· 调查报告与分析 ·

痢疾疫情预测中残差自相关性识别与修正*

申铜倩[1], 刘文东[2], 胡建利[2], 朱叶飞[2], 鲍昌俊[2], 汤奋扬[2], 彭志行[1], 陈峰[1]

关键词: 痢疾; 时间序列; 残差自相关性; 科克兰内 – 奥克特法
中图分类号: R 181.2　**文献标志码**: A　**文章编号**: 1001-0580(2013)09-1368-02　DOI: 10.11847/zgggws2013-29-09-38

痢疾是《中华人民共和国传染病防治法》中规定的乙类传染病, 是夏秋季常见的急性肠道传染病, 其流行范围广、传播快、发病率高, 危害人类健康[1]。研究痢疾的发病规律与影响因素对痢疾的疫情进行预测和分析可以为痢疾的预防控制、资源配置提供科学依据。痢疾发病数的时间序列中包含大量的信息, 是建模和预测的主要依据。时间序列资料往往具有长期趋势、季节性、周期性、短期波动和不规则变动等特点[2-3], 同时其另一个重要特点就是序列自相关性。忽略序列自相关性会导致参数估计不再具有方差最小性, 有可能使本来无效应的因素误判为有效应[4]。本研究通过对 1990—2009 年江苏省传染病疫情直报系统中南京市痢疾发病资料进行分析, 旨在识别序列自相关性, 并对其进行修正, 使拟合效果达到最优, 为痢疾的防治提供依据。结果报告如下。

资料与方法　(1) 资料来源: 资料来源于江苏省传染病疫情直报系统中 1990—2009 年南京市痢疾发病人数资料, 同时查阅南京市统计年鉴[5]收集居民年人均消费水平(万元/人)和年平均温度(℃)作为影响因素, 探讨其与痢疾发病人数之间的关系。(2) 统计分析: 采用 Eviews 5 软件进行统计分析。残差自相关性的识别包括图示法和 D – W 检验法。①图示法: 计算 ε_t 与 ε_{t-1}, 然后绘制 ε_t 与 ε_{t-1} 的坐标图, 若 ε_t 与 ε_{t-1} 的图形存在系统反映, 则可以判断随机项 ε_t 可能存在自相关。②D – W 检验: D – W 检验是 Durbin 和 Watson 于 1951 年提出的一种检验序列自相关的方法, 通过建立 D – W 检验统计量 d 判断序列自相关性[6]。本研究采用科克兰内 – 奥克特法对模型进行修正。其思想是进行一系列地反复迭代计算, 寻找出一个更好的估计值 p, 或者直到消除序列自相关为止, 所以又称为迭代法。基本步骤如下: ①利用样本数据资料, 运用普通最小二乘法估计模型, 得样本回归函数 $\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t$。②计算残差 $e_t = Y_t - \hat{Y}_t = Y_t - (\hat{\beta}_1 + \hat{\beta}_2 X_t)$。③由残差一阶自回归函数 $e_t = \rho e_{t-1} + \varepsilon_t$ 用普通最小二乘法求 $\hat{\rho}_1 = \frac{\sum e_t e_{t-1}}{\sum e_{t-1}^2}$。④利用 $\hat{\rho}_1$ 建立一阶差分函数 $Y_t - \hat{\rho}_1 Y_{t-1} = \beta_1(1 - \hat{\rho}_1) + \beta_2(X_t - \hat{\rho}_1 X_{t-1}) + \varepsilon_t$, 用 OLS 估计并检验残差序列 $\hat{e}_t$(ε_t 的估计值)的自相关性。如果无自相关性, 则计算结束, 求出参数; 如果有自相关, 则继续计算求出 ρ 的第二次估计值 $\hat{\rho}_2 = \frac{\sum \hat{e}_t \hat{e}_{t-1}}{\sum \hat{e}_{t-1}^2}$, 再次利用 $\hat{\rho}_2$ 写出原模型的新的一阶差分函数 $Y_t - \hat{\rho}_2 Y_{t-1} = \beta_1(1 - \hat{\rho}_2) + \beta_2(X_t - \hat{\rho}_2 X_{t-1}) + \varepsilon_t$, 这样经过反复的计算估计, 可求得若干 $\hat{\rho}$ 值, 直到 $\hat{\rho}$ 值满足给定的精度为止, 一般经过这样的计算, $\hat{\rho}$ 有较高的精度, 对自相关的修正效果较好[7-8]。

结　果　结果表明, 南京市痢疾疫情随时间推移总体呈下降趋势, 在 1994 和 2004 年出现 2 个高峰期。对因变量(痢疾发病人数)和自变量(人均年消费水平、年平均温度)分别取对数值, 然后进行单变量和多变量回归分析, 3 个模型均存在残差一阶自相关, 采用科克兰内 – 奥克特法进行修正。由图示法检验, 模型 A、B、C 的趋势已经被充分提取, 残差是白噪声序列, 不存在自相关性。拟合效果分别如图 1 ~ 3 所示。对比图 1、2(单变量拟合图)和图 3(多变量拟合图)结合表 1 的结果, 得出多变量拟合效果优于单变量拟合。校正 3 个模型, 消除残差自相关性后拟合结果为: 痢疾发病人数与人均年消费水平呈负相关, 即随着消费水平的提高, 痢疾发病人数减少; 与年平均气温呈正相关, 即随着温度的升高, 痢疾发病人数增多。多变量拟合优于单变量拟合, $\bar{R}^2$ 得到改善, 拟合结果更接近实际情况, 残差平方和(sum of squares for error, SSE)最小。本研究搜集到 2010 年的年平均温度为 16.2 ℃, 年平均消费水平为 1.959 7 万元/人, 利用模型 C 来预测 2010 年的痢疾发病人数为 916 例, 实际发病人数为 920 例, 绝对误差为 4, 相对误差为 0.004 3, 拟合效果较好, 所以此模型可以用来预测痢疾的发生发展。

* **基金项目**: 国家"十二五"科技重大专项(2011ZX10004 – 902); 江苏省自然科学基金重点项目(BK2010079); 江苏省科教兴卫工程(ZX201109); 江苏省高校优势学科建设资助项目
作者单位: 1. 南京医科大学公共卫生学院流行病与卫生统计学系, 江苏 南京 211166; 2. 江苏省疾病预防控制中心
作者简介: 申铜倩(1989 –), 女, 山东聊城人, 硕士研究生, 研究方向: 传染病预测预警。
通讯作者: 陈峰, E-mail: fengchen@njmu.edu.cn

中国公共卫生 CPH

表 1　模型拟合结果

变量	模型 A				模型 B				模型 C			
	β	$S_{\bar{x}}$	t 值	P 值	β	$S_{\bar{x}}$	t 值	P 值	β	$S_{\bar{x}}$	t 值	P 值
常数项	7.59	0.10	79.43	<0.001								
年人均消费水平	−1.05	0.20	−5.29	<0.001					−1.07	0.15	−6.91	0.000
年平均温度					3.40	0.47	7.24	0.000	2.71	0.03	91.59	0.000
$AR(1)$	0.50	0.18	2.71	0.017	1.08	0.07	14.84	0.000	0.45	0.19	2.40	0.031
$\bar{R}^2$	0.90				0.88				0.91			
SSE	0.42				0.69				0.37			
DW 值	2.20				2.06				2.22			

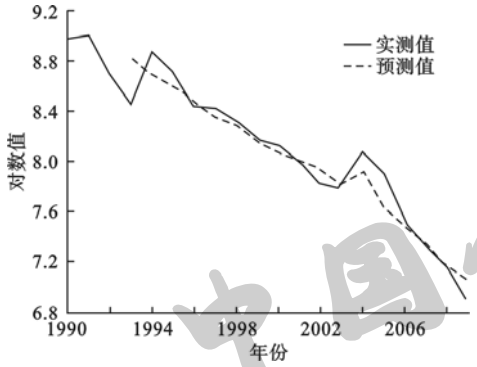

图 1　模型 A 拟合效果图

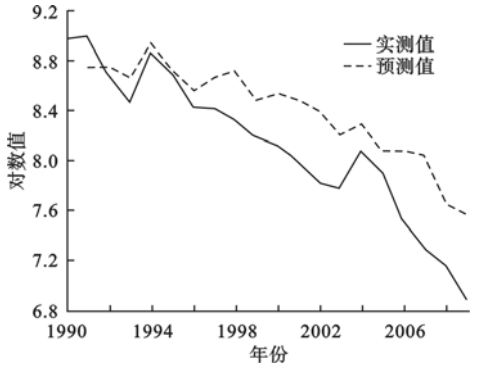

图 2　模型 B 拟合效果图

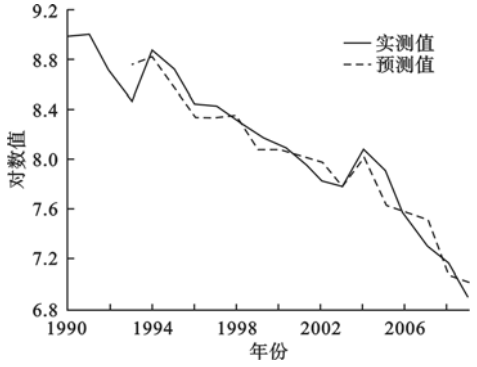

图 3　模型 C 拟合效果图

讨　论　本研究结果表明，痢疾的发病人数与居民消费水平呈负相关，与年平均温度呈正相关。这些因素对痢疾报告发病率的影响可能与其直接或间接的影响了病原体的繁殖生长以及人们的生活水平、生活方式等发生变化有关[9]。痢疾发病受气候影响较大，夏秋季高发，本研究采用指标为年平均气温，可能会低估气温的效应，因此将来可以考虑以月发病数为单位进行分析，以切实评估气候对痢疾发病的影响强度。随着南京市肠道传染病防治管理力度的逐年增强以及健康教育的广泛开展，居民生活水平逐年提高，卫生防病意识也逐步增强，痢疾发病总体呈下降趋势。本研究表明，南京市痢疾疫情在 1994 和 2004 年出现 2 个高峰期，但由于有关部门及时开展了防治工作，因此避免了痢疾对居民健康和经济损失造成的威胁。

痢疾的发病除了受自然和社会因素影响外，还与人体自身免疫力、病原、相关社会保健水平、卫生习惯及意识等因素有关；卫生统计工作与疾病预防控制相结合，纳入更多影响因素，因此采用组合预测的方法寻找最优的预测模型是未来的工作方向。本研究根据现有的痢疾发病时间序列数据定量挖掘该段时间内痢疾发病的演变规律，并预测其短期未来的发展趋势，同时为了提高预测精度，通过新的痢疾发病数据加以补充，采用科克兰内－奥克特法对模型进行自相关性的修正，效果较好；且此方法软件操作较为简单，可加以推广应用。

参考文献

[1]　于新源，刘晓辉. 大连开发区 2005—2009 年细菌性痢疾疫情分析[J]. 中国公共卫生，2011，27(10)：1357.
[2]　Qian Y, Su J, Shi P, et al. Attempted early detection of influenza A(H1N1) pandemic with surveillance data of influenza-like illness and unexplained pneumonia[J]. Influenza Other Resp, 2011, 5(6): 479－486.
[3]　朱猛，祖荣强，霍翔，等. 时间序列分析在流感疫情预测预警中的应用[J]. 中华预防医学杂志，2011，45(12)：1108－1111.
[4]　余松林，彭晓武. 广义加性模型配合时间序列资料时消除残差自相关性的一种方法[J]. 中国卫生统计，2010，27(5)：450－454.
[5]　南京市统计局. 南京统计年鉴[M]. 南京：南京市统计局，2010：4，15.
[6]　刘晓冬，姜宝法. 残差自回归模型在中国人口出生率预测中的应用[J]. 中国卫生统计，2011，28(3)：309－311.
[7]　Bramness JG, Walby FA, Tverdal A, et al. The sales of antidepressants and suicide rates in Norway and its counties 1980－2004[J]. Journal of Affective Disorders, 2007, 102(1): 1－9.
[8]　Quantitative Micro Software. Eviews 5 Command and Programming Reference[M]. USA: Quantitative Micro Software, 2004: 286－295.
[9]　贾蕾，曹卫华，贺雄，等. 北京市痢疾发病率影响因素分析[J]. 中国公共卫生，2007，23(8)：1004－1006.

收稿日期：2013－03－28　　（郭薇编辑　张翠校对）