DOI: 10.13203/j.whugis1998.03.023
第 23卷第 3期 武汉测绘科技大学学报 Vol.23 No.3
1998年 9月 Journal of Wuhan Technical University of Surveying and Mapping Sept. 1998

教育测量数据的多元自建模模型分析

尚 钢

(武汉工业大学研究生处,武汉市珞狮路 14号,430070)

摘 要 将关于形状不变的自建模回归模型推广到多元自变量,解决了计算问题,并结合教育测量数据给出了实例。该模型既可以反映多条曲线走势的共性,又可以反映每条曲线之间的相互关系及其个性。

关键词 教育测量;多元自变量;自建模模型

分类号 O212

教育测量一般指对学生的学习能力、学业成绩、兴趣爱好、思想品德及教育措施等教育现象进行数量化测定的一门教育科学。传统教育测量包含学科、智力、品德、人格等多方面测量内容。最常见的形式莫过于考试,但它的内涵远不止于考试。

面对浩繁的测量评估数据,如何对它进行深加工,以揭示深层次的问题,把握它们的共性与个性,已成为亟待探讨的课题。把握共性,寻找一般的平均走向,莫过于回归。线性回归方法简单,但未必切合实际;非线性回归可以拟合曲线走向,但其函数形式仍需事先给定;非参数回归的函数形式不必事先给定,模型形式与计算都复杂多了,但是只能拟合一条曲线,可以反映共性,不能反映个性。目前国内统计界对于非参数回归模型的理论有比较深刻的研究。

Lawton等人第一次提出一种既能反映共性又能反映个性的自建模回归模型(Self-Modeling Regression, SEMOR)。Kneip等就模型识别问题作了深入探讨,并提出迭代算法[1],Kneip等又对此模型进行了改进[2]。但上述研究中,自建模回归模型的自变量仍是一元的。

本文将自建模回归模型的自变量推广至多元,实现程序计算,为拟合教育测量数据以及其它实际问题数据提供一般方法,以揭示其中的共性与个性规律。本文提供的方法不止用于教育测量数据。

1 一元自变量的自建模回归模型

如果考察第 i 个公司在第 j 个时刻的产出,第 i 个股票在第 j 个时刻的价格,第 i 个学生第 j 门功课的成绩,可以建立回归模型:

$$y_{ij} = f_i(t_{ij}) + X_{ij} \quad i = 1,\cdots,n;\ j = 1,\cdots,T_i$$

这里自变量 t_{ij} 是一元的,回归函数 f_i 未知,X_{ij} 是测量误差,它属于非参数回归模型,但是它有 $i = 1,\cdots,n$ 条曲线,每个个体是一条曲线。这样的模型意义不大,它只是几条互不关联的曲线而已。

自建模回归模型考虑的是:

$$y_{ij} = f(t_{ij},\theta_i) + X_{ij}$$

$$i = 1,2,\cdots,n;\ j = 1,2,\cdots,T_i$$

这就把 n 条曲线通过参数 θ_i 联系起来。它实际是要估计一个函数族。求解这样的模型需要一定的假设。最常用的假设是形状不变(SIM),准确地说是形状类型不变,但形状参数可变。这样可产生基本类似的各个个体形状曲线。Kneip与Engel考虑的形状不变模型是:

$$f(t,\theta_i) = f_i(t) = \theta_i^{(1)}\, h\left(\frac{t-\theta_i^{(3)}}{\theta_i^{(2)}}\right) + \theta_i^{(4)}$$

如果把函数 h 取作合适的形式,那么非线性回归Gompertz模型与Logistic模型等都是它的特例。

作变换 $X = (t-\theta_i^{(3)})/\theta_i^{(2)}$,再将 X 换回 t,可得模型的新形式 $f_i(\theta_i^{(2)}t + \theta_i^{(3)}) = \theta_i^{(1)}\, h(t) + \theta_i^{(4)}$,$i = 1,2,\cdots,n$;$t \in [a^0, a^1]$。在满足正则条件

$$\frac{1}{n}\sum_{i=1}^{n}\theta_i^{(1)} = 1, \quad \frac{1}{n}\sum_{i=1}^{n}\theta_i^{(2)} = 1$$

$$\frac{1}{n}\sum_{i=1}^{n}\theta_i^{(3)} = 0, \quad \frac{1}{n}\sum_{i=1}^{n}\theta_i^{(4)} = 0$$

下,对 f_i 两端求和,可得模型简化形式:

$$h(t) = \frac{1}{n}\sum_{i=1}^{n}(\theta_i^{(2)}t + \theta_i^{(3)})$$

这样可以把 φ 看作各个个体曲线的平均,称作结构平均。下面叙述一元自变量的自建模回归模型的求解计算方法。

1)使用核函数方法估计回归函数 f_i:

收稿日期:1998-03-23。尚钢,男,38岁,副教授,现从事教育管理研究。

$$\hat{f}_i(t)=\sum_{j=1}^{T_i}\frac{1}{b}\int_{S_{i(j-1)}}^{S_{ij}}K\left(\frac{t-v}{b}\right)\mathrm{d}v\,y_{ij}$$

$$i=1,\cdots,n;\ b>0$$

其中,$S_{ij}=(t_{ij}+t_{i(j+1)})/2$;$K$是指定的核函数;$b$是窗宽。

2) 计算 h_0的迭代初值 $\hat{h}_0(t)$:

$$\hat{h}_0(t)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}_i(\hat{\theta}_{i0}^{(2)}t+\hat{\theta}_{i0}^{(3)})\quad t\in[a_0,a_1]$$

其中,参数$\hat{\theta}_{i0}^{(2)}$、$\hat{\theta}_{i0}^{(3)}$的选取方法为:先在每条曲线上找两个"结构点",一般取极值点,记作 f_{1i}、f_{2i},$i=1,2,\cdots,n$ 于是有 x_1、x_2,使

$$f_{1i}=\theta_i^{(2)}x_1+\theta_i^{(3)},\quad f_{2i}=\theta_i^{(2)}x_2+\theta_i^{(3)}$$

求和得:

$$x_1=\bar{f}_1=\frac{1}{n}\sum_{i=1}^{n}f_{1i},\quad x_2=\bar{f}_2=\frac{1}{n}\sum_{i=1}^{n}f_{2i}$$

这样 $\theta_i^{(2)}$与$\theta_i^{(3)}$的迭代初值可以确定为:

$$\hat{\theta}_{i0}^{(2)}=(\hat{f}_2-\hat{f}_1)/[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_2-\hat{f}_1)]$$

$$\hat{\theta}_{i0}^{(3)}=\hat{f}_1-\hat{\theta}_{i0}^{(2)}\cdot\frac{1}{n}\sum_{i=1}^{n}\hat{f}_1$$

3)迭代 $h=1,2,\cdots,h^*$ 次,以选取较好的$\theta_i^{(2)}$与$\theta_i^{(3)}$的估计,并改进 h的估计。每次迭代都是选取新的参数估计θ_{ih},使满足:

$$\int_{a_0}^{a_1}[\hat{f}_i(\theta_{ih}^{(2)}t+\theta_{ih}^{(3)})-\theta_{ih}^{(1)}\hat{h}_{h-1}(t)-\theta_{ih}^{(4)}]^2\mathrm{d}t=$$

$$\min_{\theta}\int_{a_0}^{a_1}[\hat{f}_i(\theta^{(2)}t+\theta^{(3)})-\theta^{(1)}\hat{h}_{h-1}(t)-\theta^{(4)}]^2\mathrm{d}t$$

然后将估计量正规化,即对 $k=1,4$,取

$$\hat{\theta}_{ih}^{(k)}=\theta_{ih}^{(k)}/\frac{1}{n}\sum_{i=1}^{n}\theta_{ih}^{(k)}$$

对 $j=2,3$,取 $\hat{\theta}_{ih}^{(j)}=\theta_{ih}^{(j)}-\hat{\theta}_{ih}^{(k)}\cdot\frac{1}{n}\sum_{i=1}^{n}\theta_{ih}^{(j)}$,令

$$\hat{h}_h(t)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}_i(\hat{\theta}_{ih}^{(2)}t+\hat{\theta}_{ih}^{(3)})\quad t\in[a_0,a_1]$$

当初值点取在结构点时,只迭代几次即可取得满意的收敛效果。

2 多元自变量的自建模回归模型

一元自变量的自建模回归模型考虑的是因变量 y与自变量 t的关系,一般 t取作时刻。可是在实际问题中更多的是希望揭示因变量 y与多元自变量 $X=(x_1,\cdots,x_p)$之间的关系。当然多元线性回归模型 $y=X\beta+\varepsilon$已被广泛地应用,现在希望引进多元自变量的非参数回归模型和多元自变量的自建模回归模型,自变量是多元的,同时又能揭示 n条曲线间的共性与个性。

我们找到了解决办法,这需要借鉴单指标回归模型或投影寻踪的技术。单指标回归模型与投影寻踪回归(Projection Pursuit Regression)的思想几乎是一样的。投影寻踪回归的模型为:

$$y=G(X)+\varepsilon$$

这里 X是多元自变量;G是待估的多元函数,是从多元到一元的映射;ε是随机误差。不过它的投影逼近解是一元函数 $g(X'\beta)$:

$$\hat{G}(X)=\hat{g}(X'\beta)$$

函数 $\hat{g}(\cdot)$的估计采取加权核估计,参数即投影方向 β的选取采用交叉核实的方法。这就与单指标回归模型 $y=g(X'\beta)+\varepsilon$思想一致了。

多自变量的自建模回归模型为:

$$y_{ij}=f_{pi}(X_{ij})+\varepsilon_{ij}=f_p(X_{ij},\theta_i)+\varepsilon_{ij}=f(X_{ij}'\beta,\theta_i)+\varepsilon_{ij}$$

$$i=1,2,\cdots,n;\ j=1,2,\cdots,T_i$$

这里 $X_{ij}=x_{ij1},x_{ij2},\cdots,x_{ijp}$是 p元自变量;$\beta=(\beta_1,\cdots,\beta_p)$是待估参数;$f_p$表示 p元函数;f表示一元函数;θ_i是模型参数。常取形状不变模型:

$$y_{ij}=\theta_i^{(1)}h\left(\frac{X_{ij}'\beta-\theta_i^{(3)}}{\theta_i^{(2)}}\right)+\theta_i^{(4)}+\varepsilon_{ij}$$

$$i=1,2,\cdots,n;\ j=1,2,\cdots,T_i$$

参数 θ_i的约束条件与上节一元模型相同。

下面分析模型的解法。

在一元自建模回归模型解法的第一步,由估计 $f_i(t)$改进为估计 $f_i(X_{ij}'\beta)$,这里参数 β与函数 f_i都是待估的。这可以使用核函数加权回归与交叉核实相结合的方法。

对于任意实数 u,任给定参数 β,当 X、y均视作随机变量时,回归模型 $y_i=f_i(X_i'\beta)+\varepsilon$的解应为 $f_i(u)=E(y_i|X_i'\beta=u)$。由于含有参数 β,应该记作 $f_i(u|\beta)$。可用核函数作出 $f_i(u|\beta)$的估计:

$$\hat{f}_i(u|\beta)=\left\{\sum_{j=1}^{T_i}y_{ij}K(u-X_{ij}'\beta)\right\}/\left\{\sum_{j=1}^{T_i}K(u-X_{ij}'\beta)\right\}$$

可以使用交叉核实确定最佳参数 β。在上述估计计算时,每次省掉一对数据 (X_{ik},y_{ik}),$k=1,\cdots,T_i$,对于每一 β,得到一系列估计:

$$\hat{f}_{ik}(u|\beta)=\left\{\sum_{j\neq k}y_{ij}K(u-X_{ij}'\beta)\right\}\Big\backslash\left\{\sum_{j\neq k}K(u-X_{ij}'\beta)\right\}$$

定义交叉核实函数:

$$S_i(\beta)=\int_A\sum_{k=1}^{T_i}[y_{ik}-\hat{f}_{ik}(u|\beta)]^2\mathrm{d}u$$

这里 A是 u的取值区间。然后选取 β使交叉核实函数极小化:

$$S_i=\min_{\beta}S_i(\beta)$$

一旦 U确定,就在统计量中将第 k 对数据 (X_{ik}, y_{ik})放回去。这就得到了 $i=1,2,\cdots,n$ 条曲线的估计 $\hat{f}_i(u)$,它成功地将多元自变量问题转化为一元自变量的曲线。其余问题与第一节相同,使用迭代方法既求得结构参数 $\theta_i^{(1)},\cdots,\theta_i^{(4)}$,又求得曲线的结构平均 h。

3　教育测量数据实例

以硕士研究生入学考试的 5门课程考分 y_{ij} $(i=1,2,\cdots,5)$ 为例,希望找到 y_{ij}与学生在大学阶段若干门相应课程考分 X_{ij}之间的关系及这 5门课程考分之间的关系。这就用上了多自变量的自建模回归模型。具体数据构造分 5块,每块考虑一门考研成绩,各统计 50名学生,自变量个数取 3,这样总的数据阵是 250行、4列。第 1列为研究生入学成绩;后 3列为大学相应成绩,第 1个 50行数据为英语及相关成绩,第 2个 50行数据为政治成相关成绩,等等。这样,第 1列是 $y_{ij}, i=1,2,\cdots,50$;以后 3列是 $X_{ij}, i=1,2,\cdots,5, j=1,2,\cdots,50$。

由于研究生入学考试的每门课程与大学相关的课程并不一样,例如英语、政治、数学的各自相关课程显然不一样,这就要求作投影寻踪回归和交叉核实时的系数 U选取不一样,即在前一节中将 U全部改为 $U_i, i=1,2,\cdots,n$。这并未给以后的计算带来麻烦。分别对 5个数据块作多元非参数回归,使用投影寻踪化为一元问题,使用交叉核实估计参数 U,得到 5条非参数回归曲线:

$$y_{ij}=\hat{f}_i(X_{ij}'\hat{U}_i)+\bar{X}_{ij}\qquad i=1,\cdots,5$$

实际计算显示这 5条曲线基本相似,反映了考研成绩与本科相关成绩高度相关。于是进一步采用形状不变的自建模回归模型。令

$$t_{ij}=X_{ij}'U_i\qquad i=1,\cdots,5;\ j=1,\cdots,50$$

则模型已化为一元问题,取迭代初值 $\hat{h}_0(t)$:

$$\hat{h}_0(t)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}_i(\hat{\theta}_{i0}^{(2)}t+\hat{\theta}_{i0}^{(3)})$$

经过若干次迭代,得到结构平均函数 $h(t)$和结构参数 $\theta_i^{(1)},\cdots,\theta_i^{(4)}$ 的估计。$h(t)$只是一条曲线,于是由自建模回归模型,5条曲线之间有了相互转化的关系:

$$f_i(\theta_i^{(2)}t+\theta_i^{(3)})=\theta_i^{(1)}h(t)+\theta_i^{(4)}\qquad i=1,\cdots,5$$

据该模型,就可以利用某生的大学成绩与模型参数比较准确地预报他的考研成绩,或显示他实际考研水平发挥如何。

参　考　文　献

1　Kneip A, Gasser T. Convergence and Consistency Results for Self-Modeling Nonlinear Regression. Ann. Statist, 1988, 16 (1): 82~ 112

2　Kneip A, Engel J. Model Estimation in Nonlinear Regression Under Shape Invarianee. Ann. Statist, 1995, 23 (2): 551~ 570

3　Hardle W, et al. Optimal Smoothing in Single Index Models. Ann. Statist, 1993, 21 (1): 157~ 178

4　童恒庆.经济回归模型及计算.武汉: 湖北科学技术出版社,1997

Analysis of Education Measure Data with Multi-variable Self-modeling Regression Model

Shang Gang

(Graduate Office, Wuhan University of Technology, 14 Luoshi Road, Wuhan, China, 430070)

Abstract　In this paper, the self-modeling regression model with shape invariance and multi-variable is built up, its calculation is considered, and its application example with the education measure data is given. In this model, the general characters of the regression curves can be displayed, and the individual characters of the curves as well as the relationship between the curves can be showed.

Key words　education measure; multi-variable; self-modeling regression model