山东大学学报(理学版)2025 年 2 月 第 60 卷 第 2 期
Journal of Shandong University(Natural Science), Vol.60, No.2, 2025

E-mail:xblxb@sdu.edu.cn
http://lxbwk.njournal.sdu.edu.cn
Tel:+86-531-88366917

文章编号:1671-9352(2025)02-0096-09 DOI:10.6040/j.issn.1671-9352.0.2024.021

删失分位数回归模型中的多变点估计

李学文[1],冯可馨[2],王小刚[2*]
(1.北方民族大学商学院, 宁夏 银川 750021; 2.北方民族大学数学与信息科学学院, 宁夏 银川 750021)

摘要:针对删失分位数回归模型中的变点个数、位置及模型参数同时估计问题,基于线性化技术得到参数的有效估计,消除目标函数不可导与非凸的困难。该方法能捕捉响应变量受到某一协变量的影响而存在的多个变点,能更好理解复杂非线性关系的同时保持较快的收敛速度,兼顾灵活性与可解释性。数值模拟验证估计方法在不同分位点、同(异)方差情形下具备有效性和稳健性,实证分析发现存在 2 个变点,并对其进行解释。

关键词:多变点估计;删失数据;分位数回归模型;线性化技术
中图分类号:O212.2 **文献标志码**:A
引用格式:李学文,冯可馨,王小刚.删失分位数回归模型中的多变点估计[J]. 山东大学学报(理学版),2025,60(2):96-104.

Estimation of multiple change points for censored quantile regression model

LI Xuewen[1], FENG Kexin[2], WANG Xiaogang[2*]
(1. School of Business, North Minzu University, Yinchuan 750021, Ningxia, China; 2. School of Mathematics and Information Science, North Minzu University, Yinchuan 750021, Ningxia, China)

Abstract: To simultaneously estimate the number of change points, the location of change points and the model parameters in censored quantile regression model, a linearization technique is employed to obtain estimators for above parameters. This approach overcomes the issues of non-differentiability and non-convexity objective function at the change points. It is capable of capturing the relationship between response and covariate of interest that changes across multiple change points. Furthermore, the proposed estimators strike a balance between flexibility and interpretability, making them become a useful tool for identifying and explaining change points. Simulation studies show that the estimators demonstrate robustness in both homoscedastic and heteroscedastic conditions across various quantile levels. An empirical analysis reveals the existence of two change points and their change point effects.
Key words: estimation multiple change point; censored data; quantile regression model; linearization technique

0 引言

分位数回归模型放松了对误差分布的限制,解决了强影响值的影响,能更加稳健地反映分布尾部的特征,受到了众多学者的关注[1]。在研究中常会遇到某种因素导致数据存在删失,需要开展删失数据分位数回归模型的研究。文献[2]针对响应变量存在固定删失时提出了删失 Tobit 回归模型估计方法,并拓展至删失分位数回归模型中[3], 一些学者采用重新分配删失权重方法提出了递归加权分位数回归估计[4],在删失数据中提出分位数回归模型估计[5],结合工具变量对删失分位数回归模型提出有效估计[6],基于神经网络方法与数据增强方法构造删失分位数回归模型估计[7-8]。删失分位数回归模型隐含回归函数保持线性结构,

收稿日期: 2024-01-19; **网络出版时间**: 2024-10-24 13:58:28
基金项目:宁夏自然科学基金重点项目(2023AAC02043); 全国统计科学研究项目(2023LY070); 宁夏高等教育一流学科建设基金资助项目(NXYLXK2017B09)
第一作者:李学文(1980—),女,讲师,博士,研究方向为风险管理. E-mail:421630014@qq.com
*通信作者**:王小刚(1980—),男,教授,博士,研究方向为复杂数据统计推断、风险管理. E-mail:wangxg9102@163.com

但新技术实施、经济金融危机蔓延等外部冲击可能使得模型结构发生改变,从而产生变点,此时该模型无法准确衡量非线性特征,因此有必要研究含变点的删失分位数回归模型。

以往变点研究多从单变点开始,变点估计方法有格点搜索法[9-10]、光滑化方法[11-12]与线性化技术[13]等。格点搜索法在变点位置估计时具有一定灵活性,但变点位置估计迭代速度有待提高,且变点位置估计的实际意义解释性不强,光滑化方法可以得到相合估计,导致效率损失。线性化方法将目标函数中不可导项进行线性化转换,简单易行且算法收敛速度快。文献[14]借鉴线性化技术在分位数回归模型中得到模型参数和变点估计。上述研究大多基于单变点展开,通常假设变点个数是外生的,考虑到风险冲击下模型结构可能存在不只一个变点,因此有必要研究变点个数的估计方法。线性化技术结合广义贝叶斯信息准则可得到变点个数估计[15-17],文献[18]在删失数据分段线性 Tobit 回归模型中通过贝叶斯信息准则得到变点个数、位置及模型参数估计。线性化技术弥补了存在变点时目标函数不可导和非凸困难,方法灵活,得到了广泛应用。

研究删失分位数回归模型中的变点个数及位置估计能更好地解释模型结构变动,提供有效的多变点个数估计方法,对分析变点效应的内在机制、揭示数据本质特征等方面有着重要的理论意义和实践价值。本文的研究价值主要有以下 3 个方面:一是在删失分位数回归模型中提出了未知的变点个数估计方法,结合变点位置估计与检验方法可完全解决删失分位数回归模型中的变点问题,丰富和完善了变点估计理论与应用范围;二是分段线性模型结构简化了非线性特征,保证了模型的易解释与灵活性;三是内生的多变点假设更符合真实情况,能有效解释交叉效应,为实证研究提供了支撑。

1　单变点删失分位数回归模型估计与模拟

1.1　单变点删失分位数回归模型

设$(T_i,x_i,\boldsymbol{z}_i)$,$i=1,2\cdots,n$,是总体$(T,x,\boldsymbol{z})$中独立同分布的样本序列,$T_i$ 是响应变量,x_i、$\boldsymbol{z}_i$ 分别是 1 维和 p 维协变量,T_i 在未知位置 t 处存在变点,τ 为分位数$(0<\tau<1)$。单变点删失分位数回归模型如下:

$$Y_i=\min\{T_i,C\},\quad \delta_i=I(T_i<C),$$

$$T_i=\alpha+\beta x_i+\gamma(x_i-t)_+ +\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i+e_i,$$

$$Q_{T_i}(\tau\mid x_i,z_i)=\alpha+\beta x_i+\gamma(x_i-t)_+ +\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i,$$

其中,C 和 δ 为删失常量和删失指标,$(x_i-t)_+=(x_i-t)I(x_i>t)$,$I(\cdot)$为示性函数,$e_i$ 为误差项且 $E(e_i)=0$,$\mathrm{Var}(e_i)<\infty$,当 $\gamma\neq0$ 时模型存在变点。令 $\boldsymbol{\xi}=(\alpha,\beta,\gamma,\boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}}$ 表示模型参数,$\boldsymbol{\theta}=(\boldsymbol{\xi}^{\mathrm{T}},t)^{\mathrm{T}}$ 表示所有参数,$\boldsymbol{\nu}_i(t)=(1,x_i,(x_i-t)_+,\boldsymbol{z}_i)^{\mathrm{T}}$。为了方便,记 $g(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))=\min\{C,g^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))\}$,其中 $g^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))=\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\nu}_i(t)$。

若 $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\xi}}^{\mathrm{T}},\hat{t})^{\mathrm{T}}$ 表示变点及模型参数估计,则 $\hat{\boldsymbol{\theta}}$ 可由最小化目标函数(1)得到,即

$$\hat{\boldsymbol{\theta}}=\arg\min_{\boldsymbol{\theta}}\sum_{i=1}^{n}\rho_\tau(y_i-\min\{C,g^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))\})=\arg\min_{\boldsymbol{\theta}}\sum_{i=1}^{n}\rho_\tau(y_i-g(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))),\tag{1}$$

其中损失函数为 $\rho_\tau(u)=u(\tau-I(u<0))$。

1.2　估计方法

由于目标函数(1)在变点处不可导且非凸,因此利用线性化技术将不可导的示性函数在初始值 $t^{(0)}$ 处线性展开,即

$$(x_i-t)_+\approx(x_i-t^{(0)})_+ +(-1)I(x_i>t^{(0)})(t-t^{(0)})。$$

目标函数(1)中的 $g^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))$可写成

$$g^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))=\alpha+\beta x_i+\gamma q_i+\gamma_1 p_i+\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i,$$

其中,$\gamma_1=\gamma(t-t^{(0)})$,$q_i=(x_i-t^{(0)})_+$,$p_i=-I(x_i>t^{(0)})$,通过最小化新的目标函数(2)来估计 $\hat{\boldsymbol{\xi}}^{(1)}$

$$\sum_{i=1}^{n}\rho_\tau(y_i-\min\{C,\alpha+\beta x_i+\gamma q_i+\gamma_1 p_i+\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i\})。\tag{2}$$

变点 t 可迭代得到,即 $t^{(1)}=t^{(0)}+\gamma_1^{(1)}/\gamma^{(1)}$,重复上述过程直至收敛,参数估计见算法 1。

算法 1 单变点迭代算法

步骤 1 给定初始参数 $\hat{\boldsymbol{\xi}}^{(0)}=(\hat{\alpha}^{(0)},\hat{\beta}^{(0)},\hat{\gamma}^{(0)},\gamma_1^{(0)},\hat{\boldsymbol{\zeta}}^{(0)\mathrm{T}})^{\mathrm{T}}$ 及 $\hat{t}^{(0)}$,为保证收敛不妨令 $\gamma_1^{(0)}=0.01$。

步骤 2 第 k 步,固定 $\hat{t}^{(k)}$ 最小化目标函数(3)更新回归系数估计

$$\hat{\boldsymbol{\xi}}^{(k)}(t)=\arg\min_{\boldsymbol{\xi}}\sum_{i=1}^{n}\rho_\tau(y_i-\min(C,\alpha+\beta x_i+\gamma q_i^{(k)}+\gamma_1 p_i^{(k)}+\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i)), \tag{3}$$

其中,$q_i^{(k)}=(x_i-t^{(k)})_+$,$p_i^{(k)}=-I(x_i>t^{(k)})$。

步骤 3 更新变点估计,$\hat{t}^{(k+1)}=\hat{t}^{(k)}+\gamma_1^{(k+1)}/\hat{\gamma}^{(k+1)}$。

步骤 4 重复步骤 2、3,直至算法收敛。

变点 t 的标准误差可通过 delta 方法得到,

$$\mathrm{SE}(\hat{t})=\left[\mathrm{Var}(\gamma_1)+\mathrm{Var}(\gamma_1)\left(\frac{\gamma_1}{\hat{\gamma}}\right)^2-2\left(\frac{\gamma_1}{\hat{\gamma}}\right)\mathrm{Cov}(\hat{\gamma},\gamma_1)\right]^{1/2}/|\hat{\gamma}|。 \tag{4}$$

算法 1 收敛时,系数 $\gamma_1\to 0$,式(4)可简化为 $\mathrm{SE}(\hat{t})=\mathrm{SE}(\gamma_1)/|\hat{\gamma}|$,变点的置信区间估计为$[\hat{t}-\phi_{\alpha/2}\mathrm{SE}(\hat{t}),\hat{t}+\phi_{\alpha/2}\mathrm{SE}(\hat{t})]$,$\phi_{\alpha/2}$是标准正态分布的$(1-\alpha/2)$分位数。

1.3 数值模拟

为了验证单变点估计算法的有效性和稳健性,设置同方差与异方差情形的数据生成过程。

同方差情形:$T_i=\alpha+\beta x_i+\gamma(x_i-t)_+ +\lambda\omega_i+e_i$, $i=1,2,\cdots,n$。

异方差情形:$T_i=\alpha+\beta x_i+\gamma(x_i-t)_+ +\lambda\omega_i+(0.5+0.2x_i)e_i$, $i=1,2,\cdots,n$。

假设 $x_i\sim U(-2,4)$, $t=1.5$, $\omega_i\sim N(1,2^2)$, $(\alpha,\beta,\gamma,\lambda)^{\mathrm{T}}=(1,2,-3,4)^{\mathrm{T}}$, $e_i\sim N(0,1)$, $C=35$,模型的删失度为 35%。设置 $n=200,500$, $\tau=0.25,0.50,0.75$,重复次数 500 次,模拟结果见表 1。

表 1 多变点模拟结果

Table 1 Simulation results of single change point

(n,τ)	COEF	同方差				异方差			
		BIAS	SD	ESE	CP	BIAS	SD	ESE	CP
(200,0.25)	α	0.021	0.197	0.199	0.922	−0.035	0.268	0.275	0.826
	β	0.006	0.076	0.073	0.934	−0.028	0.088	0.097	0.906
	γ	−0.019	0.212	0.228	0.932	0.023	0.344	0.309	0.922
	ζ	0.003	0.038	0.036	0.944	−0.009	0.033	0.034	0.924
	t	−0.013	0.174	0.177	0.932	−0.018	0.206	0.186	0.898
(200,0.50)	α	0.009	0.194	0.192	0.932	0.005	0.210	0.216	0.930
	β	0.003	0.070	0.072	0.952	0.029	0.163	0.152	0.938
	γ	0.005	0.198	0.209	0.948	−0.006	0.269	0.275	0.940
	λ	−0.001	0.034	0.036	0.958	0.004	0.037	0.044	0.966
	t	−0.007	0.165	0.163	0.950	−0.023	0.174	0.158	0.912
(200,0.75)	α	0.019	0.210	0.208	0.908	0.014	0.234	0.238	0.900
	β	0.004	0.076	0.079	0.934	−0.039	0.162	0.176	0.874
	γ	−0.013	0.220	0.236	0.910	0.044	0.395	0.343	0.902
	λ	0.005	0.037	0.041	0.950	−0.008	0.047	0.048	0.930
	t	−0.022	0.190	0.185	0.902	−0.039	0.196	0.194	0.900
(500,0.25)	α	−0.015	0.167	0.161	0.930	−0.032	0.171	0.162	0.920
	β	0.005	0.063	0.064	0.946	−0.011	0.081	0.079	0.940
	γ	0.007	0.152	0.163	0.936	−0.031	0.213	0.219	0.922
	λ	−0.008	0.026	0.029	0.958	−0.015	0.031	0.033	0.950
	t	−0.003	0.116	0.118	0.920	−0.013	0.115	0.115	0.910

表 1(续)

(n,τ)	COEF	同方差				异方差			
		BIAS	SD	ESE	CP	BIAS	SD	ESE	CP
(500,0.50)	α	−0.004	0.116	0.114	0.938	0.005	0.132	0.129	0.932
	β	0.001	0.041	0.043	0.964	0.006	0.032	0.031	0.930
	γ	0.001	0.123	0.126	0.954	0.007	0.196	0.187	0.960
	λ	−0.001	0.022	0.021	0.960	−0.005	0.026	0.029	0.952
	t	−0.002	0.102	0.098	0.952	−0.006	0.106	0.099	0.916
(500,0.75)	α	0.021	0.190	0.189	0.916	0.039	0.149	0.144	0.898
	β	0.008	0.077	0.075	0.942	0.013	0.121	0.108	0.930
	γ	0.005	0.159	0.167	0.930	−0.018	0.199	0.204	0.918
	λ	−0.004	0.025	0.028	0.946	0.011	0.031	0.029	0.920
	t	−0.017	0.129	0.120	0.916	−0.020	0.137	0.108	0.912

表 1 中包含偏差(bias, BIAS)、标准差(standard deviation, SD)、标准误差(estimate standard error, ESE)、95%置信区间覆盖率(coverage probability, CP)模拟结果。当 $n=200$、$\tau=0.50$ 且误差为同方差时,BIAS和 SD 均很小,SD 与 ESE 均很接近,CP 在 95%附近,表明估计结果有效。当 $n=200$、$\tau=0.50$ 且误差为异方差时,BIAS 和 SD 仍然不大,SD 与 ESE 接近,CP 比同方差时略差,表明估计结果在异方差情况下较为稳定,但估计效果不如同方差好。当 $n=200$、$\tau=0.25$ 和 0.75 时,BIAS 和 SD 比 $\tau=0.50$ 时估计结果稍差,但在可接受范围之内,SD 与 ESE 仍较为接近,说明估计具有稳健性。当 $n=500$ 时,估计的 BIAS、SD、ESE 减小,CP 增大,SD 与 ESE 仍很接近。

总体上看,当保持其他因素不变时,同方差的估计效果比异方差效果好,$\tau=0.50$ 时的估计效果优于 τ 为 0.25 和 0.75。随着样本量增加,BIAS、SD 和 ESE 都呈减小趋势,CP 呈现增大趋势,且 SD 与 ESE 较为接近,表明估计方法的有效性与稳健性。

2　多变点删失分位数回归模型参数估计与模拟

2.1　多变点删失分位数回归模型

令 $t_1,t_2,\cdots,t_m$ 分别表示 m 个未知变点,$\boldsymbol{\xi}=(\alpha,\beta,\gamma_1,\gamma_2,\cdots,\gamma_m,\boldsymbol{\zeta}^{\mathrm{T}})^{\mathrm{T}}$ 为未知参数,$\boldsymbol{\theta}=(\boldsymbol{\xi}^{\mathrm{T}},t_1,t_2,\cdots,t_m)^{\mathrm{T}}$,$\boldsymbol{\nu}_i(t)=(1,x_i,(x_i-t_1)_+,\cdots,(x_i-t_m)_+,\boldsymbol{z}_i)^{\mathrm{T}}$,令

$$f(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))=\min\{C,f^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))\},$$

$$f^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))=\alpha+\beta x_i+\sum_{k=1}^{m}\gamma_k(x_i-t_k)_+ +\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i。$$

类似地,通过最小化目标函数(5)得到参数估计

$$\hat{\boldsymbol{\theta}}_n(\tau)=\arg\min_{\boldsymbol{\theta}}\sum_{i=1}^{n}\rho_\tau(y_i-f(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau)))。\tag{5}$$

2.2　估计方法

为了得到多变点个数与位置估计,本节分为两部分加以解决。首先,假设变点个数 m 已知但变点位置未知,此时采用单变点估计方法解决,即

$$(x_i-t_k^{(1)})_+\approx(x_i-t_k^{(0)})_+ +(-1)I(x_i>t_k^{(0)})(t_k^{(1)}-t_k^{(0)}),$$

目标函数(5)中的$f^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))$可写成

$$f^*(\boldsymbol{\nu}_i;\boldsymbol{\theta}(\tau))=\alpha+\beta x_i+\sum_{k=1}^{m}\gamma_k P_{ik}+\sum_{k=1}^{m}\delta_k Q_{ik}+\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i,\tag{6}$$

其中,$P_{ik}=(x_i-t_k^{(0)})_+$,$Q_{ik}=-I(x_i>t_k^{(0)})$,$\delta_k=\gamma_k(t_k-t_k^{(0)})$。除变点以外的参数估计可通过最小化如下新的目标函数(7)得到

$$\sum_{i=1}^{n}\rho_\tau(y_i-\min(C,\alpha+\beta x_i+\sum_{k=1}^{m}\gamma_k P_{ik}+\sum_{k=1}^{m}\delta_k Q_{ik}+\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i))。\tag{7}$$

变点 t_k 由下述公式迭代得到 $t_k^{(1)}=t_k^{(0)}+\delta_k^{(1)}/\gamma_k^{(1)}$,重复上述过程直至收敛,变点 m 已知时模型参数与变点位置估计方法见算法 2。

算法 2 变点个数已知时的迭代算法

步骤 1 给定变点初始值$(t_1^{(0)},t_2^{(0)},\cdots,t_m^{(0)})$。

步骤 2 对第 s 步,通过最小化目标函数(7)得到参数估计

$$\hat{\boldsymbol{\xi}}^{(s)}=(\hat{\alpha}^{(s)},\hat{\beta}^{(s)},\gamma_1^{(s)},\gamma_2^{(s)},\cdots,\gamma_m^{(s)},\delta_1^{(s)},\delta_2^{(s)},\cdots,\delta_m^{(s)},\hat{\boldsymbol{\zeta}}^{(s)})^{\mathrm{T}}。$$

步骤 3 更新变点估计 $\hat{t}_k^{(s+1)}$,即 $\hat{t}_k^{(s+1)}=\hat{t}_k^{(s)}+\gamma_k^{(s+1)}/\delta_k^{(s+1)}$。

步骤 4 重复步骤 2、3,直至算法收敛。

其次,当变点个数未知时,设定初始变点个数 M 远大于真实值 m,赋值$(t_1,t_2,\cdots,t_M)^{\mathrm{T}}$ 均匀地分布在定义域内。利用算法 2 进行估计,每次迭代可获得新的变点位置估计$(t_1',t_2',\cdots,t_M')^{\mathrm{T}}$ 与系数估计$(\gamma_1',\gamma_2',\cdots,\gamma_M')^{\mathrm{T}}$,需要注意的是以下 2 类变点估计须剔除。

1) 当 t_k 处不存在变点,$\gamma_k\to 0$,迭代后将无实际意义须剔除。

2) 当迭代后$(t_1',t_2',\cdots,t_M')^{\mathrm{T}}$ 估计中若干变点位置相差较小时,保留一个,其余剔除。

剔除上述变点后,假设还剩余 M_1 个变点。为了进一步得到变点个数估计,利用贝叶斯信息准则将变点个数$(t_1',t_2',\cdots,t_{M_1}')^{\mathrm{T}}$ 估计问题转化为变量选择问题,得到 M_2 个变点,即为变点个数估计。贝叶斯信息准则(Bayesian information criterion, BIC)为

$$\mathrm{BIC}=\log(\hat{\sigma}^2)+e_{\mathrm{df}}\frac{\log(n)}{n}C_n,$$

其中,$\hat{\sigma}^2$ 为残差的方差估计,n 表示样本容量,e_{df}为模型自由度,$e_{\mathrm{df}}=1+2M_2$,M_2 表示模型的变点个数,$C_n=\log[\log(n)]$。变点个数未知时的估计方法见算法 3。

算法 3 变点个数未知时的迭代算法

步骤 1 给定初始变点集,变点个数为 M,真实变点个数 m,且 $m\ll M$。

步骤 2 通过目标函数(7)得到模型参数估计。

步骤 3 剔除不符合要求的变点后,变点个数为 M_1。

步骤 4 通过贝叶斯信息准则得到 M_2 个变点,即变点个数估计。

步骤 5 利用估计的变点个数,通过算法 2 得到变点位置与模型参数估计。

变点 t_k 的标准误差可通过 delta 方法计算[16],即

$$\mathrm{SE}(\hat{t}_k)=\left[\mathrm{Var}(\hat{\delta}_k)+\mathrm{Var}(\hat{\delta})\left(\frac{\hat{\delta}_k}{\hat{\gamma}_k}\right)^2-2\left(\frac{\hat{\delta}_k}{\hat{\gamma}_k}\right)\mathrm{Cov}(\hat{\gamma}_k,\hat{\delta}_k)\right]^{1/2}/|\hat{\gamma}_k|。\tag{8}$$

当算法 3 收敛时,$\hat{\delta}_k\to 0$,式(8)变为 $\mathrm{SE}(\hat{t}_k)=\mathrm{SE}(\hat{\delta}_k)/|\hat{\gamma}_k|$,$\hat{t}_k$ 的置信区间为 $[\hat{t}_k-\phi_{\alpha/2}\mathrm{SE}(\hat{t}_k),\hat{t}_k+\phi_{\alpha/2}\mathrm{SE}(\hat{t}_k)]$。

2.3 数值模拟

本节先给出变点个数模拟结果,然后给出模型参数与变点位置模拟结果。

1) 变点个数模拟

假设 $x_i\sim U(-2,4)$, $\omega_i\sim N(1,2^2)$,模型参数$(\alpha,\beta,\gamma_1,\gamma_2,\lambda)^{\mathrm{T}}=(1,2,-3,4,1)^{\mathrm{T}}$, 变点$(t_1,t_2)^{\mathrm{T}}=(0,2)^{\mathrm{T}}$, 固定删失值 $C=10$, 删失度为 20%。数据生成过程如下

$$Y_i=\min\{T_i,C\},\quad \delta_i=I(T_i<C),$$

$$T_i=\alpha+\beta x_i+\gamma_1(x_i-t_1)_+ +\gamma_2(x_i-t_2)_+ +\lambda\omega_i+e_i,\quad i=1,2,\cdots,n,$$

分别在误差项为 $e\sim N(0,1)$ 和 $e\sim(0.5+0.2x)N(0,1)$ 进行模拟,设置重复次数为 500,在样本量分别为 200、500 的情况下估计变点个数和变点位置,结果见表 2,其中准确率记为 AR(accuracy rate),定义为 $\sum I(\hat{K}=2)/500$。

表 2　变点个数估计的准确率及变点位置估计

Table 2　Accuracy rate of the number and location of change points estimation

(n,τ)	同方差					异方差				
	t_1		t_2		AR	t_1		t_2		AR
	BIAS	SD	BIAS	SD		BIAS	SD	BIAS	SD	
(200,0.25)	−0.331	0.186	−0.395	0.162	0.54	0.863	0.359	0.242	0.205	0.40
(200,0.50)	0.178	0.073	0.101	0.104	0.75	−0.246	0.187	−0.197	0.149	0.55
(200,0.75)	0.266	0.165	−0.028	0.150	0.56	0.785	0.277	−0.298	0.215	0.45
(500,0.25)	0.145	0.083	0.181	0.112	0.75	0.452	0.230	−0.218	0.136	0.42
(500,0.50)	0.115	0.069	0.052	0.099	0.78	0.236	0.159	0.146	0.134	0.66
(500,0.75)	−0.215	0.096	−0.163	0.108	0.60	−0.394	0.227	0.183	0.191	0.48

由表 2 可知,当 $n=200$、$\tau=0.50$ 时,变点个数估计的准确率 AR 为 75%;当 $n=500$、$\tau=0.50$ 时,AR 达到 78%, 估计结果较为稳健;同等样本量下,$\tau=0.25$ 与 $\tau=0.75$ 时变点个数估计的准确率均低于 $\tau=0.50$。另外,变点位置估计的 BIAS 和 SD 均不大, 且随着样本量的增加而减小。综上可知,变点个数估计有效且稳健。

2) 变点位置、模型参数模拟

在两变点删失回归模型中分别对同方差及异方差情形进行模拟。

同方差情形: $T_i=\alpha+\beta x_i+\gamma_1(x_i-t_1)_+ +\gamma_2(x_i-t_2)+\lambda\mu_i+e_i,\ i=1,2,\cdots,n$。

异方差情形: $T_i=\alpha+\beta x_i+\gamma_1(x_i-t_1)_+ +\gamma_2(x_i-t_2)+\lambda\mu_i+(0.5+0.2x_i)e_i,\ i=1,2,\cdots,n$。

其他参数设置与前述相同,模拟结果如表 3 所示。

表 3　多变点模拟结果

Table 3　Simulation results of multiple change points

(n,τ)	COEF	同方差				异方差			
		BIAS	SD	ESE	CP	BIAS	SD	ESE	CP
(200,0.25)	α	0.068	0.373	0.342	0.922	0.088	0.523	0.549	0.820
	β	0.018	0.119	0.118	0.94	−0.024	0.252	0.198	0.870
	γ_1	−0.036	0.558	0.487	0.884	0.054	0.766	0.678	0.810
	γ_2	−0.051	0.564	0.547	0.918	−0.063	0.781	0.639	0.846
	λ	0.002	0.037	0.038	0.942	0.012	0.063	0.071	0.890
	t_1	−0.052	0.288	0.225	0.874	0.08	0.342	0.337	0.798
	t_2	0.058	0.247	0.183	0.846	0.067	0.294	0.325	0.822
(200,0.50)	α	−0.016	0.326	0.332	0.934	0.022	0.476	0.461	0.902
	β	0.005	0.106	0.11	0.946	0.009	0.147	0.153	0.930
	γ_1	0.031	0.533	0.473	0.922	0.057	0.679	0.551	0.904
	γ_2	−0.005	0.554	0.522	0.934	−0.008	0.669	0.628	0.924
	λ	−0.001	0.035	0.037	0.952	0.004	0.052	0.053	0.940
	t_1	0.026	0.256	0.217	0.908	−0.021	0.398	0.261	0.878
	t_2	0.02	0.212	0.175	0.922	0.039	0.218	0.302	0.916
(200,0.75)	α	0.025	0.404	0.363	0.93	−0.061	0.548	0.659	0.870
	β	−0.012	0.129	0.122	0.932	0.041	0.201	0.193	0.906
	γ_1	0.037	0.534	0.514	0.916	0.04	0.698	0.665	0.890
	γ_2	0.066	0.563	0.57	0.924	−0.073	0.651	0.693	0.880
	λ	0.002	0.04	0.041	0.95	0.009	0.066	0.073	0.920
	t_1	0.064	0.282	0.24	0.9	0.073	0.362	0.419	0.868
	t_2	−0.036	0.209	0.186	0.912	0.051	0.343	0.314	0.810
(500,0.25)	α	0.027	0.221	0.207	0.928	−0.096	0.318	0.395	0.862
	β	−0.004	0.073	0.071	0.936	0.019	0.152	0.159	0.910
	γ_1	0.012	0.321	0.306	0.93	0.04	0.478	0.465	0.830
	γ_2	−0.018	0.333	0.335	0.934	0.031	0.395	0.354	0.908
	λ	0.001	0.023	0.024	0.952	0.007	0.079	0.066	0.910
	t_1	0.012	0.16	0.135	0.906	−0.053	0.242	0.237	0.816
	t_2	0.012	0.125	0.111	0.908	0.027	0.274	0.283	0.852

表 3(续)

(n,τ)	COEF	同方差				异方差			
		BIAS	SD	ESE	CP	BIAS	SD	ESE	CP
(500,0.50)	α	0.014	0.187	0.195	0.936	0.037	0.222	0.236	0.904
	β	0.004	0.063	0.067	0.948	0.01	0.133	0.126	0.942
	γ_1	0.005	0.285	0.284	0.934	0.036	0.346	0.339	0.916
	γ_2	−0.003	0.305	0.314	0.938	0.015	0.428	0.398	0.914
	λ	0.001	0.022	0.023	0.956	−0.003	0.056	0.061	0.956
	t_1	−0.007	0.142	0.125	0.922	−0.014	0.181	0.209	0.904
	t_2	0.009	0.108	0.102	0.928	0.022	0.133	0.116	0.896
(500,0.75)	α	0.021	0.232	0.214	0.934	−0.033	0.349	0.353	0.886
	β	0.006	0.075	0.073	0.938	0.011	0.196	0.201	0.922
	γ_1	0.031	0.329	0.317	0.926	0.043	0.338	0.395	0.860
	γ_2	−0.005	0.359	0.347	0.932	−0.009	0.381	0.333	0.908
	λ	−0.003	0.024	0.025	0.95	0.011	0.059	0.066	0.920
	t_1	−0.018	0.149	0.14	0.928	−0.030	0.242	0.137	0.860
	t_2	0.01	0.123	0.114	0.916	0.027	0.274	0.204	0.862

由表 3 可知,在 $n=200$、$\tau=0.50$、误差为同方差时,变点位置、模型参数估计的 BIAS 和 SD 都很小,SD 与 ESE 均很接近,CP 接近 95%,表明估计方法有效;异方差情形时 BIAS 较小,SD 与 ESE 接近,CP 较好,但比同方差的结果稍差,表明估计稳健。另外,当 $\tau=0.25,0.75$ 时,估计的 BIAS、SD 和 ESE 比 $\tau=0.50$ 时的大,同时 CP 结果略小,表明 $\tau=0.50$ 比其他分位数时的估计结果更好。当样本量增大到 500 时,估计的 BIAS、SD、ESE 呈现减小趋势,CP 呈现增大趋势,SD 与 ESE 接近,表明估计有效。

3 居民健康状况实证分析

基于本文方法对我国居民健康状况进行实证分析,数据来源于 2018 年北京大学中国健康与养老追踪调查数据(China Health and Retirement Longitudinal Study, CHARLS)。考虑到数据中包含受访户拒绝回答、不知道如何回答及空缺值等情况,对数据进行预处理后得到 14 720 个数据样本。为更准确地度量居民健康状况的影响因素及影响程度,研究中剔除了健康状况最差的个体数据。另外,为了测度不同健康水平下居民健康状况的影响因素动态特征,采用本文模型进行实证分析。

已有研究表明,健康状况随着年龄增加而变差,男性的健康状况略好于女性的[19]。本文将自报健康状况、5 种残疾问题(包含躯体残疾、智力缺陷等)、14 种慢性病(包含高血压、糖尿病等)等多个变量按简单随机平均加权作为客观健康状况,设置响应变量 T 为拟合后的客观健康状况,T 的值越小代表客观健康状况越好。因年龄、性别、居住环境、受教育程度以及婚姻状况对健康状况均有较大影响,故将其设置为解释变量,即 x(2018 年的年龄)、z_1(性别:1 男,2 女)、z_2(居住位置:1 城镇中心,2 城乡结合,3 农村)、z_3(受教育程度:1 未受过教育,2 未读完小学,3 私塾毕业,4 小学毕业,5 初中毕业,6 高中毕业,7 中专毕业,8 大专毕业,9 本科毕业,10 硕士毕业,11 博士毕业)、z_4(婚姻状况:1 已婚与配偶居住,2 已婚但因工作等其他原因暂时未和配偶一起居住,3 分居,4 离异,5 丧偶,6 从未结婚)。居民健康状况描述性统计分析结果见表 4。

表 4 居民健康状况数据描述性统计

Table 4 Descriptive statistical of resident health status data

健康状况数据	最小值	0.25 分位数	中位数	均值	0.75 分位数	最大值
客观健康状况 (T)	0.00	1.00	2.00	2.13	3.00	17.00
2018 年年龄 (x)	21.00	52.00	58.00	59.74	66.00	95.00
性别 (z_1)	1.00	1.00	2.00	1.52	2.00	2.00
居住位置 (z_2)	1.00	2.00	3.00	2.51	3.00	3.00
教育程度 (z_3)	1.00	2.00	4.00	3.92	5.00	11.00
婚姻状况 (z_4)	1.00	1.00	1.00	1.55	1.00	6.00

由表 4 可知,客观健康状况的最大值为 17,分位数为 0.75 时为 3,故在本实证中个人健康状况较好的比例非常大,设置 $C=10$。客观健康状况均值与中位数接近,年龄平均值为中年水平,女性与男性数据比例相当,居住位置农村占比较大,教育程度平均值为小学毕业,婚姻状况已婚与配偶居住比例大。

为了研究不同客观健康状况下的影响因素与影响程度,假定存在多个变点,不妨设 $M=5$,模型如下:

$$T_i=\alpha+\beta x+\sum_{i=1}^{M}\gamma_i(x-t_i)_+ +\boldsymbol{\zeta}^{\mathrm{T}}\boldsymbol{z}_i+e_i。$$

表 5 给出了 τ 分别为 0.25、0.5、0.75 下的估计结果,第 2—4 列表示 $\tau=0.25$ 时的模型参数估计、标准差及 p 值。从表 5 可见,除分位数为 0.75 时的居住位置估计符号与其他分位数时不同之外,其他估计结果符号相同,估计结果接近。表 5 可得以下结论。

表 5　居民健康状况数据估计结果

Table 5　Estimation results of resident health status data

系数	$\tau=0.25$			$\tau=0.50$			$\tau=0.75$		
	估计值	标准差	p 值	估计值	标准差	p 值	估计值	标准差	p 值
α	14.066	0.905	0.000	12.882	0.821	0.000	12.885	0.872	0.000
β	0.018	0.059	0.035	0.049	0.034	0.015	0.100	0.039	0.003
γ_1	0.068	0.061	0.041	0.028	0.041	0.049	0.007	0.042	0.085
γ_2	0.022	0.070	0.039	0.127	0.107	0.023	0.205	0.183	0.026
ζ_1	0.101	0.158	0.021	0.152	0.169	0.036	0.300	0.212	0.017
ζ_2	0.125	0.108	0.029	0.021	0.105	0.017	−0.127	0.131	0.034
ζ_3	−0.251	0.053	0.000	−0.285	0.049	0.000	−0.377	0.056	0.000
ζ_4	0.090	0.071	0.053	0.057	0.076	0.045	0.105	0.089	0.028
t_1	51.201	0.317	0.028	55.239	0.290	0.037	56.752	0.416	0.041
t_2	74.953	0.285	0.045	76.995	0.256	0.041	78.101	0.383	0.059

1) 在保持其他变量不变的情况下,以分位数为 0.50 为例, 性别与客观健康状况呈现正相关关系,即男性健康状况好于女性。城镇居民、高教育程度、已婚共同居住的健康状况更好。

2) 对于不同健康水平的居民来说,按照分位数为 0.25、0.50、0.75 划分为健康水平高、中及低三类。对于健康水平高的群体来说,男性比女性的健康水平要减少 0.101 个单位,居住位置从农村到城镇每增加一个单位健康水平只增加 0.125 个单位,教育程度每增加一级健康水平减少 0.251 个单位,婚姻情况增加一个单位健康水平增加 0.09 个单位。

3) 随着年龄增加,健康状况存在 2 个变点。对客观健康状况高的群体,无变点时 $\beta=0.018$,即年龄与客观健康状况存在弱的正相关关系;在第 1 个变点时,年龄系数为 $\beta+\gamma_1=0.086$,即年龄与客观健康状况的正相关关系增强;在第 2 个变点时,年龄的系数变为 $\beta+\gamma_1+\gamma_2=0.108$,即年龄与客观健康状况的正相关关系进一步增强。上述结果表明随着年龄增大,客观健康状况数值增大速度变快,健康状况越差,符合实际情况。当 $\tau=0.50$、0.75时,2 个变点前后年龄的估计逐渐增大,即随年龄增长健康状况变差。

4　结语与展望

本文在删失分位数回归模型中基于线性化技术提出了一种解决多变点个数、位置及模型参数的估计方法, 数值模拟结果验证了同(异)方差、不同分位数、不同样本量下该方法具备有效性和稳健性, 对居民健康状况的实证分析表明客观健康状况与年龄存在非线性关系,在不同分位数下均存在 2 个变点。

本文方法可从以下方面进行拓展:一是多变点个数估计时,虽然线性化技术迭代速度快,结合贝叶斯信息准则可有效得到变点个数估计,但目标函数在变点处不可导使得参数估计的大样本性质难以得到,可考虑使用其他方法得到变点位置估计,推导多变点的大样本性质;二是放松对单个协变量存在的限制,研究多个协变量存在多变点情况,即

$$T_i=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}_i+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_i+\boldsymbol{\gamma}^{\mathrm{T}}(\boldsymbol{Z}_i-\boldsymbol{t})_+ +e_i,$$

其中,$\boldsymbol{Z}_i$ 是有变点的协变量,$\boldsymbol{\beta}$ 是相应的参数向量,$\boldsymbol{t}$ 为多个变点,利用本文方法不难将模型拓展至多协变量

存在多变点的模型,即

$$(\boldsymbol{Z}_i-\boldsymbol{t})_+=(\boldsymbol{Z}_i-\boldsymbol{t}^{(0)})_+-\boldsymbol{I}(\boldsymbol{Z}_i>\boldsymbol{t}^{(0)})(\boldsymbol{t}-\boldsymbol{t}^{(0)}),$$

此时变点位置和变点个数估计方法与算法 1—3 类似,不再赘述。

参考文献:

[1] KOENKER R, BASSETT G. Regression quantiles[J]. Econometrica, 1978, 46(1):211-244.

[2] POWELL J L. Least absolute deviations estimation for the censored regression model[J]. Journal of Econometrics, 1984, 23(3):303-325.

[3] POWELL J L. Censored regression quantiles[J]. Journal of Econometrics, 1986, 32(1):143-155.

[4] PORTNOY S. Censored quantile regression[J]. Journal of American Statistical Association, 2003, 98(464):1001-1012.

[5] FRUMENTO P, BOTTAI M. An estimating equation for censored and truncated quantile regression[J]. Computational Statistics and Data Analysis, 2016, 113:53-63.

[6] CHEN Songnian. Sequential estimation of censored quantile regression models[J]. Journal of Econometrics, 2018, 207(1):30-52.

[7] JIA Y C, JEONG J H. Deep learning for quantile regression under right censoring: DeepQuantreg[J]. Computational Statistics and Data Analysis, 2022, 165:107323.

[8] NARISETTY N,KOENKER R. Censored quantile regression survival models with a cure proportion[J]. Journal of Econometrics, 2022, 261(1):192-203.

[9] ZHANG Feipeng, LI Qunhua. A continuous threshold expectile model[J]. Computational Statistics and Data Analysis, 2017, 116:49-66.

[10] WANG Xiaogang, ZHOU Xiaoying, LI Bing, et al. A bent line Tobit regression model with application to household financial assets[J]. Journal of Statistical Planning and Inference, 2022, 221:69-80.

[11] ZHOU Xiaoying, ZHANG Feipeng. Bent line quantile regression via a smoothing technique[J]. Statistical Analysis and Data Mining, 2020, 13(3):216-228.

[12] 王小刚,李冰. 基于核函数方法的逐段线性 Tobit 回归模型估计[J]. 山东大学学报(理学版),2020,55(6):1-9.
WANG Xiaogang, LI Bing. Piecewise linear Tobit regression model estimation based on kernel function method[J]. Journal of Shandong University(Natural Science), 2020, 55(6):1-9.

[13] MUGGEO V M R. Estimating regression models with unknown break-points[J]. Statistics in Medicine, 2003, 22(19):3055-3071.

[14] YAN Yanyang, ZHANG Feipeng, ZHOU Xiaoying. A note on estimating the bent line quantile regression model[J]. Computational Statistics, 2017, 32(2):661-630.

[15] MUGGEO V M R, ADELFIO G. Efficient change point detection for genomic sequences of continuous measurements[J]. Bioinformatics, 2011, 27(2):161-166.

[16] 龙振环,张飞鹏,周小英. 带多个变点的逐段连续线性分位数回归模型及应用[J]. 数量经济技术经济研究,2017,34(8):150-161.
LONG Zhenhuan, ZHANG Feipeng, ZHOU Xiaoying. A continuous piecewise linear quantile regression with multiple change points[J]. The Journal of Quantitative and Technical Economics, 2017, 34(8):150-161.

[17] ZHONG Wei, WAN Chuang, ZHANG Wenyang. Estimation and inference for multi-kink quantile regression[J]. Journal of Business and Economic Statistics, 2021, 40(3):1123-1139.

[18] 王小刚,李冰. 含多个结构突变的分段线性 Tobit 回归模型及应用[J]. 统计与决策,2021,37(19):21-25.
WANG Xiaogang, LI Bing. Piecewise linear Tobit regression model with multiple change points and application[J]. Statistics and Decision, 2021, 37(19):21-25.

[19] 曾毅,沈可. 中国老年人口多维度健康状况分析[J]. 中华预防医学,2010,34(2):108-114.
ZENG Yi, SHEN Ke. Main dimensions of health status among the Chinese elderly[J]. Chinese Journal of Preventive Medicine, 2010, 34(2):108-114.

(编辑:李艺)