·基础理论与方法·

应用Empower Stats软件分析阈值效应

林林 陈常中 余晓丹

【导读】 生物医学研究中许多研究因素在一定范围内对结果变量无影响或有正效作用，超过某一阈值后，作用大小或/和方向可发生变化，称为阈值效应。在分析研究因素(x)对结果变量(y)的作用有无阈值效应时，可先通过平滑曲线拟合观察是否有分段线性关系，然后采用分段回归模型、LRT检验和Bootstrap重抽样法进行阈值效应分析。美国X & Y Solutions软件公司开发的Empower Stats软件，设有阈值效应分析模块，可输入阈值后按所给阈值分段模拟数据，也可以不输入阈值，由软件自动确定最佳阈值模拟数据，并计算阈值置信区间。

【关键词】 阈值效应；分段式回归模型；LRT检验；Bootstrap重抽样

The analysis of threshold effect using Empower Stats software *LIN Lin*[1], *CHEN Chang-zhong*[2], *YU Xiao-dan*[3]. 1 *Department of Epidemiology, School of Binzhou Medical College, Yantai* 264003, China; 2 *Dana-Farber Cancer Institute, Harvard Medical School, Boston*, USA; 3 *Shanghai Key Laboratory of Children's Environmental Health, Xinhua Hospital Affiliated to School of Medicine, Shanghai Jiaotong University*

Corresponding author: *YU Xiao-dan*, Email: xdyu1108@163.com

【Introduction】 In many studies about biomedical research factors influence on the outcome variable, it has no influence or has a positive effect within a certain range. Exceeding a certain threshold value, the size of the effect and/or orientation will change, which called threshold effect. Whether there are threshold effects in the analysis of factors (x) on the outcome variable (y), it can be observed through a smooth curve fitting to see whether there is a piecewise linear relationship. And then using segmented regression model, LRT test and Bootstrap resampling method to analyze the threshold effect. Empower Stats software developed by American X & Y Solutions Inc has a threshold effect analysis module. You can input the threshold value at a given threshold segmentation simulated data. You may not input the threshold, but determined the optimal threshold analog data by the software automatically, and calculated the threshold confidence intervals.

【Key words】 Threshold effect; Segmented regression model; Likelihood ratio test; Bootstrap resampling method

生物医学研究中许多研究因素对结果变量的影响不是简单的直线关系，即在一定范围内无作用或有正效作用，超过某一阈值后作用大小或/和方向发生变化，称为阈值效应。以下介绍阈值效应分析原理及Empower Stats软件进行阈值效应分析的应用。

基本原理

1. 阈值效应检验：如果要检验研究因素x对结果变量y的作用，在k点存在转折。可以首先定义2个自变量x_1、x_2（当$x<k$，$x_1=x$，$x_2=0$；$x\geq k$，$x_1=0$，$x_2=x$），然后采用如下回归方程拟合数据[1]

$$y=\alpha+\beta_1 x_1+\beta_2 x_2+\varepsilon \quad (1)$$

当$x<k$时，式(1)简化为$y=\alpha+\beta_1 x_1+\varepsilon$；当$x\geq k$时，式(1)简化为$y=\alpha+\beta_2 x_2+\varepsilon$。因此$\beta_1$、$\beta_2$分别表示$x<k$与$x\geq k$两段的回归系数，即$y$随$x$变化的斜率。如果这两段斜率的差异无统计学意义，式(1)则简化为

$$y=\alpha+\beta x+\varepsilon \quad (2)$$

统计检验可以用*Wald*检验比较$\beta_2-\beta_1$是否等于零，或用似然比检验(LRT)[2]比较式(1)与式(2)；如*Wald*检验结果$\beta_2-\beta_1\neq 0$，或LRT检验结果式(1)显著优于式(2)，则表示x在阈值k前后对y的效应不同。

2. 确定阈值：一般通过平滑曲线拟合判断是否存在阈值[3]，观察危险因素x与结局变量y有无分段式的关系，然后采用最大似然法确定阈值。所谓最大似然法，即从数据中寻找一个k值，其所得出的式(1)给出最大的似然值。具体操作分两步。

第一步：从x的5%百分位数开始，按5%逐步递增，到95%百分位数，共19个点，分别赋k值为这19个点的x值，找出哪个百分位数给出最大的似然值，

DOI: 10.3760/cma.j.issn.0254-6450.2013.011.021

作者单位：264003 烟台，滨州医学院流行病学教研室（林林）；美国哈佛大学医学院Dana-Farber癌症研究所（陈常中）；上海交通大学医学院附属新华医院 上海市环境与儿童健康重点实验室（余晓丹）

通信作者：余晓丹，Email：xdyu1108@163.com

记为p_1，并分别找出$p_1-4\%$与$p_1+4\%$百分位数所对应的x值，记为k_{min}、k_{max}，应将k值缩小到该范围内。

第二步：用递归方法在k_{min}至k_{max}之间的所有观察到的x取值内，找出哪个x作为k值给出最大似然值。具体方法是首先比较该范围内的Q1（25%百分位点）、Q2（50%百分位点）与Q3（75%百分位点），找出哪个位点作为k值所给出的模型似然值最大，然后把范围缩小到该位点前后25%范围内，这样每次递归剔除50%的x取值。最终得出能给出最大的模型似然值的k值。

3. 确定阈值置信区间（*CI*）：采用Bootstrap重抽样方法确定阈值*CI*[4]。即从现有数据中随机抽样，重新抽取一个同样样本量的数据i。抽出的个体再放回，这样在抽取的数据中，原数据中有些个体可能未被抽中，也有些个体可能被抽出多次。对抽取的数据进行阈值分析，记录所得阈值k_i。这样重复随机抽样1000次，计算出1000个阈值k_i。再计算这1000个k_i的2.5%与97.5%百分位数，即为所观察阈值的95%*CI*。

4. Empower Stats软件在阈值分析上的应用：上述确定阈值及其*CI*的过程复杂，工作量大，且需很高的编程技巧。美国X & Y Solutions软件公司开发的Empower Stats软件，设计有阈值效应分析模块，用户不需自己编程就可方便快捷完成上述分析。在分析研究因素（x）对结果变量（y）的作用有无阈值效应时，可先通过该软件的平滑曲线拟合模块，观察x与y的曲线拟合是否有分段线性关系，还可从曲线上观察出阈值。如有分段线性关系，再使用阈值效应模块。可以输入阈值，该软件按所给阈值分段拟合数据；也可不输入阈值，让软件按前述方法自动确定阈值，并计算阈值的*CI*。

实例分析

【例1】　锰摄入过量可对机体产生不良影响。出生前低浓度的锰暴露对新生儿生长的影响尚不清楚。2008—2009年Yu等[5]对上海市1377对母婴进行多中心研究，测定胎儿脐血中锰浓度、新生儿出生身长和BMI等指标。鉴于血锰浓度呈强偏态分布，对锰浓度做常用对数变换后，再进行分析。研究者使用Empower Stats软件，首先绘制对数锰浓度（log Mn）与新生儿身长、BMI的曲线拟合图，发现新生儿BMI与log Mn曲线分两段，第一段随着log Mn增加呈下降趋势，后一段则呈上升趋势（图1），根据曲线观察并结合其他数据分析结果，log Mn定义转折点为0.7 μg/L。第二步，调用阈值效应分析模块分析胎儿脐带血中的锰浓度与新生儿BMI关系（表1）。结果显示，当脐血log Mn＜0.7时（即锰浓度＜5.0 μg/L），回归系数为-0.233；当脐血log Mn≥0.7时（即锰浓度≥5.0 μg/L），回归系数为0.115，两回归系数差别的*Wald*检验，$P<0.001$。如果不用分段线性模型，用一条线拟合数据，得出回归系数为0.049，$P=0.070$，但该直线回归系数显然不能正确反映胎儿脐血锰浓度与新生儿BMI关系。两模型比较的LRT检验$P<0.001$。

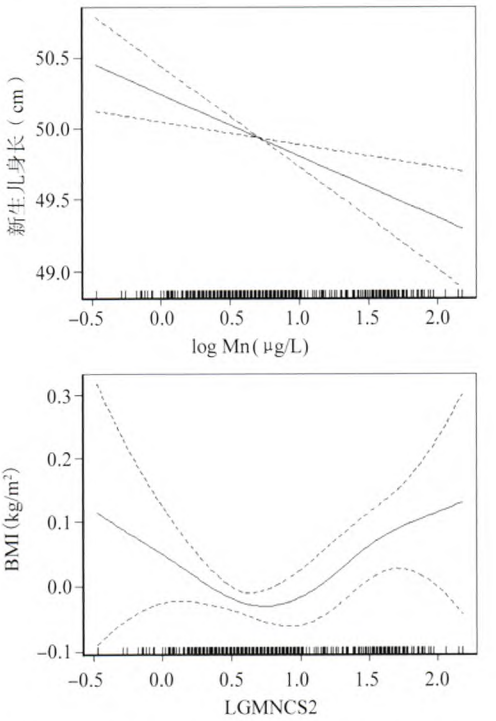

图1　孕妇产前锰暴露与新生儿生长的剂量-反应关系平滑曲线拟合

表1　新生儿脐血log Mn与BMI关系的阈值效应分析

项　目	BMI		
	β	95%*CI*	*P*值
模型Ⅰ			
一条直线回归系数	0.049	(-0.004, 0.101)	0.070
模型Ⅱ			
折点（*k*）	0.700		
＜*k*时回归系数1	-0.233	(-0.407, -0.059)	0.009*
＞*k*时回归系数2	0.115	(0.049, 0.180)	＜0.001*
回归系数2与1之差	0.348	(0.143, 0.552)	＜0.001*
对数似然比检验		*P*＜0.001*	
折点处*y*预测值		2.670(2.65, 2.70)	

注：*$P<0.05$

【例2】　黄爱群等[6]1996年对安徽省795名农村人群采用问卷调查分析家庭成员间年龄与SBP/DBP的关系。Empower Stats阈值效应分析模块输出结果见图2、表2。当年龄＜39.6岁时，年龄对SBP升高无显著作用（回归系数＝0.1，$P=0.274$）；当年龄≥39.6岁时，年龄每增加1岁，SBP升高1.3 mm Hg（回归系数＝1.3，$P<0.001$）。比较

分段线性模型与一条线模型的似然比检验结果$P<0.001$，表明年龄对SBP升高作用的阈值为39.6（95%*CI*：18.5～64.2）岁。同理，当年龄<44.6岁时，年龄每增加1岁，DBP升高的风险增加0.1 mm Hg（回归系数=0.1，$P=0.007$），而当年龄≥44.6岁时，年龄每增加1岁，DBP升高0.4 mm Hg（回归系数=0.4，$P<0.001$）。比较分段线性模型与一条线模型的似然比检验结果$P=0.003$，表明年龄对DBP明显升高作用阈值为44.6（95%*CI*：18.3～64.3）岁（表2）。

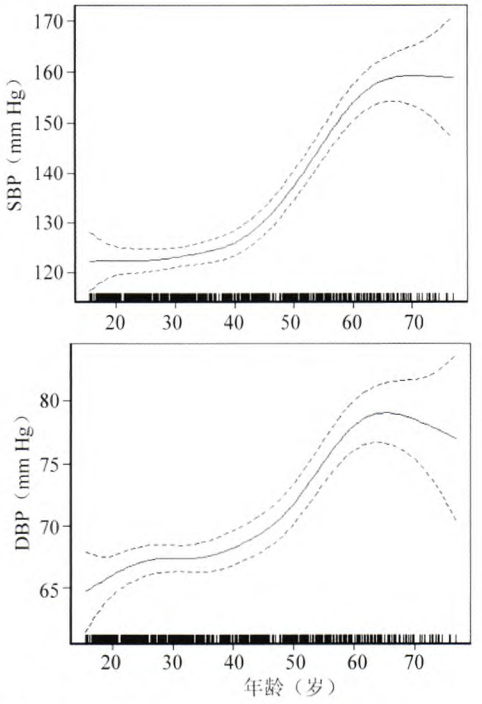

图2 年龄与SBP/DBP关系的平滑曲线拟合

讨 论

现代医学研究中常应用阈值效应，但经常被忽视。如本文例1，如不采用分段模型，可得出新生儿BMI随脐血锰浓度增加的线性关系（$\beta=0.049$，$P=0.07$）；而采用分段线性模型则更确切反映两者的分段关系，即当脐血锰浓度<5.0 μg/L，新生儿BMI随脐血锰浓度增加而减小；但脐血锰浓度≥5.0 μg/L时，新生儿BMI随脐血锰浓度增加而增加。阈值效应分析也可用于确定高危人群。如本文例2中结果显示年龄对SBP升高作用的阈值为39.6（95%*CI*：18.5～64.2）岁；年龄对DBP明显升高作用的阈值为44.6（95%*CI*：18.3～64.3）岁。根据该结果，可选定>39岁为高血压重点监测人群。

确定阈值最确切的方法是分别计算每个x（除去最小和最大的值以便分段模拟数据）作为k值给出的模型似然值进行比较，但计算量大且费时，当样本量大又需要做Bootstrap重抽样计算*CI*时，其可行性差。递归法首先比较Q1（25%百分位点）、Q2（50%百分位点）、Q3（75%百分位点）3个点给出的模型似然值，把范围缩小到最大点前后25%内，每次递归缩小一半，最终找出k值。直接用递归法效率最高，当曲线拟合波动性小，k值比较居中时，效果较好；但当k值较偏向两端，曲线拟合波动较大时，不容易找出最佳k值。本文提出的分两步方法，兼顾到k值偏向两端与曲线拟合波动较大的情况，同时又充分利用到递归法的高效率。本文主要介绍两段线性模型分析x与y的关系，所用原理与方法可扩展到分析x与y之间呈三段或多段线性关系的数据。

使用传统的如SAS、R等统计分析软件确定阈值及其*CI*，均需要通过复杂的编程才能实现。使用Empower Stats软件，用户只要输入结果变量（y）、研究因素（x）、要调整的协变量（如需要调整其他变量的作用）、分层变量（如需要分层分析），软件自动编写R程序，实现上述分析过程，输出内容中不仅包括所编写的R程序及其运算过程的详细记录，并把最终分析结果整理成表格（如例1所示），方便用户阅读。该软件大大提高了用户数据分析能力与分析效率。

表2 年龄与SBP/DBP的阈值效应分析

项 目	SBP			DBP		
	β	95%*CI*	*P*值	β	95%*CI*	*P*值
模型Ⅰ						
一条直线回归系数	0.8	(0.7,0.9)	<0.001[a]	0.3	(0.2,0.3)	<0.001[a]
模型Ⅱ						
折点（k）	39.6	(18.5,64.2)		44.6	(18.3,64.3)	
<k时回归系数1	0.1	(-0.1,0.3)	0.274	0.1	(0.0,0.2)	0.007[a]
>k时回归系数2	1.3	(1.1,1.5)	<0.001[a]	0.4	(0.3,0.6)	<0.001[a]
回归系数2与1之差	1.1	(0.8,1.5)	<0.001[a]	0.3	(0.1,0.5)	0.003[a]
对数似然比检验		$P<0.001$[a]			$P=0.003$[a]	
折点处y预测值	124.63	(122.03,127.23)		69.33	(67.84,70.83)	

注：[a]同表1

参 考 文 献

[1] Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag, 2001.

[2] Chen XR. Advanced Mathematical Statistics. Hefei: China Science and Technology University Press, 2009. (in Chinese)
陈希孺. 高等数理统计学. 合肥：中国科学技术大学出版社，2009.

[3] Cleveland WS. Robust locally weighted fitting and smoothing scatterplots. J Am Stat Assoc, 1979, 74: 829-836.

[4] Bradley E, Robert T. An introduction to the bootstrap. New York: Chapman & Hall Ltd, 1993.

[5] Yu XD, Cao LL, Yu XG, et al. Elevated cord serum manganese level is associated with a neonatal high ponderal index. Environ Res, 2012, 121: 79-83.

[6] Huang AQ, Liu X, Zhang WB, et al. The analysis of the risk factors about hypertension among rural population. Chin J Dis Control Prev, 1999, 3 (2): 82-83. (in Chinese)
黄爱群，刘学，张文兵，等. 农村地区居民高血压危险因素分析. 疾病控制杂志，1999，3(2)：82-83.

（收稿日期：2013-07-10）
（本文编辑：张林东）