doi:10. 16055/j. issn. 1672-058X. 2022. 0002. 007

响应变量缺失下变系数模型的分位数回归

叶　瑶, 袁德美
(重庆工商大学 数学与统计学院, 重庆 400067)

摘　要:针对响应变量随机缺失情况下变系数分位数回归模型的非参数估计问题,提出了将 B 样条和逆概率加权相结合的估计方法。缺失数据在统计工作中难免会遇到,首先用 logistic 模型产生响应变量的缺失概率,然后对变系数模型的系数函数采用 B 样条逼近技术,利用缺失概率构建逆概率加权分位数回归的损失函数,得到模型的未知系数函数估计;在模拟研究中,将得到的估计与直接使用完全数据的估计方法进行对比,发现在响应变量随机缺失下,将 B 样条和逆概率加权相结合的变系数模型分位数回归在有限样本情况下表现良好,模拟研究结果表明该方法有效;最后将所研究的方法运用到挪威公共道路管理局收集的奥斯陆地区相关数据中,研究了空气中二氧化氮浓度与道路车流量和风速之间的关系,得出合理的结论,进一步证明了该方法的合理性。

关键词:响应变量缺失;B 样条;逆概率加权;分位数回归
中图分类号:O212. 7　　**文献标志码**:A　　**文章编号**:1672-058X(2022)02-0046-07

0　引　言

变系数模型最早由 Hastie 等[1]提出,是重要的非参数模型,可以看作是非参数模型和半参数模型的推广,该模型具有稳健性,且模型假设少,具有更强的适应性。它能很好地分析回归系数随着时间等诸因素的变化。假设 Y 为响应变量, $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}\in R^p$ 为 p 维协变量,则变系数模型有如下形式:

$$Y=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}(U)+\varepsilon \tag{1}$$

其中,U 为光滑变量,$\boldsymbol{\beta}(\cdot)=(\beta_1(\cdot),\beta_2(\cdot),\cdots,\beta_p(\cdot))^{\mathrm{T}}\in R^p$ 为未知的关于 U 的函数向量,模型误差 ε 与$(\boldsymbol{X},U)$ 相互独立。

在统计工作中,由于调查人员的疏忽,需要的信息暂时还没办法获得,获取成本很高,因此难免会遇到缺失数据。Rubin 对缺失机制的问题构造了 3 种概念:完全随机缺失(MCAR)指是数据的缺失与完全观测数据和不完全观测数据都无关;随机缺失(MAR)是数据缺失与完全观测数据有关,与不完全观测数据无关;不完全缺失(MNAR)是指缺失数据与不完全观测数据本身有关。近几十年,对于缺失数据下变系数模型的研究也有很大的进展。李志强等[2]采用最小二乘方法,分别用拟似然借补估计和加权拟似然借补估计来处理缺失数据,发现两种方法估计的精度差别不大;Cai 等[3]用局部多项式估计方法估计系数函数,并推导出系数函数的标准误差公式,提出一种基于非参数最大似然比的拟合优度检验技术。这些研究都是基于最小二乘方法展开的,最小二乘方法计算简单,但前提条件比较苛刻,

收稿日期:2021-03-05;**修回日期**:2021-05-18.
基金项目:重庆工商大学数理统计团队(ZDPTTD201906).
作者简介:叶瑶(1996—),女,重庆长寿人,硕士研究生,从事非参数统计研究.
通讯作者:袁德美(1966—),男,四川南部人,教授,从事概率论极限理论、统计抽样分布渐近理论研究. Email:yuandemei@163. com.

对离群点敏感。为了提高估计的效率,Koenker 等[4]提出了分位数回归模型,它是根据被解释变量的条件分位数对解释变量进行回归,不需要对误差项的分布作假设,适应性更强;Honda[5]将系数函数用局部多项式方法展开,研究了变系数分位数回归模型的参数估计问题,证明了其渐进性质;Guo 等[6]用局部复合分位数方法研究了误差项带有异方差的变系数模型,提出的新方法可以对系数函数和异方差同时进行估计。

对于缺失数据的处理,一个简单的方法是直接用完全数据进行估计,但是在数据存在随机缺失的情况下,这样做会使得偏差很大。学者们提出了许多关于缺失数据处理的方法,比如逆概率加权、回归借补法、多重插补以及增强的逆概率加权方法等,为处理缺失数据提供很多的参考和帮助。Horvitz 等[7]在 1952 年提出了逆概率加权方法;Tan 和 Sun 等[8-9]用此方法研究了当协变量存在缺失时,变系数复合分位数回归模型的参数估计问题;此外,Sun 等[9]对比了分别用非参数和参数方法估计缺失概率时的差别。当影响因素对缺失机制的影响不清楚时,学者们大多会选择非参数估计,但非参数估计可能会涉及"维数灾祸"的问题,参数估计就成为一个很好的选择。比较常用的参数估计模型是 logistics 模型,其形式为 $\text{logit}\,(\pi(U,\boldsymbol{X}))=\alpha_0+\alpha_1 U+\alpha_2^{\mathrm{T}}\boldsymbol{X}$,其中 $\text{logit}(u)=\log\left(\frac{u}{1-u}\right)$,$\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\boldsymbol{\alpha}_2^{\mathrm{T}})$ 为未知参数向量。

除了用局部多项式近似未知的系数函数,还可以用样条估计方法近似逼近系数函数。1946 年,Schoenberg 提出了 B 样条估计方法,Zhao 等[10]用 B 样条方法逼近半参数变系数部分线性模型的系数函数部分,结合 SCAD 惩罚方法,研究了模型的变量选择问题;Du 等[11]研究了 B 样条近似的变系数部分线性模型的变量选择,与 Zhao 等[10]不同的是他们选择了分位数回归方法,在选择重要变量的同时又估计了未知系数;Jin 等[12]将 B 样条和逆概率加权相结合,考虑了当协变量缺失时,变系数部分线性模型的复合分位数回归,并结合了自适应 Lasso 方法,研究了变量选择问题。

相较于均值回归模型,分位数回归模型不仅可以避免均值回归因异常值或极端值对于统计推断的影响,而且挖掘出的信息更加丰富,获取的数据对响应变量的描述更完整。对于缺失数据下的变系数模型,使用局部多项式近似系数函数的研究已经有很多,但该方法存在以下问题:第一,计算速度比较慢;第二,涉及窗宽的选择,且对窗宽很敏感;第三,涉及拟合多项式阶数的选取。所以这里选择 B 样条近似逼近系数函数。B 样条可以将式(1)转换为样条基的线性组合,操作起来方便,且只需要简单节点数的选择,对节点数目不太敏感。对于基于 B 样条的回归问题,鲜有学者将其与缺失数据结合起来,而缺失数据在统计工作中难免会遇到。因此,本文针对响应变量随机缺失,研究了变系数分位数回归模型的参数估计问题。首先用 logistic 模型产生响应变量的缺失概率,然后对变系数模型的系数函数采用 B 样条逼近技术,利用缺失概率构建逆概率加权分位数回归损失函数,得到模型的未知系数函数估计,通过模拟研究证明所研究方法的有效性,最后通过实例分析进一步证明方法的合理性。

1　变系数分位数回归的 IPW 估计

假设 $\{(Y_i,\boldsymbol{X}_i,U_i,\delta_i)\}_{i=1}^n$ 是来自式(1)的独立样本,即

$$Y_i=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}(U_i)+\varepsilon_i,i=1,2,\cdots,n, \tag{2}$$

其中,$\boldsymbol{\beta}(\cdot)=(\beta_1(\cdot),\beta_2(\cdot),\cdots,\beta_p(\cdot))^{\mathrm{T}}\in R^p$ 为系数函数;$\boldsymbol{X}_i\in R^p$,$U_i\in\mathbf{R}$,$(\boldsymbol{X}_i,U_i)$ 可以完全观测到;Y_i 随机缺失,当 $\delta_i=1$ 时,Y_i 可以观测到;当 $\delta_i=0$ 时,Y_i 缺失。大多数的文献中都假定缺失机制是随机缺失的,同时这也符合普遍的实际情况。这里假设 Y_i 是随机缺失的,则有以下关系:

$$P(\delta=1\mid\boldsymbol{X},U,Y)=P(\delta=1\mid\boldsymbol{X},U)$$

设样条节点集合为 $\{\xi_i\}_{i=1}^{2(m+1)+k_n}$,且满足 $\xi_1=\cdots=\xi_{m+1}=a<\xi_{(m+1)+1}<\cdots<\xi_{(m+1)+k_n}<b=\xi_{(m+1)+k_n+1}=\cdots=\xi_{2(m+1)+k_n}$。记 $\boldsymbol{B}(u)=(B_1(u),\cdots,B_{K_j}(u))^{\mathrm{T}}$,是一组 $m+1$ 阶的 B 样条基函数,其中 $K_j=m+k_n+1$,表示渐进 $\beta_j(U_i)$ 所需要的 B 样条的数量,$\{\xi_i\}_{i=(m+1)+1}^{(m+1)+k_n}$ 称为内部节点,k_n 为内部节点个数,则对于任意 $u\in[a,b]$,$\beta_j(u)$ 可以用 B 样条基函数 $\boldsymbol{B}(u)$ 近似逼近:

$$\beta_j(u)=\sum_{s=1}^{K_j}\gamma_{j,s}B_s(u)=\boldsymbol{B}^{\mathrm{T}}(u)\,\boldsymbol{\gamma}_j$$

其中,$\boldsymbol{\gamma}_j=(\gamma_{j1},\gamma_{j2},\cdots,\gamma_{jK_j})^{\mathrm{T}}$,$j=1,2,\cdots,p$。

当数据未发生缺失时,$\boldsymbol{\gamma}$ 的 B 样条分位数回归估计 $\hat{\boldsymbol{\gamma}}=(\hat{\boldsymbol{\gamma}}_1^{\mathrm{T}},\hat{\boldsymbol{\gamma}}_2^{\mathrm{T}},\cdots,\hat{\boldsymbol{\gamma}}_p^{\mathrm{T}})^{\mathrm{T}}$,可由式(3)得到:

$$\hat{\boldsymbol{\gamma}} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau (Y_i - \boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\gamma}) \tag{3}$$

其中 $\rho_\tau(u) = u[\tau - I(u<0)]$ 是 τ 分位损失函数，$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{ip})$，$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}}, \cdots, \boldsymbol{\gamma}_p^{\mathrm{T}})^{\mathrm{T}}$ 为 B 样条系数，$\boldsymbol{W}_i = (X_{i1}\boldsymbol{B}^{\mathrm{T}}(U_i), X_{i2}\boldsymbol{B}^{\mathrm{T}}(U_i), \cdots, X_{ip}\boldsymbol{B}^{\mathrm{T}}(U_i))^{\mathrm{T}}$，$\hat{\boldsymbol{\gamma}}_j = (\hat{\gamma}_{j1}, \hat{\gamma}_{j2}, \cdots, \hat{\gamma}_{jK_j})^{\mathrm{T}}$，则 $\boldsymbol{\beta}(u)$ 的第 j 个基于 B 样条的分位数回归估计为 $\hat{\beta}_j(u) = \boldsymbol{B}^{\mathrm{T}}(u)\hat{\boldsymbol{\gamma}}_j$。

当数据发生缺失时，直接使用完全数据，用 B 样条逼近系数函数得到 $\boldsymbol{\gamma}$ 的估计 $\hat{\boldsymbol{\gamma}}^{CC}$，可通过式(4)得到：

$$\hat{\boldsymbol{\gamma}}^{CC} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \delta_i \rho_\tau (Y_i - \boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\gamma}) \tag{4}$$

从而得到 $\boldsymbol{\beta}(u)$ 的第 j 个基于完全数据的分位数回归估计为 $\hat{\beta}_j^{CC}(u) = \boldsymbol{B}^{\mathrm{T}}\hat{\boldsymbol{\gamma}}_j^{CC}$。

在实际情况中，缺失概率一般来说是未知的，缺失概率在很多文献中也被称为倾向得分。对于缺失概率的估计，许多学者选择用非参数估计，比较常用的非参数估计是 Nadaraya-Watson 估计，形式如下：

$$\hat{\pi}(u) = \frac{\sum_{i=1}^{n} \delta_i K_h(U_i - u)}{\sum_{i=1}^{n} K_h(U_i - u)}$$

其中，$K_h(\cdot) = K(\cdot/h)/h$ 是带宽为 h 的核函数，维数为 p，(U_i, δ_i)，$i = 1, 2, \cdots, n$，为来自 (U, δ) 的一个随机样本。显然，这种非参数估计在 U 为高维的情况下可能出现“维数灾祸”的问题，因此本文考虑使用参数估计方法来估计缺失概率。假设

$$P(\delta = 1 \mid \boldsymbol{X}, U, Y) = P(\delta = 1 \mid \boldsymbol{X}, U) = P(\delta = 1 \mid \boldsymbol{Z}_i) = \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha}) \tag{5}$$

其中 $\boldsymbol{Z}_i = (\boldsymbol{X}_i, U_i)^{\mathrm{T}}$，假设

$$\pi(\boldsymbol{Z}_i, \boldsymbol{\alpha}) = \frac{\exp(\alpha_0 + \alpha_1 U_i + \boldsymbol{\alpha}_2^{\mathrm{T}} \boldsymbol{X}_i)}{1 + \exp(\alpha_0 + \alpha_1 U_i + \boldsymbol{\alpha}_2^{\mathrm{T}} \boldsymbol{X}_i)}$$

其中，$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \boldsymbol{\alpha}_2^{\mathrm{T}})$ 是未知参数向量，根据式(5)可得到似然函数：

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})^{\delta_i} [1 - \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})]^{1-\delta_i}$$

通过最大化似然函数可以得到 $\boldsymbol{\alpha}$ 的估计 $\hat{\boldsymbol{\alpha}}$，令 $U_B(\boldsymbol{\alpha}) = \frac{\delta_i - \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})}{\pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})[1 - \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})]} \frac{\partial \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$，则 $\hat{\boldsymbol{\alpha}}$ 满足 $U_B(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} U_{Bi}(\boldsymbol{\alpha}) = 0$。

简记 $\pi_i = \pi(\boldsymbol{Z}_i, \boldsymbol{\alpha})$，$\hat{\pi}_i = \pi(\boldsymbol{Z}_i, \hat{\boldsymbol{\alpha}})$，基于 B 样条逼近技术，利用逆概率加权方法得到 $\boldsymbol{\gamma}$ 的估计 $\hat{\boldsymbol{\gamma}}^{W}$ 可由式(6)得到：

$$\hat{\boldsymbol{\gamma}}^{W} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} \rho_\tau (Y_i - \boldsymbol{W}_i^{\mathrm{T}} \boldsymbol{\gamma}) \tag{6}$$

从而得到当缺失概率已知时，$\boldsymbol{\beta}(u)$ 的第 j 个基于 B 样条的逆概率加权分位数回归估计为 $\hat{\boldsymbol{\beta}}_j^{W}(u) = \boldsymbol{B}^{\mathrm{T}}(u)\hat{\boldsymbol{\gamma}}_j^{W}$。

节点数目在未知系数函数的光滑程度和将拟合的数据之间起到一个平衡的作用。对于节点数 k_n 的选择，可以极小化某个准则函数来实现，常用的准则函数有 AIC 信息准则、BIC 信息准则、交叉验证和广义交叉验证。He 等[18]发现，在选择 B 样条的节点数时，用信息准则选取要优于用交叉验证和广义交叉验证，而 AIC 信息准则容易出现过拟合的现象，即节点数目选得过大。所以选择极小化 BIC 准则来选择节点数目：

$$BIC(k_n) = \log\left(\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}_i} \rho_\tau (Y_i - \boldsymbol{W}_i^{\mathrm{T}} \hat{\boldsymbol{\gamma}}) \right) + \frac{\log n}{n} \times k \tag{7}$$

其中，$k = \sum_{j=1}^{p} K_j$ 表示式(2)中待估参数的个数。

2 数值模拟

2.1 计算过程

关于变系数分位数回归模型的非参数估计，很多软件都可以实现，目前比较常用的方法是调用 R 软件中的 quantreg 包实现。假设 $\{(Y_i, \boldsymbol{X}_i, U_i, \delta_i)\}_{i=1}^{n}$ 是来自模型式(1)的独立样本，不妨假设前 n_1 个 Y_i 可以观测到，则后 $n - n_1$ 个 Y_i 缺失，模型的非参数函数估计过程如下：

(1) 使用 logistic 模型以及缺失机制，用 glm 函数产生缺失概率 $\hat{\pi}_i$，并根据缺失机制得到完全数据 $\{(Y_i, \boldsymbol{X}_i, U_i, \delta_i)\}_{i=1}^{n_1}$；

(2) 假设 B 样条函数的阶数为 M，$B_{i,q}(u)$ 表示第 i 个 q 阶 B 条基函数，$\{\xi_i\}_{i=1}^{2M+k_n}$ 为样条节点，定义

$$B_{i,1}(u) = \begin{cases} 1, \xi_i \leqslant u \leqslant \xi_{i+1} \\ 0, \text{otherwise} \end{cases}$$

$$B_{i,q}(u) = \frac{(u - \xi_i) B_{i,q-1}(u)}{\xi_{i+q-1} - \xi_i} + \frac{(\xi_{i+q} - u) B_{i+1,q+1}(u)}{\xi_{i+q} - \xi_{i+1}}$$

其中，$i = 1, \cdots, 2M + k_n - q$。上述公式为 Cox-de-Boor 递推公式。根据递推公式产生 3 次 B 样条基函数，然后根据式(7)的 BIC 准则实现节点数的选择；

(3) 用式(1)中产生的 $\{(Y_i,\boldsymbol{X}_i,U_i,\delta_i)\}_{i=1}^{n_1}$,根据式(4)实现基于完全数据的变系数模型的分位数回归估计 $\hat{\boldsymbol{\beta}}^{CC}$;

(4) 使用$\{(Y_i,\boldsymbol{X}_i,U_i,\delta_i)\}_{i=1}^{n}$,调用 R 软件中的 quantreg 包,根据式(6)实现基于逆概率加权的分位数回归,得到 $\hat{\boldsymbol{\beta}}^{W}$ 的估计。

2.2　模拟准备

考虑模型:

$$Y=\beta_1(U)X_1+\beta_2(U)X_2+0.5\varepsilon$$

其中,$X_1\sim N(0,1)$,$X_2\sim N(0,1)$,$U\sim U(0,1)$,$\varepsilon\sim N(0,1)$,系数函数 $\beta_1(u)=\cos 2\pi u$,$\beta_2(u)=\sin 2\pi u$,假设 X_1,X_2,U,ε 相互独立。

用 logistic 模型产生缺失数据,考虑如下两种缺失机制:

$$\pi(\boldsymbol{Z},\boldsymbol{\alpha})=\frac{\exp(2-0.5x_1+0.5x_2+0.5U)}{1+\exp(2-0.5x_1+0.5x_2+0.5U)}$$

$$\pi(\boldsymbol{Z},\boldsymbol{\alpha})=\frac{\exp(0.5+0.5x_1+0.5x_2+U)}{1+\exp(0.5+0.5x_1+0.5x_2+U)}$$

随机生成 n 个服从二项分布的 δ,生成的概率为上述两种缺失概率,根据程序模拟出的两种缺失概率的比例约为 11%和 33%。

对于 $\tau=0.25,0.5,0.75$,抽取的样本容量为 $n=100,200$ 时,分别模拟 300 次,取 300 次模拟结果的平均值作为最终的估计值,计算 $\boldsymbol{\beta}$ 的以下 3 种估计:

(1) 数据没有发生缺失时的估计 $\hat{\boldsymbol{\beta}}$,作为其他模拟结果的参考;

(2) 直接使用完全数据得到的估计 $\hat{\boldsymbol{\beta}}^{CC}$;

(3) 基于逆概率加权方法得到的估计 $\hat{\boldsymbol{\beta}}^{W}$;

为验证估计 $\hat{\boldsymbol{\beta}}(\cdot)$ 的性质,引入平均均方差的平方根(RASE),其值为

$$F_{\mathrm{RASE}}=\sqrt{\frac{1}{n_{\mathrm{grid}}}\sum_{s=1}^{n_{\mathrm{grid}}}\sum_{j=1}^{p}[\hat{\beta}_j(u_s)-\beta_j(u_s)]^2}$$

其中,n_{grid} 为估计 $\beta(u)$ 时格子点的数量。

2.3　模拟结果与结论

根据以上模拟前的准备,按照计算过程中的思路,用 R 软件模拟得到如下结果(表 1,表 2,图 1,图 2):

表 1　$\boldsymbol{\beta}(U)$ 3 种估计的 RASE 值(缺失概率 1)

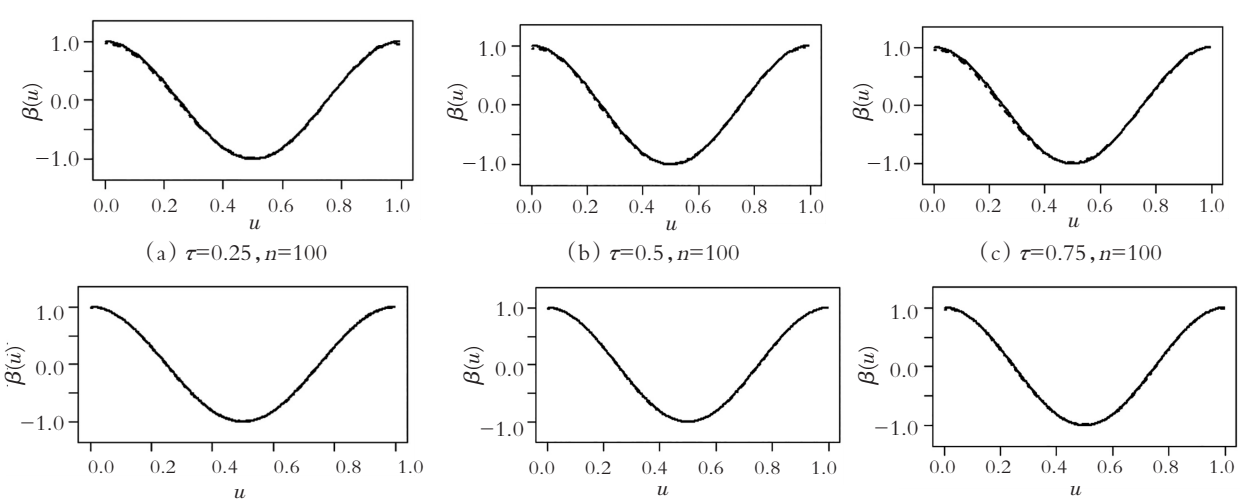
Table 1　RASE values of three estimators of $\boldsymbol{\beta}(U)$ (missing probability 1)

n	估　计	$\tau=0.25$	$\tau=0.5$	$\tau=0.75$
100	$\hat{\boldsymbol{\beta}}$	0.036	0.037	0.037
	$\hat{\boldsymbol{\beta}}^{CC}$	0.048	0.037	0.044
	$\hat{\boldsymbol{\beta}}^{W}$	0.037	0.032	0.046
200	$\hat{\boldsymbol{\beta}}$	0.020	0.017	0.019
	$\hat{\boldsymbol{\beta}}^{CC}$	0.025	0.020	0.022
	$\hat{\boldsymbol{\beta}}^{W}$	0.018	0.017	0.020

表 2　$\boldsymbol{\beta}(U)$ 3 种估计的 RASE 值(缺失概率 2)

Table 2　RASE values of three estimators of $\boldsymbol{\beta}(U)$ (missing probability 2)

n	估　计	$\tau=0.25$	$\tau=0.5$	$\tau=0.75$
100	$\hat{\boldsymbol{\beta}}$	0.036	0.037	0.037
	$\hat{\boldsymbol{\beta}}^{CC}$	0.050	0.042	0.044
	$\hat{\boldsymbol{\beta}}^{W}$	0.051	0.038	0.038
200	$\hat{\boldsymbol{\beta}}$	0.021	0.017	0.019
	$\hat{\boldsymbol{\beta}}^{CC}$	0.032	0.020	0.028
	$\hat{\boldsymbol{\beta}}^{W}$	0.032	0.017	0.024



(a) τ=0.25, n=100　(b) τ=0.5, n=100　(c) τ=0.75, n=100

(a) τ=0.25, n=200　(b) τ=0.5, n=200　(c) τ=0.75, n=200

图 1　3 个分位点 IPW 估计拟合效果(缺失概率 1)

Fig. 1　The fitting effect of IPW estimation for three subsites (missing probability 1)

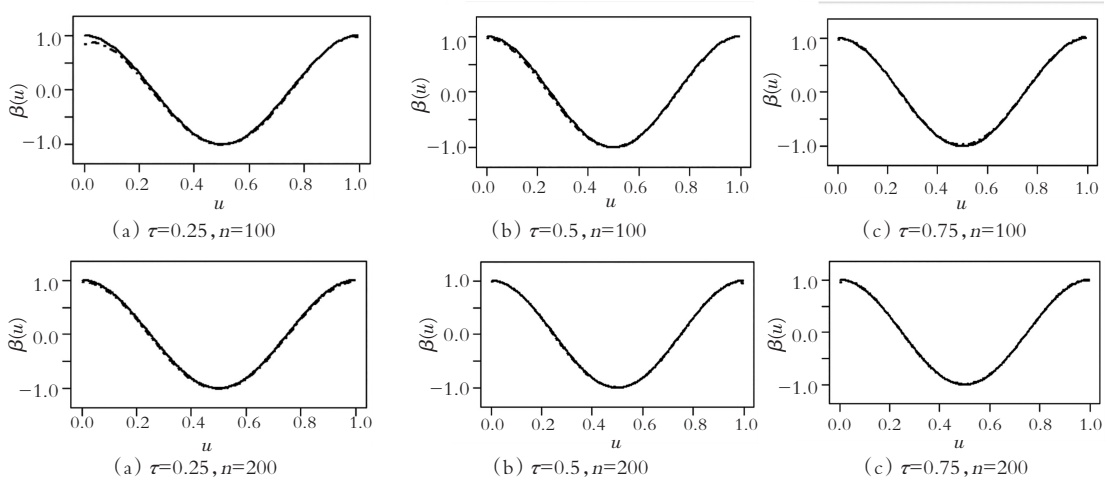


图 2 三个分位点 IPW 估计拟合效果(缺失概率 2)

Fig. 2 The fitting effect of IPW estimation for three subsites (missing probability 2)

图 1,图 2 黑色实线代表真实曲线,虚线代表拟合曲线。其中,从左到右依次代表 $\tau=0.25,0.5,0.75$ 时,$\beta_1(U)$ 的逆概率加权估计拟合曲线。从上到下依次代表 $n=100,200$ 时的逆概率加权估计拟合曲线。

表 1 和表 2 分别展示了两种缺失机制下 $\boldsymbol{\beta}$ 的 3 种估计的平均均方差平方根值,从模拟结果可以得到以下结果:

(1) 利用完全数据分位数回归估计,结果整体上比相应的基于逆概率加权的分位数回归方法的 RASE 值偏大,说明基于逆概率加权的分位数回归较基于完全数据的分位数回归估计效果好;

(2) 对于同一个估计,随着样本量的增加,对应的模拟结果数值也会减小,说明样本量的增加也会减小 3 种估计的平均均方差;

(3) 对比两种缺失机制,第二种缺失机制的 RASE 值普遍大于第一种缺失机制的 RASE 值,说明缺失比例越高,估计效果越差。

图 1 和图 2 展示了当 $n=100,200$ 时,3 个分位点下分别模拟 300 次 $\beta_1(U)$ 的 IPW 估计拟合效果图。从图中也可以看出,样本量越大,拟合效果越好,且缺失比例越大,拟合效果越差。总之,在响应变量随机缺失下,将 B 样条和逆概率加权相结合的变系数模型分位数回归在有限样本情况下表现良好。

3 实例分析

接下来用本文所研究的利用 B 样条结合逆概率加权的方法研究空气污染的数据集。二氧化氮是形成光化学烟雾和酸雨的主要物质之一,它会直接攻击人体的肺部,对人体造成很大的危害,尤其是老人或者呼吸系统有疾病的人,当二氧化氮的累计浓度增大时,会与大气中的臭氧发生化学反应,造成臭氧的损耗,而臭氧就像一把保护伞可以保护我们免受短波紫外线的伤害,所以研究二氧化氮浓度与哪些因素有关很有必要。许多学者拿该组数据研究变系数模型,比如 Tang 和 Sun 等[8-9]。

该组数据来源于 StatLib,是由挪威公共道路管理局收集的关于挪威奥斯陆地区的相关数据,时间跨度为 2001-10—2003-08,目的是研究空气污染与道路车流量和气象因子之间的关系。该数据集的样本量为 500 个,其中包括的变量有每小时二氧化氮(NO_2)浓度的对数(粒子)、每小时车流量的对数、离地面 2 m 的温度(℃)、风速(m/s)、离地面 25 m 与离地面 2 m 的温度差(℃)、风向(0°~360°之间)、每天的时刻、从 2001 年 10 月 1 日到观测时间隔的天数。

参照 Sun 等[9],设响应变量 Y 为每小时的对数二氧化氮浓度,协变量 x_1 为每小时对数车流量,x_2 为风速,t 为每天的时刻,利用本文提出的方法研究 Y 与 x_1,x_2 以及 t 之间的关系,建立以下模型:

$$Y=\beta_1(T)X_1+\beta_2(T)X_2+\varepsilon$$

首先,对响应变量和协变量进行标准化,将它们化成均值为 0,方差为 1 的数据。为了验证所研究估计方法的效率,假设数据为随机缺失(MAR),产生 δ 的缺失机制为 $\pi(\boldsymbol{Z},\boldsymbol{\alpha})=\dfrac{\exp(0.5+0.5x_1+0.5x_2+U)}{1+\exp(0.5+0.5x_1+0.5x_2+U)}$,响应变量的缺失约为 33%。这里选用三次 B 样条基函数逼近未知系数函数,节点为均匀节点,假设两个系数函数的节点数目是一样的,用 BIC 准则选取的内部节点数量为 4,在 $\tau=0.5$ 时,模拟 300 次,取 300 次结果的平均值,得到了基于 B 样条的变系数模型的逆概率加权分位数回归估计,见图 3。

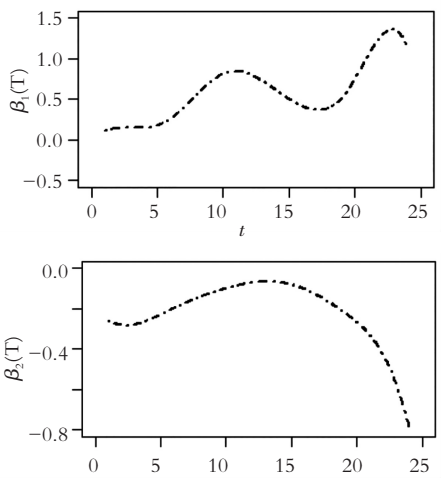


图 3　系数函数 $\beta_1(T)$ 和 $\beta_2(T)$ 估计

Fig. 3　Estimation of coefficient function $\beta_1(T)$ and $\beta_2(T)$

从图中可以看出,二氧化氮浓度与每小时的对数车流量成正相关,说明每小时经过的车辆越多,二氧化氮浓度越高;风速对二氧化氮浓度的影响为负,即在其他条件不变时,风速越大,二氧化氮浓度越低,这与实际情况是一致的;化石燃料的燃烧会带来大量的二氧化氮,所以车流量越大污染越严重;当风速大时,大气流动性越强,二氧化氮进入大气后会被稀释,浓度就会降低。

近几年,环境保护已经成为热点讨论话题,针对实例分析结果,提出以下几个建议:从根源上减少汽车尾气排放,在技术上,可以对汽车尾气处理技术做出改进,对尾气处理技术不达标的汽车进行淘汰或者改造,对汽车的发动机做出改进,使得燃烧更充分等;对城市的交通规划做出改进,尤其是在上下班高峰时,避免因道路拥挤造成发动机燃油燃烧不充分,燃油燃烧不充分时排放的二氧化氮是正常行驶情况下的好几倍,合理的交通规划可以缓解城市拥堵的现象;从民众层面上讲,政府要主动积极宣传生态环保、绿色出行理念,增强居民环境保护意识和综合素质,大力推进新能源汽车的使用,鼓励市民使用公交、地铁出行;发动群众力量植树造林,尽可能减小汽车尾气带来的危害。同时,还需要立法执法,比如严格要求汽车的排放标准以及燃油标准,限制排量大的汽车的出行时间和区域;杜绝任何人或者组织实施任何破坏环境的活动,做到有法可依,执法必严。

4　结束语

本文研究了变系数模型,用 B 样条技术逼近变系数模型的系数函数,基于逆概率加权,构造了响应变量缺失下分位数回归估计的损失函数,得到了未知系数函数的估计。在模拟研究中,用没有数据缺失时的系数函数估计作为参考,与直接使用完全数据的估计方法进行对比,模拟研究证明了该方法的有效性。将该方法应用到挪威公共道路管理局收集的关于奥斯陆地区的相关数据中,研究了空气中二氧化氮浓度与道路车流量和风速之间的关系,发现二氧化氮浓度与每小时的对数车流量成正相关,与风速成负相关关系,与实际情况一致,进一步证明了所研究方法的合理性,并给出了部分建议。

本文考虑的模型是变系数模型,还可以将模型推广到部分线性变系数模型、单指标变系数模型等;对于缺失数据问题,除了本文研究的响应变量缺失,还可以考虑协变量缺失或者响应和协变量同时缺失时的情形;对于 B 样条逼近技术中节点数的选择,假设每个系数函数估计所需要的节点数是相同的,还可以考虑节点数不同的情况。这些问题都是后续研究的重要内容。

参考文献(References):

[1] HASTIE T, TIBSHIRANI R. Varying-coefficient models[J]. Journal of the Royal Statistical Society, Series B: (Methodological), 1993, 55(4): 757—779.

[2] 李志强, 薛留根. 缺失数据下广义变系数模型的均值借补估计 [J]. 数理统计与管理, 2007, 32(3): 444—448.
LI Zhi-qiang, XUE Liu-gen. Mean borrowing and complementing estimation of generalized varying coefficient model with missing data [J]. Mathematical Statistics and Management, 2007, 32(3):444—448.

[3] CAI Z, FAN J, LI R. Efficient estimation and inferences for varying-coefficient models [J]. Journal of the American Statistical Association, 2000, 95(451):888—902.

[4] KOENKER R, BASSETT G. Regression quantiles [J]. Econometrica, 1978, 46(1): 33—50.

[5] HONDA T. Quantile regression in varying coefficient models[J]. Journal of Statistical Planning and Inference, 2004, 121(1): 113—125.

[6] GUO J, TIAN M, ZHU K. New efficient and robust estimation in varying-coefficient models with heteroscedasticity[J]. Statistica Sinica, 2012, 22(3): 1075—1101.

[7] HORVITZ D G, THOMPSON D J. Generalization of sampling without replacement from a finite universe [J]. Journal of the American Statistical Association, 1952, 47(260): 663—685.

[8] TANG L, ZHOU Z. Weighted local linear CQR for varying-coefficient models with missing covariates [J]. Test, 2015, 24(3): 583—604.

[9] SUN J, SUN Q. An improved and efficient estimation method for varying-coefficient model with missing covariates [J]. Statistics & Probability Letters, 2015, 107: 296—303.
[10] ZHAO P, XUE L. Variable selection for semiparametric varying coefficient partially linear models [J]. Statistics & Probability Letters, 2009, 79(20): 2148—2157.
[11] DU J, ZHANG Z, SUN Z. Variable selection for partially linear varying coefficient quantile regression model [J]. International Journal of Biomathematics, 2013, 6(3): 1—14.
[12] JIN J, MA T, DAI J, et al. Penalized weighted composite quantile regression for partially linear varying coefficient models with missing covariates[J]. Computational Statistics, 2021, 36: 541—575.
[13] JIN J, HAO C, MA T. B-spline estimation for partially linear varying coefficient composite quantile regression models [J]. Communications in Statistics: Theory and Methods, 2019, 48(21): 5322—5335.
[14] BO K, LI R, ZOU H. New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models[J]. The Annals of Statistics, 2011, 39(1): 305—332.
[15] CAI Z, XIAO Z. Semiparametric quantile regression estimation in dynamic models with partially varying coefficients [J]. Journal of Econometrics, 2012, 167(2): 413—425.
[16] WENG H, GUO S, CHEN M, et al. On locally weighted estimation and hypothesis testing of varying-coefficient models with missing covariate[J]. Journal of Statistical Planning and Inference, 2009, 139(9): 2933—2951.
[17] HAN P, KONG L, ZHAO J, et al. General framework for quantile estimation with incomplete data [J]. Journal of the Royal Statistical Society: Series B, 2019, 81(2): 305—333.
[18] HE X, SHI P. Convergence rate of B-spline estimators of monparametric conditional quantile functions[J]. Journal of Nonparametric Statistics, 1994, 3(3-4): 299—308.

Quantile Regression of Varying Coefficient Model with Missing Response Variables

YE Yao, YUAN De-mei

(School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: Aiming at the problem of non parameter estimation of varying coefficient quantile regression model in the case of random missing response variables, an estimation method combining B-spline and inverse probability weighting is proposed. Missing data is inevitable in statistical work. Firstly, the missing probability of response variable is generated by logistic model. Then, the coefficient function of varying coefficient model is approximated by B-spline. The loss function of inverse probability weighted quantile regression is constructed by using missing probability, and the estimation of unknown coefficient function is obtained. In the simulation study, the estimation is compared with the estimation method using complete data directly. It is found that the quantile regression of variable coefficient model combining B-spline and inverse probability weighting performs well in the case of limited samples in the case of random missing of response variables, and simulation results show that the method is effective. Finally, the research method is applied to the relevant data of Oslo collected by the Norwegian Public Roads Administration, and the relationship between the concentration of nitrogen dioxide in the air and the road traffic flow and wind speed is studied. The reasonable conclusion is drawn, which further proves the rationality of the proposed method.

Key words: missing response variables; B-spline; inverse probability weighting; quantile regression

责任编辑:李翠薇

引用本文/Cite this paper:

叶瑶,袁德美.响应变量缺失下变系数模型的分位数回归[J].重庆工商大学学报(自然科学版),2022,39(2):46—52.

YE Yao, YUAN De-mei. Quantile regression of varying coefficient model with missing response variables [J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(2):46—52.