Advances in Applied Mathematics 应用数学进展, 2024, 13(8), 3651-3665
Published Online August 2024 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2024.138348

Hans汉斯

基于纵向数部分线性模型的分位数模型平均

蒲莉丽

西南大学数学与统计学院，重庆

收稿日期：2024年7月1日；录用日期：2024年7月26日；发布日期：2024年8月2日

摘　要

本文针对纵向数据部分线性回归模型中的参数估计与非参数估计问题，基于分位数回归估计方法提出了一种稳健的模型平均估计量。为了提高估计效率，采用工作相关矩阵分解和估计方程平滑法处理纵向数据的组内相关性，并通过局部线性估计方法处理模型的非参数部分，给出了模型参数与非参数估计的Newton-Raphson迭代算法。数值模拟表明，新的估计方法具有良好的估计性能。将新估计方法应用到空气质量数据的预测分析中，证明了该方法在实际应用中也具有可行性。

关键词

纵向数据，部分线性模型，局部线性估计，分位数回归，模型平均

Quantile Model Averaging Based on Longitudinal Partial Linear Model

Lili Pu

School of Mathematics and Statistics, Southwest University, Chongqing

Received: Jul. 1st, 2024; accepted: Jul. 26th, 2024; published: Aug. 2nd, 2024

Abstract

Based on the problem of parameter estimation and non-parametric estimation in the partial linear regression model of longitudinal data, this paper proposes a robust model average estimator based on the quantile regression estimation method. In order to improve the estimation efficiency, the working correlation matrix decomposition and estimation equation smoothing method are used to deal with the intra-group correlation of longitudinal data, and the non-parametric part of the model is processed by the local linear estimation method. The Newton-Raphson iterative algorithm for model parameter and non-parametric estimation is given. Numerical simulation shows that the new estimation method has good estimation performance. The new estimation method is

文章引用: 蒲莉丽. 基于纵向数部分线性模型的分位数模型平均[J]. 应用数学进展, 2024, 13(8): 3651-3665.
DOI: 10.12677/aam.2024.138348

applied to the prediction and analysis of air quality data, which proves that the method is also feasible in practical applications.

Keywords

Longitudinal Data, Partially Linear Models, Local Linear Estimation, Quantile Regression, Model Averaging

1. 引言

假设有 n 个观测样本，记 T_{ij} 为第 i 个样本第 j 次观测的时间变量。 $\boldsymbol{X}_{ij}$ 与 Y_{ij} 分别为第 i 个样本第 j 次观测的协变量矩阵和响应变量， m_i 为第 i 个样本重复观测的次数，则有纵向数据样本集 $\left\{\left(\boldsymbol{X}_{ij},T_{ij},Y_{ij}\right),i=1,\cdots,n;j=1,\cdots,m_i\right\}$。一类重要的纵向数据部分线性回归模型为下列形式：

$$Y_{ij}=\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}+g\left(T_{ij}\right)+\varepsilon_{ij} \tag{1}$$

其中 $\boldsymbol{\beta}$ 为 p 维分位数回归系数向量， $g(\cdot)$ 是定义在有界闭区间 $[0,1]$ 上的未知光滑函数， ε_{ij} 为模型的连续误差项。

Zeger 和 Diggle [1]率先对模型(1)进行了研究，并将其应用于 HIV 研究中 CD4 细胞数的时间趋势的估计问题中。Lin 和 Carroll [2]提出了 profile-kernel 估计方法，估计兴趣参数。Hu 等[3]将 Zeger 和 Diggle 提出的后移算法与 Lin 和 Carroll 提出的 profile-kernel 估计方法进行了比较。Fan 和 Li [4]使用局部多项式回归方法估计非参数函数，提出了两种简单有效的参数分量的估计方法：差分估计和 profile 最小二乘估计。田萍[5]详细总结了前人关于模型(1)在随机点列情况下所做过的工作，并讨论了其在固定设计点列下的情况。Tian 和 Xue [6]在部分线性回归模型框架内对纵向数据进行经验似然推断。王明辉和尹居良[7]研究了模型(1)的分位数估计问题，并证明了参数估计量的渐近正态性。刘会明[8]在模型(1)框架下，对局部多项式进行改进并提出了全局拟似然方法，给出了回归系数和非参数函数的拟似然估计。

在纵向数据分析中，分位数回归也是一种被广泛使用的估计方法。为了降低分位数回归推断的效率损失，Jung [9]首次提出了一种中位数回归的准似然方法，该方法将纵向数据重复测量之间的相关性纳入其中。Fu 和 Wang [10]结合重复测量之间的相关性，引入了诱导平滑方法来获得参数估计及其方差的方法。Leng 和 Zhang [11]通过组合多组无偏估计方程，提出了一种在考虑重复测量之间的相关性的同时，产生更有效估计的分位数回归模型。Fu 和 Wang、Leng 和 Zhang 都将诱导平滑方法扩展到了分位数回归，从而获得了允许应用 Newton-Raphson 迭代的平滑目标函数，后者还给出了参数的估计及其协方差矩阵的 Sandwich 估计。Lu 和 Fan [12]在他们的基础上，提出了一个结合了纵向数据重复测量之间的相关性结构的纵向数据分位数回归模型。模型平均方法能够有效平衡估计的方差和偏差。在分位数回归框架下，Lu 和 Su [13]将 Jackknife 模型平均(JMA)方法应用到线性分位数回归，证明了该估计量的渐近最优性。对于部分线性模型的模型平均预测分析。Zhang 和 Wang [14]提出了具有独立误差的部分线性模型最佳模型平均。Fang [15]提出了一种用于二分反应的半参数模型平均预测方法。胡国治和曾婕[16]构造了基于部分线性分位数回归模型的模型平均估计量并探究了其大样本性质。这些半参数模型平均的例子都是在独立假

设下进行的。在纵向数据中，Hu 等[17]研究了纵向数据变系数部分线性模型下模型平均估计量的渐近分布。Li 等[18]考虑过具有相关误差的半参数模型的模型平均，提供了一种最佳模型平均方法来改进纵向数据部分线性模型的预测。

受上述文献启发，针对模型(1)的参数与非参数估计，本文的创新点在于将分位数回归方法和模型平均估计拓展到其中的，采用局部线性估计对模型(1)的非参数部分拟合，对模型(1)的参数估计部分，提出一种稳健的分位数回归估计量，给出了估计的 Newton-Raphson 迭代算法及协方差矩阵的 Sandwich 估计提高了估计效率，利用 JMA 估计模型平均的最优权重。并通过数值模拟与实例分析，证明该方法优于纵向数据部分线性回归模型分位数估计、纵向数据线性回归模型分位数估计、纵向数据线性回归模型、分位数模型平均估计。

2. 分位数平均估计模型模型

2.1. 构建候选子模型

定义纵向数据样本集$\left\{\left(\boldsymbol{X}_i,\boldsymbol{T}_i,\boldsymbol{Y}_i\right)\right\}_{i=1}^{n}$，其中$\boldsymbol{Y}_i=\left(Y_{i1},\cdots,Y_{im_i}\right)^{\mathrm{T}}$是第 i 个个体 $m_i\times 1$维响应变量，$\boldsymbol{X}_i=\left(\boldsymbol{X}_{i1},\cdots,\boldsymbol{X}_{im_i}\right)^{\mathrm{T}}$是第 i 个个体 $m_i\times p$ 维协变量矩阵，$\boldsymbol{X}_{ij}=\left(X_{ij1},\cdots,X_{ijp}\right)^{\mathrm{T}}$是第 i 个个体第 j 次观测的 p 维协变量，$j=1,\cdots,m_i$，$\boldsymbol{T}_i=\left(T_{i1},\cdots,T_{im_i}\right)^{\mathrm{T}}$是第 i 个个体 $m_i\times 1$维时间变量，为不失一般性，假定$T_{ij}\in[0,1]$。本文主要考虑下列纵向数据部分线性分位数回归模型：

$$\mu_{\tau ij}=Q_\tau\left(Y_{ij}\mid\boldsymbol{X}_{ij}\right)=\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_\tau+g_\tau\left(T_{ij}\right)\tag{2}$$

其中$\boldsymbol{\beta}_\tau$为 p 维分位数回归系数向量，$g_\tau(\cdot)$是定义在有界闭区间$[0,1]$上的未知光滑函数，$\tau\in(0,1)$为感兴趣的分位数水平。记$\boldsymbol{g}_\tau\left(\boldsymbol{T}_i\right)=\left(g_\tau\left(T_{i1}\right),\cdots,g_\tau\left(T_{im_i}\right)\right)^{\mathrm{T}}$，$\boldsymbol{\mu}_{\tau i}=\left(\mu_{\tau i1},\cdots,\mu_{\tau im_i}\right)^{\mathrm{T}}$，则有下列纵向数据部分线性分位数回归模型：

$$\boldsymbol{\mu}_{\tau i}=Q_\tau\left(\boldsymbol{Y}_i\mid\boldsymbol{X}_i\right)=\boldsymbol{X}_i\boldsymbol{\beta}_\tau+\boldsymbol{g}_\tau\left(\boldsymbol{T}_i\right)\tag{3}$$

为探究上述模型的模型平均估计量，我们采用嵌套的方式构建候选子模型，记$\boldsymbol{X}_{ij}^{(s)}=\left(X_{ij1},\cdots,X_{ijs}\right)^{\mathrm{T}}$是第 s 个候选子模型第 i 个个体第 j 次观测的协变量，$s=1,\cdots,p$，此时第 s 个候选子模型 M_s 为：

$$\boldsymbol{\mu}_{\tau i}^{(s)}=Q_\tau\left(\boldsymbol{Y}_i\mid\boldsymbol{X}_i^{(s)}\right)=\boldsymbol{X}_i^{(s)}\boldsymbol{\beta}_\tau^{(s)}+\boldsymbol{g}_\tau^{(s)}\left(\boldsymbol{T}_i\right)\tag{4}$$

其中$\boldsymbol{X}_i^{(s)}=\left(\boldsymbol{X}_{i1}^{(s)},\cdots,\boldsymbol{X}_{im_i}^{(s)}\right)^{\mathrm{T}}$是 $m_i\times s$ 维协变量矩阵，$\boldsymbol{\beta}_\tau^{(s)}$为 s 维分位数回归系数向量，$g_\tau^{(s)}(\cdot)$是定义在有界闭区间$[0,1]$上的未知光滑函数。

2.2. 子模型的参数估计

为获得第 s 个候选子模型 M_s 中未知回归系数向量$\boldsymbol{\beta}_\tau^{(s)}$与未知光滑函数$g_\tau^{(s)}(\cdot)$的估计值，我们采用两阶段估计法。第一阶段，不考虑纵向数据组内相关性，获得$\boldsymbol{\beta}_\tau^{(s)}$与$g_\tau^{(s)}(\cdot)$的初始估计值$\tilde{\boldsymbol{\beta}}_\tau^{(s)}$与$\tilde{g}_\tau^{(s)}(\cdot)$。

对于候选子模型中未知光滑函数$g_\tau^{(s)}(\cdot)$的估计，我们采用局部线性估计的非参数估计方法，在T_{ij}的一个小区域内对$g_\tau^{(s)}\left(T_{ij}\right)$进行一阶泰勒展开，即

$$g_\tau^{(s)}\left(T_{ij}\right)\approx g_\tau^{(s)}(t)+\dot{g}_\tau^{(s)}(t)\left(T_{ij}-t\right)=\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\theta}_\tau^{(s)}$$

其中$\dot{g}_\tau^{(s)}(\cdot)$是$g_\tau^{(s)}(\cdot)$的一阶导数，$\boldsymbol{Z}_{ij}=\left(1,T_{ij}-t\right)^{\mathrm{T}}$是局部线性估计中的设计矩阵，$\boldsymbol{\theta}_\tau^{(s)}=\left(g_\tau^{(s)}(t),\dot{g}_\tau^{(s)}(t)\right)^{\mathrm{T}}$是相应的参数向量。记$\boldsymbol{D}_{ij}^{(s)}=\left(\boldsymbol{X}_{ij}^{(s)\mathrm{T}},\boldsymbol{Z}_{ij}^{\mathrm{T}}\right)^{\mathrm{T}}$，$\boldsymbol{\eta}_\tau^{(s)}=\left(\boldsymbol{\beta}_\tau^{(s)\mathrm{T}},\boldsymbol{\theta}_\tau^{(s)\mathrm{T}}\right)^{\mathrm{T}}$，可以通过最小化下列目标函数得到$\boldsymbol{\eta}_\tau^{(s)}$的估计值

$$\tilde{\boldsymbol{\eta}}_{\tau}^{(s)}=\underset{\boldsymbol{\eta}_{\tau}^{(s)}}{\operatorname{argmin}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_{\tau}\left(Y_{ij}-\boldsymbol{D}_{ij}^{(s)\mathrm{T}}\boldsymbol{\eta}_{\tau}^{(s)}\right)K_{h_1}\left(T_{ij}-t\right)$$

其中，$\rho_{\tau}(u)=u\left(\tau-\mathrm{I}_{(u\le 0)}\right)$为给定$\tau$分位数下模型的损失函数，$K_{h_1}(\cdot)=K(\cdot/h_1)$，$K(\cdot)$是核函数，$h_1$为给定的窗宽。从而可以得到$g_{\tau}^{(s)}(t)$的初始估计值$\tilde{g}_{\tau}^{(s)}(t)=(0_{1\times s},1,0)\tilde{\boldsymbol{\eta}}_{\tau}^{(s)}$。

尽管在上述估计中我们得到了$\boldsymbol{\beta}_{\tau}^{(s)}$的估计值，但由于该估计值是局部最优估计，其收敛速率未达到$\sqrt{n}$。因此，我们利用$g_{\tau}^{(s)}\left(T_{ij}\right)$的初始估计值$\tilde{g}_{\tau}^{(s)}\left(T_{ij}\right)$，得到$\boldsymbol{\beta}_{\tau}^{(s)}$的全局最优初始估计

$$\tilde{\boldsymbol{\beta}}_{\tau}^{(s)}=\underset{\boldsymbol{\beta}_{\tau}^{(s)}}{\operatorname{argmin}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_{\tau}\left[Y_{ij}-\tilde{g}_{\tau}^{(s)}\left(T_{ij}\right)-\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\boldsymbol{\beta}_{\tau}^{(s)}\right]$$

第二阶段，在考虑纵向数据组内相关性的基础上，利用$\boldsymbol{\beta}_{\tau}^{(s)}$与$g_{\tau}^{(s)}(\cdot)$在第一阶段获得的初始估计值$\tilde{\boldsymbol{\beta}}_{\tau}^{(s)}$与$\dot{g}_{\tau}^{(s)}(\cdot)$，得到其最终估计值。令$\boldsymbol{\varepsilon}_i^{(s)}=\left(\varepsilon_{i1}^{(s)},\cdots,\varepsilon_{im_i}^{(s)}\right)^{\mathrm{T}}$，其中$\varepsilon_{ij}^{(s)}=Y_{ij}-g_{\tau}^{(s)}\left(T_{ij}\right)-\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\boldsymbol{\beta}_{\tau}^{(s)}$是候选子模型的连续误差项，具有未知条件密度函数$f_{ij}^{(s)}(\cdot)$。将$\boldsymbol{\varphi}_{\tau}\left(\boldsymbol{\varepsilon}_i^{(s)}\right)$的协方差矩阵表示为

$$\boldsymbol{V}_i^{(s)}=\mathrm{Cov}\left(\boldsymbol{\varphi}_{\tau}\left(\boldsymbol{\varepsilon}_i^{(s)}\right)\right)=\mathrm{Cov}\begin{pmatrix}\tau-\mathrm{I}_{\left(\varepsilon_{i1}^{(s)}\le 0\right)}\\ \vdots\\ \tau-\mathrm{I}_{\left(\varepsilon_{im_i}^{(s)}\le 0\right)}\end{pmatrix}$$

且令$\boldsymbol{\Gamma}_i^{(s)}=\mathrm{diag}\left[f_{i1}^{(s)}(0),\cdots,f_{im_i}^{(s)}(0)\right]$。Jung [9]提出了下列估计方程，用于计算考虑响应变量组内相关性后$\boldsymbol{\beta}_{\tau}^{(s)}$的估计值

$$\boldsymbol{U}_0\left(\boldsymbol{\beta}_{\tau}^{(s)}\right)=\sum_{i=1}^{n}\boldsymbol{X}_i^{(s)\mathrm{T}}\boldsymbol{\Gamma}_i^{(s)}\left(\boldsymbol{V}_i^{(s)}\right)^{-1}\boldsymbol{\varphi}_{\tau}\left(\boldsymbol{Y}_i-\boldsymbol{g}_{\tau}^{(s)}(\boldsymbol{T}_i)-\boldsymbol{X}_i^{(s)}\boldsymbol{\beta}_{\tau}^{(s)}\right)=0 \tag{5}$$

其中$\varphi_{\tau}(u)=\rho_{\tau'}(u)=\tau-1_{(u\le 0)}$。在求解估计方程(5)时$\boldsymbol{\Gamma}_i^{(s)}$的元素$f_{ij}^{(s)}(0)$的估计值$\hat{f}_{ij}^{(s)}(0)$可以参考Hendricks和Knenker [19]得到

$$\hat{f}_{ij}^{(s)}(0)=2h_n\left[\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\left(\hat{\boldsymbol{\beta}}_{\tau+h_n}^{(s)}-\hat{\boldsymbol{\beta}}_{\tau-h_n}^{(s)}\right)\right]^{-1}$$

当然，在某些情况下，当$f_{ij}^{(s)}(0)$的值难以估计时，$\boldsymbol{\Gamma}_i^{(s)}$可简单视为一个单位矩阵，此时$\boldsymbol{\beta}_{\tau}^{(s)}$的估计效率略有损失，本文就将$\boldsymbol{\Gamma}_i^{(s)}$视为一个单位矩阵。

由于在进行参数估计时，协方差矩阵$\boldsymbol{V}_i^{(s)}$的估计非常复杂，很难正确指定，于是Lu和Fan [12]提出了以下估计方程

$$\boldsymbol{U}_1\left(\boldsymbol{\beta}_{\tau}^{(s)}\right)=\sum_{i=1}^{n}\boldsymbol{X}_i^{(s)\mathrm{T}}\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\boldsymbol{\varphi}_{\tau}\left(\boldsymbol{Y}_i-\boldsymbol{g}_{\tau}^{(s)}(\boldsymbol{T}_i)-\boldsymbol{X}_i^{(s)}\boldsymbol{\beta}_{\tau}^{(s)}\right)=0 \tag{6}$$

其中$\boldsymbol{\Sigma}_i^{(s)}(\rho)$是$\boldsymbol{\varphi}_{\tau}\left(\boldsymbol{\varepsilon}_i^{(s)}\right)$的工作协方差矩阵，$\boldsymbol{\Sigma}_i^{(s)}(\rho)=\boldsymbol{A}_i^{(s)1/2}\boldsymbol{C}_i^{(s)}(\rho)\boldsymbol{A}_i^{(s)1/2}$，$\boldsymbol{A}_i^{(s)}=\mathrm{diag}\left[\sigma_{i1}^{(s)},\cdots,\sigma_{im_im_i}^{(s)}\right]$是一个$m_i\times m_i$维对角矩阵，$\sigma_{ijj}^{(s)}=\mathrm{Var}\left(\varphi_{\tau}\left(\varepsilon_{ij}^{(s)}\right)\right)$，$\boldsymbol{C}_i^{(s)}(\rho)$是$\boldsymbol{\varphi}_{\tau}\left(\boldsymbol{\varepsilon}_i^{(s)}\right)$的工作相关矩阵，$\rho$是相关指数参数。假定工作协方差矩阵$\boldsymbol{\Sigma}_i^{(s)}(\rho)$具有一般的平稳自相关结构，则工作相关矩阵$\boldsymbol{C}_i^{(s)}(\rho)$具有如下形式

$$\boldsymbol{C}_i^{(s)}(\rho)=\begin{pmatrix}1 & \rho_1 & \rho_2 & \cdots & \rho_{m_i-1}\\ \rho_1 & 1 & \rho_2 & \cdots & \rho_{m_i-2}\\ \vdots & \vdots & \vdots & & \vdots\\ \rho_{m_i-1} & \rho_{m_i-2} & \rho_{m_i-3} & \cdots & 1\end{pmatrix}$$

对于所有 $i=1,\cdots,n$， ρ_l 的估计值

$$\hat{\rho}_l=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m_i-l}\tilde{y}_{ij}^{(s)}\tilde{y}_{i,j+l}^{(s)}\Big/n\left(m_i-l\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m_i}\tilde{y}_{ij}^{(s)2}\Big/nm_i}$$

其中 $l=1,\cdots,m_i-1$，Hendricks 和 Knenker [19]定义

$$\tilde{y}_{ij}^{(s)}=\frac{\varphi_\tau\left(Y_{ij}-g_\tau^{(s)}\left(T_{ij}\right)-\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\boldsymbol{\beta}_\tau^{(s)}\right)}{\sqrt{\delta_{ijj}^{(s)}}}$$

为了估计 $\sigma_{ijj}^{(s)}=\mathrm{Var}\left(\varphi_\tau\left(\varepsilon_{ij}^{(s)}\right)\right)$，我们令 $\tilde{\boldsymbol{\varepsilon}}_i^{(s)}=\left(\tilde{\varepsilon}_{i1}^{(s)},\cdots,\tilde{\varepsilon}_{im_i}^{(s)}\right)^{\mathrm{T}}$， $\tilde{\varepsilon}_{ij}^{(s)}=Y_{ij}-\tilde{g}_\tau^{(s)}\left(T_{ij}\right)-\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\boldsymbol{\beta}_\tau^{(s)}$。此时应用 $\boldsymbol{\varphi}_\tau\left(\tilde{\boldsymbol{\varepsilon}}_i^{(s)}\right)=\left(\varphi_\tau\left(\tilde{\varepsilon}_{i1}^{(s)}\right),\cdots,\varphi_\tau\left(\tilde{\varepsilon}_{im_i}^{(s)}\right)\right)^{\mathrm{T}}$， $\varphi_\tau\left(\tilde{\varepsilon}_{ij}^{(s)}\right)=\tau-1_{\left(\tilde{\varepsilon}_{ij}^{(s)}\le 0\right)}$，可得到

$$\hat{\delta}_{ijj}^{(s)}=P\left(\tilde{\varepsilon}_{ij}^{(s)}\le 0\right)\left(1-P\left(\tilde{\varepsilon}_{ij}^{(s)}\le 0\right)\right)$$

求解估计方程(6)的主要困难在于不可微的非凸和非连续目标函数 $\boldsymbol{U}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)$，为了克服这些问题，Fu 和 Wang [10]将诱导平滑方法扩展到对估计方程(7)的求解中。令 $\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)=E_{\boldsymbol{H}^{(s)}}\left[\boldsymbol{U}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)+\boldsymbol{\Omega}_1^{(s)1/2}\boldsymbol{H}^{(s)}\right]$，其中 $\boldsymbol{H}^{(s)}\sim N\left(0,\boldsymbol{I}_s\right)$， $\boldsymbol{I}_s$ 为 $s\times s$ 维单位矩阵， $\boldsymbol{\Omega}_1^{(s)}$ 为参数估计器更新后的协方差矩阵的估计。经过一系列代数计算，可得到下列平滑估计方程

$$\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)=\sum_{i=1}^{n}\boldsymbol{X}_i^{(s)\mathrm{T}}\left[\boldsymbol{\Sigma}_i^{(s)}\left(\rho\right)\right]^{-1}\tilde{\boldsymbol{\varphi}}_\tau\left(\boldsymbol{Y}_i-\tilde{\boldsymbol{g}}_\tau^{(s)}\left(\boldsymbol{T}_i\right)-\boldsymbol{X}_i^{(s)}\boldsymbol{\beta}_\tau^{(s)}\right)=0 \tag{8}$$

其中， $\tilde{\boldsymbol{\varphi}}_\tau\left(\tilde{\boldsymbol{\varepsilon}}_i^{(s)}\right)=\left(\tau-1+\Phi\left(\frac{\tilde{\varepsilon}_{i1}^{(s)}}{\tilde{r}_{i1}}\right),\cdots,\tau-1+\Phi\left(\frac{\tilde{\varepsilon}_{im_i}^{(s)}}{\tilde{r}_{im_i}}\right)\right)^{\mathrm{T}}$， $\tilde{r}_{ij}=\sqrt{\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\boldsymbol{\Omega}_1^{(s)}\boldsymbol{X}_{ij}^{(s)}}$。此时 $\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)$ 对 $\boldsymbol{\beta}_\tau^{(s)}$ 的微分可以很容易的计算出来，并且可以用 $\frac{\partial\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)}{\partial\boldsymbol{\beta}_\tau^{(s)}}$ 近似地代替 $\frac{\partial\boldsymbol{U}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)}{\partial\boldsymbol{\beta}_\tau^{(s)}}$ 为

$$\frac{\partial\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)}{\partial\boldsymbol{\beta}_\tau^{(s)}}=-\sum_{i=1}^{n}\boldsymbol{X}_i^{(s)\mathrm{T}}\left[\boldsymbol{\Sigma}_i^{(s)}\left(\rho\right)\right]^{-1}\tilde{\boldsymbol{\Lambda}}_i^{(s)}\boldsymbol{X}_i^{(s)}$$

其中 $\tilde{\boldsymbol{\Lambda}}_i^{(s)}$ 为 $m_i\times m_i$ 维对角矩阵，第 j 个对角元为 $\phi\left(\tilde{\varepsilon}_{ij}^{(s)}\Big/\tilde{r}_{ij}\right)\Big/\tilde{r}_{ij}$。

此时， $\boldsymbol{\beta}_\tau^{(s)}$ 及其协方差矩阵 $\boldsymbol{\Omega}_1^{(s)}$ 的平滑估计量可以从下列 Newton-Raphson 迭代算法中获得：

步骤 1：给定 $\boldsymbol{\beta}_\tau^{(s)}$ 和协方差矩阵 $\boldsymbol{\Omega}_1^{(s)}$ 的初始值，分别为 $\hat{\boldsymbol{\beta}}_\tau^{(s)}\left(0\right)=\tilde{\boldsymbol{\beta}}_\tau^{(s)}$， $\hat{\boldsymbol{\Omega}}_1^{(s)}\left(0\right)=\frac{1}{n}\boldsymbol{I}_s$。

步骤 2：使用第 r 次迭代出的 $\hat{\boldsymbol{\beta}}_\tau^{(s)}\left(r\right)$ 和 $\hat{\boldsymbol{\Omega}}_1^{(s)}\left(r\right)$，通过下列公式更新 $\hat{\boldsymbol{\beta}}_\tau^{(s)}\left(r+1\right)$ 和 $\hat{\boldsymbol{\Omega}}_1^{(s)}\left(r+1\right)$

$$\hat{\boldsymbol{\beta}}_\tau^{(s)}\left(r+1\right)=\hat{\boldsymbol{\beta}}_\tau^{(s)}\left(r\right)+\left[\frac{\partial\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)}{\partial\boldsymbol{\beta}_\tau^{(s)}}\right]_r^{-1}\times\left[\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)\right]_r$$

$$\hat{\boldsymbol{\Omega}}_1^{(s)}\left(r+1\right)=\left[-\frac{\partial\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)}{\partial\boldsymbol{\beta}_\tau^{(s)}}\right]_r^{-1}\times\left[\mathrm{Cov}\left(\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)\right)\right]_r\times\left[-\frac{\partial\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)}{\partial\boldsymbol{\beta}_\tau^{(s)}}\right]_r^{-1}$$

其中，$\mathrm{Cov}\left(\tilde{\boldsymbol{U}}_1\left(\boldsymbol{\beta}_\tau^{(s)}\right)\right)=\sum_{i=1}^{n}\boldsymbol{X}_i^{(s)\mathrm{T}}\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\tilde{\boldsymbol{\varphi}}_\tau\left(\tilde{\boldsymbol{\varepsilon}}_i^{(s)}\right)\tilde{\boldsymbol{\varphi}}_\tau^{\mathrm{T}}\left(\tilde{\boldsymbol{\varepsilon}}_i^{(s)}\right)\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\boldsymbol{X}_i^{(s)}$，$[\,]_r$ 表示方括号内的表达式在 $\boldsymbol{\beta}_\tau^{(s)}=\hat{\boldsymbol{\beta}}_\tau^{(s)}(r)$ 的计算结果。

步骤 3：重复步骤 2 直至收敛，并记最终的收敛值为 $\hat{\boldsymbol{\beta}}_\tau^{(s)}$，$\hat{\boldsymbol{\Omega}}_1^{(s)}$。

上述估计方法提供了 $\boldsymbol{\beta}_\tau^{(s)}$ 及其协方差矩阵 $\boldsymbol{\Omega}_1^{(s)}$ 的有效估计 $\hat{\boldsymbol{\beta}}_\tau^{(s)}$ 和 $\hat{\boldsymbol{\Omega}}_1^{(s)}$。接下来，利用 $\hat{\boldsymbol{\beta}}_\tau^{(s)}$ 估计候选子模型中的 $g_\tau^{(s)}(\cdot)$。令 $\check{\boldsymbol{\varepsilon}}_i^{(s)}=\left(\check{\varepsilon}_{i1}^{(s)},\cdots,\check{\varepsilon}_{im_i}^{(s)}\right)$，$\check{\varepsilon}_{ij}^{(s)}=Y_{ij}-\boldsymbol{X}_{ij}^{(s)\mathrm{T}}\hat{\boldsymbol{\beta}}_\tau^{(s)}-g_\tau^{(s)}\left(T_{ij}\right)$，利用局部线性估计得到的一阶泰勒展开，参考 Lv 等[20]的估计方程(4)，提出下列平滑估计方程用于估计 $\boldsymbol{\theta}_\tau^{(s)}$

$$\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)=\sum_{i=1}^{n}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{K}_i(t;h_2)\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\tilde{\boldsymbol{\varphi}}_\tau\left(\boldsymbol{Y}_i-\boldsymbol{X}_i^{(s)}\hat{\boldsymbol{\beta}}_\tau^{(s)}-\boldsymbol{Z}_i\boldsymbol{\theta}_\tau^{(s)}\right)=0 \tag{9}$$

其中 $\tilde{\boldsymbol{\varphi}}_\tau\left(\check{\boldsymbol{\varepsilon}}_i^{(s)}\right)=\left(\tau-1+\Phi\left(\frac{\check{\varepsilon}_{i1}^{(s)}}{\check{r}_{i1}}\right),\cdots,\tau-1+\Phi\left(\frac{\check{\varepsilon}_{im_i}^{(s)}}{\check{r}_{im_i}}\right)\right)^{\mathrm{T}}$，$\check{r}_{ij}=\sqrt{\boldsymbol{Z}_{ij}^{\mathrm{T}}\boldsymbol{\Omega}_2^{(s)}\boldsymbol{Z}_{ij}}$，$\boldsymbol{Z}_i=\left(\boldsymbol{Z}_{i1},\cdots,\boldsymbol{Z}_{im_i}\right)^{\mathrm{T}}$，$\boldsymbol{K}_i(t;h_2)=\mathrm{diag}\left(K_{h_2}\left(T_{i1}-t\right),\cdots,K_{h_2}\left(T_{im_i}-t\right)\right)$。同理，我们可以得到 $\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)$ 关于 $\boldsymbol{\theta}_\tau^{(s)}$ 的一阶微分为

$$\frac{\partial\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)}{\partial\boldsymbol{\theta}_\tau^{(s)}}=-\sum_{i=1}^{n}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{K}_i(t;h_2)\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\check{\boldsymbol{\Lambda}}_i^{(s)}\boldsymbol{Z}_i$$

其中 $\check{\boldsymbol{\Lambda}}_i^{(s)}$ 为 $m_i\times m_i$ 维对角矩阵，第 j 个对角元为 $\phi\left(\check{\varepsilon}_{ij}^{(s)}\big/\check{r}_{ij}\right)\Big/\check{r}_{ij}$。

此时，$\boldsymbol{\theta}_\tau^{(s)}$ 及其工作协方差矩阵 $\boldsymbol{\Omega}_2^{(s)}$ 的光滑估计量可以从下列 Newton-Raphson 迭代算法中获得：

步骤 1：给定 $\boldsymbol{\theta}_\tau^{(s)}$ 和协方差矩阵 $\boldsymbol{\Omega}_2^{(s)}$ 的初始值，分别为 $\hat{\boldsymbol{\theta}}_\tau^{(s)}(0)=\left((0_{1\times s},1,0)\tilde{\boldsymbol{\eta}}_\tau^{(s)},(0_{1\times s},0,1)\tilde{\boldsymbol{\eta}}_\tau^{(s)}\right)^{\mathrm{T}}$，$\hat{\boldsymbol{\Omega}}_2^{(s)}(0)=\frac{1}{n}\boldsymbol{I}_2$。

步骤 2：使用第 r 次迭代出的 $\hat{\boldsymbol{\theta}}_\tau^{(s)}(r)$ 和 $\hat{\boldsymbol{\Omega}}_2^{(s)}(r)$，通过下列公式更新 $\hat{\boldsymbol{\theta}}_\tau^{(s)}(r+1)$ 和 $\hat{\boldsymbol{\Omega}}_2^{(s)}(r+1)$

$$\hat{\boldsymbol{\theta}}_\tau^{(s)}(r+1)=\hat{\boldsymbol{\theta}}_\tau^{(s)}(r)+\left[-\frac{\partial\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)}{\partial\boldsymbol{\theta}_\tau^{(s)}}\right]_r^{-1}\times\left[\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)\right]_r$$

$$\hat{\boldsymbol{\Omega}}_2^{(s)}(r+1)=\left[-\frac{\partial\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)}{\partial\boldsymbol{\theta}_\tau^{(s)}}\right]_r^{-1}\times\left[\mathrm{Cov}\left(\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)\right)\right]_r\times\left[-\frac{\partial\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)}{\partial\boldsymbol{\theta}_\tau^{(s)}}\right]_r^{-1}$$

其中，$\mathrm{Cov}\left(\tilde{\boldsymbol{U}}_2\left(\boldsymbol{\theta}_\tau^{(s)}\right)\right)=\sum_{i=1}^{n}\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{K}_i(t;h_2)\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\tilde{\boldsymbol{\varphi}}_\tau\left(\check{\boldsymbol{\varepsilon}}_i^{(s)}\right)\tilde{\boldsymbol{\varphi}}_\tau^{\mathrm{T}}\left(\check{\boldsymbol{\varepsilon}}_i^{(s)}\right)\left[\boldsymbol{\Sigma}_i^{(s)}(\rho)\right]^{-1}\boldsymbol{K}_i(t;h_2)\boldsymbol{Z}_i$。

步骤 3：重复步骤 2 直至收敛，并记最终的收敛值为 $\hat{\boldsymbol{\theta}}_\tau^{(s)}$，$\hat{\boldsymbol{\Omega}}_2^{(s)}$。

最后，我们得到 $\hat{g}_\tau^{(s)}(t)=(1,0)\hat{\boldsymbol{\theta}}_\tau^{(s)}$，那么第 s 个候选子模型 M_s 下 $\hat{\boldsymbol{\mu}}_{\tau i}^{(s)}$ 的估计值

$$\hat{\boldsymbol{\mu}}_{\tau i}^{(s)}=\boldsymbol{X}_i^{(s)}\hat{\boldsymbol{\beta}}_\tau^{(s)}+\hat{\boldsymbol{g}}_\tau^{(s)}\left(\boldsymbol{T}_i\right) \tag{10}$$

其中 $\hat{\boldsymbol{g}}_\tau^{(s)}\left(\boldsymbol{T}_i\right)=\left(\hat{g}_\tau^{(s)}\left(T_{i1}\right),\cdots,\hat{g}_\tau^{(s)}\left(T_{im_i}\right)\right)^{\mathrm{T}}$。

2.3. Jackknife 权重选择

设 $\boldsymbol{\omega}=\left(\omega_1,\cdots,\omega_p\right)^{\mathrm{T}}$ 是集合 $\mathcal{W}=\left\{\omega\in[0,1]^p:0\le\omega_s\le1,\sum_{s=1}^{p}\omega_s=1\right\}$ 中的权重向量，则 $\boldsymbol{\mu}_{\tau i}$ 的模型平均估计量

$$\hat{\boldsymbol{\mu}}_{\tau i}\left(\boldsymbol{\omega}\right)=\sum_{s=1}^{p}\omega_s\hat{\boldsymbol{\mu}}_{\tau i}^{(s)} \tag{11}$$

设 $\hat{\boldsymbol{\beta}}_{\tau i}^{(s)}$，$\hat{g}_{\tau i}^{(s)}(\cdot)$ 表示第 s 个子模型 M_s 中删除第 i 个个体的观测数据 $\left\{\boldsymbol{X}_i,\boldsymbol{T}_i,\boldsymbol{Y}_i\right\}$ 后 $\boldsymbol{\mu}_{\tau i}^{(s)}$ 的 Jackknife 估计值，则将留一交叉验证的准则函数定义为

$$\boldsymbol{CV}_n\left(\boldsymbol{\omega}\right)=\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(\boldsymbol{Y}_i-\sum_{s=1}^{p}\omega_s\left[\boldsymbol{X}_i^{(s)}\hat{\boldsymbol{\beta}}_{\tau i}^{(s)}+\hat{\boldsymbol{g}}_{\tau i}^{(s)}\left(\boldsymbol{T}_i\right)\right]\right) \tag{12}$$

此时，通过选择 $\boldsymbol{\omega}\in\mathcal{W}$ 以最小化上述准则函数，获得权重向量 $\boldsymbol{\omega}$ 的 Jackknife 选择 $\hat{\boldsymbol{\omega}}=\left(\hat{\omega}_1,\cdots,\hat{\omega}_p\right)$，即

$$\hat{\boldsymbol{\omega}}=\underset{\boldsymbol{\omega}\in\mathcal{W}}{\operatorname{argmin}}\,\boldsymbol{CV}_n\left(\boldsymbol{\omega}\right) \tag{13}$$

最后，通过 $\hat{\boldsymbol{\omega}}$ 可以得到 $\boldsymbol{\mu}_{\tau i}$ 的 Jackknife 模型平均(JMA)估计量

$$\hat{\boldsymbol{\mu}}_{\tau i}\left(\hat{\boldsymbol{\omega}}\right)=\sum_{s=1}^{p}\hat{\omega}_s\hat{\boldsymbol{\mu}}_{\tau i}^{(s)} \tag{14}$$

3. 数值模拟

3.1. 数据生成

本节利用 R 软件，根据模型(15)生成三种不同类型的数据，对比 Lu 和 Fan [12]提出的纵向数据线性回归模型的分位数回归估计(LQR)、基于 Lu 和 Fan [12]提出的纵向数据线性回归模型分位数回归估计的 Jackknife 模型平均估计(MLQR)、基于本文子模型估计方法的纵向数据部分线性回归模型分位数回归估计(BQR)、本文提出的纵向数据部分线性回归模型分位数回归模型平均估计(MBQR)这四种估计方法在样本外预测误差方面的表现情况。

将非参数核函数固定为 Epanechnikov 核，即 $K\left(u\right)=0.75\left(1-u^2\right)_+$，最优窗宽 h_1 和 h_2 的选择采用五折交叉验证标准。以 h_1 为例，设 $T-T^v$ 和 T^v 分别为 $v=1,\cdots,5$ 的交叉验证训练集和测试集，其中 T 是完整的数据集。对于每个 h_1 和 v，我们使用训练集 $T-T^v$ 找到 $\boldsymbol{\eta}_\tau^{(s)}$ 的估计 $\tilde{\boldsymbol{\eta}}_\tau^{(s)}$，然后，利用下列五折交叉验证标准：

$$\mathrm{CV}\left(h_1\right)=\sum_{v=1}^{5}\sum_{\left(Y_{ij},\boldsymbol{D}_{ij}^{(s)}\right)\in T^v}\rho_\tau\left(Y_{ij}-\boldsymbol{D}_{ij}^{(s)\mathrm{T}}\tilde{\boldsymbol{\eta}}_\tau^{(s)}\right)$$

通过使用网格搜索方法，我们可以找到最优窗宽 $h_{1opt}=\min_{h_1}\mathrm{CV}(\mathrm{h}_1)$。

对模型(16)式纵向数据部分线性分位数回归模型中参数向量 $\boldsymbol{\beta}$ 与非参数平滑函数 $g\left(t\right)$ 的设定如下：选择 $n=30,50,100$ 代表小样本、中等样本和大样本量，数据模拟次数 $M=100$。$p=5$，m_i 服从二项分布 $\mathrm{B}\left(10,0.8\right)$，参数向量 $\boldsymbol{\beta}=\left(-2,-1,1,3,0\right)^{\mathrm{T}}$，非参数平滑函数 $g\left(t\right)$ 分别取 $\sin\left(2\pi t\right)$，$\cos\left(5\pi t\right)$，e^{-2t}，$5t^2+1$，解释变量 $\boldsymbol{X}_{ij}$ 服从多元正态分布 $N\left(0,\boldsymbol{\Sigma}\right)$，其中，$\boldsymbol{\Sigma}=\left(\Sigma_{ij}\right)_{p\times p}$，$\Sigma_{i,j}=\rho^{|i-j|}$，$1\le i,j\le p$，$\rho=0.5$，时间变量 T_{ij} 服从 $[0,1]$ 上的均匀分布。在此设定下，根据模型(1.1)中误差项的分布，生成下述三种模拟数据：

模拟 1：误差项 $\boldsymbol{\varepsilon}_i \sim N\left(\mu_i, \boldsymbol{\Phi}_i\right)$。其中，$\mu_i$ 是一个 m_i 维向量，每个元素值都等于负标准正态分布的 τ 分位数；$\boldsymbol{\Phi}_i$ 为一阶自回归 $AR(1)$矩阵，相关系数 $\rho_{ij}=0.85$。

模拟 2：误差项 $\boldsymbol{\varepsilon}_i \sim t\left(v_i, \mu_i, \boldsymbol{\Phi}_i\right)$。其中，自由度 $v_i=3$；μ_i 是一个 m_i 维向量，每个元素值都等于负标准 t 分布的 τ 分位数；$\boldsymbol{\Phi}_i$ 与模拟 1 相同。

模拟 3：在模拟 1 生成机制的基础上，随机将样本数据中百分之五的响应变量用异常值进行替换，异常值为原值的 0.1 倍。

选取分位数 $\tau=0.5,0.75$，使用 R 软件中的 Mass、quantreg、sampling 等程序包可以实现上述模拟数据的生成。根据四种估计方法在三种不同数据模型下的样本外预测误差来展示估计方法的表现情况，样本外预测误差的计算公式如下：

$$\mathrm{FPE}\left(\hat{\boldsymbol{\mu}}_{\tau i}\right)=\frac{1}{M}\sum_{r=1}^{M}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\rho_\tau\left(Y_{ij}-\hat{\mu}_{\tau ij}\right)$$

其中 $\hat{\mu}_{\tau ij}$ 代表 $\mu_{\tau ij}=Q_\tau\left(Y_{ij}\mid \boldsymbol{X}_{ij}\right)$ 的估计量，数据集 $\left\{\left(\boldsymbol{X}_{ij}, Y_{ij}, T_{ij}\right), i=1,\cdots,n; j=1,\cdots,m_i\right\}$ 代表预测集。

3.2. 结果分析

利用 R 软件自行编写算法程序，将上述模拟数据带入程序中，得到模拟结果如下表 1~3 所示：

表 1 四种估计方法在模拟 1 下针对不同的非参数平滑函数 $g(t)$ 和样本容量 n 的预测误差。

Table 1. Out-of-sample forecast error under simulation 1
表 1. 模拟 1 下样本外预测误差

$g(t)$	n	FPE ($\tau=0.5$)				FPE ($\tau=0.75$)			
		MBQR	BQR	MLQR	LQR	MBQR	BQR	MLQR	LQR
$\sin(2\pi t)$	30	0.5224*	0.5339	0.5624	0.5609	0.4107*	0.4116	0.4296	0.4309
	50	0.6117*	0.6156	0.6357	0.6367	0.4381*	0.4396	0.4596	0.4609
	100	0.5503*	0.5606	0.5767	0.5745	0.5025*	0.5048	0.5320	0.5286
$\cos(5\pi t)$	30	0.5538*	0.5740	0.6143	0.6517	0.5194*	0.5206	0.5413	0.5478
	50	0.6325*	0.6381	0.6345	0.6389	0.4558*	0.4563	0.4723	0.4740
	100	0.5791*	0.5802	0.5873	0.5886	0.5071*	0.5092	0.5404	0.5399
e^{-2t}	30	0.6139*	0.6287	1.7130	1.7247	0.4798*	0.4854	2.2677	2.2773
	50	0.5781*	0.5830	1.5226	1.5236	0.4601*	0.4706	1.9695	1.9773
	100	0.6393*	0.6495	1.6368	1.6377	0.4876*	0.6011	2.0626	2.0652
$5t^2+1$	30	0.5392*	0.5531	1.4624	1.4663	0.4738*	0.5070	1.5009	1.5012
	50	0.7352*	0.8391	1.3237	1.3261	0.4095*	0.4202	1.5545	1.5597
	100	0.6546*	0.6649	1.3695	1.3665	0.4712*	0.4795	1.6653	1.6717

注：*表示四种估计方法中预测误差的最小值。

表 2 四种估计方法在模拟 2 下针对不同的非参数平滑函数 $g(t)$ 和样本容量 n 的预测误差。

Table 2. Out-of-sample forecast error under simulation 2
表 2. 模拟 2 下样本外预测误差

$g(t)$	n	FPE ($\tau=0.5$)				FPE ($\tau=0.75$)			
		MBQR	BQR	MLQR	LQR	MBQR	BQR	MLQR	LQR
$\sin(2\pi t)$	30	0.7480*	0.7554	0.7710	0.7727	0.5669*	0.5676	0.5933	0.6043
	50	0.9442*	0.9593	0.9768	0.9799	0.5751*	0.5771	0.5874	0.5885
	100	0.7355*	0.7378	0.7679	0.7698	0.6406*	0.6444	0.6583	0.6586
$\cos(5\pi t)$	30	0.6530*	0.6551	0.6670	0.6623	0.5927*	0.6087	0.6168	0.6309
	50	0.8775*	0.8783	0.8953	0.8940	0.5192*	0.5196	0.5222	0.5206
	100	0.8327*	0.8405	0.8426	0.8428	0.6976*	0.6999	0.7478	0.7469
e^{-2t}	30	0.9116*	0.9137	1.6703	1.6777	0.6090*	0.6145	1.9911	1.9982
	50	0.8357*	0.8557	1.7444	1.7472	0.6564*	0.6752	1.8838	1.8796
	100	0.8457*	0.8698	1.7498	1.7514	0.6737*	0.6823	2.1490	2.1453
$5t^2+1$	30	0.9263*	0.9531	1.6178	1.6295	0.6872*	0.7040	1.3731	1.3829
	50	0.7935*	0.8615	1.4238	1.4211	0.5417*	0.5717	1.7032	1.6960
	100	0.8614*	0.8684	1.5538	1.5530	0.5779*	0.5869	1.6428	1.6380

注：*表示四种估计方法中预测误差的最小值。

表 3 四种估计方法在模拟 3 下针对不同的非参数平滑函数 $g(t)$ 和样本容量 n 的预测误差。

Table 3. Out-of-sample forecast error under simulation 3
表 3. 模拟 3 下样本外预测误差

$g(t)$	n	FPE ($\tau=0.5$)				FPE ($\tau=0.75$)			
		MBQR	BQR	MLQR	LQR	MBQR	BQR	MLQR	LQR
$\sin(2\pi t)$	30	0.6307*	0.6359	0.6670	0.6673	0.5980*	0.6055	0.6090	0.6112
	50	0.6474*	0.6691	0.6782	0.6784	0.4589*	0.4643	0.4832	0.4842
	100	0.6332*	0.6547	0.6618	0.6622	0.4975*	0.4976	0.4985	0.4993
$\cos(5\pi t)$	30	0.6501*	0.6666	0.6722	0.6809	0.5288*	0.5355	0.5573	0.5742
	50	0.6249*	0.6260	0.6297	0.6294	0.6299*	0.6424	0.6788	0.6775
	100	0.6813*	0.7117	0.7213	0.7205	0.5386*	0.5391	0.5527	0.5526
e^{-2t}	30	0.6173*	0.6222	1.5898	1.5989	0.5783*	0.5868	1.8352	1.8387
	50	0.7603*	0.7629	1.7070	1.7090	0.4736*	0.5139	1.8773	1.8813
	100	0.7722*	0.7906	1.6655	1.6642	0.5410*	0.5606	2.0703	2.0668

续表

	30	0.6870*	0.7053	1.3539	1.3505	0.6128*	0.6255	1.8885	1.8715
$5t^2+1$	50	0.7890*	0.8474	1.4400	1.4387	0.4714*	0.4809	1.5493	1.5523
	100	0.6687*	0.6705	1.4034	1.4013	0.4693*	0.4763	1.5419	1.5421

注：*表示四种估计方法中预测误差的最小值。

分析上述模拟结果，以表 1 中 $\tau=0.5$，$n=100$ 时四种估计方法的样本外预测误差为例，可以发现在四种非参数光滑函数下，BQR 的样本外预测误差值优于 MLQR、LQR 的样本外预测误差值，说明部分线性回归模型分位数回归估计方法的预测精度优于线性回归模型下的分位数回归估计与分位数回归 Jackknife 模型评价估计方法。而 MBQR 的样本外预测误差值优于 BQR 的样本外预测误差值，说明了提出的纵向数据部分线性回归模型分位数回归模型平均估计方法在一定条件下是可以达到提高样本预测精度的目的。

分析表 1~3 的全部表格结果可以发现，三角函数、指数函数和幂函数三种不同类型的非参数函数在随机误差分布分别为正态分布、t 分布和数据中存在异常值的情况下的研究结果表明，本文提出的纵向数据部分线性回归模型分位数回归模型平均估计方法(MBQR)在数值模拟中的样本外预测误差都表现良好。

以模拟 1 中 $\tau=0.5$，$n=100$ 为例，绘制出了在四种非参数光滑函数下 MBQR 估计方法得出的子模型权重分布情况如图 1 所示：

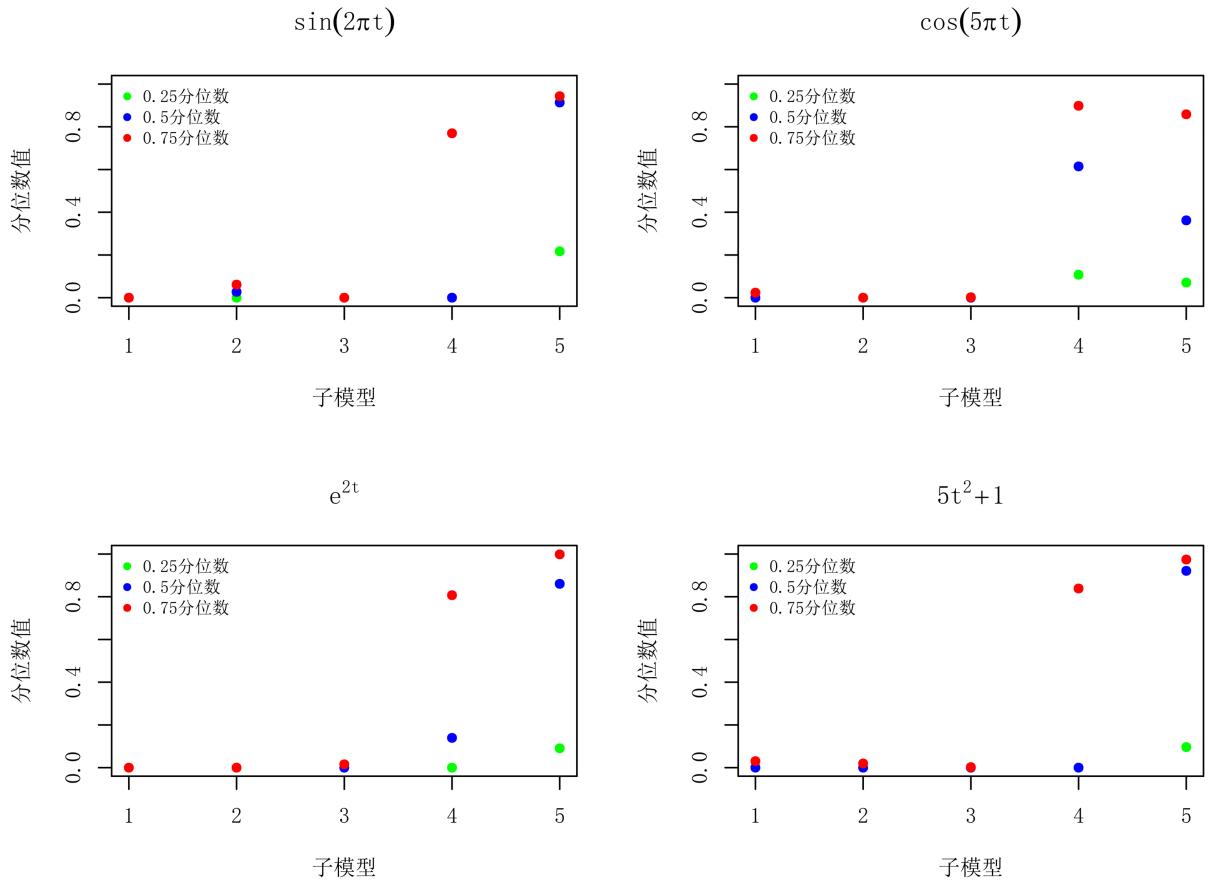

Figure 1. Weight distribution of sub-models under four non-parametric smooth functions
图 1. 四种非参数光滑函数下子模型的权重分布情况图

分析图 1 可以发现在四种非参数光滑函数下高权重主要分布在第 4 和第 5 个子模型中，其余子模型权重都较低，说明这两个子模型能够涵盖大部分模型信息，根据我们在数据生成过程中将参数向量设置为 $\boldsymbol{\beta}=(-2,-1,1,3,0)^{\mathrm{T}}$ 可以发现，MBQR 下得到的权重分布情况是合理且符合实际的。总体而言，本文提出的估计方法在有限样本情况下具有良好的效果，这说明本文估计方法是可行的。

4. 实例分析

随着我国经济的迅猛发展，工业化、城市化进程的加速，带来了大量的污染气体排放和资源消耗，空气质量问题成为城市居民日常生活中不可忽视的困扰。我们通过收集和分析空气质量监测数据，可以更准确地评估不同地区、不同时间段的空气质量状况，了解污染物的排放来源和影响范围。同时，利用空气质量指数(AQI)等综合性评价指标，预测分析空气质量问题的变化趋势，为相关政策制定和环境治理提供科学依据。因此，对空气质量数据的分析是解决空气污染问题的重要手段。

本节将本文提出的方法应用到来自“中国环境检测总站”的空气质量面板数据，数据选取了北京、上海等 50 个城市在 2020 年 1 月到 2020 年 12 月的空气质量月度数据，包含 600 组观测值以及七个变量。表 4 列出了空气质量数据集中的变量名称及其含义：

表 4 空气质量数据的变量信息。

Table 4. Variable information of instance data
表 4. 实例数据的变量信息

变量	名称	含义
$\boldsymbol{Y}$	AQI 指数	空气质量指数
$\boldsymbol{X}_1$	$PM_{2.5}$	空气中细颗粒物质量浓度
$\boldsymbol{X}_2$	PM_{10}	空气中可吸入颗粒物质量浓度
$\boldsymbol{X}_3$	SO_2	空气中二氧化硫质量浓度
$\boldsymbol{X}_4$	NO_2	空气中二氧化氮质量浓度
$\boldsymbol{X}_5$	CO	空气中一氧化碳质量浓度
$\boldsymbol{X}_6$	O_3	空气中臭氧质量浓度

通常一个地区的 AQI 空气质量指数是由当地一段时间内空气中 $PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、SO_2 质量浓度、NO_2 质量浓度、CO 质量浓度、O_3 质量浓度等六项参数通过计算而得出的空气污染程度及空气质量状况的表述。在本文的实证分析中，我们将 AQI 空气质量指数作为模型(1)中的响应变量 Y，将其余六项参数作为模型(1.1)中的协变量矩阵 X 或 T。为了确定模型(1)中参数部分的协变量矩阵 X 与非参数部分的协变量 T，分别作出 AQI 空气质量指数与其余六项参数的散点图，如图 2 所示。

分析图 2 可以发现，AQI 空气质量指数与 $PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、NO_2 质量浓度、CO 质量浓度、O_3 质量浓度有明显的线性关系，与 SO_2 质量浓度没有明显线性关系。因此，将 $PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、NO_2 质量浓度、CO 质量浓度、O_3 质量浓度作为参数部分的协变量 $\boldsymbol{X}$，SO_2 质量浓度作为非参数部分的协变量 $\boldsymbol{T}$。将 50 个城市每月的空气质量数据作为一个观测点，为了探究数据点之间的相关性，我们从 50 个城市中随机选取了 10 个城市绘制出来每个城市之间 AQI 空气质量指数的相关系数图如图 3 所示。

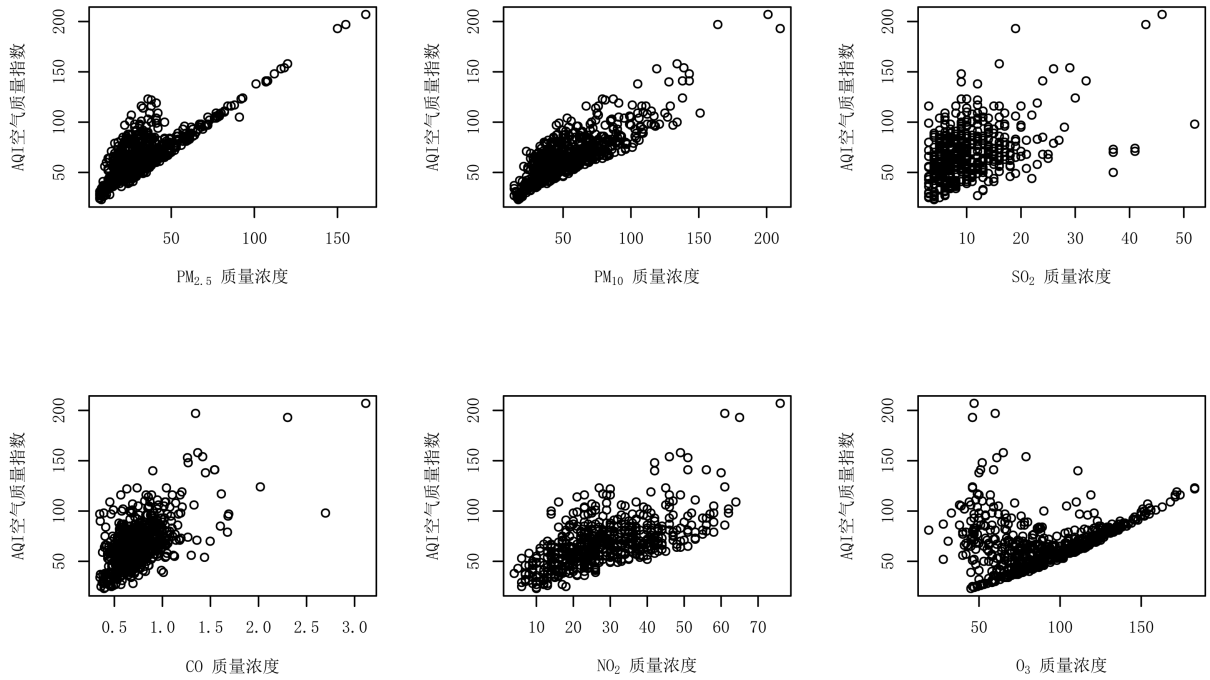

Figure 2. Scatter plot of AQI air quality data and other six parameters
图 2. AQI 空气质量数据与其余六项参数数据的散点图

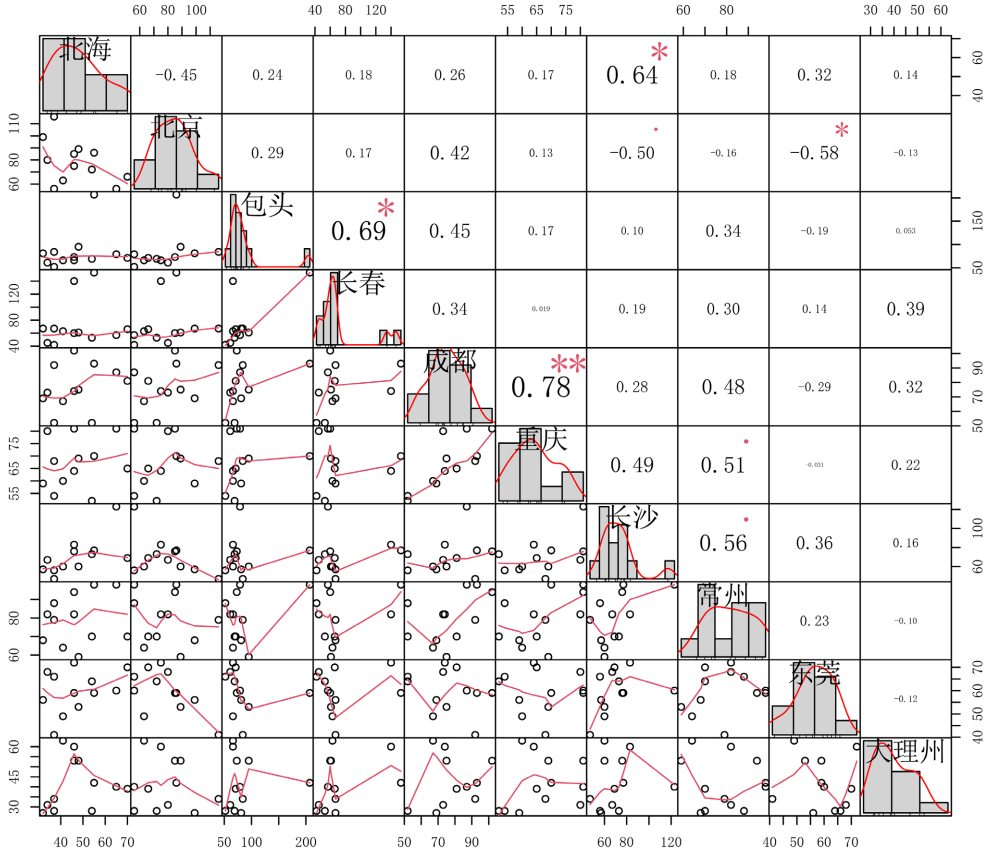

Figure 3. Correlation coefficient diagram of AQI air quality index between cities
图 3. 城市与城市之间 AQI 空气质量指数的相关系数图

分析图 3 可以发现大多数城市之间的 AQI 空气质量指数相关性并不高，极少一部分城市因为地域相近所以 AQI 空气质量指数之间存在相关性。由此，我们针对 50 个城市的空气质量数据假定 50 个城市单个城市月与月之间的数据具有相关性，城市与城市之间的数据相互独立，适用于本文提出的纵向数据部分线性回归模型的估计方法。

为了体现本文方法与其它常用方法在实际应用中的差异，将本文提出的纵向数据部分线性回归模型分位数回归模型平均估计(MBQR)与 Lu 和 Fan [14]提出的纵向数据线性回归模型的分位数回归估计(LQR)、基于 Lu 和 Fan [14]提出的纵向数据线性回归模型分位数回归估计的 Jackknife 模型平均估计(MLQR)、基于本文提出的子模型估计方法的纵向数据部分线性回归模型分位数回归估计(BQR)三种估计方法同时应用到空气质量数据的估计中。在空气质量数据线性回归模型的建模中，我们将 AQI 空气质量指数作为线性回归模型中的响应变量 $\boldsymbol{Y}$，其余六项参数作为线性回归模型中的协变量矩阵 $\boldsymbol{X}$。将 50 个城市的空气质量数据以 7:3 的比例选择城市分别作为训练集和测试集，进行 100 次随机选择，得到四种估计方法分别在分位数 $\tau=0.5,0.75$ 下的样本外预测误差如表 5 所示。

表 5 空气质量数据下四种不同估计方法的预测误差。

Table 5. Prediction error of example data
表 5. 实例数据的预测误差

估计方法	MBQR	BQR	MLQR	LQR
FPE ($\tau=0.5$)	2.7020*	2.7178	2.7434	2.7698
FPE ($\tau=0.75$)	2.2418*	2.2518	2.2823	2.3166

注：*表示四种估计方法中预测误差的最小值。

分析表 5 中的样本外预测误差可以看出，本文估计方法 MBQR 相较于其余三种估计方法在不同分位数情况下的预测误差都最小，说明本文提出的估计方法在实际数据的预测分析中也能表现良好。我们还绘制出了上述空气质量数据在 $\tau=0.5,0.75$ 时，MBQR 估计方法得出的子模型权重分布情况如图 4 所示：

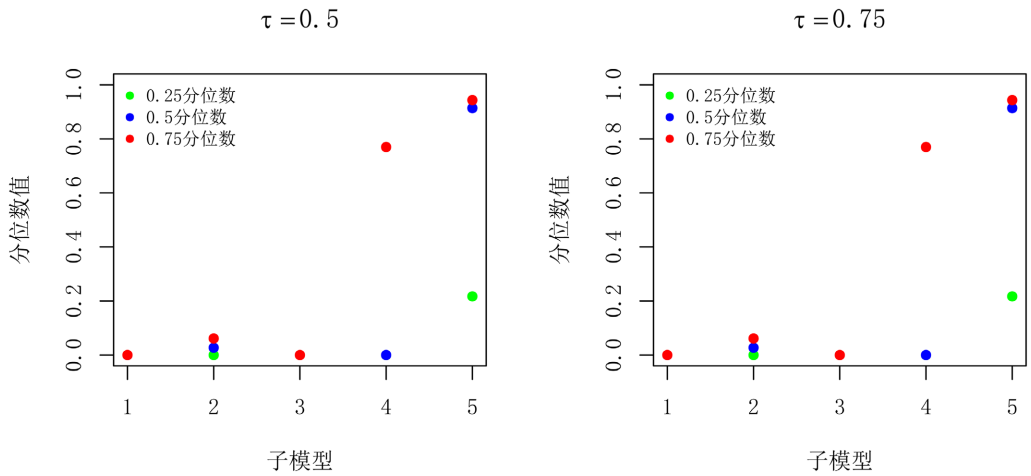

Figure 4. Weight distribution of sub-models under MBQR estimation method
图 4. MBQR 估计方法下子模型的权重分布情况图

分析图 4 可以发现，高权重主要分布在第 5 个子模型中，其余子模型权重都较低，说明第 5 子模型能够涵盖大部分模型信息，根据我们对空气质量数据七个指标的实际情况下分析 AQI 空气质量指数是由

当地一段时间内空气中 $PM_{2.5}$ 质量浓度、PM_{10} 质量浓度、SO_2 质量浓度、NO_2 质量浓度、CO 质量浓度、O_3 质量浓度等六项参数综合计算而得出的关于空气污染程度及空气质量状况的表述，MBQR 下得到的权重分布情况是合理且符合实际的。总体而言，本文提出的 MBQR 估计方法为实际数据中的纵向数据分析提供了一种更有效的新途径。

5. 结语

本文的创新点在于将模型平均方法与分位数回归估计方法相结合，提出了针对纵向数据部分线性回归模型的新估计方法，用工作相关矩阵分解和估计方程平滑法来处理纵向数据的组内相关性，用局部线性估计来处理部分线性回归模型中的非参数部分，给出了模型参数估计的 Newton-Raphson 迭代算法。并通过数值模拟和实例分析验证了该估计方法的优良性。

参考文献

[1] Zeger, S.L. and Diggle, P.J. (1994) Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters. Biometrics, **50**, 689-699. https://doi.org/10.2307/2532783

[2] Lin, X. and Carroll, R.J. (2001) Semiparametric Regression for Clustered Data Using Generalized Estimating Equations. *Journal of the American Statistical Association*, **96**, 1045-1056. https://doi.org/10.1198/016214501753208708

[3] Hu, Z. (2004) Profile-Kernel versus Backfitting in the Partially Linear Models for Longitudinal/Clustered Data. *Biometrika*, **91**, 251-262. https://doi.org/10.1093/biomet/91.2.251

[4] Fan, J. and Li, R. (2004) New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of the American Statistical Association*, **99**, 710-723. https://doi.org/10.1198/016214504000001060

[5] 田萍. 纵向数据半参数回归模型的估计理论[D]: [博士学位论文]. 北京: 北京工业大学, 2006.

[6] Tian, R. and Xue, L. (2017) Generalized Empirical Likelihood Inference in Partial Linear Regression Model for Longitudinal Data. *Statistics*, **51**, 988-1005. https://doi.org/10.1080/02331888.2017.1355370

[7] 王明辉, 尹居良. 纵向数据下部分线性模型的估计与性质[J]. 数理统计与管理, 2018, 37(5): 850-863.

[8] 刘会明. 纵向数据部分线性模型的有效估计[D]: [硕士学位论文]. 上海: 上海师范大学, 2020.

[9] Jung, S. (1996) Quasi-likelihood for Median Regression Models. *Journal of the American Statistical Association*, **91**, 251-257. https://doi.org/10.1080/01621459.1996.10476683

[10] Fu, L. and Wang, Y. (2012) Quantile Regression for Longitudinal Data with a Working Correlation Model. *Computational Statistics & Data Analysis*, **56**, 2526-2538. https://doi.org/10.1016/j.csda.2012.02.005

[11] Leng, C. and Zhang, W. (2012) Smoothing Combined Estimating Equations in Quantile Regression for Longitudinal Data. *Statistics and Computing*, **24**, 123-136. https://doi.org/10.1007/s11222-012-9358-0

[12] Lu, X. and Fan, Z. (2014) Weighted Quantile Regression for Longitudinal Data. *Computational Statistics*, **30**, 569-592. https://doi.org/10.1007/s00180-014-0550-x

[13] Lu, X. and Su, L. (2015) Jackknife Model Averaging for Quantile Regressions. *Journal of Econometrics*, **188**, 40-58. https://doi.org/10.1016/j.jeconom.2014.11.005

[14] Zhang, X. and Wang, W. (2019) Optimal Model Averaging Estimation for Partially Linear Models. *Statistica Sinica*, **29**, 693-718. https://doi.org/10.5705/ss.202015.0392

[15] Fang, F., Li, J. and Xia, X. (2022) Semiparametric Model Averaging Prediction for Dichotomous Response. *Journal of Econometrics*, **229**, 219-245. https://doi.org/10.1016/j.jeconom.2020.09.008

[16] 胡国治, 曾婕. 部分线性分位数回归模型的平均估计[J]. 安庆师范大学学报(自然科学版), 2023, 29(1): 32-36.

[17] Hu, G., Cheng, W. and Zeng, J. (2019) Focused Information Criterion and Model Averaging for Varying-Coefficient Partially Linear Models with Longitudinal Data. *Communications in Statistics—Simulation and Computation*, **50**, 2399-2417. https://doi.org/10.1080/03610918.2019.1609029

[18] Li, N., Fei, Y. and Zhang, X. (2024) Partial Linear Model Averaging Prediction for Longitudinal Data. *Journal of Systems Science and Complexity*, **37**, 863-885.

[19] Hendricks, W. and Koenker, R. (1992) Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity. *Journal of the American Statistical Association*, **87**, 58-68.

https://doi.org/10.1080/01621459.1992.10475175

[20] Lv, J., Guo, C. and Wu, J. (2018) Smoothed Empirical Likelihood Inference via the Modified Cholesky Decomposition for Quantile Varying Coefficient Models with Longitudinal Data. *TEST*, **28**, 999-1032. https://doi.org/10.1007/s11749-018-0616-0