广义可加模型及其 SAS 程序实现

冯国双[1]　陈景武[2]

回归分析中，非参数回归以其适用性强，对模型假定要求不严等优点，扩展了参数回归的应用范围，增强了模型的适应性[1]。但非参数回归也有其局限性，当模型中的解释变量个数较多而样本含量并不是很大时，非参数回归拟合的效果并不尽如人意，容易引起方差的急剧增大。这种由于维度的增加而使方差急剧扩大的问题通常被称为“维度的祸害(curse of dimensionality)”。而且非参数回归多是建立在核估计和光滑样条基础上的，其解释性也是一个问题。为了解决这些问题，Stone(1985)提出了可加模型(additive models)[2]，这种模型对多变量回归方程估计一个可加近似值，可加近似值有两个优点：①由于每一个个体的可加项是以单变量平滑估计的，因而“维度的祸害”可以避免。②个体项的估计解释了应变量如何随着自变量的变化而变化的。为了使可加模型扩展到更广范围的分布族，Hastie 和 Tibshirani(1990)又提出了广义可加模型(generalized additive models, GAM)[3]。它使反应变量的均值通过一个非线性连接函数而依赖于可加解释变量，同时还允许响应概率分布为指数分布族中的任意一员。许多广泛应用的统计模型均属于广义可加模型，包括带正态误差的经典线性模型、二分类数据的非参数 logit 模型、Poisson 数据的非参数对数线性模型等。

可加模型和广义可加模型

设 Y 为反应变量，$X_1, X_2, X_3, \cdots, X_P$ 为解释变量，经典的线性回归模型一般可表示为如下形式：

$E(Y \mid X_1, X_2, K, X_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + K + \beta_P X_P$　　其中，$\beta_0, \beta_1, \beta_2, \cdots, \beta_P$ 的参数估计一般通过最小二乘法来获得。

可加模型推广了线性模型，其形式为：

$E(Y \mid X_1, X_2, K, X_p) = s_0 + s_1(X_1) + s_2(X_2) + K + s_P(X_P)$

式中，$s_i(X_i)$，$i = 1, 2, \cdots$，称为光滑函数，它满足 $E\, s_i(X_i) = 0$。这一函数并不给定一个参数形式，而是以非参数形式来估计。

广义可加模型与广义线性模型相似，它包括一个随机成分(random component)，一个可加成分(additive component)以及一个联系于这两个成分的连接函数(link function)。

反应变量 Y，即随机成分，服从下面的指数分布族：

$$f_Y(y; \theta; \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

其中，θ 被称为自然参数，ϕ 被称为尺度参数。

可加成分为 $\eta = s_0 + \sum_{i=1}^{p} s_i(X_i)$

连接函数 $g(\cdot)$ 将随机成分与可加成分联系成 $g(\mu) = \eta$。例如，普通的可加模型中，$\eta = g(\mu) = \mu$，是恒等函数；而二分类数据的非参数 logit 模型中，$\eta = g(\mu) = \log \frac{\mu}{1-\mu}$，为 logit 函数。

广义可加模型和广义线性模型可用于相似的情形，但它们的目的不同。广义线性模型强调模型中参数的估计和推断，而广义可加模型更加注重对数据进行非参数性的探索。当研究目的是想对数据进行探索性分析或显示反应变量和解释变量之间关系时，用广义可加模型更为合适。

广义可加模型的 SAS 实现

广义可加模型在 SAS 中可通过 GAM 程序来实现[4]，在 SAS 8.1 版本中 GAM 程序是作为一个试验性程序嵌入的，SAS 8.2 及以后版本中已经成为一个正式程序。GAM 程序建立在非参数回归和平滑技术的基础上，提供了一批功能强大的数据分析工具。当要采用非参数方法来分析一个反应变量与多个解释变量之间关系或反应变量不服从正态分布时，便可以采用 GAM 程序来实现。

GAM 程序的主要语句如下；

PROC GAM<option>；

CLASS variables;

MODEL dependent=<PARAM(effects)>

　　Smoothing effects</options>；

SCORE data=SAS-data-set out=SAS-data-set;

OUTPUT<out=SAS-data-set> keyword<···keyword></option>；

BY variables;

1. 北京大学临床肿瘤学院流行病研究室(100036)

2. 山东省潍坊医学院卫生统计教研室(261042)

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

ID variables;

FREQ variable;

上述语句中，CLASS、OUTPUT、BY、ID、FREQ与其他SAS程序一致。这里主要介绍一下MODEL语句和SCORE语句。

利用MODEL语句可以拟合不同形式的模型，如参数模型、半参数模型、非参数模型、可加模型等。表1表明了对于反应变量 Y 和解释变量 x、x_1、x_2 如何指定不同的模型。

表1 拟合GAM模型的不同语句

模型类型	语句	数学形式
参数(Parametric)模型	Model y= *param*(x)	$E(y)=\beta_0+\beta_1 x$
非参数(Nonparametric)模型	Model y= *spline*(x)	$E(y)=\beta_0+s(x_2)$
半参数(Semiparametric)模型	Model y= *param*(x_1)*spline* x_2)	$E(y)=\beta_0+\beta_1 x_1+s(x_2)$
可加(Additive)模型	Model y= *spline*(x_1) *spline*(x_2)	$E(y)=\beta_0+s_1(x_1)+s(x_2)$
薄板样条(Thin-plate spline)模型	Model y= *spline*(x_1, x_2)	$E(y)=\beta_0+s(x_1, x_2)$

在MODEL语句中，有几个较为重要的选项这里稍做介绍：

DIST=distribution-id选项用指定模型所选用的分布，分布族可以选择GAUSSIAN或LOGISTIC。在PROC GAM过程中，对于每一个分布族，只有典型连接才被执行。即当选择GAUSSIAN时，其典型连接函数为恒等函数(identity function)，这时广义可加模型就是可加模型。当选择LOGISTIC时，其典型连接函数为 $g(\mu)=\log\frac{\mu}{1-\mu}=\eta$。除此之外，不能选择其他连接函数。

METHOD=GCV选项指定光滑参数的值由广义交叉确认(generalized cross validation, GCV)来选择。在PROCGAM中，利用GCV来作为选择光滑参数的标准。

SCORE语句用于输出预测值。其中，data=SAS-data-set指明要预测的数据集，out=SAS-data-set指明预测值输出的数据集。

GAM拟合广义可加模型举例

本文以Sockett等(1987)的一项研究为例[5]来说明如何利用GAM实现广义可加模型。该研究目的是为了探索血清C肽水平与年龄、碱缺乏之间的关系。反应变量为C肽浓度的对数，解释变量为年龄(age)和碱缺乏(basedeficit)。

对反应变量的正态性进行检验，结果发现反应变量不服从正态分布(Shapiro-Wilk值为0.944，P=0.035)。从图1和图2的散点图也可以发现，解释变量与反应变量之间很难确定是一种什么关系，因而用线性回归分析可能并不恰当，而且本例中共有2个解释变量，故采用GAM程序拟合广义可加模型。

```
data dgam;
input age basedeficit cpeptide;
loqcp=log(cpeptide);
cards;
5.2-8.1 4.8
```

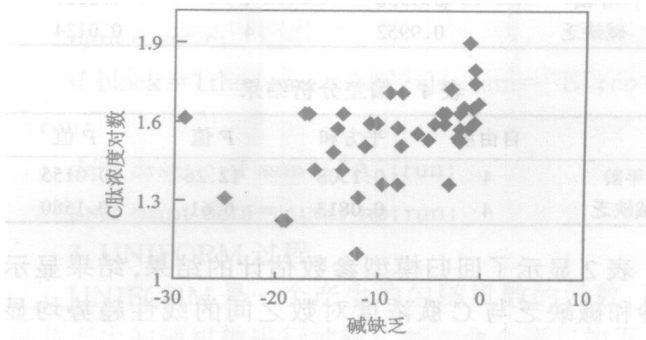

图1 碱缺乏与C肽浓度对数的散点图

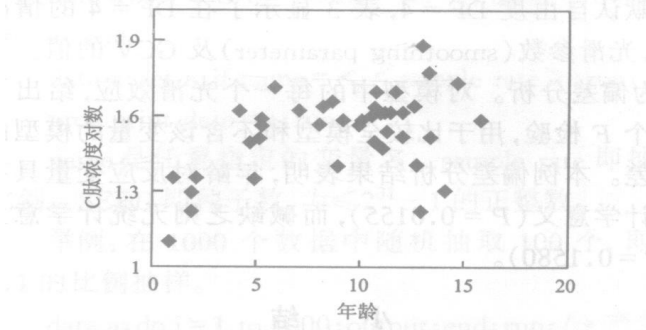

图2 年龄与C肽浓度对数的散点图

```
8.8-16.1 4.1
……
10.8-13.5 5.1
;
proc gam;
model logcp=spline(age) spline(basedeficit);
score data=dgam out=pred;
proc print data=pred;
run;
```

本例中包含两个解释变量，MODEL语句表明采用光滑样条拟合年龄和碱缺乏对反应变量logcp的效应，其自由度默认为4。SCORE语句表示对数据集dgam中的变量进行预测，预测值存放于数据集pred中(由于本例数据较多，故没有显示SCORE语句指定的预测值结果)程序输出结果主要包括两部分。第一部分描述了模型的基本情况，如拟合的算法、采用的分布及连接函数等，本例采用的是后退拟合(back-fit-

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ting)算法(Friedman and Stuetzle, 1981), 拟合的最终残差平方和为 0.4181。采用的分布为 GAUSSIAN 分布, 相应的连接为恒等连接函数。第二部分是统计分析结果, 它包含了普通回归模型的参数估计以及采用广义可加模型的参数估计, 结果见表 2、3 和 4。

表 2　回归模型分析结果

	参数估计	标准误	*t* 值	*P* 值
年龄	0.0144	0.0044	3.28	0.0024
碱缺乏	0.0081	0.0024	3.35	0.0020

表 3　光滑成分的统计量拟合

	光滑参数	自由度	GCV
年龄	0.9956	4	0.0117
碱缺乏	0.9952	4	0.0124

表 4　偏差分析结果

	自由度	平方和	*F* 值	*P* 值
年龄	4	0.1508	12.26	0.0155
碱缺乏	4	0.0813	6.61	0.1580

表 2 显示了回归模型参数估计的结果, 结果显示年龄和碱缺乏与 C 肽浓度对数之间的线性趋势均显著。在 GAM 中, 若没有指定 METHOD=GCV 选项, 则默认自由度 DF=4, 表 3 显示了在 DF=4 的情况下, 光滑参数(smoothing parameter)及 GCV 的值。表 4 为偏差分析。对模型中的每一个光滑效应, 给出了一个 *F* 检验, 用于比较全模型和不含该变量的模型的偏差。本例偏差分析结果表明, 年龄对反应变量具有统计学意义($P=0.0155$), 而碱缺乏则无统计学意义($P=0.1580$)。

小　结

广义可加模型具有非参数模型的诸多优点, 如放宽了线性条件的要求, 适用于任意分布的资料等。当反应变量与解释变量之间的具体依存关系不明确、反应变量的分布不易判定或不符合所要求的分布, 而解释变量的个数大于 1 时, 可以采用广义可加模型。如本例中, 反应变量不服从正态分布, 且很难判断反应变量与解释变量之间确切的依存关系, 传统的线性模型假定条件不满足, 故可采用广义可加模型。从分析结果可以看出, 采用广义可加模型分析, 碱缺乏实际上无统计学意义。而如果采用参数回归分析, 则年龄和碱缺乏均具有统计学意义, 从而得出错误的结论。

关于非参数回归的理论和应用, 国内医学领域已有介绍[6-8], 但大多是针对一个解释变量的情形, 当模型中的解释变量个数增多时, 其拟合效果会由于"维度的祸害"问题而随之降低。在以往关于非参数回归应用的文章中, 由于其计算繁琐, 因而多采用编程实现, 这在一定程度上限制了实际中的推广应用。SAS8.0及以前的版本中, SAS/INSIGHT 可进行光滑样条回归、kernel 回归等的拟合, 但仅限于一个解释变量。

SAS8.1及以后版本中的 GAM 程序则克服了上述缺点。GAM 程序具有一系列的优点: 提供可加模型的非参数估计; 可对多个解释变量进行非参数估计; 可拟合广义半参数可加模型和广义可加模型; 可通过指定模型自由度或光滑参数选择特定的模型, 等等。相比于以往利用编程实现而言, 应用 GAM 程序更为简洁方便, 且结果解释也较为容易, 有利于实际中的推广应用。

参　考　文　献

1. 陈长生, 徐勇勇, 夏结来. 医学研究的非参数同归分析方法. 中国卫生统计, 2002, 19(1): 56-59.
2. Stone CJ. Additive Regression and Other Nonparametric Models. Annals Of Statistics, 1985, 13: 689-705.
3. Hastie TJ, Tibshirani RJ. Generalized Additive Models (with di Sus Sion). Statistical Science, 1990, 1: 297-328.
4. SAS Institute Inc. SAS/STAT 9.1 User's Guide, SAS Institute Inc, 2004, 1557-1607.
5. Sockett EB, Daneman D, Clarson C, et al. Factors Affecting and Patterns Of Residual Insulin Secretion During the First Year of Type I (Insulin Dependent) Diabetes Mellitus in Children. Diabet, 1987, 30: 453.459.
6. 张宏培, 常荣芬. 非参数 Monotonic 同归及其应用. 中国卫生统计, 1997, 14(3): 1-4.
7. 何大卫, 徐勇勇. 非参数同归在医学中的应用. 山西医学院学报, 1995, 26(2): 94-96.
8. 陈长生, 徐勇勇, 夏结来. 光滑样条非参数回归方法及其医学应用. 中国卫生统计, 1999, 16(6): 342-345.

?1994-2014 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net