适应性回归分析(Ⅲ)——构建具有混合结构的回归模型

罗艳虹[1,2],胡良平[2,3]*

(1.山西医科大学公共卫生学院卫生统计学教研室,山西 太原 030001;
2.世界中医药学会联合会临床科研统计学专业委员会,北京 100029;
3.军事科学院研究生院,北京 100850
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是通过两个实例介绍适应性回归样条算法对具有混杂结构的数据集进行回归建模的实践。当一个数据集包含多个不同的回归模型时,只要给定分类变量的具体取值,SAS 中的"ADAPTIVEREG 过程"可以比较准确地发掘出其内在规律,并以图形方式呈现模型对资料的拟合效果,还能呈现由"基函数"及其组合所构造出的"回归模型"。图形呈现的结果确实可以给分析者提供一些有价值的"分析线索"或"积极暗示",缩小"探索性研究的空间"。

【关键词】 混合结构;适应性;样条;回归分析;基函数;失拟;拟合优度
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.02.003

Adaptive regression analysis(Ⅲ)——building the regression models with the mixed structures

Luo Yanhong[1,2], *Hu Liangping*[2,3]*

(1. *Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan* 030001, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*
* *Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【Abstract】 This paper was to introduce the application of the building regression model adopted the adaptive regression spline algorithm on the data sets with mixed structure through two examples. When a data set contained several distinct regression models, the ADAPTIVEREG procedure in SAS software could discover the internal rules as long as the specific values of the classification variables were given, and could present the fitted results of the model with graphs and the regression model constructed with the basis functions and their combinations. The information presented in the graphs could provide some valuable analytical clues or positive hints for the analysts, which greatly reduced the space of the exploratory research.

【Keywords】 Mixed structure; Adaptability; Spline; Regression analysis; Basis function; Lack of fit; Goodness of fit

1 基于多个回归模型采用 ADAPTIVEREG 过程实现一次性拟合

1.1 问题与模型

【实例 1】假定在自变量的定义域内,可以构造出三个回归模型,它们分别为指数回归模型、对数回归模型和直线回归模型[1-2]。这三个回归模型的表达式见式(1)。

$$y=\begin{cases}\exp[5(x-0.3)^2], & \text{如果 } c=0\\ \log(x-x^2), & \text{如果 } c=1\\ 7x, & \text{如果 } c=2\end{cases} \quad (1)$$

是否可以通过一个 SAS 过程步来一次性拟合出上述三个回归模型?

1.2 利用 ADAPTIVEREG 过程实现上述要求

1.2.1 所需要的 SAS 程序

```
data Mixture;
drop i;
do i=1 to 1000;
    X=ranuni(1);
    C=int(3*ranuni(1));
if C=0 then Y=exp(5*(X-0.3)**2)+rannor(1);
else if C=1 then Y=log(X*(1-X))+rannor(1);
else Y=7*X+rannor(1);
output;
end;
run;
```

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

```
ods graphics on;
proc adaptivereg data = Mixture plots = fit;
class c;
model y = c x;
run;
```

1.2.2 SAS 数据步程序说明

在 SAS 数据步程序中,拟产生 1 000 个观测(即样本含量 N = 1 000);将自变量 X 设置为在“0 ~ 1”区间上变化且服从均匀分布的随机变量;首先将分类变量 C 设置为在“0 ~ 1”区间上变化且服从均匀分布的随机变量,然后将变量 C 乘以 3,最后再将其取整(这样做的目的是使“C”成为随机变量,而不是一般变量,也就是说,它在数据集中的取值仍为 0 ~ 2,但不是按确定性的顺序出现的,而是随机出现的);接下来按式(1)进行计算,得出 C 在取不同值条件下的因变量 y 的数值。应注意:在因变量 y 的每个数值上,还加上了一个服从 N(0,1)分布的随机变量的数值,其意义在于:因变量 y 也是一个随机变量,而不是一个一般变量。

1.2.3 SAS 过程步程序说明

调用“ADAPTIVEREG 过程”,在过程步语句中,要求绘制图形;使用“CLASS 语句”,指定分类变量为“C”;在“MODEL 语句”中,包含了两个自变量,一个为变量 C、另一个为变量 X。

1.2.4 SAS 主要输出结果及解释

拟合统计量

GCV	1.08046
GCV R - Square	0.90279
Effective Degrees of Freedom	25
R - Square	0.90740
Adjusted R - Square	0.90628
Mean Square Error	1.04064
Average Square Error	1.02711

以上为拟合统计量的计算结果,R^2 和调整 R^2 分别为 0.90740 和 0.90628,说明模型对资料的拟合效果比较好。

向后选择后的回归样条模型

名称	系数	父级	变量	结点	水平
Basis0	5.3829		Intercept		
Basis1	-4.3871	Basis0	C		10
Basis3	32.7761	Basis0	C		1
Basis5	20.2859	Basis4	X	0.7665	
Basis7	-11.4183	Basis2	X	0.7665	
Basis8	-7.0758	Basis2	X	0.7665	
Basis9	58.4911	Basis3	X	0.5531	
Basis10	-71.6388	Basis3	X	0.5531	
Basis11	-69.0764	Basis3	X	0.04580	
Basis13	-119.71	Basis3	X	0.9526	
Basis15	66.5733	Basis1	X	0.9499	
Basis17	6.6681	Basis1	X	0.5143	
Basis19	-185.21	Basis1	X	0.9890	

以上为“向后选择后的回归样条模型”中“各基函数”及其回归系数,以“基函数”为新“自变量”的适应性回归模型比式(1)更复杂。

ANOVA 分解

功能性成分	基数	DF	变化量(若忽略)	
			失拟	GCV
C	2	4	1112.50	1.1519
C X	10	20	3773.94	3.7690

以上为“方差分析分解”的计算结果,对因变量 y 影响较大的是“C”与“X”之间的交互作用项,其次是变量 C。

变量重要性

变量	基数	重要性
C	12	100.00
X	10	50.68

以上为两个变量“C”与“X”对因变量 y 的重要性的计算结果,可以看出:变量 C 对因变量 y 的影响最大,其次是变量 X。

拟合结果用图示法呈现,见图 1。

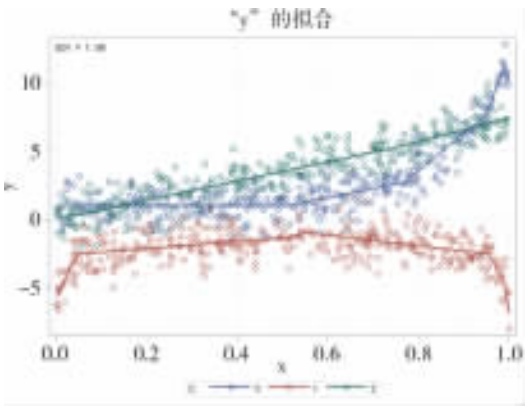

图 1 ADAPTIVEREG 过程按式(1)拟合的结果

由图 1 可知:自上而下有三条线,第 1 条为“直线”,对应式(1)中第 3 式;第 2 条为“指数曲线”,对应式(1)中第 1 式;第 3 条为“对数曲线”,对应式(1)中第 2 式。

2 基于具有混合结构的数据集采用 ADAPTIVEREG 过程实现一次性拟合

2.1 问题与数据结构

【实例 2 】假定有一个具有混合结构的数据集。见表 1。

表 1 一个具有三类(C = 1、2、3)结构不同的混合型数据集

x	y	C	x	y	C	x	y	C
1	0.555	1	1	−115	2	1	−29.998	3
2	1.295	1	2	−70	2	2	−29.993	3
3	2.751	1	3	−35	2	3	−29.982	3
4	11.116	1	4	−10	2	4	−29.950	3
5	24.879	1	5	5	2	5	−29.865	3
6	43.476	1	6	10	2	6	−29.632	3
7	69.297	1	7	5	2	7	−29.000	3
8	97.037	1	8	−10	2	8	−27.282	3
9	114.631	1	9	−35	2	9	−22.611	3
10	121.645	1	10	−70	2	10	−9.914	3
11	124.412	1	11	−115	2	11	24.598	3
12	125.619	1	12	−170	2	12	118.413	3

注:在 C = 1、2、3 类的数据集中,x 的取值均为 1 ~ 12,但 y 的取值是不同的

【问题】试在每一类中,构建 y 依赖 x 变化而变化的回归模型。

2.2 试采用 ADAPTIVEREG 过程直接拟合该数据集

2.2.1 创建 SAS 数据集

所需要的 SAS 数据步程序如下:

```
Data a1;
    INPUT x y c @@;
CARDS;
此处输入表 1 中 12 行 6 列数据;
;
RUN;
```

2.2.2 调用 ADAPTIVEREG 过程建模

所需要的 SAS 过程步程序如下:

```
ods graphics on;
proc adaptivereg data = a1 plots = fit;
    class c;
model y = c x;
run;
```

2.2.3 显示 SAS 主要分析结果

拟合统计量

GCV	132.61471
GCV R − Square	0.97285
Effective Degrees of Freedom	21
R − Square	0.99501
Adjusted R − Square	0.99302
Mean Square Error	33.15368
Average Square Error	23.02339

以上为“拟合统计量”的计算结果。由 R^2 和调整 R^2 的计算结果可知,模型对资料的拟合效果比较好。

向后选择后的回归样条模型

名称	系数	父级	变量	结点	水平
Basis0	87.2987		Intercept		
Basis4	-11.0742	Basis1	x	10.0000	
Basis6	-39.7160	Basis2	x	6.0000	
Basis7	-107.33	Basis0	c		2
Basis9	44.6297	Basis7	x	10.0000	
Basis10	9.5651	Basis7	x	10.0000	
Basis11	-22.0541	Basis2	x	9.0000	
Basis13	15.7856	Basis8	x	6.0000	
Basis15	-25.0254	Basis2	x	3.0000	
Basis17	49.1853	Basis7	x	11.0000	
Basis19	-13.9451	Basis8	x	8.0000	

以上为“向后选择后的回归样条模型”的计算结果,需要用到 19 个“基函数”。

ANOVA 分解

功能性成分	基数	DF	变化量(若忽略)	
			失拟	GCV
C	1	2	12807	1565.97
C X	9	18	163389	5296.08

以上为“方差分析分解”的计算结果,说明变量 C 和“C”与“X”的交互作用项对因变量 y 的影响很大。

变量重要性

变量	基数	重要性
C	9	100.00
X	10	93.90

以上是对两个变量的重要性所做的评价,两个变量对于因变量 y 的影响都很大。

由于模型的表达式非常复杂且不直观,SAS 采用图形方式呈现模型拟合结果。见图 2。

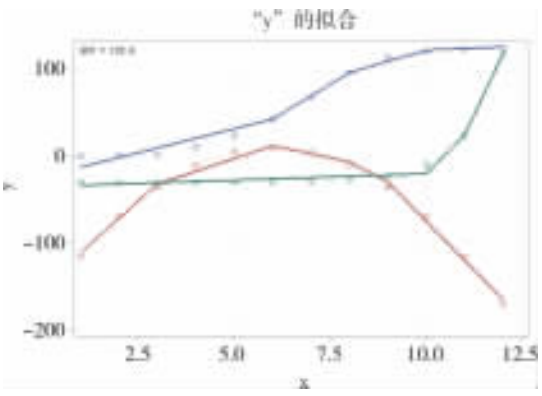

图 2 采用 ADAPTIVEREG 过程拟合表 1 资料的结果以图形呈现

在图 2 中,可以比较清楚地看出:在“C = 1”类中,y 与 x 之间呈现“Logistic 曲线”关系;在“C = 2”类中,y 与 x 之间呈现“抛物线”关系;在“C = 3”类中,y 与 x 之间呈现“指数曲线”关系。

2.3 数据结构的揭秘

在表 1 的“C = 1”类中,(x,y)的两列数据来自文献[3],该资料描述的是“某县疟疾发病的季节性特点”,即某县 1961 年 - 1996 年疟疾的月累计发病率(x 代表 1 月 - 12 月,y 代表“累计发病率”,单位为“1/10 万”)。绘制该资料的散布图,呈现“Logistic 曲线”变化趋势,适合拟合“Logistic 曲线回归模型”。

在表 1 的“C = 2”类中,(x,y)的两列数据中的“x”保持不变,而“y”列数据是采用如下的式(2)计算出来的:

$$y = -5 \times (x-6)^2 + 10 \quad (2)$$

式(2)表达的是一个 y 关于 x 的“二次抛物线模型”。

在表 1 的“C = 3”类中,(x,y)的两列数据中的“x”保持不变,而“y”列数据是采用如下的式(3)计算出来的:

$$y = e^{(x-7)} - 30 \quad (3)$$

式(3)表达的是一个 y 关于 x 的“指数曲线模型”。

结合上面图 1 中呈现的“三条曲线”及其解释,不难发现:适应性回归样条算法给出的结果与数据所代表的真实模型是基本吻合的。

3 讨论与结论

3.1 讨论

“实例 1”与“实例 2”看起来有所不同,前者似乎

是从“模型”出发,产生“数据”,再用“ADAPTIVEREG 过程”去拟合数据;而后者似乎是从“数据”出发,采用“ADAPTIVEREG 过程”去拟合数据,再交代各类数据所代表的“模型”。其实,二者在本质上是完全一样的。对于“ADAPTIVEREG 过程”而言,它并不知晓正在拟合的“数据”究竟包含了“哪几种模型”或存在“哪些客观规律”,只是基于“特定类中两变量之间的数量关系”并依据“适应性回归样条算法”去逐一构造“基函数”,在“失拟(LOF)”和“广义交叉验证(GCV)”等的“拟合优度评价指标”的“监控”之下,找到“基函数”及其组合。

3.2 结论

适应性回归样条算法(由 ADAPTIVEREG 过程实现)确实具有一定的揭示“混杂结构数据集”中隐藏的数据规律的“能力”;然而,它给出的基于“基函数”的“回归模型”过于复杂且很不直观;通过“图形方式”呈现的结果虽然很直观,但很不“精确”,不同的分析者可能会给出不同的“解读结果”。但是,图形呈现的结果确实可以给分析者提供一些有价值的“分析线索”或“积极暗示”,有利于分析者缩小“探索性研究的空间”[4]。

参考文献

[1] SAS Institute Inc. STAT SAS 9.3 User's Guide[M]. Cary, NC: SAS Institute Inc, 2011: 69-106.
[2] Friedman J. Multivariate adaptive regression splines[J]. Ann Stat, 1991, 19(1): 1-67.
[3] 胡良平. 科研设计与统计分析[M]. 北京: 军事医学科学出版社, 2012: 403-411.
[4] 谷恒明, 胡良平. 基于经典统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 7-11.

(收稿日期:2019-04-10)
(本文编辑:陈 霞)

(上接第 104 页)

3.2 结论

适应性回归样条算法(即由“ADAPTIVEREG 过程”来实现)并不是“万能的”,它仅适合于数据结构中确实包含了“具有某种联系的变量集合”,而并不适合于“因变量与自变量之间不存在任何数量联系”的数据结构。

参考文献

[1] SAS Institute Inc. STAT SAS 9.3 User's Guide[M]. Cary, NC: SAS Institute Inc, 2011: 69-106.
[2] Friedman J. Multivariate adaptive regression splines[J]. Ann Stat, 1991, 19(1): 1-67.
[3] 胡良平. 回归建模的基础与要领(Ⅲ)——变量状态与相互间关系[J]. 四川精神卫生, 2018, 31(6): 498-502.
[4] 谷恒明, 胡良平. 基于经典统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 7-11.
[5] 胡良平, 黄国平. 医学科研设计方法与关键技术[M]. 成都: 四川大学出版社, 2017: 1-16.

(收稿日期:2019-04-10)
(本文编辑:陈 霞)