文章编号:1671-9352(2025)03-0069-08 DOI:10.6040/j.issn.1671-9352.0.2024.084

基于线性化技术的变点分位数回归模型的估计与应用

周小英[1,2],吉晨[1],涂晓艺[1*]
(1.海南师范大学数学与统计学院, 海南 海口 571158; 2.海南师范大学数据科学与智慧教育教育部重点实验室, 海南 海口 571158)

摘要:构建变点分位数回归模型,该模型由 1 条直线和 1 条二次曲线在变点处相交而成,可以灵活处理变点数据,还能捕捉响应变量分布的全貌。由于变点参数的存在,使得模型的损失函数是非凸的,给估计参数带来了挑战。为了解决这个问题,基于线性化技术将损失函数线性化,利用迭代算法,同时得到变点参数和其他参数的估计,给出估计量的区间估计。数值模拟结果表明,本文的估计方法具有良好的相合性和有效性,人均国内生产总值与电力质量数据的实证分析也验证了所提模型和方法的可行性和实用性。

关键词:变点;线性-二次分位数回归模型;线性化技术;电力质量
中图分类号:O213.9 **文献标志码**:A
引用格式:周小英,吉晨,涂晓艺. 基于线性化技术的变点分位数回归模型的估计与应用[J]. 山东大学学报(理学版),2025,60(3):69-76.

Estimation and application of the change-point quantile regression model based on linearization technique

ZHOU Xiaoying[1,2], JI Chen[1], TU Xiaoyi[1*]
(1. School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, Hainan, China; 2. Key Laboratory of Data Science and Smart Education, Ministry of Education, Hainan Normal University, Haikou 571158, Hainan, China)

Abstract: The change-point quantile regression model constructed by the intersection of a straight line and a quadratic curve at a change point. This model can flexibly handle change point data and capture the overall distribution of the response variable. Due to the presence of the change point parameter, the model's loss function is non-convex, which is a challenge for parameter estimation. To address this issue, the loss function is linearized based on the linearization technique combining with an iterative algorithm, which can simultaneously estimate the change point and other parameters. The interval estimation theory for the estimators is also derived. Numerical simulation results indicate that the proposed estimation method exhibits good consistency and effectiveness. Empirical analysis of per capita GDP and power quality data further verifies the feasibility and practicality of the proposed model and method.

Key words: change point; linear quadratic quantile regression model; linearization technique; power quality

0 引言

数据结构随着大数据时代的到来日益复杂,变点现象也在各个领域中广泛存在。例如,在能源经济领域中,在某个变点前后,人均国内生产总值随着电力质量的提高先缓慢增长而后迅速增长;在社会领域中,在某个变点前后,月薪随着每周的工作时间的增加先缓慢提高而后迅速上涨,因此,选择正确的模型分析变点数据是十分重要的。学者常用均值回归的变点模型来研究这种非线性现象。已有大量学者对分段线性回归模型的变点检测和估计问题进行了研究,详见文献[1-11]等。

收稿日期:2024-03-14; **网络出版时间**:2024-12-17 14:42:52
基金项目:国家自然科学地区基金资助项目(72263007); 2023 年海南省研究生创新科研课题项目(Qhys2023-384)
第一作者:周小英(1989—),女,副教授,博士,研究方向为变点数据建模与应用. E-mail:zhouxy213@163.com
*通信作者:涂晓艺(2000—),女,硕士研究生,研究方向为应用统计. E-mail:txy18181281029@163.com

基于均值回归的分段线性模型能够灵活处理非线性的变点数据,但是这类模型只考虑了回归函数在变点前后都呈现线性关系的情况,而且对于一些实际数据,回归函数在变点前后分别呈现线性函数和二次函数的关系。为了进一步对这类数据进行建模,Pastor 等[12]提出线性-二次逻辑斯谛回归模型及其估计方法。Zhang 等[13]在广义框架下更加全面地研究了线性-二次模型,以适用所有指数分布族的变点数据。值得注意的是,Zhang 等[13]研究的广义线性-二次回归模型只能得到 1 条回归曲线,能提供响应变量的信息是有限的,不能捕捉响应变量的所有信息。在现实中,该模型既不能得到具有异方差、高尖峰、厚尾特点的金融数据的更多尾部信息,又因均值回归通常需要随机误差的正态性假设而较大地限制实际数据。

对此,分位数回归模型作为替代模型,既能得到协变量对响应变量整体条件分布的影响,又能更加灵活地处理异方差数据,还能使响应变量的非正态误差和异常值具有较强的稳健性,详见文献[14]。对此,Li 等[15]提出了折线分位数回归模型,该模型研究了给定分位数水平时回归函数在变点前后都呈现线性关系的情况;但是在文献[13]中,对于一些实际数据,回归函数在变点前后分别呈现线性函数和二次函数的关系,因此,本文提出了线性-二次分位数回归模型,该模型不仅可以灵活处理具有变点效应的数据,还可以全面捕捉响应变量的条件分布,对异常值数据也有较强的稳健性。由于模型中未知变点参数的存在导致了模型的目标函数非凸,因此不能直接使用传统的优化方法得到模型的参数估计。

基于以上情形,本文的创新之处在于:

1) 提出了 1 种线性化方法,既能解决目标函数非凸的问题,又能同时估计变点和其他参数;

2) 给出了参数的区间估计方法。

1 线性-二次分位数回归模型

1.1 广义线性-二次回归模型

Zhang 等[13]提出了在广义框架下的线性-二次回归模型。令 X 为解释变量,Y 为响应变量,$\boldsymbol{Z}$ 为 p 维协变量。在给定解释变量 X 和协变量 $\boldsymbol{Z}$ 的条件下,Y 的分布属于如下的指数分布族:

$$f_Y(Y|X,\boldsymbol{Z})=\exp\{[\eta Y-b(\eta)]/a(\phi)+c(Y,\phi)\},$$

其中:$a(\phi)$ 为广义线性模型中用于构建概率密度函数的 1 个组成部分;$b(\eta)$ 为积累函数;$c(Y,\phi)$ 为正则化函数。指数分布族假定 $\mu=E(Y|X,\boldsymbol{Z})=b(\eta)$,广义线性模型是通过该假设将 Y 的条件均值与 $\boldsymbol{Z}$ 联系起来,即

$$g(\mu)=\eta=\beta_1 X+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z},$$

其中,$g(\mu)$ 为自然连接函数。数据的分布类型不同使得自然连接函数的表达式也不同。例如,均值回归的自然连接函数的表达式为 $g(\mu)=\mu$;逻辑斯谛回归模型的自然连接函数的表达式为 $g(\mu)=\log(\mu/(1-\mu))$。

带有变点的广义线性-二次回归模型如下:

$$\eta=\beta_0+\beta_1 X+\beta_2(X-t)_+^2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}, \tag{1}$$

其中:$\boldsymbol{\theta}=(\beta_0,\beta_1,\beta_2,\boldsymbol{\gamma}^{\mathrm{T}},t)^{\mathrm{T}}$ 为模型中的未知参数;t 为变点位置;$(X-t)_+^2=(X-t)^2I(X>t)$,$I(\cdot)$ 为示性函数。当解释变量 $X\leqslant t$ 时,$(X-t)_+^2=0$,此时模型为 $\eta=\beta_0+\beta_1 X+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}$;当解释变量 $X>t$ 时,$(X-t)_+^2=(X-t)^2$,此时模型为 $\eta=\beta_0+\beta_1 X+\beta_2(X-t)^2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}$。

1.2 线性-二次分位数回归模型

虽然模型(1)可以灵活处理具有线性-二次形式的数据,但是该模型只能得到 1 条回归曲线,提供的响应变量的信息有限,不能捕捉响应变量的所有信息。对此,本文提出线性-二次分位数回归模型。设 Y 为响应变量,X 为解释变量,$\boldsymbol{Z}$ 为 p 维协变量,构建的线性-二次分位数模型如下:

$$V_\tau(Y|X,\boldsymbol{Z})=\beta_0+\beta_1 X+\beta_2(X-t)_+^2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}, \tag{2}$$

其中:t 为未知变点位置;$V_\tau(Y|X,\boldsymbol{Z})$ 为响应变量 Y 关于解释变量 X 和协变量 $\boldsymbol{Z}$ 的条件 τ 分位数;模型待估参数为 $\boldsymbol{\theta}=(\beta_0,\beta_1,\beta_2,\boldsymbol{\gamma}^{\mathrm{T}},t)^{\mathrm{T}}$;$(X-t)_+^2=(X-t)^2I(X>t)$,$I(\cdot)$ 为示性函数。当解释变量 $X\leqslant t$ 时,$(X-t)_+^2=0$,此时模型为 $V_\tau(Y|X,\boldsymbol{Z})=\beta_0+\beta_1 X+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}$;当解释变量 $X>t$ 时,$(X-t)_+^2=(X-t)^2$,此时模型为 $V_\tau(Y|X,\boldsymbol{Z})=\beta_0+\beta_1 X+\beta_2(X-t)_+^2+\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}$,因此,模型(2)在变点 t 之前为线性变化,在变点 t 之后为二次关系,并且模型(2)可以在任意分位数下对数据进行全面分析,在一定程度上弥补了模型(1)的不足。

本文感兴趣的是对模型(2)进行参数估计。假设有独立数据集$\{(Y_i,X_i,\boldsymbol{Z}_i),\ i=1,2,\cdots,n\}$,参数$\boldsymbol{\theta}$的估计可以通过式(3)得到

$$\boldsymbol{\theta}=\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\,\rho_\tau\{Y_i-\beta_0-\beta_1X_i-\beta_2(X_i-t)_+^2-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i\},\tag{3}$$

其中$\rho_\tau(u)=u|\tau-I(u<0)|$为$\tau$分位数的损失函数。值得注意的是,由于变点的存在,损失函数(3)关于变点参数t非凸,因此不能直接应用传统的优化方法求解。对此,本文提出一种新颖的线性化方法求解模型(2)的参数估计。

目标函数(3)中的$(X_i-t)_+^2$关于变点t连续,线性化方法就是在$t^{(0)}$处将$\beta_2(X_i-t)_+^2$进行一阶泰勒展开,得到

$$\beta_2(X_i-t)_+^2\approx\beta_2(X_i-t^{(0)})_+^2-2\beta_2(X_i-t^{(0)})_+(t-t^{(0)})。$$

为了方便表达,令$u_i^{(0)}=(X_i-t^{(0)})_+^2$,$v_i^{(0)}=-2(X_i-t^{(0)})_+$,$\beta_3=\beta_2(t-t^{(0)})$。对于给定的$t^{(0)}$,$u_i$和$v_i$都可以看作为新的协变量,$\beta_3$为新的回归系数。这样,参数$\boldsymbol{\zeta}=(\beta_0,\beta_1,\beta_2,\beta_3,\boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$的估计可以通过式(4)得到

$$\hat{\boldsymbol{\zeta}}=\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\sum_{i=1}^{n}\rho_\tau\{y_i-\beta_0-\beta_1X_i-\beta_2u_i^{(0)}-\beta_3v_i^{(0)}-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i\},\tag{4}$$

这是标准的线性分位数回归模型损失函数的极小化问题,可用现有的理论和方法快速得到参数$\boldsymbol{\zeta}$的估计$\hat{\boldsymbol{\zeta}}=(\hat\beta_0,\hat\beta_1,\hat\beta_2,\hat\beta_3,\hat{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$,再由$\beta_3=\beta_2(t-t^{(0)})$得到变点的估计$\hat t=t^{(0)}+\hat\beta_3/\hat\beta_2$。重复上述过程,直到所有参数收敛。现将迭代算法总结如下:

第 1 步:设定变点初值$t^{(0)}$,通过线性化方法,得到参数$\boldsymbol{\zeta}$的估计$\hat{\boldsymbol{\zeta}}^{(0)}=(\hat\beta_0^{(0)},\hat\beta_1^{(0)},\hat\beta_2^{(0)},\hat\beta_3^{(0)},\hat{\boldsymbol{\gamma}}^{(0)\mathrm{T}})^{\mathrm{T}}$和变点估计$\hat t^{(1)}=t^{(0)}+\hat\beta_3^{(1)}/\hat\beta_2^{(1)}$。

第 2 步:对于第k次迭代$\hat t^{(k)}$,可通过式(5)得到$\hat{\boldsymbol{\zeta}}^{(k)}$

$$\hat{\boldsymbol{\zeta}}^{(k)}=\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\sum_{i=1}^{n}\rho_\tau\{Y_i-\beta_0-\beta_1X_i-\beta_2u_i^{(k-1)}-\beta_3v_i^{(k-1)}-\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Z}_i\},\tag{5}$$

其中,$u_i^{(k-1)}=(X_i-t^{(k-1)})_+^2$;$v_i^{(k-1)}=-2(X_i-t^{(k-1)})_+$。

第 3 步:更新变点的估计,得到$\hat t^{(k+1)}=\hat t^{(k)}+\hat\beta_3^{(k+1)}/\hat\beta_2^{(k+1)}$。

第 4 步:重复第 2、3 步,直到所有的参数收敛。

通过上述步骤,最终得到模型(2)中参数$\boldsymbol{\theta}$的估计并记为$\hat{\boldsymbol{\theta}}$,下面给出其区间估计。注意到,参数$\boldsymbol{\zeta}$的渐近性质可以直接从现有的标准线性分位数回归模型理论中获得,变点参数t的标准误差可以根据 delta 方法得到

$$\mathrm{SE}(\hat t)=\mathrm{SE}(\hat\beta_3/\hat\beta_2)=[\mathrm{Var}(\hat\beta_3)+(\hat\beta_3/\hat\beta_2)^2\mathrm{Var}(\hat\beta_2)-2(\hat\beta_3/\hat\beta_2)\mathrm{Cov}(\hat\beta_2,\hat\beta_3)]^{1/2}/|\hat\beta_2|。$$

在实际中,$\hat\beta_3$趋近于 0,上式可近似为$\mathrm{SE}(\hat t)=\mathrm{SE}(\hat\beta_3)/|\hat\beta_2|$,$\hat t$的$100(1-\alpha)\%$置信区间估计为

$$[\hat t-z_{\alpha/2}\mathrm{SE}(\hat t),\hat t+z_{\alpha/2}\mathrm{SE}(\hat t)],$$

其中$z_{\alpha/2}$为标准正态分布的$(1-\alpha/2)$分位数。

2　数值模拟

为了评估线性-二次分位数模型的适用性和参数估计方法的有效性,本文进行了大量的数值模拟试验,所有数值模拟试验均基于 R 语言完成。

现考虑如下 2 种模型形式:

1) 同方差情况:$Y=\beta_0+\beta_1X+\beta_2(X-t)_+^2+\upsilon W+e$;

2) 异方差情况:$Y=\beta_0+\beta_1X+\beta_2(X-t)_+^2+\upsilon W+(1+0.3W)e$。

其中,X服从均匀分布$U(-4,2)$;W服从二项分布$B(1,0.5)$;误差项e满足其τ分位数为 0,即$e=\tilde e-Q_\tau(\tilde e)$,$Q_\tau(\tilde e)$为$\tilde e$的$\tau$分位数。对于上面 2 种模型,本文均考虑了以下 3 种误差项:(ⅰ)$\tilde e\sim N(0,1)$;(ⅱ)$\tilde e\sim 0.9N(0,1)+0.1t_5$;(ⅲ)$\tilde e\sim t_5$,其中$N(0,1)$为标准正态分布;$t_5$为自由度为 5 的$t$分布。模型的回归参数设置为$(\beta_0,\beta_1,\beta_2,\upsilon)^{\mathrm{T}}=(1,-1,2,1)^{\mathrm{T}}$,变点参数设置为$t=-1$。对于每种模拟情况,样本设置为 400,模拟次数为 1 000 次。

为了节省空间,本文只展示分位数水平 $\tau=0.1,0.3,0.5,0.7,0.9$ 的情况,并计算了真实参数与估计值的偏差(the empirical bias, Bias)、1 000 次估计的标准误差(the empirical standard error, SD)、1 000 次估计的标准差的平均值(the average estimated standard error, ESE)、1 000 次均方误差(the mean square error, MSE)和置信度为 95%下的置信区间覆盖率(95% coverage probability, CP),详见表 1—3。

从表 1—3 中可以看出,3 种分布的各个参数的 Bias 基本接近 0,说明本文估计量具有相合性;各参数的 SD 与 ESE 值都很接近,说明本文估计方法的渐近正态性是有效的;在置信度为 95%时,CP 在 95%值的附近上下波动,说明了本文估计量的有效性。

从误差分布类型上看,误差项来自标准正态分布的数值模拟结果最优,其次为混合分布,再次为 t_5 分布,说明这 3 种分布数据对应的估计精度依次递减。在实际情况中,如果待估参数存在很大的噪声,那么该参数的估计就可能存在比较大的干扰,导致模型不能很好地估计出参数的真实信息。

从是否为同方差的角度看,以分位数水平 $\tau=0.9$ 为例,在同方差情况下,β_1 的 Bias 呈现为标准正态分布最小,混合分布次之,t_5 分布最大;SD 与 ESE 的差值接近于 0;β_1 的 MSE 呈现为标准正态分布最小,混合分布次之,t_5 分布最大;CP 在 0.90 到 0.95 间波动。在异方差情况下,β_1 的 Bias 呈现为混合分布最小,t_5 分布次之,标准正态分布最大;SD 与 ESE 的差值接近于 0;β_1 的 MSE 呈现为标准正态分布最小,混合分布次之,t_5 分布最大;CP 在 0.95 附近上下波动。相同分位数水平下的同方差的估计结果优于异方差,虽然不同分位数的估计结果存在一定差异,但是整体规律相似,分析步骤相同。

综上所述,本文通过大量的数值模拟证实,线性-二次分位数回归模型无论在不同的误差分布类型角度,还是在是否为同方差的角度,都可以准确估计不同分位数水平下的模型参数,并且估计结果与理论相符。

表 1 $\tilde{e}\sim N(0,1)$ 的模拟结果

Table 1 Simulation results of $\tilde{e}\sim N(0,1)$

分位数水平 τ	各项指标	同方差					异方差				
		β_0	β_1	β_2	v	t	β_0	β_1	β_2	v	t
0.1	Bias	0.006	0.006	−0.011	0.008	0.000	0.010	0.004	−0.012	−0.009	0.001
	SD	0.321	0.120	0.144	0.175	0.117	0.355	0.139	0.163	0.201	0.134
	ESE	0.319	0.122	0.141	0.169	0.116	0.366	0.142	0.164	0.197	0.135
	MSE	0.103	0.014	0.021	0.031	0.014	0.126	0.019	0.027	0.040	0.018
	CP	0.917	0.925	0.905	0.916	0.912	0.935	0.933	0.933	0.928	0.945
0.3	Bias	−0.001	−0.002	−0.006	0.000	0.000	0.010	0.004	−0.004	−0.012	0.002
	SD	0.254	0.099	0.112	0.133	0.094	0.279	0.108	0.128	0.153	0.105
	ESE	0.247	0.094	0.109	0.131	0.090	0.280	0.109	0.123	0.153	0.103
	MSE	0.065	0.010	0.013	0.018	0.009	0.078	0.012	0.016	0.024	0.011
	CP	0.929	0.927	0.930	0.938	0.928	0.943	0.945	0.934	0.948	0.944
0.5	Bias	−0.009	−0.003	−0.011	−0.003	−0.006	0.009	0.003	−0.002	0.001	0.004
	SD	0.239	0.089	0.107	0.128	0.086	0.265	0.102	0.121	0.146	0.098
	ESE	0.236	0.090	0.105	0.126	0.086	0.265	0.102	0.119	0.145	0.098
	MSE	0.057	0.008	0.012	0.016	0.007	0.070	0.010	0.015	0.021	0.010
	CP	0.938	0.926	0.936	0.941	0.933	0.949	0.956	0.942	0.956	0.950
0.7	Bias	0.000	−0.001	−0.007	−0.005	−0.001	0.001	0.001	−0.007	0.007	0.000
	SD	0.256	0.096	0.108	0.134	0.090	0.273	0.105	0.130	0.153	0.105
	ESE	0.249	0.095	0.110	0.132	0.091	0.278	0.107	0.125	0.153	0.103
	MSE	0.065	0.009	0.012	0.018	0.008	0.074	0.011	0.017	0.024	0.011
	CP	0.931	0.926	0.942	0.942	0.940	0.942	0.946	0.939	0.942	0.949
0.9	Bias	0.010	0.003	−0.002	0.011	0.005	0.033	0.011	0.001	0.001	0.010
	SD	0.317	0.121	0.144	0.170	0.117	0.353	0.137	0.159	0.192	0.130
	ESE	0.312	0.119	0.137	0.165	0.114	0.364	0.142	0.161	0.196	0.134
	MSE	0.100	0.015	0.021	0.029	0.014	0.126	0.019	0.025	0.037	0.017
	CP	0.923	0.915	0.901	0.916	0.906	0.940	0.945	0.939	0.939	0.941

表 2　$\tilde{e}\sim 0.9N(0,1)+0.1t_5$ 的模拟结果

Table 2　Simulation results of $\tilde{e}\sim 0.9N(0,1)+0.1t_5$

分位数水平 τ	各项指标	同方差					异方差				
		β_0	β_1	β_2	υ	t	β_0	β_1	β_2	υ	t
0.1	Bias	0.003	0.004	−0.006	−0.001	0.003	0.029	0.009	−0.003	−0.004	0.007
	SD	0.335	0.127	0.143	0.178	0.120	0.381	0.144	0.167	0.207	0.138
	ESE	0.324	0.123	0.144	0.172	0.118	0.376	0.146	0.167	0.202	0.139
	MSE	0.112	0.016	0.021	0.032	0.014	0.146	0.021	0.028	0.043	0.019
	CP	0.914	0.910	0.916	0.907	0.910	0.929	0.940	0.934	0.937	0.950
0.3	Bias	0.011	0.006	−0.008	0.007	0.001	−0.014	−0.003	−0.012	0.007	−0.005
	SD	0.248	0.094	0.114	0.136	0.091	0.267	0.104	0.125	0.154	0.101
	ESE	0.248	0.094	0.109	0.132	0.090	0.281	0.109	0.127	0.155	0.104
	MSE	0.062	0.009	0.013	0.018	0.008	0.071	0.011	0.016	0.024	0.010
	CP	0.933	0.943	0.925	0.934	0.939	0.951	0.945	0.948	0.949	0.956
0.5	Bias	0.014	0.005	−0.005	−0.004	0.002	0.009	0.003	−0.007	−0.003	0.000
	SD	0.242	0.093	0.101	0.125	0.085	0.265	0.103	0.123	0.148	0.099
	ESE	0.235	0.090	0.104	0.125	0.086	0.266	0.103	0.119	0.146	0.098
	MSE	0.059	0.009	0.010	0.016	0.007	0.070	0.011	0.015	0.022	0.010
	CP	0.941	0.925	0.953	0.945	0.954	0.955	0.953	0.936	0.945	0.948
0.7	Bias	0.024	0.006	−0.004	−0.002	0.004	0.012	0.006	−0.002	0.014	0.004
	SD	0.266	0.100	0.110	0.138	0.093	0.275	0.109	0.124	0.153	0.103
	ESE	0.249	0.095	0.110	0.132	0.090	0.280	0.109	0.126	0.154	0.104
	MSE	0.071	0.010	0.012	0.019	0.009	0.076	0.012	0.015	0.024	0.011
	CP	0.927	0.930	0.938	0.935	0.944	0.941	0.942	0.936	0.960	0.947
0.9	Bias	0.023	0.006	−0.001	−0.001	0.007	0.010	0.004	−0.011	0.007	0.001
	SD	0.334	0.128	0.143	0.174	0.120	0.385	0.150	0.171	0.203	0.140
	ESE	0.324	0.123	0.143	0.171	0.118	0.370	0.144	0.166	0.201	0.137
	MSE	0.112	0.016	0.021	0.030	0.014	0.148	0.023	0.029	0.041	0.020
	CP	0.907	0.910	0.922	0.905	0.916	0.924	0.927	0.919	0.935	0.921

表 3　$\tilde{e}\sim t_5$ 的模拟结果

Table 3　Simulation results of $\tilde{e}\sim t_5$

分位数水平 τ	各项指标	同方差					异方差				
		β_0	β_1	β_2	υ	t	β_0	β_1	β_2	υ	t
0.1	Bias	0.025	0.008	−0.014	0.012	0.003	0.017	0.002	−0.019	−0.007	0.000
	SD	0.435	0.166	0.192	0.230	0.157	0.479	0.182	0.217	0.274	0.174
	ESE	0.426	0.162	0.189	0.225	0.155	0.504	0.196	0.229	0.271	0.188
	MSE	0.190	0.028	0.037	0.053	0.025	0.229	0.033	0.047	0.075	0.030
	CP	0.918	0.913	0.919	0.926	0.919	0.944	0.949	0.943	0.939	0.947
0.3	Bias	0.005	0.001	−0.011	0.004	−0.003	−0.004	−0.002	−0.002	0.004	0.002
	SD	0.290	0.109	0.126	0.144	0.105	0.311	0.122	0.136	0.167	0.113
	ESE	0.274	0.104	0.121	0.145	0.099	0.308	0.119	0.139	0.170	0.115
	MSE	0.084	0.012	0.016	0.021	0.011	0.097	0.015	0.019	0.028	0.013
	CP	0.917	0.930	0.934	0.940	0.928	0.951	0.940	0.945	0.946	0.944
0.5	Bias	0.003	0.002	−0.002	0.004	0.002	0.008	0.002	−0.002	−0.002	0.003
	SD	0.246	0.097	0.109	0.133	0.090	0.275	0.108	0.122	0.156	0.102
	ESE	0.248	0.095	0.110	0.132	0.091	0.281	0.109	0.125	0.155	0.104
	MSE	0.060	0.009	0.012	0.018	0.008	0.075	0.012	0.015	0.024	0.010
	CP	0.940	0.938	0.941	0.930	0.942	0.951	0.946	0.953	0.945	0.946
0.7	Bias	0.009	0.002	−0.002	−0.002	0.003	0.005	0.002	0.001	0.002	0.006
	SD	0.270	0.106	0.122	0.147	0.101	0.309	0.120	0.139	0.171	0.114
	ESE	0.273	0.104	0.120	0.144	0.099	0.311	0.121	0.138	0.170	0.115
	MSE	0.073	0.011	0.015	0.022	0.010	0.096	0.014	0.019	0.029	0.013
	CP	0.944	0.928	0.938	0.931	0.940	0.939	0.949	0.946	0.947	0.955
0.9	Bias	0.024	0.007	−0.009	0.011	0.009	0.021	0.010	−0.020	−0.009	0.003
	SD	0.442	0.170	0.198	0.249	0.164	0.500	0.197	0.224	0.270	0.183
	ESE	0.426	0.162	0.187	0.225	0.154	0.507	0.199	0.232	0.276	0.188
	MSE	0.196	0.029	0.039	0.062	0.027	0.250	0.039	0.051	0.073	0.033
	CP	0.908	0.907	0.902	0.893	0.906	0.942	0.951	0.936	0.938	0.947

3 实证分析

人均国内生产总值和电力质量数据来源于《2019 年全球竞争力报告》①,其中,人均国内生产总值是了解和把握一个国家或者地区经济发展情况的重要指标之一,也是评价该国或地区潜力和前景的有效指标。电力是现代社会经济发展和人们生活不可或缺的基础设施之一,对工业生产、商业经济、医疗保健、创新技术等领域发展有着不可替代的作用。电力质量数据是电力输配损耗,即供电源与配电点之间的传输损耗,以及分配给消费者的损耗(包括窃电损耗)。在《2019 年全球竞争力报告》中,为了允许不同性质和量级的指标聚合,电力质量得分使用了最大-最小转换。转换公式为

$$s=\left(\frac{v-w}{f-w}\right)\times 100,$$

其中:v 为电力质量的原始值;w 为最低可接受值,设定为 4;f 为可能出现的最佳结果,设定为 100。如果某国的 v 值低于 w 值,则其得分为 0;如果某国的 v 值高于 f 值,则其得分为 100。我国在 2019 年电力质量得分为 99。根据国家统计局显示,2019 年我国电力用电总量为 74 866.1 亿 kW·h,同比增长 4.7%。随着人民生活水平和国民经济的快速发展,保障生产生活用电可靠,对促进经济社会的发展极为重要。

目前,很多学者都进行了关于经济和电力的研究。Ferguson 等[16]研究发现人均国内生产总值和电力使用具有强相关性,Wolde-Rufael[17]研究发现电力质量对经济增长有着重要作用。已有学者通过不同的变点模型研究发现,人均国内生产总值和电力质量之间存在变点,在变点之前人均国内生产总值提高比较缓慢,但在变点之后,人均国内生产总值迅速增加,详见文献[18-20]等。本文从《2019 年全球竞争力报告》获取了 141 个国家人均国内生产总值(美元)和电力质量数据,以此来探讨它们之间的复杂关系。

图 1 展示了人均国内生产总值和电力质量数据的分布情况。为了方便画图和表示,本文把人均国内生产总值(美元)除以 10 000 表示,记为万美元。

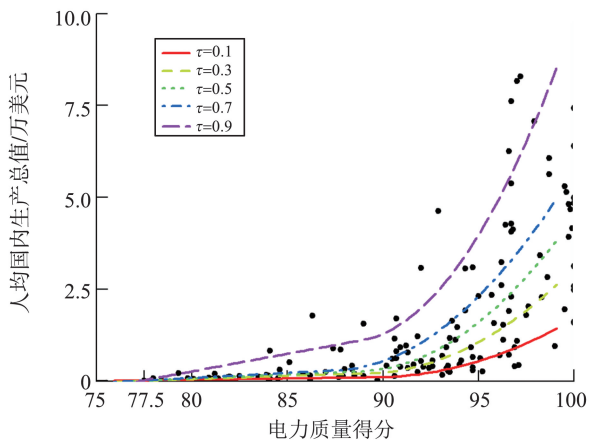

图 1 人均国内生产总值和电力质量的数据拟合结果

Fig.1 Fitting results of gross domestic product per capita and power quality

从图 1 中可以发现,人均国内生产总值和电力质量得分之间并不是简单的线性关系,数据中可能存在变点。电力质量得分在达到某个数值前,人均国内生产总值随着电力质量缓慢增长,而在这个数值之后,人均国内生产总值迅速增加。为了准确分析出它们之间的关系,本文构建的线性-二次分位数回归模型如下:

$$v_\tau(Y_i\mid X_i)=\beta_0+\beta_1X_i+\beta_2(X_i-t)_+^2,$$

其中:解释变量 X_i 为第 i 个国家电力质量得分;响应变量 Y_i 为第 i 个国家的人均国内生产总值;模型回归系数为 β_0、β_1 和 β_2;t 为变点位置;τ 为分位数水平;$v_\tau(Y_i\mid X_i)$ 为响应变量 Y_i 关于解释变量 X_i 的第 τ 分位值。运用 R 软件和第 2 章的参数求解理论计算得到模型回归参数,如表 4 所示。

① https://cn.weforum.org/publications/global-competitiveness-report-2019/

表 4　参数估计结果
Table 4　The estimating results

τ	各项指标	β_0	β_1	β_2	t
0.1	估计值	−5.166	0.070	0.141	89.630
	标准误差	7.799	0.093	0.010	0.729
	95%置信区间	[−20.452, 10.119]	[−0.112, 0.253]	[0.122, 0.160]	[88.201, 91.060]
0.3	估计值	−13.326	0.174	0.236	89.357
	标准误差	12.330	0.147	0.016	0.689
	95%置信区间	[−37.491, 10.840]	[−0.114, 0.463]	[0.206, 0.267]	[88.005, 90.708]
0.5	估计值	−17.504	0.228	0.319	88.879
	标准误差	13.532	0.162	0.017	0.560
	95%置信区间	[−44.027, 9.019]	[−0.088, 0.545]	[0.286, 0.353]	[87.782, 89.975]
0.7	估计值	−20.832	0.273	0.332	87.615
	标准误差	10.517	0.126	0.013	0.418
	95%置信区间	[−41.444, −0.220]	[0.027, 0.519]	[0.306, 0.358]	[86.796, 88.434]
0.9	估计值	−73.759	0.957	0.631	88.984
	标准误差	62.351	0.744	0.079	1.304
	95%置信区间	[−195.964, 48.447]	[−0.502, 2.416]	[0.477, 0.785]	[86.427, 91.540]

表 4 展现了模型参数的估计值、标准误差和 95%置信区间。从整体上看,人均国内生产总值和电力质量得分之间的变点出现在 87~90。当电力质量得分小于变点时,人均国内生产总值随电力质量得分的增加缓慢提高;当电力质量得分超过变点时,人均国内生产总值随电力质量得分的增加迅速提高,呈现二次关系,说明电力质量对人均国内生产总值有重要影响。从表 4 可知,不同分位数水平下的变点不同。虽然变点位置有所波动且波动幅度不一;但还是提供了科学依据,因此不用盲目地寻找变点,可以重点关注这 5 个分位数水平下的变点位置,为制定政策提供了先行准备条件。

以分位数 $\tau=0.5$ 为例,当电力质量得分低于 88.879 时,人均国内生产总值和电力质量的关系为 $Y_i=-17.504+0.228X_i$,此时电力质量得分每提升 1 个单位,人均国内生产总值增加 228 美元;当电力质量得分高于 88.879 时,人均国内生产总值和电力质量得分的关系为 $Y_i=-17.504+0.228X_i+0.319(X_i-88.879)^2$,此时电力质量对人均国内生产总值的促进作用更为明显。同理,其他分位数水平也有类似的结论。在每个分位数水平下,电力质量和人均国内生产总值都呈正相关,因此国家和政府应该充分认识到电力质量与经济的关系并采取积极的措施提高电力质量,以实现人均国内生产总值的提高。

从经济发展水平分析,人均国内生产总值较低的国家或者地区,即 0.1 分位数水平的国家或者地区,当电力质量得分低于 89.630 时,提高电力质量对人均国内生产总值的提高作用较小;当电力质量得分高于 89.630 时,电力质量水平的提高对人均国内生产总值的提高作用明显增强。对于经济发展水平较高的国家或者地区,即 0.9 分位数水平的国家或者地区,当电力质量得分低于 88.984 时,电力质量水平对人均国内生产总值提高的促进作用明显大于低分位数水平的国家或者地区;当电力质量得分高于 88.984 时,电力质量水平的提高可以极大程度地提高人均国内生产总值。

4　结论

为了灵活处理线性−二次的数据且为弥补以往模型的不足,本文构建线性−二次分位数回归模型。由于损失函数非凸,给参数估计带来了困难,因此,本文提出 1 种简单的线性化技巧,将线性−二次分位数回归模型近似转化为标准线性分位数回归模型,计算模型中的回归系数和变点参数的估计,并且基于标准线性分位数回归模型的理论和 delta 方法给出参数的区间估计。数值模拟结果表明,本文的估计方法具有良好的有效性和稳健性。人均国内生产总值与电力质量数据的实证分析也验证了所提模型和方法的可行性和实用性。

本文提出的线性−二次分位数回归模型在一定程度上丰富了现有变点模型,为研究数据呈现线性−二次分布的情况提供了新思路以及研究方法;但是模型系数以及变点的估计是建立在变点存在的基础上,而在实际应用中变点是否存在是未知的,因此,后续将扩展研究变点检测问题。

致谢:本文得到了海南省院士工作站(于长斌)和中国留学基金委资助的支持。

参考文献:

[1] HUDSON D J. Fitting segmented curves whose join points have to be estimated[J]. Journal of the American Statistical Association, 1966, 61(316):1097-1129.

[2] ROBISON D E. Estimates for the points of intersection of two polynomial regressions[J]. Journal of the American Statistical Association, 1964, 59(305):214-224.

[3] FEDER P I. The log likelihood ratio in segmented regression[J]. The Annals of Statistics, 1975, 3(1):84-97.

[4] CHAPPELL R. Fitting bent lines to data, with applications to allometry[J]. Journal of Theoretical Biology, 1989, 138(2):235-256.

[5] JONES M C, HANDCOCK M S. Determination of anaerobic threshold: what anaerobic threshold? [J]. The Canadian Journal of Statistics, 1991, 19(2):236-239.

[6] HANSEN B E. Inference when a nuisance parameter is not identified under the null hypothesis[J]. Econometrica, 1996, 64(2):413-430.

[7] BAI Jushan. Likelihood ratio tests for multiple structural changes[J]. Journal of Econometrics, 1999, 91(2):299-323.

[8] LERMAN P M. Fitting segmented regression models by grid search[J]. Journal of the Royal Statistical Society Series C: Applied Statistics, 1980, 29(1):77-84.

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

[10] LEE S, SEO M H, SHIN Y. Testing for threshold effects in regression models[J]. Journal of the American Statistical Association, 2011, 106(493):220-231.

[11] 蒋家坤,林华珍,蒋靓,等. 门槛回归模型中门槛值和回归参数的估计[J]. 中国科学(数学),2016,46(4):409-422.
JIANG Jiakun, LIN Huazhen, JIANG Liang, et al. Estimation of threshold values and regression parameters in threshold regression model[J]. SCIENTIA SINICA Mathematica, 2016, 46(4):409-422.

[12] PASTOR R, GUALLAR E. Use of two-segmented logistic regression to estimate change-points in epidemiologic studies[J]. American Journal of Epidemiology, 1998, 148(7):631-642.

[13] ZHANG Feipeng, YANG Jiejing, LIU Lei, et al. Generalized linear-quadratic model with a change point due to a covariate threshold[J]. Journal of Statistical Planning and Inference, 2022, 216:194-206.

[14] KOENKER R, BASSETT G. Regression quantiles[J]. Journal of the Econometric Society, 1978, 46(1):33-50.

[15] LI Chenxi, WEI Ying, CHAPPELL R, et al. Bent line quantile regression with application to an allometric study of land mammals' speed and mass[J]. Biometrics, 2011, 67(1):242-249.

[16] FERGUSON R, WILKINSON W, HILL R. Electricity use and economic development[J]. Energy Policy, 2000, 28(13):923-934.

[17] WOLDE-RUFAEL Y. Electricity consumption and economic growth: a time series experience for 17 African countries[J]. Energy Policy, 2006, 34(10):1106-1114.

[18] ZHOU Xiaoying, ZHANG Feipeng. A new estimation method for continuous threshold expectile model[J]. Communications in Statistics: Simulation and Computation, 2018, 47(8):2486-2498.

[19] ZHANG Feipeng, ZHENG Shenglin, ZHOU Xiaoying. Bent-cable quantile regression model[J]. Communications in Statistics: Simulation and Computation, 2023, 52(5):2000-2011.

[20] 周小英. 逐段连续线性分位数回归模型的统计推断及其应用[D]. 长沙:湖南大学,2018.
ZHOU Xiaoying. Statistical inference and application in continuous threshold linear quantile regression model[D]. Changsha: Hunan University, 2018.

(编辑:李艺)