文章编号:1006-3080(2017)01-0129-04 **DOI**:10.14135/j.cnki.1006-3080.2017.01.020

基于 Alpha 稳定分布的二元响应变量回归模型

许 哲, 钱夕元
(华东理工大学理学院,上海 200237)

摘要:Logit 模型是常用的针对二元响应变量的回归模型,当 0-1 响应变量不平衡时,Logit 模型将会带来连接函数设定错误。为了更灵活地捕捉带偏和厚尾特征,提出了以 Alpha 稳定分布作为连接函数的二元响应变量回归模型,称之为稳定分布模型。借助期望传播-近似贝叶斯计算(EP-ABC)方法,克服了 Alpha 稳定分布由于没有概率密度函数解析表达式所带来的困难,同时也解决了高维运算所导致的低接收率的问题。结果表明该模型对平衡或不平衡二元响应变量数据拟合和预测的效果均明显优于 Logit、Probit、Cloglog 和 GEV 模型。

关键词:Alpha 稳定分布模型; EP-ABC 方法; 广义线性回归模型; 不平衡数据
中图分类号:O212.1 **文献标志码**:A

Alpha-Stable Distribution Based Regression for Binary Response Data

XU Zhe, *QIAN Xi-yuan*
(*School of Science*, *East China University of Science and Technology*, *Shanghai* 200237, *China*)

Abstract: Logit model is the most popular binary regression models for modelling binary response data. When dealing with unbalanced data, Logit model will cause link misspecification. A more flexible model of alpha-stable model, is introduced to fit unbalanced data by setting alpha-stable distribution as the link function. For model estimation, since alpha-stable distribution admits no closed-form expression for the density, we employ expectation propagation with approximate Bayesian computation (EP-ABC) algorithm. It overcomes the difficulties that high dimensionality results in low acceptance rate through data partitioning. According to the simulation results, alpha-stable model performs better than Logit, Probit, Cloglog or GEV model in fitting both balanced and unbalanced data.

Key words: Alpha-stable model; EP-ABC algorithm; generalized regression model; unbalanced data

在针对二元响应变量建立回归模型时,由于 Logit 模型潜在变量的概率响应曲线是对称的,一般性假设响应变量中的各类样本数目基本均衡,而当 0-1 响应变量样本数差异明显时,继续使用 Logit 模型,将会产生连接函数设定错误(Link misspecification),为参数的估计带来较大的偏差和均方误差[1]。同时,Cloglog 模型只有固定的偏度,GEV 模型的形状参数同时控制偏度和尾部厚度,偏度特征与尾部特征相互制约,使得模型拟合的灵活性并没有完全放松[2]。

因此,针对不平衡的二元响应变量,应当选择更加灵活的连接函数来捕捉数据的偏度特征,用以

收稿日期:2016-04-19
基金项目:国家高技术发展研究"863"计划项目(2015AA20107);上海市经信委"软件和集成电路产业发展专项资金"(140304)
作者简介:许 哲(1991-),女,江苏南京人,硕士生,研究方向为统计计算。E-mail:belle_910803@163.com
通信联系人:钱夕元,E-mail:xyqian@ecust.edu.cn

达到更好的拟合效果。鉴于 Alpha 稳定分布可以很好地刻画数据偏度和尾部特征,本文采用 Alpha 稳定分布作为连接函数构建了二元响应变量回归模型,称为稳定分布模型。

1 稳定分布模型

假定二元响应变量 y_i 是取值为 0、1 的随机变量,$y_i=1$ 表示事件在第 i 个样本下发生,而 $y_i=0$ 表示事件在第 i 个样本下不发生。因此,判断事件在第 i 个样本下发生概率的表达式为

$$p_i = \Pr(y_i = 1) = F(\boldsymbol{x}_i^{\mathrm{T}}\beta_c) \tag{1}$$

其中:$\boldsymbol{x}_i$ 为 k 维自变量向量;$\boldsymbol{\beta}_c$ 为 k 维回归系数;F 为累积分布函数,而 F^{-1} 称为连接函数。Logit 模型中 $F^{-1}(p_i)=\lg(p_i/(1-p_i))$;在 Probit 模型中 $F^{-1}(p_i)=\Phi^{-1}(p_i)$,$\Phi$ 表示标准正态分布的累积分布函数;Cloglog 模型中 $F^{-1}(p_i)=\lg(-\lg(1-p_i))$;GEV 模型中 $F^{-1}(p_i)=G^{-1}(p_i)$,G 表示标准 GEV 分布的累积分布函数。

除了高斯分布($\alpha=2$)、柯西分布($\alpha=1,\beta=0$)和 Lévy 分布($\alpha=1/2,\beta=1$)等少数几种情况外,Alpha 稳定分布的概率密度函数和累积分布函数没有解析表达式,通常可以用如下的特征函数来表示:

$$\Phi_X(t)=\begin{cases}\exp\left\{i\delta t-\gamma^{\alpha}|t|^{\alpha}\left\{1+i\beta(\tan\frac{\pi\alpha}{2})\cdot \operatorname{sgn}(t)(|\gamma t|^{1-\alpha}-1)\right\}\right\}, & \alpha\neq 1\\ \exp\left\{i\delta t-\gamma|t|\left\{1+i\beta\frac{2}{\pi}\cdot \operatorname{sgn}(t)\lg|\gamma t|\right\}\right\}, & \alpha=1\end{cases} \tag{2}$$

其中:$\operatorname{sgn}(t)=t/|t|$;$i^2=-1$;特征参数 $\alpha\in(0,2]$ 刻画了分布的尾部特征,α 越小拖尾越厚;对称参数 $\beta\in[-1,1]$ 决定了分布的偏度,$\beta=0$ 时称为对称稳定分布,$\beta>0$ 时分布右偏;$\gamma>0$ 为尺度参数;$\delta\in\mathbb{R}$ 为分布的位置参数。

依据特征函数的性质,稳定分布拥有连续且单调的概率密度函数和累积分布函数[3],因此,根据式(1),稳定分布模型可以表示为

$$p_i = \Pr(y_i = 1) = S(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_c\,|\,\alpha,\beta) \tag{3}$$

其中,$S(x|\alpha,\beta)$ 表示标准 Alpha 稳定分布($\gamma=1$,$\delta=0$)在 x 处的累积概率,连接函数 $F^{-1}(p_i)=S^{-1}(p_i)$。α 和 β 参数赋予了该模型在偏度或拖尾厚度上更多的变化,使得稳定分布模型对于平衡或不平衡的数据集都能进行良好的拟合。

由于 $S(x|\alpha,\beta)$ 没有解析表达式,直接对稳定分布模型的参数进行极大似然估计比较困难,而基于特征函数的参数估计计算过程又较为复杂,因此本文借助于期望传播-近似贝叶斯计算(Expectation Propagation-Approximate Bayesian Computation,EP-ABC)方法进行参数估计。

2 EP-ABC 方法

EP-ABC 方法是综合了 ABC 方法和 EP 方法而形成的一种贝叶斯推断方法[4]。

ABC 方法的提出是为了解决在似然函数难以计算或表达情况下贝叶斯估计的困难[5]。它主要通过对模型的模拟产生人工数据集,以选定的概要统计量 $s(\cdot)$(Summary statistics)作为度量,比较人工数据集与实际观测数据在这个度量上的距离,如果它们足够接近(小于设定的误差 ε),则认为用于模拟人工数据集的参数来自与真实后验分布相同的分布,这样后验分布就可以直接通过数值模拟进行估计,从而省去了似然函数的计算。ABC 方法成功的关键取决于概要统计量 $s(\cdot)$ 和误差 ε 的选取,这对稳定分布模型参数估计构成了一定的挑战;另外,标准 ABC 方法计算还存在效率不高的缺点。

EP 方法是寻找一个目标分布 $p(\theta|y)$ 的近似分布 $q(\theta)$ 用以对目标分布进行估计。一般地,假设将数据分成 K 个部分,则第 i 个分块的似然函数为 $p(y_i|\theta)$,再设先验为 $p(\theta)$,则目标后验分布可表示为 $p(\theta|y)\propto p(\theta)\prod_{i=1}^{K}p(y_i|\theta)$,相应地,近似分布也可因式分解为 $q(\theta)=p(\theta)\prod_{k=1}^{K}q_k(\theta)$,通过依次更新位点 $q_k(\theta)$,使得近似分布与目标分布更为接近。EP 算法的核心是构造适当的混合分布 $h_k(\theta)=p(y_k|\theta)q_{-k}(\theta)$(Hybrid distribution),寻找与混合分布 KL 散度(Kullback-Leibler divergence)最小的分布来更新位点 $q_k(\theta)$,继而达到对目标后验分布的最优近似[6],其中 $q_{-k}(\theta)=q(\theta)/q_k(\theta)$ 称为空心分布(Cavity distribution)。

对于稳定分布模型,似然函数没有解析的表达式,这会使混合分布也同样不能被解析表达,最小化 KL 散度的运算是十分困难的。EP-ABC 方法就是利用 ABC 方法替代最小化 KL 散度的过程来估计混合分布的矩。令概要统计量取 $s(y_i)=y_i$,则在 EP-ABC 方法下,参数为 $\theta=(\beta_c,\alpha,\beta)$ 的稳定分布模型联合后验分布表示为

$$p(\theta|y) \propto p(\beta_c|\alpha,\beta)p(\alpha|\beta)p(\beta)\cdot \prod_{i=1}^{K}\left\{\int p(y_i^*|x_i,\theta)1_{\{|y_i-y_i^*|\leqslant\varepsilon\}}\mathrm{d}y_i^*\right\} \quad (4)$$

其中：$p(\beta)$为 β 的先验；$p(\beta_c|\alpha,\beta)$和 $p(\alpha|\beta)$分别是 β_c 和 α 的条件先验。

EP-ABC 方法不但克服了 EP 方法中最小化 KL 散度的计算困难，而且明显优于标准 ABC 方法。EP-ABC 方法每次只需处理一个数据点，而且对每个数据点构建一维的概要统计量 $s(y_i)=y_i$，这样 $s(\cdot)$充分包含了数据点 y_i 的信息，从而避免了高维时概要统计量选择不当所带来的估计偏差。特别地，对于二元响应变量，距离函数只有 0 和 1 两种取值，当 $|y_i-y_i^*|=0$ 时接收 y_i^*，而当 $|y_i-y_i^*|=1$ 时拒绝 y_i^*，可以有效消除误差 ε 设定不当带来的影响，这在一定程度上保证了稳定分布模型的求解精度和效率。

3　数值模拟实验

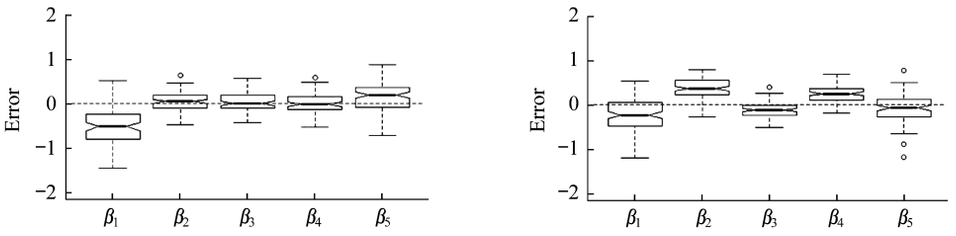
3.1　模拟数据集

在二元响应变量回归模型中，线性回归部分模拟产生 1 个截距项 X_1 和 4 个解释变量(X_2，X_3，X_4，X_5)，对应的回归系数设为 $\beta_c=(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)$，其中，设定 X_2 为服从标准正态分布的连续变量，X_3 为服从伯努利分布的二元 0-1 虚拟变量，X_4 和 X_5 为三分类的类别变量，且令样本容量为 $n=200$。分别模拟产生两组服从 n 维伯努利分布的二元响应变量：一组为基于 Cloglog 模型产生的不平衡数据集，回归系数设为 $\beta_c=(0.5,1,1,0.5,-2)$，称为模拟数据集 1；另一组为基于 Probit 模型产生的平衡数据集，回归系数设为 $\beta_c=(0.5,2,-1,1.5,-2)$，称为模拟数据集 2。

为了验证稳定分布模型对不平衡的模拟数据 1 和平衡的模拟数据集 2 都能进行很好的估计，从而进一步展示 Alpha 稳定分布连接函数的灵活性，将稳定分布模型的估计结果与 Logit、Probit、Cloglog 和 GEV 模型进行比较。这里所有的模型均采用 EP-ABC 算法估计，并且每个待估参数都选取标准正态分布作为先验分布。对于稳定分布模型，待估参数为 $\theta=(\beta_c,\alpha,\beta)$，即先验分布 $\beta_c\sim N(0_5,\mathrm{diag}(1,1,1,1,1))$，$\Phi^{-1}(\alpha/2)\sim N(0,1)$，$\Phi^{-1}(\beta+1/2)\sim N(0,1)$。

3.2　实验结果与分析

为了评价稳定分布模型对不平衡的模拟数据集 1 和平衡的模拟数据集 2 的拟合效果，利用箱线图来展示两次数值模拟实验回归系数参数 β_c 的估计误差 $|\beta_c-\beta_c^*|$ 分布情况，结果见图 1。从图中可以看出，稳定分布无论在估计平衡或不平衡数据集时，β_1，β_2，β_3，β_4，β_5 的估计误差都在相当接近 0 的范围内呈现小幅度波动，β_c 的整体估计较为精确。



(a) Simulation 1: Cloglog dataset　　(b) Simulation 2: Probit dataset

图 1　回归系数参数 β_c 的估计误差分布箱线图

Fig. 1　Estimation error of coefficient parameters β_c with Alpha-stable link model

稳定分布模型在拟合 Cloglog 模型模拟产生的不平衡数据集 1 时，模型连接函数中的特征参数 α 和对称参数 β 的估计值和 95%置信区间(HPDI)分别为 $\alpha=1.635$(1.435，1.834)，$\beta=-0.129$(−0.559，0.301)，由于 Cloglog 函数和 Alpha 稳定分布并不来自相同分布族，在 Alpha 稳定分布中与 Cloglog 函数同等的分布的真实参数未知，因此这里只通过构建 Cloglog 函数分位数和 Alpha 稳定分布分位数的 Q-Q 图来评价稳定分布模型的参数估计效果。从图 2(a)中可见，数据点在[−2，4]的区间内均落在一条直线上，可以认为以估计值 $\alpha=1.635$ 和 $\beta=-0.129$ 为形状参数的稳定分布模型对 Cloglog 连接函数已经提供了较好的拟合效果。

稳定分布模型在拟合 Probit 模型模拟产生的平衡数据集 2 时，模型连接函数中的特征参数 α 和对称参数 β 的估计值和 95%置信区间(HPDI)分别为 $\alpha=1.879$(1.838，1.920)，$\beta=-0.057$(−0.304，

0.190)。当 $n=200$ 时,真实值 $\beta=0$ 已经包含在95%的 HPDI 置信区间中,而对于 α 而言,真实值 $\alpha=2$ 是 α 取值范围(0,2]的边界,因此 α 的估计取不到真实值,$\alpha=2$ 也不可能包含在95%的 HPDI 置信区间中,尽管如此,估计值 $\alpha=1.879$ 已经相当接近真实值 $\alpha=2$,可以认为 α 的估计是比较精确的。与此同时,还可以通过构建高斯分布分位数和 Alpha 稳定分布分位数的 Q-Q 图来评价稳定分布模型对高斯分布连接函数的拟合效果,从图 2(b)中可见,数据点几乎都落在一条直线上,说明以估计值 $\alpha=1.879$ 和 $\beta=-0.057$ 为形状参数的稳定分布模型对高斯分布的连接函数拟合效果较好,α 和 β 的估计是精确的。

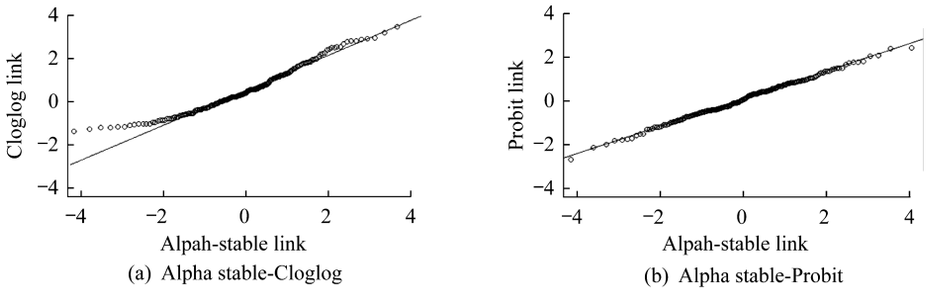


图 2 稳定分布模型对 Cloglog 模型(a)和 Probit 模型(b)近似效果的 Q-Q 图

Fig. 2 Plot of Cloglog(a) and Probit(b) quantiles against Alpha-stable quantiles

综上所述,稳定分布模型的连接函数可以灵活捕捉和描述数据峰度、尾部及偏度的特征,对平衡或不平衡的数据集都能做到较为精确的估计。

3.3 模型比较与预测

参数估计以外,本文还将稳定分布模型的估计能力与 Logit、Probit、Cloglog 和 GEV 模型进行对比。依据表 1 的模型比较结果,在 DIC 准则方面,稳定分布模型的 DIC 值都明显小于 Probit,Cloglog 和 GEV 模型,说明稳定分布模型的复杂度优于其他 3 个模型。

表 1 模型比较结果

Table 1 Model comparison results

模型	DIC		RMSE
	模拟数据集 1	模拟数据集 2	模拟数据集 1
Logit	141.040 6	122.038 8	0.479 1
Probit	151.608 6	118.305 7	0.464 5
Cloglog	144.634 6	123.813 6	0.468 5
GEV	141.084 9	136.722 1	0.484 5
稳定分布	140.995 0	116.952 4	0.459 1

为了比较不同模型对于不平衡数据集的预测能力,计算各模型 5 折交叉验证的均方根误差(RMSE)并绘制其接收者操作特性曲线(ROC),由表 1 可知稳定分布模型的均方根误差最小,图 3 显示稳定分布模型 ROC 曲线下的面积(AUC 值)最大,说明稳定分布模型具有最佳的预测能力。

由此可见,稳定分布模型的拟合和预测效果都优于其他 3 个模型,是二元响应变量回归模型在 GEV 模型基础上的合理改进。

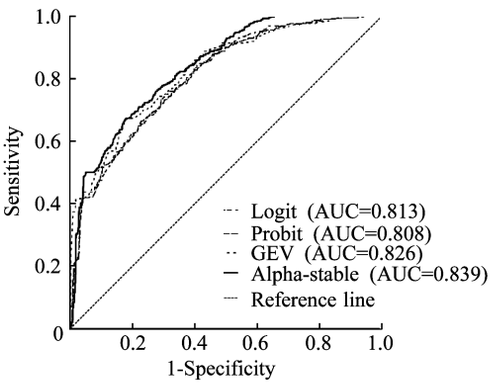


图 3 模拟数据集 1 中各模型 ROC 曲线和 AUC 值

Fig. 3 Cloglog Dataset: ROC curve and AUC value of different models on simulated dataset

4 结 论

本文对广义线性模型进行推广,给出了稳定分布模型,针对不平衡二元响应变量回归分析时,拥有更多偏度和拖尾厚度上的灵活性。针对稳定分布概率密度和累积分布函数没有解析表达式的特点,还给出了 EP-ABC 的估计方法。最后通过数值模拟实验,验证了稳定分布模型对二分类不平衡数据的估计效果与预测能力优于 Logit, Probit, Cloglog 以及 GEV 模型,该模型可以广泛地应用于经济、金融、信号处理等领域。

(下转第 142 页)

后迅速下降。转弯速度和轴质量的增加将增大Mises应力和最大剪应力的极大值,并明显增加深度小于30 mm范围内的应力值;增大曲线半径和设置合适外轨超高则能有效降低轮轨内部的Mises应力和最大剪应力。

参考文献:

[1] KALKER J J. On the rolling contact of two elastic bodies in the presence of dry friction[D]. The Netherlands: Delft University,1967.

[2] KALKER J J. Simplified Theory of Rolling Contact[M]. Netherlands:Delft University Press,1973:1-10.

[3] KALKER J J. A fast algorithm for the simplified theory of rolling contact[J]. Vehicle Systom Dynamics,1982,11:1-13.

[4] SHEN Z Y,HEDRICK J K,ELKINS J A. A comparison of alternative creep-force models for rail vehicles dynamic analysis[C]//Proceedings of 8th IAVSD-Symposium. Cambridge MA:[s. n.],1984:591-605.

[5] KALKER J J. The compution of three-dimensional rolling contact with dry friction[J]. International Journal Fornumerical Methods in Engineering,1979,14:1293-1307.

[6] KALKER J J. Variation principles of contact elastostatics [J]. J Inst Maths Applics,1977,20:199-219.

[7] PIOTROWSKI J,KIK W. A simplified model of wheel/rail contact mechanics for non-Hertzian problems and its application in rail vehicle dynamic simulations[J]. Vehicle System Dynamics,2008,46(1/2):27-48.

[8] PIOTROWSKI J,CHOLLET H. Wheel-rail contact models for vehicle system dynamics including multi-point contact[J]. Vehicle System Dynamics,2005,43(6/7):455-483.

[9] CRETU S. The influence of the rail inclination and lateral shift on pressure distribution in wheel-rail contact[J]. Acta Tribologica,2010,18:12-18.

[10] BARBINTA C I, YALDIZ S, DRAGOMIR A, *et al*. An elastic-plastic solver of the wheel-rail contact[C]//ASME 2010 10th Biennial Conference on Engineering Systems Design and Analysis. USA: American Society of Mechanical Engineers,2010:531-538.

[11] BARBINTA C I,ULIANOV C,FRANKLIN F,*et al*. Wheel-rail contact modelling and analysis,considering profiles types and lateral displacement[C]//Proceedings of the 5th TRA Conference. Paris: Gheorghe Asachi Technical University of Iasi,2014:1-10.

[12] 宋华,杨建,张月,等. 非线性稳态曲线通过时轮轨滚动接触的数值求解方法[J]. 中国铁道科学,2015,36(5):80-86.

[13] 郭俊,温泽峰,李伟,等. 切向载荷对钢轨弹塑性滚动接触应力的影响分析[J]. 润滑与密封,2010,35(7):5-8.

[14] 王少锋,周宇,许玉德,等. 轮对蠕滑条件下钢轨疲劳裂纹萌生寿命预测[J]. 同济大学学报(自然科学版),2014,42(6):894-899.

[15] 王彩芸,王文健,郭俊,等. 曲线半径对钢轨磨损影响的数值计算与试验分析[J]. 摩擦学学报,2010,30(6):584-588.

[16] ROVIRA A, RODA A, MARSHALL M B, *et al*. Experimental and numerical modelling of wheel-rail contact and wear[J]. Wear,2011,271(1/2):911-924.

[17] 黄龙文,李正美,安琦. 轮轨直线滚动过程接触力学性能计算方法研究[C]//第7届全国机械工程博士论坛论文集. 广西柳州:中国工程机械学会,2015:221-234.

[18] HUANG L W,LI Z M,LI L X,*et al*. Methods to calculate accurate wheel/rail contact positions and static contact stress levels[J]. Proceedings of the Institution of Mechanical Engineers:Part F. Journal of Rail and Rapid Transit,2016,230:138-150.

[19] 温泽峰,金学松,刘兴奇. 两种型面轮轨滚动接触蠕滑率和摩擦功[J]. 摩擦学学报,2001,21(4):208-212.

[20] 金学松,刘启跃. 轮轨摩擦学[M]. 北京:中国铁道出版社,2004.

(上接第132页)

参考文献:

[1] CZADO C,SANTNER T J. The effect of link misspecification on binary regression inference[J]. Journal of Statistical Planning and Inference,1992,33(2):213-231.

[2] WANG X,DEY D K. Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption[J]. The Annals of Applied Statistics,2010,4(4):2000-2023.

[3] NOPAN J P. Stable Distribution: Models for Heavy Tailed Data[M]. Washington D C:Birkhauser,2010.

[4] BARTHELME S, CHOPIN N. Expectation-propagation for likelihood-free inference[J]. Journal of the American Statistical Association,2014,109(505):315-333.

[5] TURNER B M, ZANDT T V. A tutorial on approximate Bayesian computation[J]. Journal of Mathematical Psychology,2012,56(2):69-85.

[6] SKAR C. The expectation propagation algorithm for use in approximate Bayesian analysis of latent gaussian models[D]. Norway:Institutt for Matematiske Fag,2010.