kalman 滤波递推公式如下:

$$\begin{aligned}&\alpha_{t/t-1}=T_t\alpha_{t-1/t-1}\\&P_{t,t/t-1}=T_tP_{t-1,t-1/t-1}T_t'+Q_t\\&K_t=P_{t,t/t-1}Z_t'(Z_tP_{t,t/t-1}Z_t'+H)^{-1}\\&\alpha_{t/t}=\alpha_{t/t-1}+K_t(y_t-Z_t\alpha_{t/t-1})\\&P_{t,t/t}=P_{t,t/t-1}-K_tZ_tP_{t,t/t-1}\end{aligned} \tag{5}$$

在实际问题中,如果对于系统初始状态的先验知识了解很少,一般可以任意选取一个向量做为初始状态向量的期望值,同时将其方差阵的对角线元素设为充分大的正数,如设置 α_0=0 和 $P_0=10^6$ I。为提高估计的精度,还可以在向前滤波之后,利用 Kalman 平滑公式对 $t=n,n-1,\cdots,1$ 进行向后滤波,得到 $\alpha_{0|n}$ 和 $P_{0,0|n}$,并将其做为下一次迭代向前滤波的初值[4]。

Kalman 平滑公式为:

$$\begin{aligned}&\alpha_{t-1/n}=\alpha_{t-1/t-1}+J_{t-1}(\alpha_{t/n}-\alpha_{t/t-1})\\&P_{t-1,t-1/n}=P_{t-1,t-1/t-1}+J_{t-1}(P_{t,t/n}-P_{t,t/t-1})J_{t-1}'\\&J_{t-1}=P_{t-1,t-1/t-1}T_t'P_{t,t/t-1}^{-1}\end{aligned} \tag{6}$$

Kalman 预测公式为:

$$\begin{aligned}&\alpha_{n+k/n}=T_n\alpha_{n+k-1/n}=T_n^k\alpha_{n/n}\\&y_{n+k/n}=Z_n\alpha_{n+k/n}\end{aligned} \tag{7}$$

其中 k=1,2,…是预测步长,预测方差随预测步长增大而增长。

1.3 超参数的估计

上述利用 Kalman 滤波递推公式求状态向量的估计量时,假定模型中的超参数是已知的,但实际上向量 Ψ 需要用极大似然估计进行求解。适当的选择参数初值 Ψ_0 并用其进行正向滤波,根据滤波的结果按下式计算对数似然函数[5]:

$$\ln L(y_t;\Psi)=-\frac{nk\log(2\pi)}{2}-\frac{1}{2}\sum_{t=1}^{n}|F_t|-\frac{1}{2}\sum_{t=1}^{n}v_t'F_t^{-1}v_t \tag{8}$$

其中 $F_t=Z_tP_{t,t|t-1}Z_t'$,v_t 为预测的误差,$v_t=y_t-Z_t\alpha_{t|t-1}$ $(t=1,2,\cdots,n)$。

多数求极大似然估计的数值搜索算法都基于 Newton's 法,本文实例中程序所用为 BFGS(Broyden-Fletcher-Goldfarb-Shannon)算法[6]。

2 我国肺结核发病情况实例研究

肺结核是由结核分枝杆菌引发的肺部感染性疾病,是严重威胁人类健康的疾病。结核分枝杆菌的传染源主要是排菌的肺结核患者,通过呼吸道传播。我国是全球 22 个肺结核流行严重的国家之一,2001~2010 年我国肺结核报告发病人数始终位居全国甲乙类传染病报告发病数的前列。

2.1 模型的设定

数据资料来源于我国卫生部网站 2005 年 1 月~2011 年 12 月的全国法定报告传染病疫情资料以及国家统计局网站 2005~2011 年人口统计资料,据此计算肺结核月发病率数据(y_t,单位:1/10 万),其中 2011 年 1~12 月发病率数据用于模型预测效果的考核样本。

首先通过移动平均比率加法模型对发病率数据 y_t 进行季节调整(季节因子记为 S_t),然后采用 HP(Hodrick-Prescott)滤波方法将季节调整后的数据 y_t^* 分解成趋势成分 T_t 和循环成分 C_t。通过对循环成分 C_t 的自相关系数与偏相关系数的分析,得到结论为可以将其作为白噪声序列而不需要用 ARMA 模型去拟合,否则系数显著性检验无法通过。因此下面将循环成分 C_t 归入不规则项 I_t。

根据以上预处理的结果,将发病率数据 y_t 分解为 T_t、S_t 和 I_t,即趋势项、季节项和不规则项。参照 Harvey(1990)的模型[7],我国肺结核发病率状态空间模型表示如下:

$$y_t=T_t+S_t+I_t,t=1,\cdots n \tag{9}$$

趋势项采用随机增长模型来描述,在不同的噪

声值下,模型既可以满足一阶差分约束,又可以满足二阶差分约束。

$$T_t=T_{t-1}+\beta_{t-1}+\eta_t$$
$$\beta_t=\beta_{t-1}+\xi_t \qquad t=1,\cdots n \tag{10}$$

其中 η_t、ξ_t 为相互独立、具有零均值及有限方差的白噪声。

季节变动 S_t 是以年为周期的季节变化,在一个周期内各季节成分之和应满足零均值条件,可用随机季节模型进行拟合。

$$S_t=-\sum_{j=1}^{s-1}S_{t-j}+\omega_t \quad t=1,\cdots,n \tag{11}$$

其中 S=4 或 12 分别对应于季度或月度资料。ω_t 为零均值、方差为常数的白噪声。

不规则项 I_t 则如前面所分析,是一个零均值及方差为 σ_ε^2 的白噪声。

综上所述,所建状态空间模型的状态方程为:

$$\alpha_t=\begin{pmatrix}T_t\\ \beta_t\\ S_t\\ S_{t-1}\\ \vdots\\ S_{t-10}\end{pmatrix}=\begin{pmatrix}1&1&0&0&\cdots&0&0\\ 0&1&0&0&\cdots&0&0\\ 0&0&-1&-1&\cdots&-1&-1\\ 0&0&1&0&\cdots&0&0\\ \vdots&\vdots&&&\ddots&&\\ 0&0&0&0&\cdots&1&0\end{pmatrix}\begin{pmatrix}T_{t-1}\\ \beta_{t-1}\\ S_{t-1}\\ S_{t-2}\\ \vdots\\ S_{t-11}\end{pmatrix}+\begin{pmatrix}\eta_t\\ \xi_t\\ \omega_t\\ 0\\ \vdots\\ 0\end{pmatrix} \tag{12}$$

量测方程为:

$$y_t=(1\ 0\ 1\ 0\ \cdots\ 0)\alpha_t+\varepsilon_t \tag{13}$$

状态噪声和量测噪声的方差 σ_η^2、σ_ξ^2、σ_ω^2 和 σ_ε^2 为模型待估计的超参数,这些参数的值越小,则意味着选择了合适的阶数,对数据的拟合越好[8]。

2.2　参数估计、模型拟合及相关解释

本文应用 MATLAB toolbox 中的 SSM (State Space Model)程序包来实现参数的估计、数据的拟合及预测,M 文件主要代码如下 (其中 fbl.dat 文件中包含数据为 2005 年 1 月~2010 年 12 月的发病率资料):

```
y =load('fbl.dat')';
time=05:1/12:11;
tr =ssm_llt;
seas=ssm_seasonal('dummy',12);
bstsm=[tr seas];
[bstsm logL]=estimate(y,bstsm,[10 1 1 1],[],
   fmin','bfgs','disp','off');
[alphahat V]=statesmo(y,bstsm);
irr=disturbsmo(y,bstsm);
ycom=signal(alphahat,bstsm);
ytr=ycom(1,:);
yseas=ycom(2,:);
figure('Name','Estimated Components');
subplot (3,1,1),plot (time(1:end-1),ytr),title
('trend');
subplot(3,1,2),plot(time(1:end-1),yseas),ti-
tle('Seasonal');
subplot (3,1,3),plot (time(1:end-1),irr),title
('Irregular');
```

其中 ytr、yseas 和 irr 即为发病率 y_t 的趋势项、季节项和不规则项。

模型的参数估计结果如表 1 所示。

表 1　模型参数估计结果

Table 1　Results of the model parameter estimation

名称	V 值
epsilon var (σ_ε^2)	0.395 7
zeta var1 (σ_η^2)	0.010 28
zeta var2 (σ_ξ^2)	3.302e-010
omega var (σ_ω^2)	1.706e-008

从 $\sigma_\xi^2\approx0$ 可以看出模型的趋势项近似满足一阶差分约束,而 $\sigma_\omega^2\approx0$ 则表明各周期间季节变动的差异不明显。发病率 y_t 的趋势项、季节项和不规则项如图 1 所示(横轴由程序中 time 向量构成,表示 2005 年起的月度时间)。

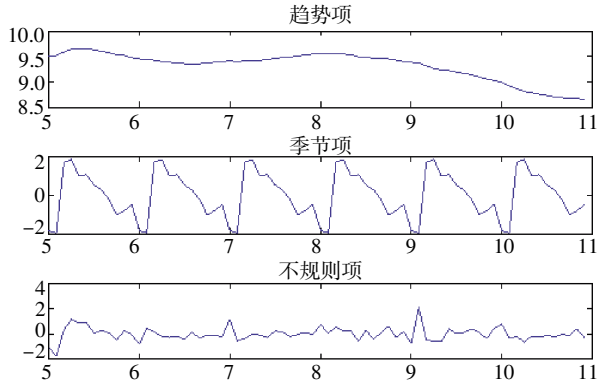

图 1　发病率 y_t 的趋势项、季节项和不规则项

Figure1　Trend, seasonal and irregular items of the incidence

从图中可以看出我国肺结核发病率呈缓慢下降的趋势,同时肺结核的发病也呈现出明显的季节特征,发病率在每年的 3、4 月份到达高峰,随后慢慢下降,在 1、2 月份到达低谷。另外由不规则项围绕零均值的波动来看,拟合的效果还是很好的。

在上述 M 文件中加入语句“[a P v F]= kalman ([y repmat (NaN,1,12)],bstsm)”对 2011 年 1~12 月我国肺结核发病率进行预测,并将预测值与实际发病率比较,结果如表 2 所示。

表 2　2011 年我国肺结核发病率的预测

Table 2　Forecasts for the incidence of tuberculosis in China in 2011　(%)

月份	发病率			
	实 际 值	预测值	绝对误差	相对误差
1	7.393 6	6.775 0	0.618 6	8.37
2	7.285 2	6.648 9	0.636 3	8.73
3	10.082 6	10.285 2	–0.202 6	2.01
4	9.600 4	10.419 4	–0.819 0	8.53
5	9.287 0	9.557 5	–0.270 4	2.91
6	8.857 7	9.603 1	–0.745 4	8.42
7	8.360 6	9.078 0	–0.717 4	8.58
8	8.545 7	8.788 0	–0.242 4	2.84
9	7.935 9	8.287 7	–0.351 7	4.43
10	7.451 1	7.492 9	–0.041 9	0.56
11	8.213 3	7.702 0	0.511 3	6.23
12	7.771 6	8.023 3	–0.251 7	3.24

从表 2 可以看出,模型的预测精度达到 90%以上,而且其精度受预测步长的影响较小,表明状态空间模型不仅可以用于短期预测,也比较适合进行中长期预测。

3　结　论

自上世纪 60 年代卡尔曼滤波提出以来,状态空间模型的研究及其在工程、信息、航天和经济等领域的应用,国内外均有大量可供检索的文献资料进行描述,但是将其应用于医学领域中对传染病进行时序分析和预测的文献却不多见。

本文较系统地研究了线性高斯状态空间模型的相关理论, 并结合 MATLAB 软件的 SSM 程序包对我国肺结核发病率进行了预测,得到以下结论:①状态空间模型是一个有力的时间序列分析建模工具,非常适用于对具有趋势、循环、季节等特征的传染病资料进行分析;②在建立模型之前,对资料进行预分析将有助于我们确定模型的阶数, 从而提高分析的准确性;③状态空间模型的相关算法较为复杂,借助 SSM 之类的程序包(其他如 SsfPACK、STAMP 等)可以提高模型的实用性;④从本文实例来看,传染病发病的状态空间模型其预测精度受预测步长的影响较小,因而除短期预测外,中长期预测也是适用的。

[参考文献]

[1] 彭志行,鲍昌俊,赵　杨,等. ARIMA 乘积季节模型及其在传染病发病预测中的应用[J]. 数理统计与管理,2008,27(2):180–186

[2] Kalman RE. A new approach to linear filtering and prediction theory[J]. Basic Eng,1960,82(1):35–45

[3] 高铁梅. 计量经济分析方法与建模——Eviews 应用及实例[M]. 北京:清华大学出版社,2006:353–379

[4] 顾　岚. 关于经济时间序列分解的状态空间方法研究[J]. 统计研究,1993,10(3):46-51

[5] 苗敬毅. 关于状态空间方法中超参数估计的札记[J]. 山西经济管理干部学院学报,2001,3(1):44–46

[6] James Durbin,Siem Jan Koopman. Time Series Analysis by State Space Methods[M]. Oxford:Oxford University Press,2001:138–152

[7] Harvey AC,Peters S. Estimation procedures for structural time series models[J]. J Forec,1990,9(1):89–108

[8] 仇伟杰. 基于状态空间模型与 KALMAN 滤波的中国电力需求分析[J]. 工业技术经济,2006,25(2):63–66

[收稿日期]　2014–01–09