Advances in Applied Mathematics 应用数学进展, 2022, 11(10), 7248-7256
Published Online October 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1110769

Hans汉斯

基于艾滋病数据的复合分位数回归分析

毛文杰，戴家佳*，秦前伟

贵州大学，数学与统计学院，贵州 贵阳

收稿日期：2022年9月19日；录用日期：2022年10月11日；发布日期：2022年10月21日

摘　要

HIV通过损害人体内的CD4细胞，减弱人的抵抗力而导致感染艾滋病。未感染的人每毫升血液大约含有1100个CD4细胞，所以可以通过测量患者的CD4细胞数对病情的好坏程度进行一定的评估。本文使用复合分位数回归方法对来自多中心艾滋病队列研究的数据进行分析，在响应变量与部分协变量同时缺失的情况下，我们提出了部分线性变系数模型的加权B样条复合分位数回归估计来描述整个时期CD4百分率变化的情况，并得到加权B样条复合分位数回归估计具有Horvitz-Thompson性质。同时，我们将所提出的估计方法与自适应LASSO惩罚方法相结合，得到患者的吸烟状态以及HIV感染时的年龄对HIV感染后患者的CD4百分率的影响不显著。

关键词

艾滋病，复合分位数，自适应LASSO，缺失值

Analysis of AIDS Data Based on Composite Quantile Regression

Wenjie Mao, Jiajia Dai*, Qianwei Qin

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Sep. 19th, 2022; accepted: Oct. 11th, 2022; published: Oct. 21st, 2022

Abstract

HIV causes AIDS by damaging the body's CD4 cells and weakening the body's resistance. An uninfected person has about 1100 CD4 cells per milliliter of blood, thus we can measure the CD4 cell count of patients to evaluate the degree of disease. In this paper, composite quantile regression was used to analyse the data from a multi-center AIDS cohort study, at the same time in the part response variables and covariate missing cases, we propose a weighted B-spline composite quan-

*通讯作者。

文章引用: 毛文杰, 戴家佳, 秦前伟. 基于艾滋病数据的复合分位数回归分析[J]. 应用数学进展, 2022, 11(10): 7248-7256. DOI: 10.12677/aam.2022.1110769

tile regression estimate of the partial linear variable coefficient model to describe the change of CD4 percentage over the whole period, and obtain that the weighted B-spline composite quantile regression estimate has the Horvitz-Thompson property. At the same time, we combined the proposed estimation method with the adaptive LASSO penalty method, and found that smoking status and age at HIV infection had no significant effect on CD4 percentage of HIV infected patients.

Keywords

AIDS, Composite Quantile Regression, Adaptive LASSO, Missing Value

1. 引言

在实际应用中由于常常遇到信息丢失，技术缺陷，预算限制或者被调查者拒绝回答某些问题等原因，往往使收集到的数据是带有缺失的。处理缺失数据，一个简单的方法是完整数据分析(complete case analysis, CC)方法。CC 方法也就是丢弃含有缺失数据的样本，仅利用可以完全观测的样本进行统计推断。CC 方法虽然简单，但其丢失掉了大量不完全样本中的信息，因而会损失估计效率。特别地，当协变量缺失时，如果选择概率与响应变量有关，CC 估计和真实的参数存在偏差，从而会误导人们的分析和判断。所以，对于缺失数据的研究是当前统计学的一个热点问题。

为了消除偏差，提高效率，Robins 等[1]提出了基于逆概率加权(inverse probability weighted, IPW)来调整 CC 估计。Horvitz 和 Thompson [2]提出逆概率加权估计具有 Horvitz-Thompson 性质，即使用估计的选择概率得到的逆概率加权估计与真实的选择概率得到的逆概率加权估计相比，前者的渐进方差小于后者的渐进方差。Tsiatis [3]考虑了对缺失协变量的线性模型的逆概率加权估计。此外，Liang [4]和 Wong 等[5]将逆概率加权扩展到具有缺失协变量的半参数模型。Xue 等[6]研究了协变量和响应变量同时随机缺失时，部分线性单指标模型的经验似然推断。Liu 等[7]基于经验似然异方差诊断技术，利用回归方法对缺失的响应变量进行插补，并引入经验似然方法来考虑具有完整数据集下模型的异方差性。

然而这些文献大多都是基于似然函数或者最小二乘的方法来进行分析的，当误差分布不服从正态分布，而是重尾分布或者含有极端异常值时，这些方法在估计精度上往往表现很差。众所周知，与普通的最小二乘回归相比，分位数回归的相对效率可以任意小。然而分位数估计只考虑了单个分位数点的损失，因此会降低估计的效率。为此，Zou 和 Yuan [8]提出了复合分位数回归(composite quantile regression, CQR)，无论误差分布如何，都总是有效的，与普通最小二乘回归相比，估计效率更好。到目前为止，在不同的模型下，学者们研究了复合分位数回归。例如，Kai 等[9]提出了基于复合分位数回归方法的部分线性变系数模型的有效估计。Sun 等[10]考虑了线性模型和变系数模型的数据驱动加权复合分位数回归估计。

直到 1983 年，人们对艾滋病感染的早期过程和该综合征的全谱都没有得到很好的了解。已完成的艾滋病流行病学调查仍主要包括对艾滋病和艾滋病相关疾病进行有限的横断面研究。若想进行深入的研究，需要对艾滋病早期病理进行大型调查以获得更加全面的数据。正是在这种背景下，多中心艾滋病队列研究在 1984~1991 年间对 283 位感染艾滋病病毒的患者每半年进行一次定期检查，记录他们的感染情况，从而获得了大量的数据。该数据集已被许多作者用于各种模型之中研究影响艾滋病病情的因素，例如，Huang 等[11]将该数据用于变系数模型，Fan 和 Li [12]用于半参数模型等。但是在他们的基础上我们还可

以进行一些补充和改进。一方面，他们都只局限于完整数据的情况下，对数据可能存在缺失的情况没有进行讨论。讨论数据是否缺失是很有必要的，因为它不仅可以充分考虑到现实生活中，数据尤其是生物医学数据可能存在缺失的情况。而且可以对比估计方法在完整数据和缺失数据情况下的估计效率；另一方面，通过观察他们的研究过程和结果可以发现，该数据中的部分变量可能存在回归参数是函数的情况。即部分回归参数是常数，部分回归参数是函数。由此看来，部分线性变系数模型较其他模型来说可能更适合用于分析该数据。综合以上几个原因，在本文中，我们将考虑数据存在缺失时，结合 B 样条近似和逆概率加权，提出了当部分协变量和响应变量同时随机缺失时，部分线性变系数模型的加权复合分位数回归估计。此外，我们通过将模型与自适应 LASSO 结合来识别模型中的重要参数，得到影响患者 CD4 浓度的重要因素。

2. 模型和估计方法

2.1. 部分线性变系数模型

变系数模型，作为线性模型的一种推广，已被广泛用于生物医学以及计量经济学等领域。近年来，人们试图平衡变系数模型的解释性和线性模型的灵活性，假设模型中的一些回归参数是常数，其余的回归参数是函数，于是提出了部分线性变化系数模型(Partially linear varying coefficient model, PLVCM)。部分线性变系数模型具有以下形式：

$$Y_i = X_i^{\mathrm{T}}\beta + Z_i^{\mathrm{T}}\alpha(U_i) + \varepsilon_i, i = 1,\cdots,n, \tag{1.1}$$

其中 Y_i 是响应变量，(X_i, U_i, Z_i) 是协变量，X_i 是已知的 p 维协变量，Z_i 是 q 维协变量。$\beta = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$ 是 p 维未知参数向量，$\alpha(\cdot) = (\alpha_1(\cdot), \alpha_1(\cdot), \cdots, \alpha_q(\cdot))^{\mathrm{T}}$ 是 q 维未知光滑系数向量，ε 是模型随机误差，满足 $E(\varepsilon_i) = 0$，$Var(\varepsilon_i) = \sigma^2$。此外，为了避免维数灾难，我们假设变量 U 的范围在一个非退化的紧区间上，不失一般性，假设它为单位区间 $U[0,1]$。

在过去的几十年里，有不少学者对部分线性变系数模型进行了相关研究。例如 Xiao 等[13]研究了模型具有非参数分量测量误差和缺失响应变量时，基于局部修正的轮廓最小二乘法，提出了参数向量和非参数函数的两种估计量。Jin 等[14]针对模型中协变量存在缺失时，基于逆概率加权和 B 样条逼近，提出了一种复合分位数回归方法。Jiang 等[15]、Jin 等[16]、Fan 等[17]等在文章中也对 PLVCM 进行了研究。

2.2. 逆概率加权复合分位数回归估计

本文假定数据缺失的机制为随机缺失，随机缺失是指数据缺失与否只依赖于可以观测到的变量，不依赖于缺失的变量本身。假设 $\{X_i, U_i, Z_i, Y_i\}_{i=1}^n$ 是来自于 $\{X, U, Z, Y\}$ 的独立同分布样本，考虑通过变换 $\{X_i^{\mathrm{T}}, U_i^{\mathrm{T}}, Z_i^{\mathrm{T}}, Y_i^{\mathrm{T}}\}^{\mathrm{T}}$ 的元素组成向量 $\{V_i^{\mathrm{T}}, \overline{V}_i^{\mathrm{T}}\}^{\mathrm{T}}$。其中 V_i^{T} 是对于所有的 i 都是可以观测到的 d 维非空向量 $(0 < d \le p+q)$，$\overline{V}_i^{\mathrm{T}}$ 表示对于某些 i 来说是缺失的向量。如果 $\overline{V}_i$ 所有的值都是可以观测到的，那么定义 $\delta_i = 1$，否则 $\delta_i = 0$。因此，随机缺失蕴含着选择概率 $P(\delta_i = 1 | X_i, U_i, Z_i, Y_i) = P(\delta_i = 1 | V_i) = \pi(V_i)$，为了方便表示，设 $\pi(V_i) = \pi_i$。本文假设 $\overline{V}_i$ 由 Y 和 X 的一个真子集所构成的，即响应变量与部分协变量同时缺失的情况。

当没有数据缺失时，可以通过最小化(1.2)式得到部分线性变系数模型的分位数估计：

$$\sum_{i=1}^{n} \rho_\tau \left(Y_i - X_i^{\mathrm{T}}\beta - Z_i^{\mathrm{T}}\alpha(U_i) \right). \tag{1.2}$$

为了增加分位数回归估计的效率，Zou 和 Yuan (2008)提出了复合分位数估计。目标损失函数的定义为：

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\rho_{\tau}\left(Y_i - X_i^{\mathrm{T}}\beta - Z_i^{\mathrm{T}}\alpha\left(U_i\right) - b_k\right), \tag{1.3}$$

其中 $\rho_{\tau_k}\left(x\right)=\tau_k xI\left(x\geq 0\right)-\left(1-\tau_k\right)xI\left(x<0\right),\tau_k=k/K+1$，$K$ 表示分位点的个数，b_k 为随机误差 ε 的 τ_k 分位点。

然而，当有数据缺失时，它们的方法不能直接应用。CC 方法是一种分析缺失数据的简单方法，即丢弃掉含有缺失数据的样本，仅利用可以完全观测的样本进行统计推断。CC 估计量被定义为：

$$\begin{aligned}&\left(\hat{b}_1^{\mathrm{CQR_{CC}}},\hat{b}_2^{\mathrm{CQR_{CC}}},\cdots,\hat{b}_K^{\mathrm{CQR_{CC}}},\hat{\beta}^{\mathrm{CQR_{CC}}}\right)\\&=\underset{b_1,b_2,\cdots,b_K,\beta}{\arg\min}\sum_{k=1}^{K}\sum_{i=1}^{n}\delta_i\rho_{\tau}\left(Y_i - X_i^{\mathrm{T}}\beta - Z_i^{\mathrm{T}}\alpha\left(U_i\right) - b_k\right).\end{aligned} \tag{1.4}$$

CC 方法虽然简单，但其丢失掉了大量不完全样本中的信息，因而会损失估计的效率。特别地，当协变量缺失时，如果选择概率与响应变量有关，CC 估计和真实的参数存在偏差，从而会误导人们的分析和判断。因此，这种方法往往不能直接使用。为了消除偏差，提高效率，我们采用逆概率加权的方法来处理缺失的数据。当选择概率 π_i 已知时，我们将逆概率加权复合分位数估计量定义为：

$$\begin{aligned}&\left(\hat{b}_1^{\mathrm{WCQR}_{\pi}},\hat{b}_2^{\mathrm{WCQR}_{\pi}},\cdots,\hat{b}_K^{\mathrm{WCQR}_{\pi}},\hat{\beta}^{\mathrm{WCQR}_{\pi}}\right)\\&=\underset{b_1,b_2,\cdots,b_K,\beta}{\arg\min}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\rho_{\tau_k}\left(Y_i - X_i^{\mathrm{T}}\beta - Z_i^{\mathrm{T}}\alpha\left(U_i\right) - b_k\right).\end{aligned} \tag{1.5}$$

现在用 B 样条逼近技术对函数系数部分进行近似表示，设 $z=\left(z_1,z_2,\cdots,z_k\right)$，$a<z_1<\cdots<z_k<b$ 为区间 $\left[a,b\right]$ 上的节点，$B\left(t\right)=\left(B_1\left(t\right),B_2\left(t\right),\cdots,B_N\left(t\right)\right)^{\mathrm{T}}$ 为 m 次 B 样条函数的基，$S\left(m,z\right)=\left\{B^{\mathrm{T}}\left(t\right)\gamma:\gamma\in R^N\right\}$ 为 m 次 B 样条函数空间，其中 $N=m+k_n+1$，k_n 为节点数。每个光滑系数函数 $\alpha_l\left(\cdot\right)$ 都可以用 B 样条函数 $S_l\left(u\right)\in S\left(h,z\right)$ 来逼近。使用 B 样条基函数来逼近每一个系数函数 $\alpha_l\left(\cdot\right)$，则 $\alpha_l\left(u\right)=B^{\mathrm{T}}\left(u\right)\gamma_l$，$\gamma_l=\left(\gamma_{l1},\gamma_{l2},\cdots,\gamma_{lN}\right)^{\mathrm{T}}$。所以当选择概率 π_i 已知时，考虑如下的目标函数：

$$\begin{aligned}&\left(\hat{b}_1^{\mathrm{WBCQR}_{\pi}},\hat{b}_2^{\mathrm{WBCQR}_{\pi}},\cdots,\hat{b}_K^{\mathrm{WBCQR}_{\pi}},\hat{\beta}^{\mathrm{WBCQR}_{\pi}}\right)\\&=\underset{b_1,b_2,\cdots,b_K,\beta}{\arg\min}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\rho_{\tau_k}\left(Y_i - X_i^{\mathrm{T}}\beta - D_i^{\mathrm{T}}\gamma - b_k\right).\end{aligned} \tag{1.6}$$

这里对系数函数部分进行了化简，其中 $Z_i=\left(Z_{i1},Z_{i2},\cdots,Z_{iq}\right)^{\mathrm{T}}$，$\gamma=\left(\gamma_1^{\mathrm{T}},\gamma_2^{\mathrm{T}},\cdots,\gamma_q^{\mathrm{T}}\right)^{\mathrm{T}}$，$D_i=\left(B^{\mathrm{T}}\left(U_i\right)Z_{i1},B^{\mathrm{T}}\left(U_i\right)Z_{i2},\cdots,B^{\mathrm{T}}\left(U_i\right)Z_{iq}\right)^{\mathrm{T}}$。因此，选择概率已知时的加权 B 样条复合分位数估计定义为(1.6)式。

然而，在实际情况下，选择概率函数通常是未知的，需要进行估计。为此，经常采用非参数平滑法。设 V 是一个 d 维向量。$\pi\left(v\right)$ 的 Nadariya-Watson 估计量可以定义为：

$$\hat{\pi}\left(v\right)=\frac{\sum_{i=1}^{n}\delta_i L_{h_0}\left(V_i-v\right)}{\sum_{i=1}^{n}L_{h_0}\left(V_i-v\right)},$$

其中 $L_{h_0}\left(\cdot\right)=L\left(\cdot/h_0\right)/h_0^d$，$L$ 是一个 d 维的核函数，h_0 是一个带宽。

但是，当 V 的维数较高时，一个完全非参数的估计可能会遇到维数灾难。在这种情况下，参数化方法可能更适用于估计 $\pi\left(\cdot\right)$。假设 $\pi\left(V\right)=\pi\left(V,\theta\right)$，$\pi\left(V,\theta\right)$ 的形式已知，且参数 θ 是有限维的。不失一般性，我们假设

$$\pi\left(V_i,\theta\right)=\frac{\exp\left(\theta_0+\theta_1Y_i+\theta_2U_i+\theta_3^{\mathrm{T}}X_i\right)}{1+\exp\left(\theta_0+\theta_1Y_i+\theta_2U_i+\theta_3^{\mathrm{T}}X_i\right)},$$

其中 $\theta=\left(\theta_0,\theta_1,\theta_2,\theta_3^{\mathrm{T}}\right)^{\mathrm{T}}$ 是一个未知的参数向量。一旦正确地指定了 $\pi\left(V_i,\theta\right)$ 的结构，我们就可以通过极大似然估计得到 θ 的一致估计量 $\hat{\theta}$。当参数估计 $\pi\left(V_i,\hat{\theta}\right)$ 可用时，加权 B 样条复合分位数估计量可以定义为

$$\begin{aligned}&\left(\hat{b}_1^{\mathrm{WBCQR}_{\hat{\pi}}},\hat{b}_2^{\mathrm{WBCQR}_{\hat{\pi}}},\cdots,\hat{b}_K^{\mathrm{WBCQR}_{\hat{\pi}}},\hat{\beta}^{\mathrm{WBCQR}_{\hat{\pi}}}\right)\\&=\underset{b_1,b_2,\cdots,b_K,\beta}{\arg\min}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\rho_{\tau_k}\left(Y_i-X_i^{\mathrm{T}}\beta-D_i^{\mathrm{T}}\gamma-b_k\right).\end{aligned}\tag{1.7}$$

接下来，我们考虑求解目标函数(1.7)，对于任意的 x，我们设 $x^+=xI\left(x\geq 0\right)$，$x^-=xI\left(x<0\right)$，所以 $x=x^+-x^-$，$\left|x\right|=x^++x^-$。我们引入松弛变量，$\beta^+=\left(\beta_1^+,\beta_2^+,\cdots,\beta_p^+\right)^{\mathrm{T}}$，$\beta^-=\left(\beta_1^-,\beta_2^-,\cdots,\beta_p^-\right)^{\mathrm{T}}$，$\alpha^+=\left(\left(\alpha_1^+\right)^{\mathrm{T}},\left(\alpha_2^+\right)^{\mathrm{T}},\cdots,\left(\alpha_q^+\right)^{\mathrm{T}}\right)^{\mathrm{T}}$，$\alpha^-=\left(\left(\alpha_1^-\right)^{\mathrm{T}},\left(\alpha_2^-\right)^{\mathrm{T}},\cdots,\left(\alpha_q^-\right)^{\mathrm{T}}\right)^{\mathrm{T}}$，$\alpha_l^+=\left(\alpha_{l1}^+,\alpha_{l2}^+,\cdots,\alpha_{lN}^+\right)^{\mathrm{T}}$，$\alpha_l^-=\left(\alpha_{l1}^-,\alpha_{l2}^-,\cdots,\alpha_{lN}^-\right)^{\mathrm{T}}$。设 $y_i-b_k-\left(X_i^{\mathrm{T}}\beta+\gamma^{\mathrm{T}}D_i\right)=\xi_{ik}^+-\xi_{ik}^-$，这里 $\xi_{ik}^+=\max\left(0,y_i-b_k-\left(X_i^{\mathrm{T}}\beta+D_i^{\mathrm{T}}\gamma\right)\right)$，$\xi_{ik}^-=\max\left(0,-\left(y_i-b_k-\left(X_i^{\mathrm{T}}\beta+D_i^{\mathrm{T}}\gamma\right)\right)\right)$。所以 1.7 式能被重写为如下线性规划问题：

$$\underset{\beta^+,\beta^-,b_k^+,b_k^-,\alpha^+,\alpha^-}{\arg\min}=\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\left(\tau_k\xi_{ik}^++\left(1-\tau_k\right)\xi_{ik}^-\right),\tag{1.8}$$

其中 $\beta^+\geq 0$，$\beta^-\geq 0$，$b_k^+\geq 0$，$b_k^-\geq 0$，$\alpha_{lj}^+\geq 0$，$\alpha_{lj}^-\geq 0$，

$$\begin{aligned}\xi_{ik}^+-\xi_{ik}^-=&Y_i-\left(b_k^+-b_k^-\right)-X_{i1}\left(\beta_1^+-\beta_1^-\right)-\cdots-X_{ip}\left(\beta_p^+-\beta_p^-\right)\\&-\dot{\pi}\left(U_i\right)\left(\alpha_1^+-\alpha_1^-\right)Z_{i1}-\cdots-\dot{\pi}\left(U_i\right)\left(\alpha_q^+-\alpha_q^-\right)Z_{iq}\end{aligned}。$$

为了调整 k_n，可以应用许多选择标准，如交叉验证、广义交叉验证、AIC 信息准则和 BIC 信息准则。对于 B 样条近似，已经有学者指出信息准则要优于交叉验证和广义交叉验证。而 AIC 信息准则通常会出现过拟合，因此，我们选取 BIC 信息准则

$$\mathrm{BIC}\left(k_n\right)=\log\left(\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\rho_{\tau_k}\left(Y_i-X_i^{\mathrm{T}}\hat{\beta}^{\mathrm{WBCQR}_{\hat{\pi}}}-\hat{\gamma}^{\mathrm{WBCQR}_{\hat{\pi}}}D_i-\hat{b}_k^{\mathrm{WBCQR}_{\hat{\pi}}}\right)\right)+\frac{\log\left(n\right)}{n}\times df_{k_n},\tag{1.9}$$

其中 df_{k_n} 是模型中待估参数的个数，n 是样本量。我们可以通过最小化(1.9)来得到适当的结点数(k_n)。

2.3. 变量选择方法

在处理实际问题时，人们通常是根据自身的经验将各种可能与响应变量有关的解释变量引入回归模型。这样往往会把一些与响应变量关系很小或者没有关系的解释变量也纳入回归模型，从而降低了估计的精度。因此，选择模型中的重要变量就成为了统计学的重要研究内容。变量选择方法可分为两类：传统的变量选择方法与惩罚函数方法。

传统的变量选择方法是指对协变量组成集合的所有子集进行比较，通过 AIC 和 BIC 等准则选出一个最优的子集来拟合回归模型。这类方法在理论上难以解释，当协变量维数较高时，计算量相当大，在实际中难于实现。近年来，通过惩罚函数的方法进行变量选择越来越受到重视。这是由于惩罚的方法可以对模型的参数进行估计，同时把较小的系数自动压缩为 0，因此在理论上具有较强的可解释性且减轻了计算的负担。

自适应 LASSO 可以看作是 LASSO 惩罚的一种拓展。事实上，这个想法是通过使用自适应权值来惩罚不同的系数。所以，模型(1.1)的自适应 LASSO 惩罚加权 B 样条复合分位数回归估计量(PWBCQR)，用 $\hat{\beta}^{\mathrm{PWBCQR}_{\pi}}$ 表示，可以通过求解 3.1 式得到

$$\arg\min = \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i}\rho_{\tau_k}\left(Y_i - X_i^{\mathrm{T}}\beta - \gamma D_i - b_k\right) + \lambda_\pi \sum_{j=1}^{p}\frac{\left|\beta_j\right|}{\left|\hat{\beta}_j^{\mathrm{WBCQR}_\pi}\right|^2}. \tag{3.1}$$

$\hat{\beta}^{\mathrm{PWBCQR}_{\hat{\pi}}}$ 通过求解 3.2 式可得

$$\arg\min = \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\rho_{\tau_k}\left(Y_i - X_i^{\mathrm{T}}\beta - \gamma D_i - b_k\right) + \lambda_{\hat{\pi}} \sum_{j=1}^{p}\frac{\left|\beta_j\right|}{\left|\hat{\beta}_j^{\mathrm{WBCQR}_{\hat{\pi}}}\right|^2}. \tag{3.2}$$

对于调优参数 $\lambda_{\hat{\pi}}$ 的选择，有学者指出，在样本量趋于无穷大时，广义交叉验证方法往往会产生过拟合的模型。因此，他们提出了一个 BIC 型选择标准，这使得我们考虑如下的 BIC 准则：

$$\mathrm{BIC}\left(\lambda_{\hat{\pi}}\right) = \log\left(\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}\rho_{\tau_k}\left(Y_i - X_i^{\mathrm{T}}\hat{\beta}^{\mathrm{PWBCQR}_{\hat{\pi}}} - \hat{\gamma}^{\mathrm{PWBCQR}_{\hat{\pi}}} D_i - \hat{b}_k^{\mathrm{PWBCQR}_{\hat{\pi}}}\right)\right) + \frac{\log(n)}{n} \times df_{\lambda_{\hat{\pi}}}, \tag{3.3}$$

其中 $df_{\lambda_{\hat{\pi}}}$ 是模型中非零参数的个数，n 是样本量。我们可以通过最小化(3.3)式得到最优的惩罚参数 $\lambda_{\hat{\pi}}$。

3. 数据分析

我们现在通过分析来自多中心艾滋病队列研究的数据集来说明本文中提出的 PWBCQR 和 WBCQR 估计方法。该数据集包含有 1984~1991 年跟踪调查期间 283 个感染 HIV 的同性恋男人的 HIV 情况。HIV 通过损害人体内的 CD4 细胞减弱人的抵抗力而导致艾滋病，未感染的人每毫升血液大约含有 1100 个 CD4 细胞。所以，通过观察患者的 CD4 细胞数可以对病情进行一定的评估。在本文中，我们尝试在有数据缺失的情况下，描述患病整个时期 CD4 百分率的变化情况，并评估吸烟状态、年龄和感染前 CD4 百分率对患者现在 CD4 百分率的影响。假设 Y 为患者现在的 CD4 百分率，X_1 为 HIV 感染前中心化的 CD4 百分率，X_2 为患者的抽烟状态，根据患者是否抽烟分别取值为 1 和 0，X_3 为 HIV 感染时患者中心化的年龄。为了演示和简单起见，我们省略了其他可用协变量的可能影响。然后，我们考虑以下模型：

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \alpha_0(t)Z + \varepsilon,$$

其中 $Z \equiv 1$，$\alpha_0(t)$ 为 CD4 百分率基准函数。

我们假设协变量 X_2、X_3 与响应变量 Y 是同时缺失的，通过使用选择概率 $p = P(\delta = 1 \mid X_1 = x_1, Z = z, T = t) = \dfrac{1}{1 + 0.4\exp(x_1 - z - t + 1.65)}$ 随机删除约 25%的观察值。由于 $\delta = 1$ 是由伯努力分布 $b(1, p)$ 随机生成的，所以我们从 200 次模拟运行中计算出最终的结果。在这里我们记 $\mathrm{BCQR5}_{\mathrm{CC}}$、$\mathrm{WBCQR5}_{\pi}$、$\mathrm{WBCQR5}_{\hat{\pi}}$ 分别表示分位点个数为 5 时的 B 样条复合分位数回归。表 1 总结了 $\mathrm{BCQR5}_{\mathrm{CC}}$、$\mathrm{WBCQR5}_{\pi}$ 和 $\mathrm{WBCQR5}_{\hat{\pi}}$ 方法的估计系数和标准差。在没有缺失数据的情况下，我们可以得到 BCQR 估计量 $\hat{\beta}^{\mathrm{BCQR5}} = (0.40611, -0.00102, -0.00422)$ 和 PBCQR5 估计量 $\hat{\beta}^{\mathrm{PBCQR5}} = (0.3177, 0, 0)$。在这里，为了比较我们所提出的变量选择方法的性能，我们使用 PBCQR5 估计量作为黄金标准，计算平均绝对误差(mean absolute error, MAE)，$\mathrm{MAE} = \dfrac{\sum_{i=1}^{200}\sum_{j=1}^{3}\left|\hat{\beta}_{ij} - \hat{\beta}_{ij}^{\mathrm{PBCQR5}}\right|}{200}$，变量选择的结果见表 2。

表 1 显示 $\mathrm{WBCQR5}_{\hat{\pi}}$ 方法下 β_1、β_2、β_3 的标准差小于 $\mathrm{WBCQR5}_{\pi}$ 方法下 β_1、β_2、β_3 的标准差，即使用估计的选择概率得到的逆概率加权估计值与真实的选择概率得到的逆概率加权估计值相比，前者的渐进方差小于后者的渐进方差。换一句话说，逆概率加权估计具有 Horvitz-Thompson 性质。

从估计值的大小来看，$\mathrm{WBCQR5}_{\hat{\pi}}$ 和 $\mathrm{WBCQR5}_{\pi}$ 的估计值与 $\mathrm{BCQR5}_{\mathrm{CC}}$ 相比，逆概率加权方法得到的估计值比 CC 方法得到的估计值更接近于完整数据时的估计值。说明逆概率加权方法得到的估计结果更符合真实的情况，且在估计精度上是高于 CC 方法的。

Table 1. The coefficients estimates and sample standard deviations (in parentheses)
表 1. 系数估计值和样本标准差(括号内)

方法	β_1	β_2	β_3
BCQR5	0.40611	−0.00102	0.00422
$BCQR5_{CC}$	0.36624 (0.02123)	−0.00154 (0.00023)	0.00864 (0.00341)
$WBCQR5_{\pi}$	0.40356 (0.03105)	−0.00109 (0.00024)	0.00375 (0.00485)
$WBCQR5_{\hat{\pi}}$	0.40443 (0.02901)	−0.00105 (0.00022)	0.00405 (0.00443)

Table 2. Variable selection results
表 2. 变量选择结果

方法	选择频率			MAE
	β_1	β_2	β_3	
$PBCQR5_{CC}$	1.000	0.21	0.22	0.03765
$PWBCQR5_{\pi}$	1.000	0.033	0.034	0.03422
$PWBCQR5_{\hat{\pi}}$	1.000	0.022	0.022	0.03354

从表 2 中，可以看到各回归系数的选择情况为：$PBCQR5_{CC}$、$PWBCQR5_{\pi}$、$PWBCQR5_{\hat{\pi}}$ 方法对 β_2 和 β_3 的选择频率都较低，其中 $PWBCQR5_{\hat{\pi}}$ 方法得到的选择频率最低，仅有 0.022；$PWBCQR5_{\pi}$ 和 $PWBCQR5_{\hat{\pi}}$ 方法得到的结果与 $PBCQR5_{CC}$ 方法相比，后者的大小是前者的 6 到 9 倍，说明逆概率加权方法对于 β_2 和 β_3 的选择频率是远低于 CC 方法的。结果表明年龄和吸烟状态对患者现在的 CD4 百分率没有显著影响。而 $PBCQR5_{CC}$、$PWBCQR5_{\pi}$、$PWBCQR5_{\hat{\pi}}$ 方法对 β_1 的选择频率都等于 1，说明变量 X_1 (HIV 感染前中心化的 CD4 百分率)对该模型具有显著性影响，且从表 1 可以得到 X_1 与 Y (患者现在的 CD4 百分率)呈正相关的关系。

Huang 等[11]在 2002 年发表的文献中指出，在显著性水平 0.05 上，吸烟状态和年龄对患者 CD4 百分率均无显著影响。通过比较我们与 Huang 等得到的结论来看，可以发现，我们的结论和 Huang 等得到的结论是一致的。由于逆概率加权方法对于 β_2 和 β_3 的选择频率是远低于 CC 方法的，所以我们还可以得到逆概率加权方法是优于 CC 方法的，逆概率加权方法得到的结果更有说服力。

另一方面，从平均绝对误差来看，$PWBCQR5_{\hat{\pi}}$ 的平均绝对误差是最小的，说明 $PWBCQR5_{\hat{\pi}}$ 方法与 $PBCQR5_{CC}$ 和 $PWBCQR5_{\pi}$ 两种方法相比有更高的精确度，得到的结果与完整数据时的结果最为接近。

综上，$WBCQR5_{\hat{\pi}}$ 和 $PWBCQR5_{\hat{\pi}}$ 的表现分别优于 $WBCQR5_{\pi}$ 和 $PWBCQR5_{\pi}$。

4. 结论

在本文中，对于多中心艾滋病队列研究数据，我们在前面学者研究的基础上，将其应用拓展到部分线性变系数模型中，并用 B 样条复合分位数回归方法对相关参数进行估计；同时，我们对数据存在响应变量与部分协变量同时缺失时的情况也进行了考虑。

首先，当响应变量和部分协变量同时随机缺失时，我们提出了部分线性变系数模型的加权 B 样条复合分位数回归(WBCQR)估计。结果显示，我们提出的逆概率加权估计量具有 Horvitz-Thompson 性质；在对方法的比较中，逆概率加权方法的估计结果比 CC 方法更符合真实的情况，在估计精度上是高于 CC 方法的。

其次，我们将所提出的估计方法与自适应 LASSO 惩罚方法相结合，研究了变量的选择过程。结果表

明年龄和吸烟状态对患者现在的 CD4 百分率没有显著影响，这与其他学者得到的结论一致。同时得到 HIV 感染前中心化的 CD4 百分率对该模型具有显著性影响，且与患者现在的 CD4 百分率呈正相关的关系。

最后，综上可得，用复合分位数回归方法对该数据进行分析是可行的，得到的结果是可信的。而且当数据存在缺失时，用复合分位数回归方法对其进行估计，得到的估计值与完整数据时的估计值很接近，说明复合分位数回归方法得到的结果是稳健的。

基金项目

本文获得贵州省数据驱动建模学习与优化创新团队项目(黔科合平台人才[2020] 5016)资助。

参考文献

[1] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994) Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American statistical Association*, **89**, 846-866. https://doi.org/10.1080/01621459.1994.10476818

[2] Horvitz, D.G. and Thompson, D.J. (1952) A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685. https://doi.org/10.1080/01621459.1952.10483446

[3] Tsiatis, A.A. (2006) Semiparametric Theory and Missing Data. Springer, New York. https://doi.org/10.1007/0-387-37345-4

[4] Liang, H. (2008) Generalized Partially Linear Models with Missing Covariates. *Journal of Multivariate Analysis*, **99**, 880-895. https://doi.org/10.1016/j.jmva.2007.05.004

[5] Wong, H., Guo, S., Chen, M. and IP, W.-C. (2009) On Locally Weighted Estimation and Hypothesis Testing of Varying-Coefficient Models with Missing Covariates. *Journal of Statistical Planning and Inference*, **139**, 2933-2951. https://doi.org/10.1016/j.jspi.2009.01.016

[6] Xue, L. and Zhang, J. (2020) Empirical Likelihood for Partially Linear Single-Index Models with Missing Observations. *Computational Statistics & Data Analysis*, **144**, Article ID: 106877. https://doi.org/10.1016/j.csda.2019.106877

[7] Liu, F., Gao, W., He, J., Fu, X. and Kang, X. (2021) Empirical Likelihood Based Diagnostics for Heteroscedasticity in Semiparametric Varying-Coefficient Partially Linear Models with Missing Responses. *Journal of Systems Science and Complexity*, **34**, 1175-1188. https://doi.org/10.1007/s11424-020-9240-7

[8] Zou, H. and Yuan, M. (2008) Composite Quantile Regression and the Oracle Model Selection Theory. *The Annals of Statistics*, **36**, 1108-1126. https://doi.org/10.1214/07-AOS507

[9] Kai, B., Li, R. and Zou, H. (2011) New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *Annals of Statistics*, **39**, 305-332. https://doi.org/10.1214/10-AOS842

[10] Sun, J., Gai, Y. and Lin, L. (2013) Weighted Local Linear Composite Quantile Estimation for the Case of General Error Distributions. *Journal of Statistical Planning and Inference*, **143**, 1049-1063. https://doi.org/10.1016/j.jspi.2013.01.002

[11] Huang, J.Z., Wu, C.O. and Zhou, L. (2002) Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements. *Biometrika*, **89**, 111-128. https://doi.org/10.1093/biomet/89.1.111

[12] 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

[13] Xiao, Y.T. and Li, F.X. (2020) Estimation in Partially Linear Varying-Coefficient Errors-in-Variables Models with Missing Response Variables. *Computational Statistics*, **35**, 1637-1658. https://doi.org/10.1007/s00180-020-00967-3

[14] Jin, J., Ma, T., Dai, J. and Liu, S.Z. (2021) Penalized Weighted Composite Quantile Regression for Partially Linear Varying Coefficient Models with Missing Covariates. *Computational Statistics*, **36**, 541-575. https://doi.org/10.1007/s00180-020-01012-z

[15] Jiang, R., Qian, W.M. and Zhou, Z.G. (2018) Weighted Composite Quantile Regression for Partially Linear Varying Coefficient Models. *Communications in Statistics-Theory and Methods*, **47**, 3987-4005. https://doi.org/10.1080/03610926.2017.1366522

[16] Jin, J., Hao, C. and Ma, T. (2019) B-Spline Estimation for Partially Linear Varying Coefficient Composite Quantile Regression Models. *Communications in Statistics-Theory and Methods*, **48**, 5322-5335.

https://doi.org/10.1080/03610926.2018.1510006

[17] Fan, Y., Härdle, W.K., Wang, W. and Zhu, L.X. (2018) Single-Index-Based CoVaR with Very High-Dimensional Covariates. *Journal of Business & Economic Statistics*, **36**, 212-226. https://doi.org/10.1080/07350015.2016.1180990