Pure Mathematics 理论数学, 2024, 14(2), 612-623
Published Online February 2024 in Hans. https://www.hanspub.org/journal/pm
https://doi.org/10.12677/pm.2024.142060

Hans汉斯

基于B样条的可加Logistic模型估计

昝思吕

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2024年1月8日；录用日期：2024年1月23日；发布日期：2024年2月29日

摘　要

非参数模型因不需要事先假定函数形式，所以相较于参数模型更有灵活性和适应性，但存在维数灾难问题。可加模型的提出能有效克服这一问题，同时又能保留非参数的优点。本文针对可加Logistic模型，采用B样条近似，结合极大似然思想得到函数的估计，并证明了其最优收敛速度。同时通过数值模拟和实证分析，比较了可加Logistic模型和Logistic模型的表现，结果说明可加Logistic模型的表现更优。

关键词

Logistic模型，可加Logistic模型，B样条

Estimation of Additive Logistic Model Based on B-Spline

Silv Zan

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jan. 8th, 2024; accepted: Jan. 23rd, 2024; published: Feb. 29th, 2024

Abstract

Nonparametric model is more flexible and adaptive than parametric model because it does not need to assume the function form in advance, but it has the problem of dimensional disaster. The additive model can effectively overcome this problem, while retaining the advantages of nonparameters. The paper uses B-spline for approximating additive Logistic model, and adopts the maximum likelihood idea to estimate the function, and proves the optimal convergence rate. Meanwhile, through numerical simulation and empirical analysis, the performance of the additive Logistic model and the logistic model was compared, and the results showed that the additive logistic model performed better.

文章引用: 昝思吕. 基于 B 样条的可加 Logistic 模型估计[J]. 理论数学, 2024, 14(2): 612-623.
DOI: 10.12677/pm.2024.142060

Keywords

Logistic Model, Additive Logistic Model, B Spline

1. 引言

传统的参数模型需事先假设模型形式再求解其中的参数，具有易解释性，但其精度依赖于模型是否假定正确。实际情况中，某些数据无法知道其具体服从的模型形式，当假定的模型形式与实际情况差异很大时，基于假定模型得到的结论可能是完全错误的。为解决这一问题，提出了非参数模型：

$$Y = m(X) + \varepsilon$$

其中，Y 为响应变量，X 为 p 维协变量，$m(\cdot)$ 为未知函数，ε 为随机误差，满足 $E(\varepsilon|X) = 0$。

要使得 $m(X)$ 在 X 处得到较精确的估计，就必须要在 X 的领域内包含足够多的数据。当 X 的维数增大时，估计所需要的数据个数就成指数倍增加。例如，一个局部领域沿着每一个坐标轴包含 a 个数据，在 p 维领域就需要 a^p 个数据点。因为高维空间会使得数据更稀疏，为获取足够多的数据进行估计，可以增大带宽或者增加样本数，但增大带宽会导致估计的偏差增大，增加样本数在许多情况下是不实际的。即，当协变量维数增加时，多元非参数回归估计的精度下降很快，这种现象叫“维数灾难”。可加模型是克服这一问题的有效方法：

$$Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

其中，Y 为响应变量，X_j 为协变量，$f_j(\cdot)$ 为关于 X_j 的光滑函数，为保证模型的可识别性，满足 $E(f_j(\cdot)) = 0$，ε 为随机误差。因为可加模型将非参数回归中的多维光滑问题削减为一维光滑问题，所以避免了多维数据中的“维数灾难”问题。

实际问题中，遇到响应变量的取值为离散值时，常使用广义线性模型：

$$g(\mu) = X^{\mathrm{T}}\beta + \alpha \tag{1}$$

其中，$\mu = E(Y)$，响应变量 Y 服从指数分布族中某一分布，$g(\cdot)$ 为链接函数，X 为 p 维协变量，β 为 p 维待估参数。

使用模型(1)仍事先假设了协变量对响应变量的影响是线性的，这不能很好地反应出变量间的其他非线性关系，所以考虑将非参数可加的思想引入到其中，得到广义可加模型，其具体形式如下：

$$g(\mu) = \sum_{j=1}^{p} f_j(X_j) + \alpha$$

其中，$\mu = E(Y)$，响应变量 Y 服从指数分布族中某一分布，$g(\cdot)$ 为链接函数，X_j 为协变量，$f_j(\cdot)$ 为未知函数，为保证模型的可识别性，满足 $E(f_j(\cdot)) = 0$，α 为常数项。

广义可加模型由 Hastie 和 Tibshirani (1990) [1]提出，其作为广义线性模型的一种推广，结合了广义线性模型和可加模型的优点：即可以解决响应变量离散以及非正态时的建模问题；不需事先假定协变量与响

应变量之间具体的关系，降低了广义线性模型中由线性假定造成的偏差。因广义可加模型的灵活性，现许多学者对其进行研究与应用。Horowitz 和 Mammen (2004) [2]提出一种两阶段方法估计广义可加模型，并证明了当可加分量函数连续两次可微时，该方法所得到的估计量具有 oracle 性质，该估计量是渐近正态分布的。Alimadad 和 Salibian-Barrera (2011) [3]基于反拟合算法讨论了广义可加模型的 outlier-稳健拟合问题，证明了所得到的估计的平均函数的统计性质。Wood (2008) [4]提出了一种计算效率高的方法用于广义可加模型的光滑度选择。Liu 等(2013) [5]提出样条回转拟合核估计方法(SBK)，该方法在弱相依条件下是 oracally 有效的，可用于分析高维时间序列。Dominici (2002)等人[6]将广义可加模型应用在环境污染对健康影响研究中。Zou (2016)等[7]使用广义可加模型和典型线性土地利用回归(LUR)模型估计 PM2.5 浓度，对比结果表明广义可加模型在年度和季节尺度上都优于 LUR 模型，广义可加模型的调整 R2 更高，RMSES 更低。

可加 Logistic 模型是广义可加模型的一种特殊情况，常用于解决分类问题：

$$\ln\left(\frac{\mu}{1-\mu}\right)=\sum_{j=1}^{p} f_j\left(X_j\right)+\alpha \tag{2}$$

其中，$\mu=P(Y=1)$，响应变量 Y 服从 $B(1,\mu)$，X_j 为协变量，$f_j(\cdot)$ 为未知函数，为保证模型的可识别性，满足 $E\left(f_j(\cdot)\right)=0$，$\alpha$ 为常数项。

Berg (2007) [8]将可加 Logistic 模型用于破产预测，通过样本外和时间外验证表明广义可加模型在所有风险水平上都显著优于线性判别分析、广义线性模型和神经网络等流行模型。张娟和张贝贝(2016) [9]使用半参数可加 logistic 模型对用户违约概率进行建模，并使用 Group LASSO 方法进行变量选择，实证研究表明该模型与线性 Logistic 回归模型相比，在判别能力和计算效率上均有较大优势。Dlamini 等人(2017) [10]使用可加 Logistic 模型研究与坦桑尼亚五岁以下儿童死亡率相关的风险因素，以指导决策者加快为人民提供更好的生活。Ana 等人(2019) [11]基于局部线性估计使用 Logistic 模型，对高血压病例进行建模分析。方匡南和陈子岚(2020) [12]提出了一种基于半监督可加 Logistic 回归的信用评分模型，并应用于个人信用贷款的违约风险评估中。

本文余下内容安排如下：第 2 节介绍对可加 Logistic 模型的估计方法。第 3 节和第 4 节分别进行数值模拟和实证分析，比较可加 Logistic 模型和 Logistic 模型的表现。第 5 节进行相关理论性质的证明。第 6 节对全文进行总结。

2. 估计方法

对于 Logistic 模型，采用的是极大似然法求解模型中的参数，即通过最大化以下的对数似然函数得到参数 β 的估计：

$$\ell(\beta)=\sum_{i=1}^{n}\left[y_i X_i^{\mathrm{T}}\beta-\log\left(1+\mathrm{e}^{X_i^{\mathrm{T}}\beta}\right)\right]$$

其中，y_i 为响应变量，X_i 为协变量，β 为待估参数，n 为样本量。$\ell(\beta)$ 是关于 β 的高阶可导连续凸函数，需采用一些数值方法求得最优解 $\beta^*=\arg\max_{\beta}\{\ell(\beta)\}$。

对于可加 Logistic 模型，仍采用极大似然法，此时目标函数变为：

$$Q(\beta)=\sum_{i=1}^{n}\left[y_i\sum_{j=1}^{p} f_j\left(x_{ij}\right)-\log\left(1+\mathrm{e}^{\sum_{j=1}^{p} f_j\left(x_{ij}\right)}\right)\right] \tag{3}$$

其中，y_i 为响应变量，$f_j(\cdot)$ 为未知函数，n 为样本量。

要求解(3)式，首先需对未知函数 $f_j(\cdot)$，$j=1,\cdots,p$ 进行估计。本文采取的是 B 样条近似，其样条基

函数形式如下：

$$B_{i,d}(x)=\frac{x-\nu_i}{\nu_{i+d}-\nu_i}B_{i,d-1}(x)+\frac{\nu_{i+d+1}-x}{\nu_{i+d+1}-\nu_{i+1}}B_{i+1,d-1}(x)$$

且

$$B_{i,0}(x)=\begin{cases}1, & \text{if } \nu_i\le x\le\nu_{i+1}\\ 0, & \text{Otherwise}\end{cases}$$

其中，d 为多项式次数，ν_i 称为节点。

设 $\nu=(\nu_1,\cdots,\nu_k)$ 为节点向量，其中 $a\le\nu_1\le\cdots\le\nu_k\le b$ 为协变量范围 $[a,b]$ 上的节点，k 为节点个数；$\varphi(x)=(\varphi_1(x),\cdots,\varphi_{K_n}(x))$ 为 d 次 B 样条基函数，其中 $K_n=d+k$。令 $\overline{\varphi}_{jw}(x)=\frac{1}{n}\sum_{i=1}^{n}\varphi_w(x_{ij})$，$B_{jw}(x)=\varphi_w(x)-\overline{\varphi}_{jw}$，其中 $j=1,\cdots,p$；$w=1,\cdots,K_n$，则 $f_j(X_j)$ 可用 B 样条函数近似，即：

$$f_j(X_j)\approx B_j(X_j)\beta_j$$

其中，$\beta_j=(\beta_{j1},\cdots,\beta_{jK_n})^{\mathrm{T}}$，$B_j(X_j)=(B_{j1}(X_j),\cdots,B_{jK_n}(X_j))$。则模型(2)可写为：

$$\ln\left(\frac{\mu}{1-\mu}\right)\approx\sum_{j=1}^{p}B_j(X_j)\beta_j+\alpha \tag{4}$$

令 $\beta=(\beta_1^{\mathrm{T}},\cdots,\beta_p^{\mathrm{T}},\alpha)^{\mathrm{T}}$，$B(x)=(B_1(X_1),\cdots,B_p(X_p),B_0)$，其中 α 为常数项，$B_0=(1,\cdots,1)_{n\times1}^{\mathrm{T}}$，$\{(y_i,x_i),i=1,\cdots,n\}$ 为样本，则(4)式可进一步简化表达为：

$$\ln\left(\frac{\mu}{1-\mu}\right)=B(x)\beta \tag{5}$$

目标函数(3)式可写为：

$$Q(\beta)=\sum_{i=1}^{n}\left[y_iB(x_i)\beta-\log\left(1+\mathrm{e}^{B(x_i)\beta}\right)\right] \tag{6}$$

通过最大化目标函数(6)式可得到 β 的估计 $\hat{\beta}$，则未知函数 $f_j(\cdot)$ 的估计为：

$$\hat{f}_j(X_j)=B_j(X_j)\hat{\beta}_j,j=1,\cdots,p$$

采用拟牛顿法来求解(6)式中的 β 的估计值，过程如下：

Step 1：取初始值 β_0，初始 n 阶阵 H_0。令 $k=0$。

Step 2：计算 $g_k=\frac{\partial Q(\beta_k)}{\partial\beta_k}$，$d_k=-H_kg_k$。令 $\beta_{k+1}=\beta_k+\alpha d_k$，其中，步长 α 由线搜索产生。若 $\|g_{k+1}\|\le10^{-6}$，算法停止：否则，转 Step 3。

Step 3：令 $H_{k+1}=H_k+\frac{(s_k-H_ky_k)(s_k-H_ky_k)^{\mathrm{T}}}{(s_k-H_ky_k)^{\mathrm{T}}y_k}$，$k=k+1$，转 Step 2。

这里的线搜索方法，采用 Armijo 步长规则：取 $\alpha_k=\beta\gamma^{m_k}$，其中，$m_k$ 为满足下式的最小非负整数 m：

$$f(x_k+\beta\gamma^{m_k}d_k)\le f(x_k)+\sigma\beta\gamma^{m_k}g_k^{\mathrm{T}}d_k$$

其中，$\beta>0$，$\sigma,\gamma\in(0,1)$。

3. 数值模拟

考虑生成符合(7)式的模拟数据：

$$E\left(y_i\right)=\frac{\mathrm{e}^{\sum_{j=1}^{p} f_j\left(x_{ij}\right)+\alpha}}{1+\mathrm{e}^{\sum_{j=1}^{p} f_j\left(x_{ij}\right)+\alpha}},\quad i=1,2,\cdots,n \tag{7}$$

$$p_i=\frac{1}{1+\exp\left(-\sum_{j}^{4} f_j\left(x_{ij}\right)+\varepsilon_i\right)}$$

其中，$f_1(x)=-2\sin(2x)$，$f_2(x)=x^2-25/12$，$f_3(x)=x$，$f_4(x)=\exp(-x)-2\sinh(5/2)/5$，协变量 $X_j\sim U(-2.5,2.5)$，$y_i\sim B(1,p_i)$，$\varepsilon_i\sim N(0,1)$，n 为样本量。

使用第 2 节提到的方法估计 $f_j(\cdot)$，$j=1,2,3,4$。考虑 3 种不同样本量的模拟数据集 $\{(x_i,y_i)\}_{i=1}^{n}$，$n=200,500,800$。对模拟数据分别建立可加 Logistic 模型(GAM)和 Logistic 模型(GLM)，得到不同样本量下这 2 种模型的函数拟合图，见图 1。以 *MSE* 指标评估函数估计的准确性，重复产生 $N=100$ 组数据集，计算 4 种函数估计的平均 *MSE*，结果记录于表 1；同时以查准率 *P*、查全率 *R* 及 *F*1 为指标评价分类正确率，计算 *N* 组数据集的平均 *P*、*R*、*F*1 值，结果记录于表 2。

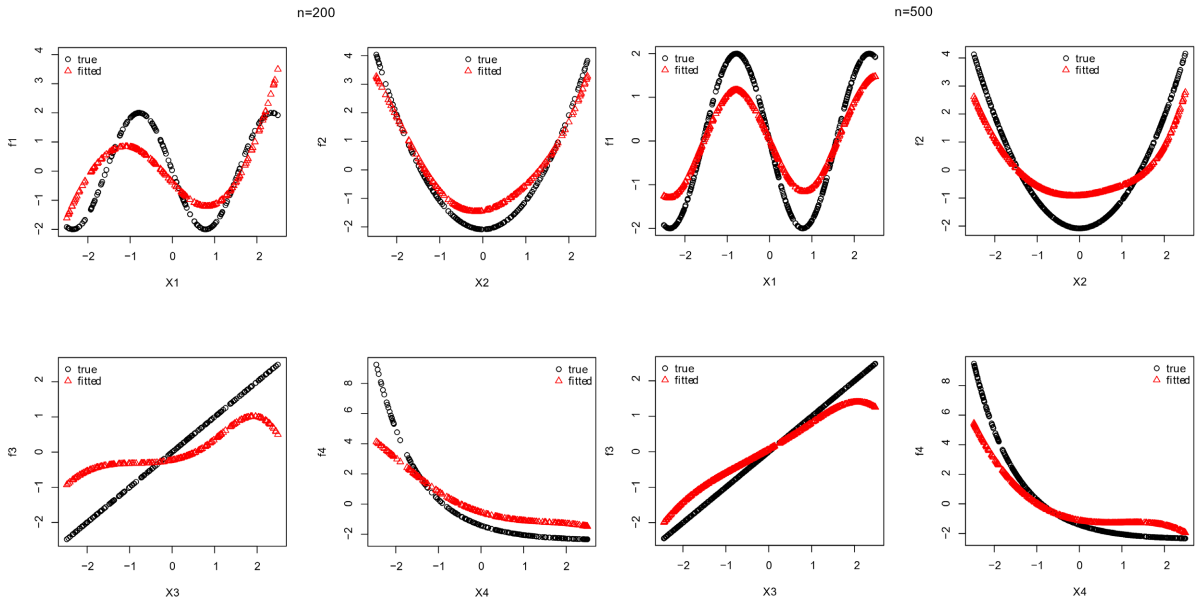
$$MSE_j=\frac{1}{N}\sum_{m=1}^{N}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_j^{(m)}\left(x_{ij}\right)-f_j^{(m)}\left(x_{ij}\right)\right)^2\right],\ j=1,\cdots,p$$

$$P=\frac{1}{N}\sum_{m=1}^{N}\frac{TP^{(m)}}{TP^{(m)}+FP^{(m)}}$$

$$R=\frac{1}{N}\sum_{m=1}^{N}\frac{TP^{(m)}}{TP^{(m)}+FN^{(m)}}$$

$$F1=\frac{2}{1/P+1/R}=\sum_{m=1}^{N}\frac{2TP^{(m)}}{2TP^{(m)}+FP^{(m)}+FN^{(m)}}$$

其中，*N* 表示重复实验次数，*TP*、*FN*、*FP*、*TN* 分别表示 4 种预测情况的样本数量：*TP* 表示真正例样本数；*FN* 表示假反例样本数；*FP* 表示假正例样本数；*TN* 表示真反例样本数。*MSE* 越小，表明函数拟合得



(a1) (b1)

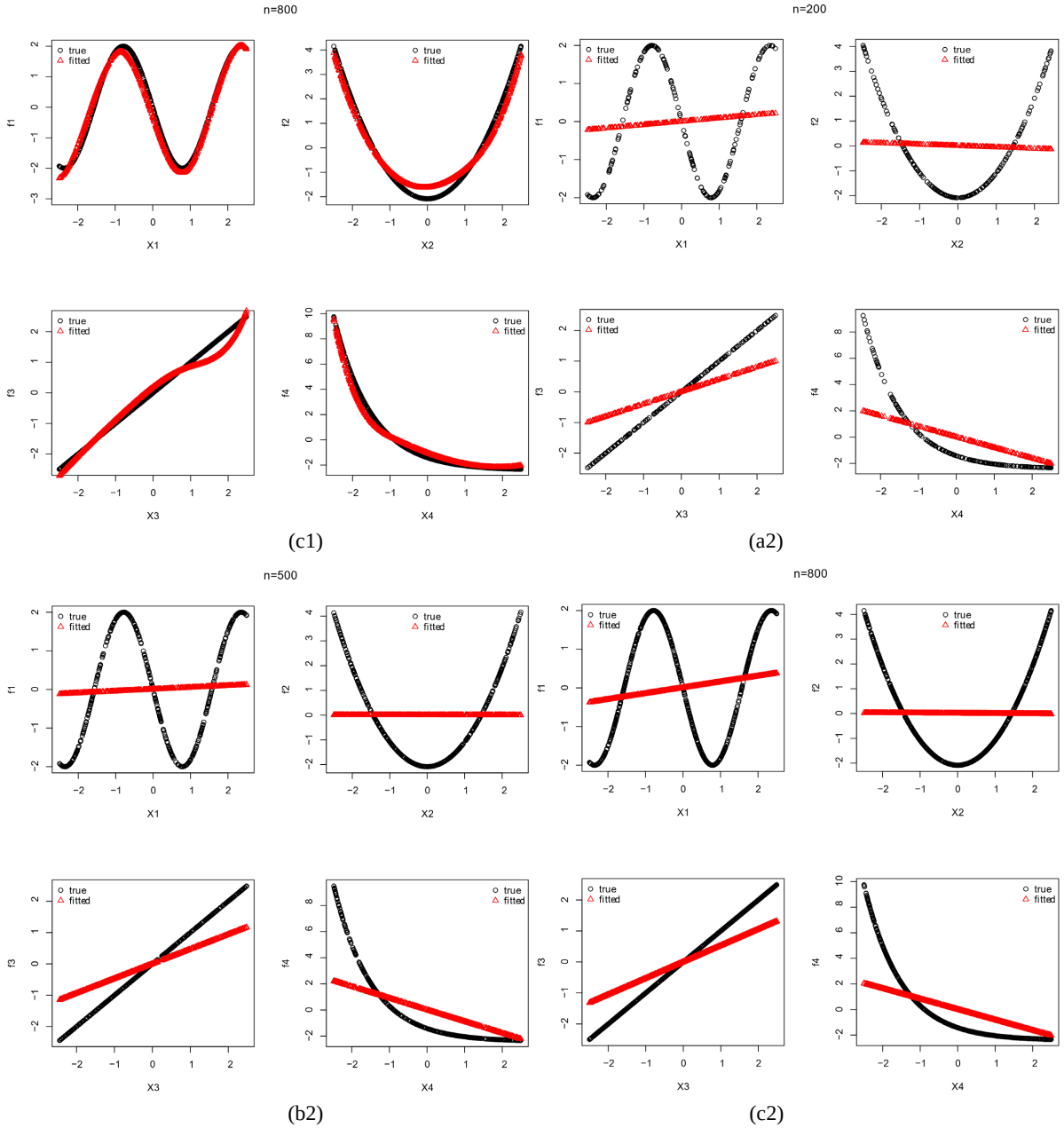


Figure 1. (a1), (b1) and (c1) are the function fitting graphs of GAM; (a2), (b2) and (c2) are the function fitting graphs of GLM

图 1. (a1)、(b1)、(c1)为 GAM 的函数拟合图；(a2)、(b2)、(c2)为 GLM 的函数拟合图

Table 1. *MSE* values fitted by two models

表 1. 2 种模型拟合的 *MSE* 值

n	模型	MSE_1	MSE_2	MSE_3	MSE_4
200	GAM	0.6675	0.6624	0.5773	1.3683
	GLM	2.0654	3.5356	0.5450	4.04240
500	GAM	0.1621	0.1945	0.1554	0.3868
	GLM	2.0355	3.4648	0.5091	4.3182

续表

800	GAM	0.0985	0.1528	0.1132	0.3419
	GLM	2.0315	3.4676	0.5347	4.4172

越好；P、R 和 $F1$ 是二分类问题常用指标，这 3 个指标越大，说明模型的预测效果越好。

由图 1 可直观地看出 2 种模型对每个函数 $f_j(\cdot), j=1,2,3,4$ 的拟合效果，可明显看出，在变量间存在非线性关系时，使用可加 Logistic 模型比使用 Logistic 模型能更好的拟合函数，这一点可进一步通过表 1 和表 2 中的数值来说明。

表 1 的数值结果显示，3 种不同的样本量下，通过可加 Logistic 模型得到的函数估计误差值均比通过 Logistic 模型得到的函数估计误差值小，这表明在变量存在非线性关系时，通过可加 Logistic 模型能得到更精确的估计。且随着样本量的增加，通过两种模型得到的函数估计误差都在减小，这表明样本量越大，估计越精确。

Table 2. Classification accuracy under two models
表 2. 2 种模型下的分类正确率

n	模型	训练集			测试集		
		P	R	$F1$	P	R	$F1$
200	GAM	0.8884	0.8574	0.8727	0.8055	0.7544	0.7791
	GLM	0.7441	0.7062	0.7246	0.7326	0.7099	0.7210
500	GAM	0.8588	0.8267	0.8425	0.8372	0.7906	0.8132
	GLM	0.7328	0.706	0.7192	0.7342	0.6977	0.7155
800	GAM	0.8536	0.8182	0.8355	0.8392	0.8044	0.8214
	GLM	0.7322	0.7050	0.7183	0.7330	0.7061	0.7193

表 2 的数值结果显示，3 种不同的样本量下，无论是训练集还是测试集上，通过可加 Logistic 模型得到的 P、R 和 $F1$ 值均比通过 Logistic 模型得到的 P、R 和 $F1$ 值大，这表明在变量间存在非线性关系时，可加 Logistic 模型较 Logistic 模型能得到更高的分类正确率。

4. 实证分析

选取鲭鱼卵密度数据集进行实例分析，该数据集记录了对欧洲西北部海岸附近的鲭鱼卵丰度进行的调查，可通过 R 语言获取。变量描述见表 3：

Table 3. Variable description of mackerel egg density dataset
表 3. 鲭鱼卵密度数据集的变量描述

变量名	变量解释
Density	卵密度
smack.lat	采样位置的纬度
smack.long	采样位置的经度
smack.depth	海底深度
Temperature	海面温度
D200	与 200-m 等高线之间的距离

以 Density 为响应变量，其余变量为协变量。Density = 0.00 时赋值为 0，表示该位置不存在鲭鱼，Density > 0.00 时赋值为 1，表示该位置存在鲭鱼。赋值后有 108 个样本对应的响应变量为 1309 个样本对应的响应变量为 0，为得到平衡数据样本，将响应变量为 1 的样本复制 3 次后，与 309 个响应变量为 0 的样本组成新的数据集，共 633 个样本。对新数据集的协变量进行函数估计，得到图 2 所示的函数估计图。

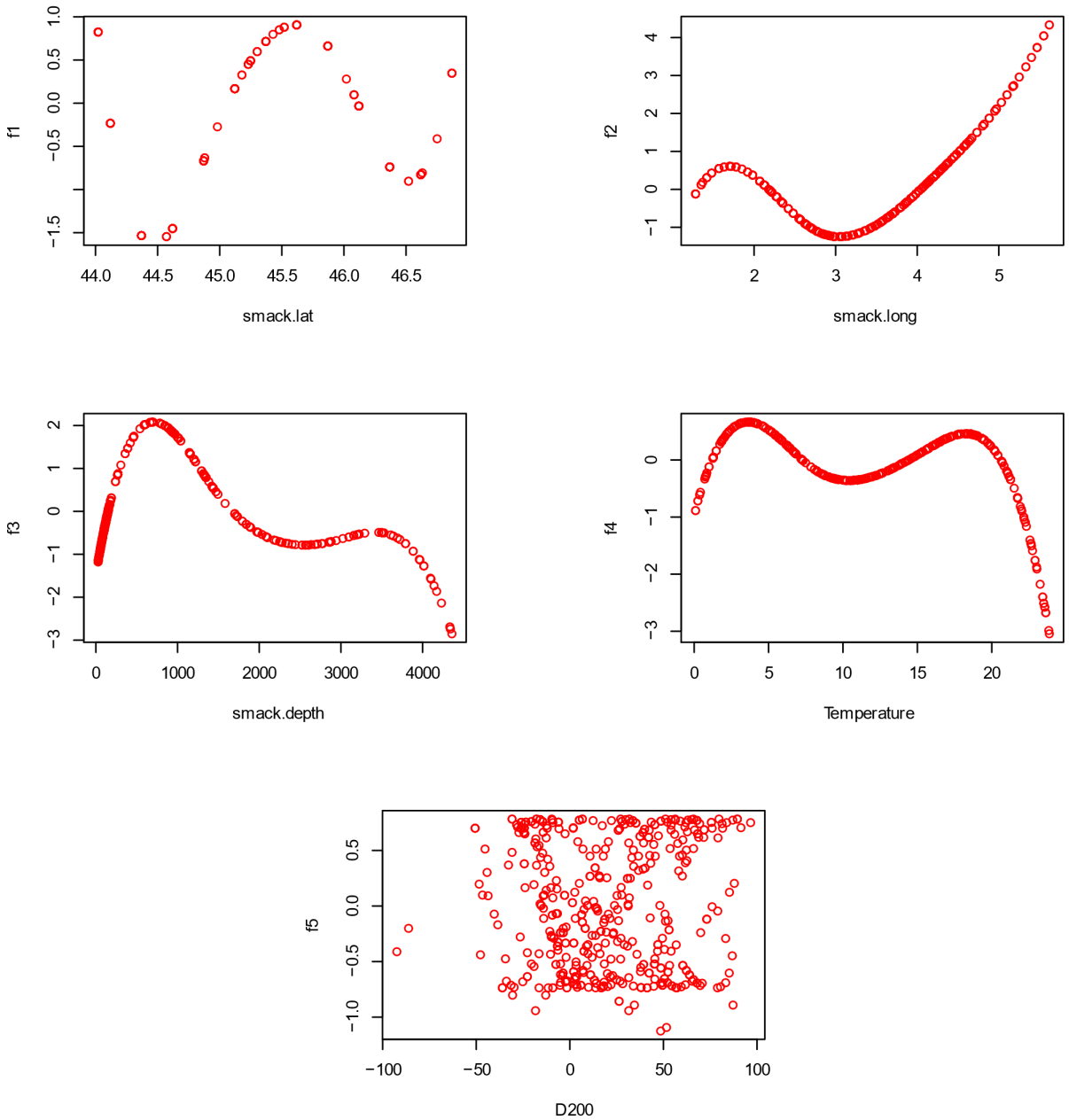


Figure 2. Function estimation graph of mackerel egg density data set
图 2. 鲭鱼卵密度数据集的函数估计图

图 2 描绘了鲭鱼卵密度数据集的 5 个协变量函数估计结果，由此可看出协变量对于响应变量的影响是非线性的，因此可知建立可加 Logistic 模型是合适的。为进一步说明当变量间存在非线性关系时，

Logistic 模型不能很好的进行判别，对鲭鱼卵密度数据分别建立可加 Logistic 模型和 Logistic 模型。对数据集进行 100 次测试集和训练集的划分，以 P、R 和 $F1$ 值为评价指标，对比两种模型在训练集和测试集上的评价指标平均值，结果见表 4。

Table 4. Classification accuracy of mackerel egg density data set under two models
表 4. 鲭鱼卵密度数据集在 2 种模型下的分类正确率

模型	训练集			测试集		
	P	R	$F1$	P	R	$F1$
GAM	0.7010	0.7507	0.7507	0.6732	0.7198	0.6958
GLM	0.6387	0.7042	0.6699	0.6287	0.6768	0.6518

表 4 的数值结果显示，使用可加 Logistic 模型在训练集和测试集上的 P、R 和 $F1$ 值都高于使用 Logistic 模型得到的 P、R 和 $F1$ 值，这表明使用可加 Logistic 模型能得到更高的分类精度。

综上，由模拟数据和实际数据可知，当事先无法确定数据服从的模型形式时，使用非参数模型能得到更正确的结论和更高的精度。

5. 主要定理及证明

假设下列条件成立：

C1：$E\left(f_j\left(X_j\right)\right)=0$，$j=1,2,\cdots,p$。

C2：假设 X_j 的密度函数是 $g_{X_j}(\cdot)$，并且存在两个常数 c_1 和 c_2 在区间 $[a,b]$ 满 $0<c_1\le g_{X_j}(x)\le c_2<\infty$ 足，$j=1,2,\cdots,p$。

C3：令 d 是非负整数，且满足 $0<d<r-1$，$r\ge 2$。定义在区间 $[a,b]$ 上的函数 f_j，其 d 阶导数满足条件 $\eta\in(0,1]$，即对 $d+\eta\ge 0.5$，$s,t\in[a,b]$ 有 $\left|f_j^{(d)}(t)-f_j^{(d)}(s)\right|\le c|t-s|^{\eta}$，$j=1,2,\cdots,p$，其中 c 是正的有限常数。

注：条件 C1~C3 是多项式样条估计的一般性假设，见 Huang [13]等，Guo [14]等。

引理 1：在条件 C1~C3 的假定下，模型(5)的参数真实值为 β^*，则 $f_j^*=B_j\left(X_j\right)\beta_j^*$。对任意的 f_j 有：

$$\left\|f_j^*-f_j\right\|=\mathrm{O}_p\left(K_n^{-r}+K_n^{1/2}n^{-1/2}\right),\ j=1,\cdots,p$$

特别的，当节点数 $K_n=\mathbf{O}\left(n^{1/(2r+1)}\right)$ 时，

$$\left\|f_j^*-f_j\right\|=\mathrm{O}_p\left(n^{-r/(2r+1)}\right),\ j=1,\cdots,p$$

证明过程参见 Huang [13]。

引理 2：定义 $D=\frac{1}{n}\sum_{i=1}^{n}B^{\mathrm{T}}(x_i)B(x_i)$，$\lambda_{\min}(D)$ 和 $\lambda_{\max}(D)$ 分别是 D 的最小和最大特征值，$K_n=\mathbf{O}\left(n^{v}\right)$，其中 $0<v<0.5$，且 $h\equiv h_n\asymp K_n^{-1}$，在条件 C1~C2 下，以概率 1 成立，有：

$$c_3h_n\le\lambda_{\min}(D)\le\lambda_{\max}(D)\le c_4h_n$$

其中，c_3 和 c_4 两个正的常数。

证明过程参见 Wei [15]。

定理 1：在条件 C1~C3 下，存在 $\hat{f}_j\left(X_j\right)=B_j\left(X_j\right)\hat{\beta}_j$，当 $K_n\to\infty$，$K_n/n\to 0$ 时，有下式成立：

$$\left\|\hat{f}_j-f_j\right\|=\mathrm{O}_p\left(K_n^{-r}+K_n^{1/2}n^{-1/2}\right)$$

特别地， $K_n = \mathbf{O}\left(n^{1/(2r+1)}\right)$ 时，有：

$$\left\| \hat{f}_j - f_j \right\| = \mathrm{O}_p\left(n^{-r/(2r+1)}\right)$$

定理 1 的证明过程见附录。

6. 结论

本文对可加 Logistic 模型中的非参数函数部分使用 B 样条逼近，结合极大似然思想，得到函数的估计，并在一定条件下，证明了该估计量的最优收敛速度。通过对模拟数据和实际数据分别建立可加 Logistic 模型和 Logistic 模型，对比 *MSE*、*P*、*R* 和 *F*1 的指标值，结果表明可加 Logistic 模型得到的 *MSE* 值更小，*P*、*R* 和 *F*1 值更大，这说明在变量间存在非线性关系时，可加 Logistic 模型较 Logistic 模型能得到更好的估计和更高的分类正确率。

参考文献

[1] Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models. Chapman and Hall, New York. https://doi.org/10.1201/9780203753781-6

[2] Horowitz, J.L. and Mammen, E. (2004) Nonparametric Estimation of an Additive Model with a Link Function. *The Annals of Statistics*, **32**, 2412-2443. https://doi.org/10.1214/009053604000000814

[3] Alimadad, A. and Salibian-Barrera, M. (2011) An Outlier-Robust Fit for Generalized Additive Models with Applications to Disease Outbreak Detection. *Journal of the American Statistical Association*, **106**, 719-731. https://doi.org/10.1198/jasa.2011.tm09654

[4] Wood, S.N. (2008) Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **70**, 495-518. https://doi.org/10.1111/j.1467-9868.2007.00646.x

[5] Liu, R., Yang, L. and Härdle, W.K. (2013) Oracally Efficient Two-Step Estimation of Generalized Additive Model. *Journal of the American Statistical Association*, **108**, 619-631. https://doi.org/10.1080/01621459.2013.763726

[6] Dominici, F., McDermott, A., Zeger, S.L. and Samet, J.M. (2002) On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health. *American Journal of Epidemiology*, **156**, 193-203. https://doi.org/10.1093/aje/kwf062

[7] Zou, B., Chen, J., Zhai, L., Fang, X. and Zheng, Z. (2016) Satellite Based Mapping of Ground $PM_{2.5}$ Concentration Using Generalized Additive Modeling. *Remote Sensing*, **9**, 1. https://doi.org/10.3390/rs9010001

[8] Berg, D. (2007) Bankruptcy Prediction by Generalized Additive Models. *Applied Stochastic Models in Business and Industry*, **23**, 129-143. https://doi.org/10.1002/asmb.658

[9] 张娟, 张贝贝. 基于 Group-LASSO 方法的广义半参数可加信用评分模型应用研究[J]. 数理统计与管理, 2016, 35(3): 517-524.

[10] Dlamini, W.J., Melesse, S.F. and Mwambi, H.G. (2017) Logistic Regression Additive Model: Application to Tanzania Demographic and Health Survey Data. *The Open Public Health Journal*, **10**, 294-302. https://doi.org/10.2174/1874944501710010294

[11] Ana, E., Chamidah, N., Andriani, P. and Lestari, B. (2019) Modeling of Hypertension Risk Factors Using Local Linear of Additive Nonparametric Logistic Regression. *Journal of Physics: Conference Series*, **1397**, Article 012067. https://doi.org/10.1088/1742-6596/1397/1/012067

[12] 方匡南, 陈子岚. 基于半监督广义可加 Logistic 回归的信用评分方法[J]. 系统工程理论与实践, 2020, 40(2): 392-402.

[13] Huang, J., Horowitz, J.L. and Wei, F. (2010) Variable Selection in Nonparametric Additive Models. *Annals of Statistics*, **38**, 2282-2313. https://doi.org/10.1214/09-AOS781

[14] Guo, J., Tang, M., Tian, M. and Zhu, K. (2013) Variable Selection in High-Dimensional Partially Linear Additive Models for Composite Quantile Regression. *Computational Statistics & Data Analysis*, **65**, 56-67. https://doi.org/10.1016/j.csda.2013.03.017

[15] Wei, F. (2012) Group Selection in High-Dimensional Partially Linear additive Models. *Brazilian Journal of Probability and Statistics*, **26**, 219-243. https://doi.org/10.1214/10-BJPS129

附　录

定理 1 的证明：根据引理 1 可知 $R=\sum_{j=1}^{p}\left[f_j^*\left(x_{ij}\right)-f_j\left(x_{ij}\right)\right]=\mathbf{O}_p\left(K_n^{-r}+K_n^{1/2}n^{-1/2}\right)$。

为证定理 1 的结论，需先证：

$$\left\|\hat{\beta}-\beta^*\right\|=\mathrm{O}_p\left(K_n^{-(2r-1)/2}+K_n/\sqrt{n}\right)\triangleq\mathrm{O}_p\left(\delta\right) \tag{8}$$

令 $\beta=\beta^*+\delta u$，u 为 pK_n 维向量。为证(8)式成立，需证对任意给定的 ε，存在足够大的常数 C，使得下式成立：

$$P\left\{\sup_{\|u\|=C}Q\left(\beta\right)<Q\left(\beta^*\right)\right\}\geq 1-\varepsilon \tag{9}$$

因为(9)式成立，意味着至少以 $1-\varepsilon$ 的概率在球 $\left\{\beta^*+\delta u:\|u\|\leq C\right\}$ 中存在一个局部最大值。因此，存在一个局部最大值使得 $\left\|\hat{\beta}-\beta^*\right\|=\mathrm{O}_p\left(\delta\right)$。

对目标函数的表达式使用泰勒展开：

$$\begin{aligned}Q\left(\beta\right)&=\sum_{i=1}^{n}\left[y_iB\left(x_i\right)\beta-\log\left(1+\mathrm{e}^{B\left(x_i\right)\beta}\right)\right]\\&=\sum_{i=1}^{n}\left[y_iB\left(x_i\right)\left(\beta^*+\delta u\right)-\log\left(1+\mathrm{e}^{B\left(x_i\right)\left(\beta^*+\delta u\right)}\right)\right]\\&\text{在 }B\left(x_i\right)\beta^*\text{ 处展开}\\&=\sum_{i=1}^{n}\left\{y_i\left[B\left(x_i\right)\beta^*+B\left(x_i\right)\delta u+o\left(B\left(x_i\right)\delta u\right)^2\right]\right.\\&\left.-\left[\log\left(1+\mathrm{e}^{B\left(x_i\right)\beta^*}\right)+\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{1+\mathrm{e}^{B\left(x_i\right)\beta^*}}B\left(x_i\right)\delta u+\frac{1}{2}\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{\left(1+\mathrm{e}^{B\left(x_i\right)\beta^*}\right)^2}\left(B\left(x_i\right)\delta u\right)^2\right]+o\left(B\left(x_i\right)\delta u\right)^2\right\}\end{aligned}$$

则有：

$$\begin{aligned}Q\left(\beta\right)-Q\left(\beta^*\right)&=\sum_{i=1}^{n}\left\{y_iB\left(x_i\right)\delta u-\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{1+\mathrm{e}^{B\left(x_i\right)\beta^*}}B\left(x_i\right)\delta u-\frac{1}{2}\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{\left(1+\mathrm{e}^{B\left(x_i\right)\beta^*}\right)^2}\left(B\left(x_i\right)\delta u\right)^2+o\left(B\left(x_i\right)\delta u\right)^2\right\}\\&\triangleq I_1+I_2+o\left(B\left(x_i\right)\delta u\right)^2\end{aligned} \tag{10}$$

其中，$I_1=\sum_{i=1}^{n}\left[y_i-\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{1+\mathrm{e}^{B\left(x_i\right)\beta^*}}\right]B\left(x_i\right)\delta u$；$I_2=-\frac{1}{2}\delta^2u^{\mathrm{T}}\sum_{i=1}^{n}\left[\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{\left(1+\mathrm{e}^{B\left(x_i\right)\beta^*}\right)^2}\cdot B\left(x_i\right)^{\mathrm{T}}B\left(x_i\right)\right]u$ 对 I_1 在 $\sum_{j=1}^{p}f_j\left(x_i\right)$ 处泰勒展开，有：

$$I_1=n\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\frac{\mathrm{e}^{B\left(x_i\right)\beta^*}}{1+\mathrm{e}^{B\left(x_i\right)\beta^*}}\right]B\left(x_i\right)\delta u=n\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\frac{\mathrm{e}^{R+\sum_{j=1}^{p}f_j\left(x_i\right)}}{1+\mathrm{e}^{R+\sum_{j=1}^{p}f_j\left(x_i\right)}}\right]B\left(x_i\right)\delta u$$

$$= n\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}} - \frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{\left(1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}\right)^2} R + o(R) \right] B(x_i)\delta u$$

令 $W_1 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}} - \frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{\left(1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}\right)^2} R \right] B(x_i)u$，则有：

$$E(W_1) = \frac{1}{n}\sum_{i=1}^{n} E\left\{\left[y_i - \frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}} - \frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{\left(1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}\right)^2} R \right] B(x_i)\right\} u$$

$$= -\frac{1}{n}\sum_{i=1}^{n} E\left[\frac{\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}}{\left(1+\mathrm{e}^{\sum_{j=1}^{p} f_j(x_i)}\right)^2} RB(x_i) \right] u$$

$$\le -\frac{C_1}{n}\sum_{i=1}^{n} E\left[RB(x_i)\right] u$$

$$\le -C_1 R\|u\| \sqrt{E\left[\frac{1}{n}\sum_{i=1}^{n} B(x_i)^{\mathrm{T}} B(x_i)\right]}$$

$$= -C_1\|u\| \mathrm{O}_p\left(K_n^{-r} + K_n^{\frac{1}{2}} n^{-\frac{1}{2}}\right) \mathrm{O}_p\left(\sqrt{K_n^{-1}}\right)$$

$$= \|u\| \mathrm{O}_p\left(K_n^{-(2r+1)/2} + n^{-1/2}\right)$$

即 $I_1 = \mathbf{O}_p\left(n\delta\left(K_n^{(-2r-1)/2} + n^{-1/2}\right)\|u\|\right)$。

类似可得：$I_2 = \mathbf{O}_p\left(n\delta^2 K_n^{-1}\|u\|^2\right) = \mathbf{O}_p\left(n\delta\left(K_n^{(-2r-1)/2} + n^{-1/2}\right)\|u\|^2\right)$。

因此，当 C 足够大时，I_2 控制 I_1，即(10)式的符号取决于 I_2。因为 $I_2 < 0$，所以 $Q(\beta) - Q(\beta^*) < 0$，因此(9)式成立，即：$\|\hat{\beta} - \beta^*\| = \mathbf{O}_p\left(K_n^{-(2r-1)/2} + K_n/\sqrt{n}\right)$ 成立。则，

$$\|\hat{f}_j - f_j\| \le \|\hat{f}_j - f_j^*\| + \|f_j^* - f_j\|$$

$$= \|B(x_i)\hat{\beta}_j - B(x_i)\beta_j^*\| + \|f_j^* - f_j\|$$

$$= \|\hat{\beta}_j - \beta_j^*\| \sqrt{B^{\mathrm{T}}(x_i)B(x_i)} + \|f_j^* - f_j\|$$

$$= \mathrm{O}_p\left(K_n^{-(2r-1)/2} + K_n/\sqrt{n}\right)\mathrm{O}_p\left(\sqrt{K_n^{-1}}\right) + \mathrm{O}_p\left(K_n^{-r} + K_n^{1/2} n^{-1/2}\right)$$

$$= \mathrm{O}_p\left(K_n^{-r} + K_n^{1/2} n^{-1/2}\right)$$

定理 1 证毕。