DOI:10.3969/j.issn.1003-5060.2023.05.022

响应变量随机删失时函数型非参数分位数回归模型的估计

杨锦涛, 凌能祥
(合肥工业大学 数学学院,安徽 合肥 230601)

摘 要:文章在响应变量随机删失时,研究了函数型非参数分位数回归模型,通过极小化逆概率加权分位数损失函数,构造模型中未知非参数函数的估计量。在一定的条件下,获得估计量的渐近正态性;通过模拟研究,验证了估计量的有效性。

关键词:函数型数据分析(FDA);分位数回归;随机删失;逆概率加权;渐近正态
中图分类号:O212.7 **文献标志码**:A **文章编号**:1003-5060(2023)05-0709-04

Estimation of nonparametric quantile regression model for functional data with censored response at random

YANG Jintao, LING Nengxiang
(School of Mathematics, Hefei University of Technology, Hefei 230601, China)

Abstract: In this paper, the nonparametric quantile regression model is presented to characterize the association between censored survival time and a set of functional predictors when response variables are censored at random, and estimates of nonparametric functions are obtained by minimizing the inverse probability weighted quantile loss function. Under some mild conditions, the asymptotic normality of the estimates is given. Simulation studies further verify the validity of the proposed model.

Key words: functional data analysis(FDA); quantile regression; censoring at random; inverse probability weighting; asymptotic normal

0 引 言

近年来,随着收集和存储数据技术的进步,人们越来越多地收集到具有函数特征的诸如曲线、曲面、图像等数据,称之为函数型数据。函数型数据广泛存在于生物学、化学计量学、计量经济学、医学、气象学、神经科学等领域,如何进行函数型数据分析(functional data analysis,FDA)受到很多学者的关注。有关FDA的背景、建模理论和方法见文献[1-3]。

分位数回归是研究解释变量与响应变量之间关系的主要统计方法之一,自文献[4]的开创性工作以来,很多学者在此方向开展了研究。这种方法与传统的最小二乘回归相比,有许多优点,对于异常值的处理比均值回归更加稳健,因此有更好的估计效率。事实上,分位数回归模型已经应用于分析函数型数据。例如,在函数型线性分位数回归模型中,文献[5]利用光滑样条基重新表示函数型协变量,并给出估计量的收敛速度。同时,主成分逼近方法被广泛应用于研究函数型回归模型[6]。此外,非参数统计方法在函数型数据方面的研究进展见文献[7]。

收稿日期:2022-03-18;**修回日期**:2022-04-12
基金项目:国家自然科学基金资助项目(72071068)
作者简介:杨锦涛(1996—),男,贵州铜仁人,合肥工业大学硕士生;
凌能祥(1968—),男,安徽合肥人,合肥工业大学教授,博士生导师.

在很多实际情况中,如抽样调查、生存分析、药物追踪测试和可靠性测试等,收集的数据可能是不完全的,如响应变量随机删失。最近,文献[8]提出函数型删失分位数模型,扩展经典截尾分位数回归中基于鞅的估计方法,以适应具有截尾响应和函数协变量的部分函数线性分位数回归模型,并给出估计量的渐近性质;文献[9]提出了响应变量删失时部分线性分位数回归模型。本文基于逆概率加权的方法,进一步研究响应变量删失时函数型非参数分位数回归模型的估计,并建立估计量的渐近正态性。

1 模型与方法

对于给定的分位数 $\tau\in(0,1)$,考虑如下的随机删失响应函数型非参数分位数回归模型:

$$\widetilde{Y}=m_\tau(X)+\varepsilon(\tau) \tag{1}$$

其中:$\widetilde{Y}$ 为响应变量;X 为半度量空间 $\mathscr{F}$ 上的函数型解释变量;$m_\tau(X)$ 为半度量空间 $\mathscr{F}$ 到实数集 $\mathbf{R}$ 上的未知的平滑函数型算子;在给定 X 的情况下,$\varepsilon(\tau)$ 的条件 τ 分位数为 0。在随机删失的情况下,实际只能观测到 $Y=\min(\widetilde{Y},C)$ 和删失示性变量 $\Delta=I(\widetilde{Y}\leqslant C)$,其中 C 为删失变量,其生存函数用 $G(\cdot)$ 表示。

考虑独立同分布样本 $\{(X_i,\widetilde{Y}_i),1\leqslant i\leqslant n\}$ 被完全收集,与核估计相似,本文提出的 $m_\tau(X)$ 的函数型分位数核估计如下:

$$\hat{m}_\tau(\chi)=\underset{m\in\mathbf{R}}{\operatorname{argmin}}\sum_{i=1}^{n}\rho_\tau(Y_i-m)K_h(d(\chi,X_i)),$$

其中:$\rho_\tau(u)=u(\tau-I)(u<0))$ 为分位数损失函数;$K_h=K(u/h)$。

然而在删失响应变量的机制下,本文得到不完全的样本 $\{(X_i,Y_i,\Delta_i),1\leqslant i\leqslant n\}$,此处 X_i 被完全观测。本文提出 $m_\tau(\chi)$ 的逆概率加权分位数估计如下:

$$\hat{m}_\tau(\chi)=\underset{m\in\mathbf{R}}{\operatorname{argmin}}\sum_{i=1}^{n}\frac{\Delta_i}{\hat{G}(Y_i)}\rho_\tau(Y_i-m)K_h(d(\chi,X_i)),$$

其中,$\hat{G}(\cdot)$ 为 $G(\cdot)$ 的 Kaplan-Meier 估计量,具体表达式[10]为:

$$\hat{G}(y)=1-\prod_{i=1}^{n}\left[\frac{n-i}{n-i+1}\mathrm{e}\right]^{D_i},$$

其中:$D_i=I\{Y_{(i)}\leqslant y,\Delta_{(i)}=0\}$;$\Delta_{(i)}$ 为对应于 $Y_{(i)}$ 的删失示性变量;$Y_{(i)}$ 为 $\{Y_{(i)},\ i=1,2,\cdots,n\}$ 的第 i 个次序统计量。

此外,本文提出的估计涉及到半度量,选取如下半度量[1]:

$$d_q^{[a,b]}(\chi_i,\chi_j)=\left(\int_a^b(\chi_i^{(q)}(t)-\chi_j^{(q)}(t))^2\mathrm{d}t\right)^{1/2}。$$

2 理论结果与假设

2.1 符号和假设条件

设 $B(\chi,h)=\{y|d(y,x)<h\}$ 为中心为 χ、半径为 h 的开球;$F_\chi(h)=P(d(X,\chi)<h)=P(X\in B(\chi,h))$ 为 X 的鞅分布函数。为了方便,设 $f_\tau(0|\chi)$ 和 $F_\tau(0|\chi)$ 分别为随机误差 $\varepsilon(\tau)$ 的条件密度函数和条件累计分布函数。设 c_1、c_2、c_3、c_4 是不依赖于 n 的正常数,它们在每次出现时可能取不同的值。

假设 1 K 为区间 $[0,1]$ 上的非负有界的核函数且 $K(1)=0$,在区间 $[0,1]$ 上 K' 存在且 $K'(t)<0$,$t\in[0,1]$ 并且 $\left|\int_0^1(K^j)'(t)\mathrm{d}t\right|<\infty$,$j=1,2$。

假设 2 存在最大值 L 和常数 $c_1>0$,使得 $P\{\widetilde{Y}>L\}>0$ 和 $P\{C\leqslant L\}=P\{C=L\}>c_1$。

假设 3 存在 $c_2>0$ 和 $\alpha>0$,对于任意 $u,v\in\mathscr{F}$,有 $|m_\tau(u)-m_\tau(v)|\leqslant c_2d(u,v)^\alpha$。

假设 4 存在 $c_3>0$ 和 $c_4>0$,且 $\varphi(h)\in(0,\infty)$,对任意 $\chi\in\mathscr{F}$,有

$$0<c_3\varphi(h)\leqslant P\{\chi\in B(\chi,h)\}\leqslant c_4\varphi(h)。$$

假设 5 当 $\chi\in\mathscr{F}$ 时,存在 1 个确定的非负有界函数 f_1 和趋向于 0 的非负的实函数 φ,满足

(1) $F_\chi(t)=\varphi(t)f_1(\chi)+o(\varphi(t))$,$t\to0$;

(2) 存在 1 个非递减的有界函数 μ_0,当 $t\in[0,1]$,$h\to0$ 且 $\int_0^1K(t)\mu_0\mathrm{d}t<\infty$ 时,有

$$\frac{\varphi(ht)}{\varphi(h)}=\mu_0(t)+o(1)。$$

假设 6 $\sup\limits_{0<z<L}|\hat{G}(z)-G(z)|=o_p\left(\frac{1}{\sqrt{n}}\right)$。

假设 1 为核函数常见的假设[11];假设 2 为删失数据分析中常见的假设[7],确保对任意个体没有删失的概率都是正的;假设 3~假设 5 为删失变量函数型数据分析中常见的假设[12];假设 6 为生存函数 $G(\cdot)$ 的常用的性质[11]。

2.2 主要结果

定理 1 设假设 1~假设 6 都成立,则有

$$\sqrt{n\varphi(h)}(\hat{m}_\tau(\chi)-m_\tau(\chi))\to N(0,\tilde{\omega}(\chi)),$$

其中

$$\tilde{\omega}(\chi)=\tau(1-\tau)\frac{M_2}{M_1^2}\frac{1}{f_1(\chi)f^2(0\mid\chi)}E\left(\frac{1}{G(Y)}\middle|\chi\right);$$

$$M_k=K^k(1)-\int_0^1(K^k)'(t)\mu_0(t)\mathrm{d}t,\ k=1,2。$$

3　模拟研究

本节通过蒙特卡洛模拟研究文中所提估计方法的实际表现。考虑如下的模型[6]：

$$\widetilde{Y}_i = m_\tau(X_i) + \varepsilon_i(\tau) \tag{2}$$

其中：$X_i(t)=1-\sin[W_i(t-\pi/3)]$，$W_i$ 互相独立服从标准正态分布（$i=1,2,\cdots,n$），t 为区间$[0,\pi/3]$上的 100 个等距值；$m_\tau(X_i)=\left[\int_0^{\pi/3} X_i'(t)\mathrm{d}t\right]^2$；$\varepsilon_1,\cdots,\varepsilon_n$ 为独立同分布的随机变量，且 $\varepsilon_i(\tau)=\varepsilon_i-F^{-1}(\tau)$，$F$ 是 ε_i 的累积分布函数。本文考虑随机误差为 $\varepsilon_i\sim N(0,0.25)$，删失变量[9] C_i 服从均匀分布 $U(0,c)$，其中 c 为控制删失比例的常数，取核函数为 $K(u)=\frac{3}{4}(1-u^2)I_{(0,1)}(u)$。

为了分析不同的分位数水平、不同删失率以及样本量对模型性能的影响，本文考虑分位数水平分别为 $\tau=0.25$、0.50、0.75，删失率分别为 20%和 40%，样本量 n 分别取 100、200、300，每个例子模拟 1 000 个数据集。此外，本文使用偏差（Bias）和根均方误差（root average squared error，RASE）评价估计量 $\hat{m}_\tau(\chi)$ 的精度。不同删失率下和不同的样本量下估计量 $\hat{m}_\tau(\chi)$ 偏差和根均方误差（RASE）见表 1 所列。

表 1　不同删失率下估计量 $\hat{m}(\chi)$ 的偏差和根均方误差

n	τ	20%		40%	
		偏差	RASE	偏差	RASE
100	0.25	0.047 1	0.027 7	0.109 2	0.054 2
	0.50	0.044 5	0.020 6	0.096 9	0.043 5
	0.75	0.045 4	0.017 7	0.086 3	0.039 6
200	0.25	0.035 1	0.017 6	0.086 3	0.039 6
	0.50	0.032 2	0.013 2	0.072 2	0.028 6
	0.75	0.029 8	0.010 2	0.063 5	0.019 1
300	0.25	0.033 1	0.015 2	0.079 3	0.034 3
	0.50	0.028 7	0.010 9	0.065 8	0.024 7
	0.75	0.024 2	0.007 4	0.056 7	0.014 8

从表 1 可以看出，估计量 $\hat{m}_\tau(\chi)$ 的准确性随样本量的增大而增加；随着删失率的增大，估计量的精确性下降，但仍是可行的；随着分位数水平的增大，估计量的精确性也随之增加。

4　定理证明

设 $r_i(\varepsilon_i)=r_i(\varepsilon_i\leqslant 0)-\tau$，对固定的随机函数 χ，设 $R_i=m_\tau(X_i)-m_\tau(\chi)$，$\theta(\alpha)=\sqrt{n\varphi(h)}(\alpha-m_\tau(\chi))$ 和 $\hat{\theta}=\theta(\alpha)$。

然后，设

$$\hat{l}_i(\theta)=\frac{\Delta_i}{\hat{G}(Y_i)}\left[\rho_\tau\left(\varepsilon_i+R_i-\frac{\theta}{\sqrt{n\varphi(h)}}\right)-\rho_\tau(\varepsilon_i+R_i)\right]K_h(d(\chi,X_i)),$$

$$l_i(\theta)=\frac{\Delta_i}{G(Y_i)}\left[\rho_\tau\left(\varepsilon_i+R_i-\frac{\theta}{\sqrt{n\varphi(h)}}\right)-\rho_\tau\varepsilon_i+R_i)\right]K_h(d(\chi,X_i))。$$

引理 1　设假设 1～假设 6 都成立，则有

$$E\left(\frac{1}{G(Y_i)}K_h^k\,d(\chi,X_i)\,\middle|\,X\right)=\varphi(h)\left[M_k f_1(\chi)E\left(\frac{1}{G(Y)}\,\middle|\,X\right)+O_{a.s}\left(\frac{g_{i,\chi}(h)}{\varphi(h)}\right)E\left(\frac{1}{G(Y)}\,\middle|\,X\right)\right],$$

$$E\left(\frac{\Delta_i}{G(Y_i)}K_h^k\,d(\chi,X_i)\right)=\varphi(h)(M_k f_1(\chi)+o(1)),k=1,2,$$

且

$$\frac{1}{n\varphi(h)}\sum_{i=1}^{n}E(K_h d(\chi,X_i)f(0\mid X_i))=M_1 f(0\mid\chi)+o(1)。$$

引理 1 的证明类似于文献[13]中引理 1 的证明。通过引理 1 和 Lindeberg-Feller 中心极限定理，得到如下引理 2。

引理 2　设假设 1～假设 6 都成立，则有

$$\frac{1}{n\varphi(h)}\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}K_h(d(\chi,X_i))r_i(\varepsilon_i)\to N(0,\sigma^2(\chi)),\ n\to\infty。$$

其中

$$\sigma^2(\chi)=\tau(1-\tau)M_2 f_1(\chi)E\left(\frac{1}{G(Y)}\,\middle|\,X\right)。$$

定理 1 的证明　与文献[12]引理 3 的证明相似，有 $\sup_{|\theta|\leqslant L}|\hat{l}(\theta)-l(\theta)|=o_p(1)$。于是有如下等式：

$$\sum_{i=1}^{n}\frac{\Delta_i}{\hat{G}(Y_i)}\left[\rho_\tau\left(\varepsilon_i+R_i-\frac{\theta}{\sqrt{n\varphi(h)}}\right)-\rho_\tau(\varepsilon_i+R_i)\right]K_h(d(\chi,X_i))=$$

$$\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}\left[\rho_\tau\left(\varepsilon_i+R_i-\frac{\theta}{\sqrt{n\varphi(h)}}\right)-\rho_\tau(\varepsilon_i+R_i)\right]K_h(d(\chi,X_i))=$$

$$\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}K_h(d(\chi,X_i))\frac{\theta}{\sqrt{n\varphi(h)}}\times(I_{\{\varepsilon_i<0\}}-\tau)+\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}K_h(d(\chi,X_i))\times$$

$$\int_0^{\frac{\theta}{\sqrt{n\varphi(h)}}}(I_{\{\varepsilon_i+R_i<s\}}-I_{\{\varepsilon_i<0\}})\mathrm{d}s \tag{3}$$

与文献[12]引理 6 的证明相似,有如下等式:

$$E\Big(\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}K_h(d(\chi,X_i))\times \int_0^{\frac{\theta}{\sqrt{n\varphi(h)}}}(I_{\{\varepsilon_i+R_i<s\}}-I_{\{\varepsilon_i<0\}})\mathrm{d}s\Big)= \frac{1}{2}\theta^2M_1f_1(\chi)f(0\mid\chi)+o(1) \tag{4}$$

根据文献[6],可以证明

$$\mathrm{Var}\Big(\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}K_h(d(\chi,X_i))\times \int_0^{\frac{\theta}{\sqrt{n\varphi(h)}}}(I_{\{\varepsilon_i+R_i<s\}}-I_{\{\varepsilon_i<0\}})\mathrm{d}s\Big)=o_p(1) \tag{5}$$

由(4)式和(5)式可得:

$$\sum_{i=1}^{n}\frac{\Delta_i}{\hat{G}(Y_i)}\Big[\rho_\tau\Big(\varepsilon_i+R_i-\frac{\theta}{\sqrt{n\varphi(h)}}\Big)-\rho_\tau(\varepsilon_i+R_i)\Big]K_h(d(\chi,X_i))=$$

$$\sum_{i=1}^{n}\frac{\Delta_i}{G(Y_i)}K_h(d(\chi,X_i))\frac{\theta}{\sqrt{n\varphi(h)}}(I_{\{\varepsilon_i<0\}}-\tau)+\frac{1}{2}\theta^2M_1f_1(\chi)f(0\mid\chi)+o_p(1)=$$

$$\theta W_n+\frac{1}{2}\theta^2M_1f_1(\chi)f(0\mid\chi)+o_p(1),$$

其中,W_n 是均值为 0、方差为

$$\sigma^2(\chi)=\tau(1-\tau)M_2f_1(\chi)E\Big(\frac{1}{G(Y)}\Big|\chi\Big)$$

的正态随机变量。根据文献[14]的理论结果,有

$\hat{\theta}=(M_1f_1(\chi)f(0\mid\chi))^{-1}W_n+o_p(1)$。

结合引理 2,得到如下结论:

$$\sqrt{n\varphi(h)}(\hat{m}_\tau(\chi)-m_\tau(\chi))\to N(0,\tilde{\omega}(\chi)),$$

其中

$$\tilde{\omega}(\chi)=\tau(1-\tau)\frac{M_2}{M_1}\times\frac{1}{f_1(\chi)f^2(0\mid\chi)}E\Big(\frac{1}{G(Y)}\Big|\chi\Big)。$$

5 结 论

本文基于随机删失逆概率加权的方法,提出响应变量随机删失时函数型非参数分位数回归模型的一种估计方法,并且给出估计量的渐近正态性,模拟实验验证了所提出方法的优越性和可行性。

[参 考 文 献]

[1] RAMSARY J O,SILVERMAN B W. Functional data analysis[M]. New York:Springer,2005:1-18.

[2] FERRATY F,VIEU P. Nonparametric functional data analysis: theory and practice [M]. New York: Springer, 2006:47-68.

[3] HORVATH L,KOKOSZKA P. Inference for functional data with applications[M]. New York: Springer Science & Business Media,2012:32-54.

[4] KOENKER R,BASSETT G,Jr. Regression quantiles[J]. Econometrica: Journal of the Econometric Society, 1978, 46(2):33-50.

[5] CARDOT H,CRAMBES C,SARDA P. Quantile regression when the covariates are functions[J]. Nonparametric Statistics,2005,17(7):841-856.

[6] TANG L J, ZHOU Z G, WU C C. Weighted composite quantile estimation and variable selection method for censored regression model[J]. Statistics & Probability Letters, 2012,82(3):653-663.

[7] LING N X,VIEU P. Nonparametric modelling for functional data:selected survey and tracks for future[J]. Statistics, 2018,52(4):934-949.

[8] JIANG F,CHENG Q,YIN G S,et al. Functional censored quantile regression[J]. Journal of the American Statistical Association,2020,115(530):931-944.

[9] 史功明,张忠占,谢田法. 响应变量删失时函数型部分线性分位数回归模型的估计[J]. 数学的实践与认识,2021, 51(3):153-165.

[10] FENG H L,LUO Q Q. A weighted quantile regression for nonlinear models with randomly censored data[J]. Communications in Statistics-Theory and Methods, 2021, 50(18):4167-4179.

[11] KAPLAN E L,MEIER P. Nonparametric estimation from incomplete observations[J]. Journal of the American Statistical Association,1958,53(282):457-481.

[12] XU D K,DU J. Nonparametric quantile regression estimation for functional data with responses missing at random [J]. Metrika,2020,83(8):977-990.

[13] LAIB N,LOUANI D. Nonparametric kernel regression estimation for functional stationary ergodic data:asymptotic properties[J]. Journal of Multivariate Analysis, 2010, 101(10):2266-2281.

[14] GEYER C J. On the asymptotics of constrained M-estimation[J]. The Annals of Statistics,1994,22(4):1993-2010.

(责任编辑 朱晓临)