适应性回归分析(Ⅱ)——排除噪声变量的干扰

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】 本文目的是通过分析一个带有 8 个噪声变量的数据集,揭示适应性回归模型的实际应用价值。在数据集包含两个自变量与因变量有密切数量联系的前提条件下,适应性回归模型受噪声变量的影响接近于零;在数据集包含一个自变量与因变量有密切数量联系的前提条件下,适应性回归模型受噪声变量的影响较大,其分析结果出现了一定程度的"失真";在数据集包含零个自变量与因变量有密切数量联系的前提条件下,适应性回归模型受噪声变量的影响非常大,其分析结果是完全不可信的。得出的结论是:适应性回归分析模型不是万能的,其结果的可信度取决于数据集中是否真正包含"客观存在的规律性"。

【关键词】 适应性;样条;回归分析;基函数;噪声变量;失拟;拟合优度;重要性
中图分类号:R195.1 **文献标识码**:A doi:10.11886/j.issn.1007-3256.2019.02.002

Adaptive regression analysis(Ⅱ)——eliminating the disturbing of the noise variables

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 revealed the practical application value of the adaptive regression model through a data set with eight noise variables. Under the premise that the data set had a close quantitative relationship between two independent variables and the dependent variable, the influence of the noise variables on the adaptive regression model was close to zero. Under the preconditions that the data set had a close quantitative relationship between an independent variable and the dependent variable, the adaptive regression model was remarkably affected by the noise variables, and the results showed a certain degree of "distortion". Under the premise that the data set did not have a close quantitative relationship between the independent variables and the dependent variable, the adaptive regression model was greatly affected by the noise variables, and the results of the analysis were completely unreliable. The conclusion was reached as follows: the adaptive regression analysis model was not universal, and the credibility of the results came from the approach mentioned above depended on whether the data set truly contained "the regularity of objective existence".

【**Keywords**】 Adaptability; Spline; Regression analysis; Basis function; Noise variable; Lack of fit; Goodness of fit; Importance

1 一个人工生成的数据集

1.1 生成数据的构想

生成包含一个因变量和 10 个连续型自变量的模拟数据集,样本含量 $N=400$。生成的方法如下:

第一,每个连续型自变量都是从一个均匀分布总体 U(0,1)中独立抽样产生的,它们分别被命名为 $x_1 \sim x_{10}$。

第二,因变量 y 仅由两个连续型自变量 x_1 和 x_2 按式(1)计算而得到:

$$y=\frac{40\exp\{8[(x_1-0.5)^2+(x_2-0.5)^2]\}}{\exp\{8[(x_1-0.2)^2+(x_2-0.7)^2]\}+\exp\{8[(x_1-0.7)^2+(x_2-0.7)^2]\}} \quad (1)$$

在给定了连续型自变量 x_1 和 x_2 的每一对数值后,将它们代入式(1),并且,基于标准正态分布 N(0,1)添加误差而生成真实模型。把样本含量设定为 $N=400$[1]。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1.2 用 SAS 生成上述数据的方法

1.2.1 生成包含 11 个变量及其 400 个观测值所需要的 SAS 程序

```
data artificial;
drop i;
array x{ 10 };
do i = 1 to 400;
do j = 1 to 10;
        x{ j } = ranuni( 1 );
end;
    y = 40 * exp( 8 *(( x1 - 0.5 ) * *2 +( x2 - 0.5 )
        * *2 ))/
        exp( 8 *(( x1 - 0.2 ) * *2 +( x2 - 0.7 )
        * *2 )) +
        exp( 8 *(( x1 - 0.7 ) * *2 +( x2 - 0.2 )
        * *2 ))) + rannor( 1 );
output;
end;
run;
```

1.2.2 输出数据集前 10 个观测所需要的 SAS 程序

```
proc print data = artificial( obs = 10 );
    var x1 - x10 y;
run;
```

1.2.3 输出数据集前 10 个观测

Obs	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	0.18496	0.97009	0.39982	0.25940	0.92160	0.96928	0.54298	0.53169	0.04979	0.06657	-0.08496
2	0.85339	0.06718	0.95702	0.29719	0.27261	0.68993	0.97676	0.22651	0.68824	0.41276	0.39759
3	0.47579	0.84499	0.63452	0.59036	0.58258	0.37701	0.72836	0.50660	0.93121	0.92912	2.07362
4	0.39104	0.47243	0.67953	0.16809	0.16653	0.87110	0.29879	0.93464	0.90047	0.56878	9.09919
5	0.51132	0.43320	0.17611	0.66504	0.40482	0.12455	0.45349	0.19955	0.57484	0.73847	8.26384
6	0.52238	0.34337	0.02271	0.71289	0.93706	0.44599	0.94694	0.71290	0.10327	0.17517	4.99275
7	0.42071	0.07174	0.35849	0.71143	0.18985	0.14797	0.56184	0.27011	0.32520	0.56918	2.61939
8	0.91744	0.52584	0.73182	0.90522	0.57600	0.18794	0.33133	0.69887	0.12156	0.18067	1.06047
9	0.42137	0.03798	0.27081	0.42773	0.82010	0.84345	0.87691	0.26722	0.30602	0.39705	4.63711
10	0.54340	0.61257	0.55291	0.73591	0.37186	0.64565	0.55718	0.87504	0.57124	0.75677	7.93675

1.3 数据结构的特点

$x_1 \sim x_{10}$都是在“0 ~ 1”之间取值且服从均匀分布的随机变量,它们之间是互相独立的;y 是在依据式(1)计算结果的基础上,添加一个服从“均值为 0、方差为 1”的标准正态分布随机变量的取值(或称为误差)。显然,11 个变量都是计量的,且 y 仅依赖于 x_1 和 x_2 两个变量,独立于“$x_3 \sim x_{10}$”这 8 个变量。

1.4 回归分析的目的

【实例 1】基于前述的数据集,试建立 y 依赖于 $x_1 \sim x_{10}$的多重回归模型。

【实例 2】基于前述的数据集,试建立 y 依赖于 x_1 和 $x_1 \sim x_{10}$的多重回归模型(即丢弃 x_2)。

【实例 3】基于前述的数据集,试建立 y 依赖于 $x_2 \sim x_{10}$的多重回归模型(即丢弃 x_1)。

【实例 4】基于前述的数据集,试建立 y 依赖于 $x_3 \sim x_{10}$的多重回归模型(即丢弃 x_1 和 x_2)。

2 利用 ADAPTIVEREG 过程建模[1-2]

2.1 对实例 1 进行适应性回归分析

2.1.1 所需要的 SAS 过程步程序

```
ods graphics on;
proc adaptivereg data = artificial plots = fit;
model y = x1 - x10;
run;
```

2.1.2 SAS 输出结果及解释

拟合统计量

GCV	1.55656
GCV R - Square	0.86166
Effective Degrees of Freedom	27
R - Square	0.87910
Adjusted R - Square	0.87503

Mean Square Error	1.40260
Average Square Error	1.35351

以上为“拟合统计量”的计算结果，模型对资料的拟合优度界值 GCV = 1.55656；R^2 和调整 R^2 分别为 0.87910 和 0.87503；均方误差和平均平方误差分别为 1.40260 和 1.35351。

向后选择后的回归样条模型

名称	系数	父级	变量	结点
Basis0	12.3031		Intercept	
Basis1	13.1804	Basis0	x1	0.05982
Basis3	-23.4892	Basis0	x2	0.1387
Basis4	-171.03	Basis0	x2	0.1387
Basis5	-86.1867	Basis3	x1	0.6333
Basis7	-436.86	Basis4	x1	0.5488
Basis8	397.18	Basis4	x1	0.5488
Basis9	11.4682	Basis1	x2	0.6755
Basis10	-19.1796	Basis1	x2	0.6755
Basis13	126.84	Basis11	x1	0.6018
Basis14	40.8134	Basis11	x1	0.6018
Basis15	22.2884	Basis0	x1	0.7170
Basis17	-53.8746	Basis12	x1	0.2269
Basis19	598.89	Basis4	x1	0.2558

以上为“向后选择后的回归样条模型”的计算结果。此结果中涉及到很多“基函数(Basis)”，而基函数中的“元素”基本上只有“x_1”“x_2”以及由它们以不同的系数联系起来的“交互作用项”。

ANOVA 分解

功能性成分	基数	DF	变化量(若忽略)	
			失拟	GCV
x_1	2	4	405.18	1.1075
x_2	2	4	947.87	2.6348
x_2 x_1	9	18	2583.21	6.6187

以上是基于“方差分析分解”的算法对所构建的模型进行逐项分解的结果。其中，涉及到“x_1”的基函数有 2 个，占用了 4 个自由度，其对应的“失拟”LOF = 405.18，GCV = 1.1075；涉及到“x_2”的基函数有 2 个，占用了 4 个自由度，其对应的 LOF = 947.87，GCV = 2.6348；涉及到“x_1”与“x_2”交互作用项的基函数有 9 个，占用了 18 个自由度，其对应的 LOF = 2583.21，GCV = 6.6187。

【说明】在上面的输出结果中，最后两列的顶端“变化量(若忽略)”，其含义是：若忽略掉各行上的“项”(第 1 行为“x_1”、第 2 行为“x_2”、第 3 行为“$x_1 \times x_2$”)，将会使“失拟(LOF)”或“广义交叉验证(GCV)”发生改变的数量大小，此“变化量”越大，表明对应行上的“项”对因变量的影响越大。

变量重要性

变量	基数	重要性
x_1	11	100.00
x_2	11	99.19

以上结果表明：x_1 与 x_2 对因变量 y 的重要性接近相等，分别为 100.00%、99.19%。

因变量 y 关于 x_1 与 x_2 的二次曲面回归模型在二维直角坐标系内以“等高线”呈现出来的图形见图 1。

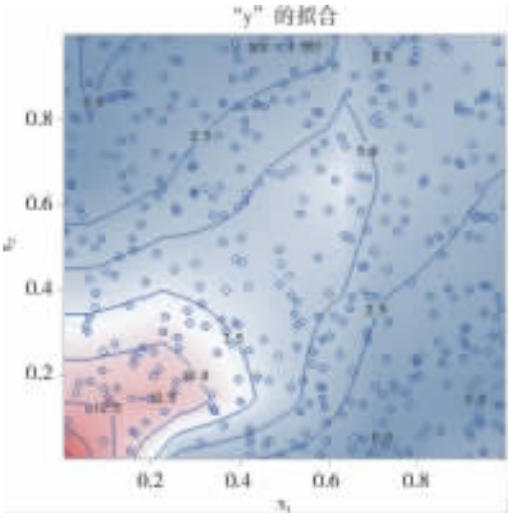

图 1 因变量 y 关于 x_1 与 x_2 的二次曲面回归模型的等高线图

图 1 是以“等高线”形式呈现式(1)所代表的二次曲面。由于式(1)属于三维空间里的二次曲面，无法采用二维平面图来呈现其立体形状。设想：采用一系列平行于二维平面的“平面”去切割三维空间里的“二次曲面”，所形成的“切口”自上而下沿垂直于纵轴 y 的方向投影到由(x_1，x_2)所形成的二维平面上，就出现了图 1 中的“曲线”。每一条曲线的高度“y”是相同的，故被称为“等高线”。等高线上标注的“数据”(例如 12.5、10.0、7.5、5.0 和 2.5 等)代表“切割平面”离“底部二维平面”的“高度”的数值。

由图 1 中多条等高线的形状可知：式(1)所代表的“二次曲面”比较复杂；若是一个“圆球”曲面模型，则其所有等高线就会形成一系列的“同心圆”。

2.2 对实例 2 进行适应性回归分析

2.2.1 所需要的 SAS 过程步程序

在前面的 SAS 过程步程序的“MODEL 语句”中,不写入“ x_2 即可。

2.2.2 SAS 输出结果及解释

下面仅给出最后一部分输出结果:

变量重要性

变量	基数	重要性
x_1	9	100.00
x_4	2	26.54
x_3	5	12.10
x_5	2	11.32
x_6	1	8.77
x_9	2	7.83
x_7	2	3.78

以上结果表明:除 x_1 真正对因变量 y 有影响外,还得出 x_4 对因变量 有较大的影响;甚至还有 x_3 和 x_5。而实际上,除 x_1 之外,其他变量对因变量 y 没有任何影响。

2.3 对实例 3 进行适应性回归分析

2.3.1 所需要的 SAS 过程步程序

在前面的 SAS 过程步程序的“MODEL 语句”中,不写入“x_1”即可。

2.3.2 SAS 输出结果及解释

下面仅给出最后一部分输出结果:

变量重要性

变量	基数	重要性
x_2	11	100.00
x_3	2	29.25
x_5	2	19.20
x_6	2	14.95
x_4	1	7.38
x_7	1	6.12
x_8	2	3.90

以上结果表明:除 x_2 真正对因变量 y 有影响外,还得出 x_3 对因变量 y 有较大的影响;甚至还有 x_5 和 x_6。而实际上,除 x_2 之外,其他变量对因变量 y 没有任何影响。

2.4 对实例 4 进行适应性回归分析

2.4.1 所需要的 SAS 过程步程序

在前面的 SAS 过程步程序的“MODEL 语句”中,不写入“x_1”和“x_2”即可。

2.4.2 SAS 输出结果及解释

下面仅给出最后一部分输出结果:

变量重要性

变量	基数	重要性
x_3	6	100.00
x_4	2	60.87
x_7	2	42.66
x_8	1	16.58

以上结果表明:在 $x_3 \sim x_{10}$这 8 个与因变量 y 毫无关系的变量中,得出:x_3 和 x_4 对因变量 y 的影响很大;x_7 和 x_8 对因变量 y 的影响也比较大。显然,这个结果是不可信的。

3 讨论与结论

3.1 讨论

基于对“实例 1”的分析结果来看,“ADAPTIVEREG 过程”对于包含多个“噪声变量”的数据结构具有很强的“甄别能力”,能够“挖掘”出“隐藏”在复杂数据结构中的“真正规律”;而基于对“实例 2”和“实例 3”的分析结果来看,“ADAPTIVEREG 过程”对于包含多个“噪声变量”的数据结构具有较强的“甄别能力”,能够“突显”出“隐藏”在复杂数据结构中的“真正规律”,但也在较大程度上受到了“噪声变量”的干扰和影响;再基于对“实例 4”的分析结果来看,“ADAPTIVEREG 过程”对于全部由“噪声变量”组成的数据结构不具有“甄别能力”。

通常,真实资料的数据结构是错综复杂的,其是否包含有变量之间的真实数量联系是未知的,比较可靠的做法是依据基本常识和专业知识尽可能找全找准与结果变量有联系的“自变量”和/或“中间变量”,并适当引入由前述提及的那些变量产生的“派生变量”[3-4]。在此基础上,尽可能使收集数据的过程受控于“标准操作规程”和“质量控制策略”[5],确保样本能很好地代表研究总体且具有足够大的样本含量。再尽可能多采用一些统计模型和技术方法去拟合数据,并基于测试数据集评估模型的拟合效果。

(下转第 109 页)

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)