适应性回归分析(Ⅳ)——与非适应性回归分析的比较

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】 本文目的是分析一个已知真实情况的资料,比较适应性回归分析与非适应性回归分析建模的效果。结论如下:当资料中存在与因变量确有关系的自变量时,ADAPTIVEREG 过程具有较好的甄别能力;REG 过程具有较好的甄别能力,但需要满足一定条件,即采用"前进法"或"逐步法"筛选自变量,同时还需要"假定模型包含截距项"。当资料中不存在与因变量确有关系的自变量时,ADAPTIVEREG 过程几乎完全失去了甄别能力;REG 过程具有较好的甄别能力,但需要满足一定条件,即采用"前进法"筛选自变量,同时还需要"假定模型包含截距项"。若研究者基于"基本常识"和"专业知识"确定的自变量都与因变量有关系,对因变量进行 Logistic 变换,并且,假定回归模型中不含截距项时,会在回归模型中保留非常多的自变量。

【关键词】 适应性回归分析;多重线性回归分析;变量变换;自变量筛选
中图分类号:R195.1 **文献标识码**:A doi:10.11886/j.issn.1007-3256.2019.02.004

Adaptive regression analysis(Ⅳ)——the comparison between the adaptive regression analysis and non-adaptive regression analysis

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 article compared the effect of adaptive regression analysis with the effect of non-adaptive regression analysis through a known real data set. The conclusions were as below: while the independent variables had a definite relationship with the dependent variable, the ADAPTIVEREG procedure had a great discriminating ability; the REG procedure had a good screening ability, but it must satisfy two conditions, the first one was that the forward selection or the stepwise selection to screen the independent variables would be used, and the second one was that the model would be assumed to contain an intercept term. While the independent variables in the data set didn't have the relationship with the dependent variable, the ADAPTIVEREG procedure almost completely lost its ability to discriminate; the REG procedure had a good screening ability, but it needed satisfy almost the same two conditions mentioned before. If the independent variables determined by researchers based on their common knowledge and the professional one were related to the dependent variable, a large number of independent variables would be retained in the regression model while the dependent variable was transformed by means of the Logistic transformation and it was assumed that there wasn't an intercept term in the regression model.

【**Keywords**】 Adaptive regression analysis; Multiple linear regression analysis; Variable transformation; Independent variable selection

1 概 述

相对于因变量是"计数变量"和"定性变量"而言,因变量为"计量变量"的回归建模方法的种类更多。其中,若按是否采用"适应性回归分析"可划分为以下两类:适应性回归分析[1-2]与非适应性回归分析[3-12]。在非适应性回归分析方法中,最常用且最有代表性的就是多重线性回归分析方法,在 SAS 软件中,可以通过 REG 过程来实现。

本文将采用 ADAPTIVEREG 过程和 REG 过程来实现对同一个数据集的回归建模,并结合数据集和数据子集的真实情况,反映并揭示两种建模思想对数据集和数据子集的建模效果。从而得出对回归建模及建模效果评价有意义的参考性建议。

2 问题与数据结构

2.1 原问题与数据集

沿用本期科研设计方法专题中第二篇文章《适应性回归分析(II)——排除噪声变量的干扰》(以

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

下简称“前文”)中的“问题与数据结构”,其数据集名为“artificial”。

2.2 新问题与数据集

在原问题中,数据集 artificial 包含一个因变量 y 及其取值,10 个在(0,1)区间上服从均匀分布的随机变量 $x_1 \sim x_{10}$ 及其取值;y 是 x_1 和 x_2 的函数,并具有“前文”式(1)的表达式。整个数据集的样本含量 N=400。随机变量 $x_3 \sim x_{10}$ 是独立于因变量 y 的,或者说,它们是与因变量 y 无关的随机变量。

2.3 扩展的数据集

2.3.1 扩展的数据集 a1

在数据集 artificial 的基础上,引入由随机变量 $x_1 \sim x_{10}$ 产生的派生变量[3],它们是由随机变量 $x_1 \sim x_{10}$ 的全部二次项组成,共 55 项(包括 10 个平方项和 45 个交叉乘积项)。连同随机变量 $x_1 \sim x_{10}$ 共有 65 个自变量,所得的数据集为 a1。所需要的 SAS 数据步程序如下:

```
data a1;
    set artificial;
z1=x1*x1;z2=x1*x2;z3=x1*x3;
z4=x1*x4;z5=x1*x5;z6=x1*x6;
z7=x1*x7;z8=x1*x8;z9=x1*x9;
z10=x1*x10;z11=x2*x2;z12=x2*x3;
z13=x2*x4;z14=x2*x5;z15=x2*x6;
z16=x2*x7;z17=x2*x8;z18=x2*x9;
z19=x2*x10;z20=x3*x3;z21=x3*x4;
z22=x3*x5;z23=x3*x6;z24=x3*x7;
z25=x3*x8;z26=x3*x9;z27=x3*x10;
z28=x4*x4;z29=x4*x5;z30=x4*x6;
z31=x4*x7;z32=x4*x8;z33=x4*x9;
z34=x4*x10;z35=x5*x5;z36=x5*x6;
z37=x5*x7;z38=x5*x8;z39=x5*x9;
z40=x5*x10;z41=x6*x6;z42=x6*x7;
z43=x6*x8;z44=x6*x9;z45=x6*x10;
z46=x7*x7;z47=x7*x8;z48=x7*x9;
z49=x7*x10;z50=x8*x8;z51=x8*x9;
z52=x8*x10;z53=x9*x9;z54=x9*x10;
z55=x10*x10;
run;
```

【说明】在上面的程序中,“z1=x1*x1”和“z11=x2*x2”分别代表 x_1 与 x_2 的平方项;“z2=x1*x2”代表 x_1 与 x_2 的交叉乘积项;$z_3 \sim z_{10}$ 代表与 x_1 有关的交叉乘积项;$z_{12} \sim z_{19}$ 代表与 x_2 有关的交叉乘积项。也就是说,$z_1 \sim z_{19}$ 或多或少与因变量 y 有一定的联系。$z_{20} \sim z_{55}$ 这 36 个变量都独立于因变量 y;另外,由前面的介绍可知,随机变量 $x_3 \sim x_{10}$ 是独立于因变量 y 的,故独立于因变量 y 的自变量共有 44 个(即 $x_1 \sim x_{10}, z_{20} \sim z_{55}$)。

2.3.2 扩展的数据集 a2

由于因变量 y 是计量变量,故可以对其进行变量变换。常规的变量变换方法有以下五种:对数变换、平方根变换、倒数变换、指数变换和 Logistic 变换[13-15]。在数据集 a1 基础上,引入对因变量 y 的上述五种变换,分别记为 y1 ~ y5,所得的数据集为 a2。所需要的 SAS 数据步程序如下:

```
data a2;
    set a1;
    y1=log(y+5);y2=sqrt(y+5);y3=1/(y+5);
    y4=exp(y+5);y5=exp(y+5)/(1+exp(y+5));
run;
```

【说明】因为在因变量 y 的取值中,出现了负数和零,不便于取对数和平方根变换,统一加上一个正数 5 即可。

数据集 a2 中包含了数据集 a1,而其又包含了数据集 artificial,故以下仅基于数据集 a2 进行计算即可。

2.3.3 数据集 a2 中变量的分类

因变量有 6 种表现形式,分别为原先的形式 y、取了不同变量变换后的形式 $y_1 \sim y_5$;自变量可分为以下两类:A 类中含有 21 个与因变量有关系的自变量,即“x_1 和 $x_2, z_1 \sim z_{19}$”;B 类中含有 44 个与因变量无关系的自变量,即“$x_3 \sim x_{10}, z_{20} \sim z_{55}$”。

3 基于 a2 数据集中 A 与 B 两类共 65 个自变量进行回归建模

3.1 采用 ADAPTIVEREG 过程对 65 个自变量进行回归建模

3.1.1 建模策略

分别选取 $y、y_1 \sim y_5$ 为因变量,利用 65 个自变量(其中,A 类 21 个自变量与因变量有关系,B 类 44 个自变量与因变量无关系),采用 ADAPTIVEREG 过程进行回归建模。

3.1.2 建模结果

建模的输出结果较多,下面仅给出自变量对因变量贡献排名前五位的自变量及反映其重要性大小的百分数。见表 1。

表 1 ADAPTIVEREG 过程对 65 个自变量进行回归建模给出前五位自变量及其重要性数值

因变量	第一(%)	第二(%)	第三(%)	第四(%)	第五(%)
y	x_1(100.00)	x_2(97.16)	z_1(27.30)	z_2(24.36)	z_{11}(17.58)
y_1	z_{11}(100.00)	z_2(95.36)	z_1(67.31)	x_2(26.99)	z_7(21.55)
y_2	z_2(100.00)	z_{11}(64.56)	z_1(59.83)	z_7(23.56)	x_1(20.80)
y_3	x_1(100.00)	z_{11}(99.70)	z_1(29.00)	z_2(15.95)	z_{42}(5.70)
y_4	z_2(100.00)	z_{39}(37.84)	z_{11}(38.81)	z_{45}(13.12)	z_{19}(10.93)
y_5	z_{42}(100.00)	z_{18}(87.75)	x_1(49.98)	z_2(31.52)	z_{33}(24.16)

由表 1 可知:分别以“y、y_1 和 y_2”为因变量时,ADAPTIVEREG 过程从 65 个自变量中提取的前五位重要的自变量全部属于 A 类中的自变量;而分别以“y_3、y_4 和 y_5”为因变量时,ADAPTIVEREG 过程从 65 个自变量中提取的前五位重要的自变量中分别有 4、3、3 个属于 A 类中的自变量,即出错数目分别为 1、2、2 个。

3.2 采用 REG 过程对 65 个自变量进行回归建模

3.2.1 建模策略

分别选取 y、y_1 ~ y_5 为因变量,利用 65 个自变量(其中,A 类 21 个自变量与因变量有关系,B 类 44 个自变量与因变量无关系),采用 REG 过程进行回归建模。具体地说,在假定模型包含截距项与不含截距项的条件下,再分别采用“前进法”“后退法”和“逐步法”筛选自变量,并记录下最终回归模型的有关重要信息。

3.2.2 建模结果

建模的输出结果较多,下面仅给出最终的回归模型中分别包含 A 类与 B 类自变量的数目。见表 2。

表 2 REG 过程对 65 个自变量进行回归建模保留 A 与 B 类自变量的数目

因变量	截距	保留 A 类自变量数目			保留 B 类自变量数目		
		前进法	后退法	逐步法	前进法	后退法	逐步法
y	有	7	6	7	0	4	0
y_1	有	7	6	7	1	4	1
y_2	有	7	6	7	0	4	0
y_3	有	5	7	5	1	4	1
y_4	有	5	5	5	0	4	0
y_5	有	7	6	5	1	8	2
y	无	9	10	9	7	17	7
y_1	无	14	13	13	15	24	13
y_2	无	14	13	13	22	23	21
y_3	无	6	8	6	1	10	1
y_4	无	7	10	5	3	4	1
y_5	无	15	16	13	20	37	20

由表 2 可知:假定回归模型中不含截距项时,保留在回归模型中的自变量数目明显增多,此时,有很多与因变量无关的自变量会被保留在最终的回归模型之中。假定回归模型中包含截距项且采用前进法或逐步法筛选自变量时,回归模型中保留与因变量无关的自变量的数目比较少,即结论的正确性较高。

4 基于 a2 数据集的 B 类中 44 个自变量进行回归建模

4.1 采用 ADAPTIVEREG 过程对 B 类中 44 个自变量进行回归建模

4.1.1 建模策略

分别选取 y、y_1 ~ y_5 为因变量,利用 B 类中 44

个自变量,采用 ADAPTIVEREG 过程进行回归建模。

4.1.2 建模结果

建模的输出结果很多,为节省篇幅,下面仅给出自变量对因变量贡献排名前五位的自变量及反映其重要性大小的百分数。见表 3。由表 3 可知:尽管 B 类中 44 个自变量独立于因变量,但 ADAPTIVEREG 过程仍以较高的"重要性"数值保留了较多的自变量。

表 3 ADAPTIVEREG 过程对 B 类中 44 个自变量进行回归建模给出前五位自变量及其重要性数值

因变量	第一(%)	第二(%)	第三(%)	第四(%)	第五(%)
y	Z_{20}(100.000)	Z_{31}(86.18)	Z_{47}(76.01)	Z_{40}(62.66)	Z_{24}(61.82)
y_1	Z_{31}(100.000)	Z_{53}(81.72)	Z_{49}(78.20)	Z_{48}(58.32)	Z_{30}(48.63)
y_2	Z_{40}(100.000)	Z_{49}(78.10)	Z_{47}(69.99)	Z_{31}(65.38)	Z_{48}(58.43)
y_3	Z_{31}(100.000)	Z_{26}(72.30)	Z_{47}(70.18)	Z_{30}(69.27)	Z_{42}(61.99)
y_4	Z_{24}(100.000)	Z_{46}(86.51)	X_7(70.54)	Z_{37}(51.15)	
y_5	Z_{32}(100.000)	Z_{42}(100.00)			

4.2 采用 REG 过程对 B 类中 44 个自变量进行回归建模

4.2.1 建模策略

分别选取 y、y_1 ~ y_5 为因变量,利用 B 类中 44 个自变量,采用 REG 过程进行回归建模。具体地说,在假定模型包含截距项与不含截距项的条件下,再分别采用"前进法""后退法"和"逐步法"筛选自变量,并记录下最终回归模型的有关重要信息。

4.2.2 建模结果

建模的输出结果较多,下面仅给出最终的回归模型中包含 B 类自变量的数目。见表 4。

表 4 REG 过程对 B 类中 44 个自变量进行回归建模保留自变量的数目

因变量	有截距时保留自变量数目			无截距时保留自变量数目		
	前进法	后退法	逐步法	前进法	后退法	逐步法
y	0	5	0	4	7	4
y_1	0	7	0	13	19	16
y_2	0	4	0	12	18	12
y_3	0	3	1	9	13	8
y_4	0	4	0	1	1	1
y_5	0	4	0	43	43	43

由表 4 可知:假定回归模型中包含截距项且采用前进法筛选自变量时,与因变量无关的自变量全部都不被保留在回归模型中,结果最可信;假定回归模型中包含截距项且采用逐步法筛选自变量时,与因变量无关的自变量几乎都不被保留在回归模型中(本例仅因变量 y_3 时有一个自变量),结果比较可信;而假定回归模型中不含截距项且采用"指数变换(因变量 y_4)"时,与因变量无关的自变量被保留在回归模型中的数目最少(本例中出现了一个);假定回归模型中不含截距项且采用 Logistic 变换(因变量 y_5)时,与因变量无关的自变量被保留在回归模型中的数目最多(本例中出现了 43 个)。

5 回归建模所需要的 SAS 程序

5.1 实现表 1 和表 3 计算的 SAS 程序

5.1.1 实现表 1 计算所需要的 SAS 过程步程序

```
proc adaptivereg data = a2;
model y = x1 - x10 z1 - z55;
quit;
```

5.1.2 实现表 3 计算所需要的 SAS 过程步程序

```
proc adaptivereg data = a2;
model y = x3 - x10 z20 - z55;
quit;
```

5.1.3 关于上述 SAS 程序的说明

将因变量 y 依次修改为“$y_1 \sim y_5$”，分别运行上面的 SAS 过程步程序。

5.2 实现表 2 和表 4 计算的 SAS 程序

5.2.1 实现表 2 计算所需要的 SAS 过程步程序

```
proc reg data = a2;
    model y = x1 - x10 z1 - z55/selection = forward sle = 0.05;
quit;
proc reg data = a2;
    model y = x1 - x10 z1 - z55/noint selection = forward sle = 0.05;
quit;
```

5.2.2 实现表 4 计算所需要的 SAS 过程步程序

```
proc reg data = a2;
    model y = x3 - x10 z20 - z55/selection = forward sle = 0.05;
quit;
proc reg data = a2;
    model y = x3 - x10 z20 - z55/noint selection = forward sle = 0.05;
quit;
```

5.2.3 关于上述 SAS 程序的说明

将“model 语句”中的“selection = ”及其后面的内容分别修改为“backward sls = 0.05;”和“stepwise sle = 0.5 sls = 0.05;”，就是采用“后退法”和“逐步法”筛选自变量；将因变量 y 依次修改为“$y_1 \sim y_5$”，分别运行上面的 SAS 过程步程序。

6 讨论与结论

6.1 讨论

统计学教科书上所讲授的、统计软件中所实现的回归分析方法，主要依据是数学原理和数理统计知识；实际工作者在运用回归分析技术时，并不知晓资料的自变量中哪些与因变量有关系、哪些与因变量无关系，全依靠回归分析技术和统计软件计算的结果来作出肯定或否定的结论。

基于本文的“数据结构”“真实情况”和“两种回归建模思路及计算结果”，可提出以下问题：在通常情况下，使用回归分析技术处理各种“真实情况未知”的试验数据时，所得到的“回归分析结果”的可信度究竟有多高？究竟应该如何提高回归分析结果的可信度？

笔者认为：对于“真实情况未知”的试验数据而言，无论采用“参数法”“半参数法”“非参数数”或所谓的“机器学习或深度学习”等方法进行回归建模，都是在“无中生有”，其整个过程都是一个“黑箱”，结果的可信度在相当大的程度上取决于资料中变量之间的真实情况。然而，当研究者对资料的真实情况一无所知时，应采取非常审慎的态度看待其分析结果。

6.2 结论

当资料中存在与因变量确有关系的自变量时，①由表 1 可知，ADAPTIVEREG 过程具有较好的甄别能力；当对因变量采取对数变换或平方根变换时，其甄别能力下降；当对因变量采取倒数变换或指数变换或 Logistic 变换时，其甄别能力下降较为明显。②由表 2 可知，REG 过程具有较好的甄别能力，但需要满足一定条件，即采用“前进法”或“逐步法”筛选自变量，同时还需要“假定模型包含截距项”。

当资料中不存在与因变量确有关系的自变量时，①由表 3 可知，ADAPTIVEREG 过程几乎完全失去了甄别能力；②由表 4 可知，REG 过程具有较好的甄别能力，但需要满足一定条件，即采用“前进法”筛选自变量，同时还需要“假定模型包含截距项”。若对因变量采取指数变换且“假定模型不含截距项”时，无论采取“前进法”“后退法”或“逐步法”筛选自变量，都具有较好的甄别能力（见表 4 中倒数第 2 行最后 3 个数据，从 44 个独立于因变量的自变量中仅错误地保留了一个自变量）。若研究者基于基本常识和专业知识确定的自变量都与因变量有关系，对因变量进行 Logistic 变换，并且，假定回归模型中不含截距项时，会在回归模型中保留非常多的自变量。此时，反映模型对资料拟合效果的 R^2 值非常接近 1、均方误差 MSE 的数值远远小于 1（因输出结果较多，未在表 2 和表 4 中呈现出来）。

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] 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.

[4] 谷恒明, 胡良平. 基于贝叶斯统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 12 - 14.

[5] 谷恒明, 胡良平. 基于机器学习统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 15 - 18.

[6] 胡良平. 基于正交化方法的回归分析[J]. 四川精神卫生, 2018, 31(3): 197 - 200.

[7] 胡良平. 稳健回归分析[J]. 四川精神卫生, 2018, 31(3): 201 - 204.

[8] 胡良平. 反应曲面回归分析[J]. 四川精神卫生, 2018, 31(3): 205 - 208.

[9] 胡良平. 加性与广义加性模型回归分析[J]. 四川精神卫生, 2018, 31(4): 289 - 295.

[10] 胡良平. 分位数模型回归分析[J]. 四川精神卫生, 2018, 31(4): 296 - 301.

[11] 胡良平. 局部模型回归分析[J]. 四川精神卫生, 2018, 31(4): 303 - 306.

[12] 胡良平. 有限混合模型回归分析[J]. 四川精神卫生, 2018, 31(4): 307 - 312.

[13] 胡良平. 提高回归模型拟合优度的策略(Ⅰ)——哑变量变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 1 - 8.

[14] 胡良平. 提高回归模型拟合优度的策略(Ⅱ)——算术均值变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 9 - 15.

[15] 胡良平. 提高回归模型拟合优度的策略(Ⅲ)——校正均值变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 16 - 20.

(收稿日期:2019 - 04 - 10)

(本文编辑:陈 霞)