·科研方法专题·

适应性回归分析(Ⅰ)——回归模型的构建与求解

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】 本文目的是介绍适应性回归模型的构建与求解方法。众所周知,在自变量数目很多时,就会出现维数灾难,此时,统计学家倾向于采用非参数回归模型取代参数回归模型。然而,当自变量数目大到一定程度时,普通的非参数回归模型也不堪重负,于是,适应性回归样条算法应运而生。此法由以下几种统计技术组成:①特殊的变量变换;②基于向前选择法构建过拟合回归模型,再基于向后选择法"修剪"回归模型;③基于"减少在向前选择的每个步骤中,检验 B、V 和 t 的组合的数目"的基本思想,实现快速算法;④借助"GCV"和"LOF"作为"拟合优度"的界值,评价已构建的回归模型的拟合效果。此法为复杂数据结构的回归建模提供了新思路。

【关键词】 适应性;样条;回归分析;基函数;结点;广义交叉验证;失拟
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.02.001

Adaptive regression analysis(Ⅰ)——the construction and solution of the regression model

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 construction and solution method of the adaptive regression models. As we all know, when the number of independent variables was large, the dimensional disaster would occur. At this time, statisticians tended to use a non-parametric regression model instead of a parametric regression model. However, when the number of independent variables was large to a certain extent, the usual non-parametric regression model was also overwhelmed, so the adaptive regression spline algorithm came into being. This method consisted of the following statistical techniques: ①the special variable transformations were used; ②the overfitted regression model was constructed based on the approach of forward selection, and then the regression model was "pruned" based on the approach of backward selection; ③based on the basic idea of "reducing the number of combinations of B, V and t in each step of the forward selection", a fast algorithm was implemented; ④using "GCV" and "LOF" as the boundary value of "goodness of fit", the fitted effect of the established regression model was evaluated. The method mentioned before provided a new way for the regression modeling of the complex data structures.

【**Keywords**】 Adaptability; Spline; Regression analysis; Basis function; Node; Generalized cross validation; Lack of fit

1 适应性回归模型

1.1 维数灾难

在回归分析中,当自变量的数目很多(往往问题本身带有很多自变量,再加上派生变量)时,模型空间就非常大,此时,建模者倾向于选择非参数模型取代参数模型。然而,当自变量数目大到一定程度时,由于自变量水平组合所形成的"试验点"在高维空间中会显得非常"稀疏",从而导致回归模型的方差迅速增大,以至于回归建模过程无法收敛或回归分析的结果失去其应有的价值,这种现象被称为"维数灾难"。

1.2 何为适应性回归分析

解决"维数灾难"问题的常用方法有以下两种:其一,将所研究的问题限于"低维空间";其二,假定建模过程具有"可加性",采用"加性模型"[1]。这两

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

种思维方法都存在一定的局限性,只是部分地或回避式地解决了“维数灾难”问题。Friedman[2]提出的“多元适应性回归样条建模技术”在一定程度上较好地解决了前述提及的难题。多元适应性回归样条建模技术被简称为“适应性回归分析方法”,此法由以下两步组成:第一步,采用“快速更新算法”创建一个“过拟合模型”,以下将被称为“向前选择”;第二步,采用“后向选择”修剪已经创建的回归模型。

1.3 适应性回归模型的形式[3]

由多元适应性回归样条算法产生的回归模型可用式(1)或式(2)表示:

$$\hat{f}(x)=\beta_0+\sum_{m=1}^{M}\beta_m B_m \quad (1)$$

$$\hat{f}(x)=\beta_0+\sum_{m=1}^{M}\beta_m\prod_{k=1}^{K_m}T_m(x_{k,m},t_{k,m}) \quad (2)$$

在式(1)和式(2)中,等号左边的“$\hat{f}$”是因变量的非参数估计值;M 是非常数“基函数”的数目;B_m 是第 m 个“基函数”;对于每个所形成的“基函数”,K_m 是交互作用的“阶数”(即“次数”);T_m 是变量变换函数,其变换方式取决于变量类型;$x_{k,m}$ 是一个变量,它代表第 m 个“基函数”中第 k 个分量;而 $t_{k,m}$ 是第 m 个“基函数”中该变量的第 k 个水平值。

式(1)或式(2)还可以用式(3)表示:

$$\hat{f}(x)=\beta_0+\sum_{i:k_m=1}f_i(x_i)+\sum_{i,j:k_m=2}f_{ij}(x_i,x_j)+\sum_{i,j,k:k_m=3}f_{ijk}(x_i,x_j,x_k)+\cdots \quad (3)$$

在式(3)中,第 2、3、4 项分别代表仅含单个自变量、含两个自变量及其交互作用项和含三个自变量及其交互作用项所形成的基函数之和。由此可知,多元适应性回归样条模型的结构非常复杂,以拟合复杂程度不同的数据结构并使之达到所期望的“拟合优度界值”。

1.4 适应性回归模型的解说

式(1)到式(3)在本质是一样的,但式(2)和式(3)的形式非常复杂,而式(1)相对简单。下面用通俗的语言解释式(1)。

所谓“多元适应性回归模型”,更确切地说,应该是“多重适应性回归模型”,因为在此回归模型中,只有一个因变量 y,其非参数估计值为“$\hat{f}$”;“多重”指有“M(≥2)”个“基函数 B_m($m=1,2,\cdots,M$)”,每一个“基函数”由若干个“原变量 x”经过某种变量变换并复合而形成。若将“基函数 B_m”简单地视为“原变量 x_m”,则式(1)就与通常的多重线性回归模型完全相同了。见式(4)。

$$\hat{y}=\hat{f}(x)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m \quad (4)$$

由此可知:构建适应性回归模型的关键在于如何构造各个“基函数”;模型求解的关键在于如何估计出式(1)或式(2)或式(3)中的回归系数。这个计算过程比较繁琐,通常需要借助统计软件(如 SAS/STAT 12.1 中的“ADAPTIVEREG”过程,此模块已嵌入 SAS 9.3 及以上版本)来完成。

2 变量变换方法

2.1 概述

在进行适应性回归建模过程中,需要对变量进行变换,而不是直接将原变量代入回归模型。对变量进行怎样的变换,取决于变量的类型。换言之,对连续型变量与分类变量将采取不同的变换方法。

2.2 连续型变量的变换

对于连续型变量,采用线性截断幂样条变换,分别见式(5)、式(6):

$$T_1(V,t)=(v-t)_+=\begin{cases}v-t,\text{如果 } v>t\\0,\text{如果 } v\leqslant t\end{cases} \quad (5)$$

$$T_2(V,t)=[-(v-t)]_+=\begin{cases}0,\text{如果 } v>t\\t-v,\text{如果 } v\leqslant t\end{cases} \quad (6)$$

其中 t 为变量 V 的结点值(或称为分割值),而 v 为其观测值。为了不用测定变量 V 的每个值,通过假设底层函数的平滑度来使用一系列的最小跨度的结点值。Friedman[2]使用以下公式来确定结点之间的合理数目(跨度大小)。对于内部结点,跨度大小由以下公式决定,见式(7):

$$-\frac{2}{5}\log_2\left[-\frac{\log(1-\alpha)}{pn_m}\right] \quad (7)$$

对于边界结点,跨度大小由以下公式决定,见式(8):

$$3-\log_2\frac{\alpha}{p} \quad (8)$$

其中,α 为决定结点密度的参数,p 为变量数,n_m 为父基函数 $B_m>0$ 的观察数目。

2.3 分类型变量的变换

对于分类变量,变量通过指示函数进行转换,分别见式(9)、式(10):

$$T_1(V,t)=(v-t)_+=\begin{cases}1,\text{如果 } v\in\{c_1,\cdots,c_t\}\\0,\text{如果 } v\notin\{c_1,\cdots,c_t\}\end{cases} \quad (9)$$

$$T_2(V,t)=[-(v-t)]_+=\begin{cases}0,\text{如果 } v\in\{c_1,\cdots,c_t\}\\1,\text{如果 } v\notin\{c_1,\cdots,c_t\}\end{cases} \quad (10)$$

其中$\{c_1,\cdots,c_t\}$为变量 V 类别的子集。这种平滑法适用于分类变量,它假设各个类别的子集具有相似的性质,类似于假设对连续变量进行局部区域的预测。

如果一个分类变量有 k 个不同的分类,那么共有($2^{k+1}-1$)种可能的子集。计算成本等于回归中所有子集的选择,对于大的 k 值来说代价较大。多元自适应回归样条算法采用逐步选择的方法选择分类,从而形成子集为{ $c_1,\cdots c_t$ }。该方法仍然是贪婪的,但它减少了计算,并产生合理的最终模型。

3 回归模型中自变量的筛选

3.1 向前选择

多元自适应样条算法的向前选择过程如下:

(1)设定初值 $B_0=1$, $M=1$;

(2)重复以下步骤,直到基函数的值达到最大,最大值为 M_{max};或者 B_m、v 和 t 三个参数的任意组合都不会使模型性能得到更好的提升。

1)设置“失拟(即模型不能表达资料变化的部分)”的界值 $LOF^*=\infty$;

2)对于筛选出的基函数 B_m, $m\in\{0,\cdots,M-1\}$ 都对变量 v 做以下操作,对于 $v\notin\{v(k,m)\mid 1\leqslant k\leqslant K_m\}$ 者除外。

①对于满足 $v:t\in\{v\mid B_m>0\}$ 的变量,每个结点值(或类别中子集)t 建立一个由当前所有选定基函数组成的模型及两个新基函数: $B_mT_1(v,t)$ 和 $B_mT_2(v,t)$;

②计算新模型 LOF 欠拟合的界值;

③如果 $LOF<LOF^*$,则更新以下变量: $LOF=LOF^*$, $m^*=m$, $v^*=v$ 和 $t^*=t$。

3)通过在模型中增加 $B_m*T_1(v^*,t^*)$ 和 $B_m*T_2(v^*,t^*)$,最大程度上更新模型。

4)设定 $M=M+2$。

每个条目最重要的部分是发现 B_m、v 和 t 之间的关系,例如在模型中添加两个相关基函数。向前选择的目标是建立一种过拟合数据的模型。线性模型的残差准则通常是残差平方和(RSS)。

3.2 向后选择

多元自适应回归样条算法的向后选择过程如下:

(1)通过设置整体的欠拟合标准来初始化: $LOF^*=\infty$;

(2)重复以下步骤,直到达到空模型为止。最后一个模型是在向后筛选过程中发现的最佳模型。

1)对于筛选的基函数 B_m, $m\in\{0,\cdots,M-1\}$:

①对于不包含 B_m,计算欠拟合标准(LOF);

②如果 $LOF<LOF^*$,模型达到最优,使 $m^*=m$;

③从当前模型中减去 B_m*。

2)设定 $M=M-1$。

向后选择的目的是“修剪”过拟合的模型,找出预测性能最好的模型。因此,使用过拟合界值来表示模型对原始数据表达的真实性是不合理的。相反,多元自适应回归样条算法使用一个类似于广义交叉验证界值的数量。更多信息参见“拟合优度界值”一节。

4 快速算法

原始的多变量自适应回归样条算法计算代价较大。为了提高计算速度,Friedman 提出了快速算法。快速算法的基本思想是减少在向前选择的每个步骤中,检验 B、V 和 t 的组合的数量。

假设有在第 k 次迭代之后形成的($2K+1$)个基,其中选择父基 B_m 来构造两个新的基。考虑一个以基为元素的队列,在队列的顶部是 B_{2k} 和 B_{2k+1} 两个新构造的基。队列的其余部分根据每个基的最小无匹配条件进行排序,排序方法见式(11):

$$J(B_i)=\min_{\substack{\text{for all eligible v}\\ \text{or all knot t}}} LOF(v,t\mid B_i),\ i=1,\cdots,2k-1 \qquad (11)$$

式(11)中,求极小值函数“min”下部有两个条件,其一,“for all eligible V”的含义是“对于所有合格的变量 V”;其二,“for all knot t”的含义是“对于所有的结点 t”。

当 k 不小时,模型中有相对较多的基,增加基的个数不太可能显著提高拟合优度。因此,在相邻迭代期间,优先队列中基的排名变化太大。候选的父基可以被限制为第一次迭代队列中的前 K 个基。第 k 次迭代之后,顶部基有新的 $J(B_i)$ 值,而底基的值不变。队列根据 $J(B_i)$ 值重新排序。这对应于 MODEL 语句中 FAST 选项的 $K=$ 选项值。

为了避免排在最后的候选基被放弃使用,并允许它们重新上升到顶部,一个自然的“老化”因素被引入到每个基。通过定义每个基函数的优先级来实现,见式(12):

$$P(B_i)=R(B_i)+\beta(k_c-k_r) \qquad (12)$$

其中 $R(B_i)$ 为队列中第 i 个基的秩, k_c 为当前迭代次数, k_r 为上次计算 $J(B_i)$ 值的迭代次数。然后根据这个优先级重新对前 K 个候选基进行排序。较大的 β 值会导致在以前的迭代中改进较小的基以更快的速度上升到列表顶部。这对应于 MODEL 语句中 FAST 选项的“BETA =”值。

对于优先级队列顶部的候选基,将重新计算($k+1$)次迭代的所有合格变量 V 的最小失拟界值 $J(B_i)$。得出的最优变量可能与前一次迭代中找到的变量相同。因此,快速多元自适应回归样条算法引入另一个因子 H 以节省计算成本。该因子指定 $J(B_i)$ 应该为所有合格变量重新计算的频率。如果 $H=1$,在考

虑父基时,每次迭代中对所有变量都进行优化。如果 $H=5$,经过 5 次迭代完成视为最优。如果小于指定 H 的迭代计数,则优化只在之前完全的优化中找到的最优变量进行。当然,有前三个候选项例外,B_{2k+1}(这是用于构建两个新基的父基 B_m)和两个新基:B_{2k} 和 B_{2k+1}。在每次迭代中执行它们的完整优化。这与 MODEL 语句中 FAST 选项的"$H=$"选项值有关。

5 拟合优度界值

与其他非参数回归过程一样,多元自适应回归样条算法可以产生复杂的模型,这些模型包含高阶交互作用项并考虑许多结点值或子集。除了基函数,向前选择和向后选择过程都是高度非线性的。考虑在偏倚与方差之间取其折中,包含多个参数的复杂模型倾向于较低偏倚而较高方差。为了选择具有良好预测性能的模型,Craven 等[4]提出了被广泛使用的广义交叉验证(GCV)界值,见式(13):

$$GCV=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i-\hat{f}_i}{1-trace(S)/n}\right]^2=\frac{RSS}{n[1-trace(S)/n]^2} \quad (13)$$

其中 y 为因变量,$\hat{f}$ 为基础光滑函数的估计值,S 为光滑矩阵,满足如下的关系式:$\hat{y}=Sy$。平滑样条的有效自由度可以定义为 trace(S)(代表光滑矩阵 S 的迹)。在多元自适应回归样条算法中,Friedman[2]使用了类似的数量作为"失拟"界值,见式(14):

$$LFO=\frac{RSS}{n\{1-[M+d(M-1)/2]/n\}^2} \quad (14)$$

其中 d 为每个非线性基函数所需要的自由度,M 为模型中线性无关基函数的总数。因为在多变量自适应回归样条算法的每个步骤中评估的任何候选模型都是一个线性模型,所以 M 实际上是冒子矩阵的迹。GCV 界值和 LOF 界值的唯一区别是额外项 d(M - 1)。相应的有效自由度被定义为 M + d (M - 1)/2。在形成新基函数时,需要考虑非线性,故引入了 d 这个数量,同时,它也作为一个平滑参数而存在。d 值越大,函数估计越平滑。Friedman[2]认为 d 值一般为"2 ~ 4"。对于结构复杂的数据,d 值可以更大。用户也可以使用交叉验证作为拟合优度界值,或使用各自的验证数据集来选择模型和单独的测试数据集来评估选定的模型。

参考文献

[1] 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.

[2] Friedman J. Multivariate adaptive regression splines[J]. Ann Stat, 1991, 19(1): 1 - 67.

[3] SAS Institute Inc. STAT SAS 9. 3 User's Guide[M]. Cary, NC: SAS Institute Inc, 2011: 69 - 106.

[4] Craven P, Wahba G. Smoothing noisy data with spline functions - estimating the correct degree of smoothing by the method of generalized cross - validation[J]. Numerical Mathematics, 1979: 31(4): 377 - 403.

(收稿日期:2019 - 04 - 10)
(本文编辑:陈 霞)

科研方法专题策划人——胡良平教授简介

胡良平, 男, 1955 年 8 月出生, 教授, 博士生导师, 曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授; 现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长,《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 48 部, 参编统计学专著 10 部;发表第一作者学术论文260余篇,发表合作论文 130 余篇, 获军队科技成果和省部级科技成果多项; 参加并完成三项国家标准的撰写工作; 参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中, 为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学, 在全国各地作统计学学术报告 100 余场, 举办数十期全国统计学培训班, 培养多名统计学专业硕士和博士研究生。近几年来, 参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的"八性"和"八思维"的统计学思想, 独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。