doi:10. 16055/j. issn. 1672-058X. 2022. 0001. 012

基于含测量误差半参数模型的糖尿病数据研究

孙　兴, 黄振生
(南京理工大学 理学院,南京 210094)

摘　要:对于糖尿病数据,利用单指标部分含扭曲测量误差的部分变系数单指标模型进行拟合,由于实验数据的维数较大,相较于传统的参数模型和非参数模型,应用半参数模型不仅可以较好地拟合数据,还可以避免“维数灾祸”问题;此外,如果在拟合时忽略误差的影响,可能导致模型估计产生偏差,因此,进一步选择体质指数(BMI)作为潜在的混淆因子,并假设响应变量和单指标变量均受到 BMI 的乘积污染;观察实验结果发现:6 种血清指标测量数据和性别的系数会随着 BMI 的变化而变化,并且对比带有测量误差和不含测量误差两种情形下的结果发现,糖尿病人定量测量值、年龄和平均血压均受到 BMI 的污染;这些结果说明选择单指标部分带有测量误差的部分变系数单指标模型对该数据集进行拟合是合理的,并且相较于不含测量误差的半参数模型,可以更好地挖掘数据中的信息。

关键词:部分变系数单指标模型;乘积扭曲测量误差;糖尿病数据;
中图分类号:O212. 7　　**文献标志码**:A　　**文章编号**:1672-058X(2022)01-0085-07

0　引　言

在医疗领域,糖尿病是备受关注的慢性病之一,也是全球严重的公共卫生安全问题之一,它除了会给患者带来痛苦之外,还会给家庭和社会带来巨大的经济负担和压力。根据国际糖尿病联盟(IDF)2017 的调研结果显示,全球共有约 4. 25 亿糖尿病患者,其中中国糖尿病患者占比超 25%[1]。研究糖尿病与体质测量数据之间的关系,可以更好地了解和预防糖尿病,因此具有较为重要的意义。研究的糖尿病数据集包含 442 个观测样本,其中响应变量是糖尿病患者的定量测量数据,协变量分别为年龄(Age)、性别(Sex)、体质指数(BMI)、平均血压(BP)和 6 种血清指标测量数据,分别以符号 TC、LDL、HDL、TCH、LTG 和 GLU 表示。

部分变系数单指标模型(PVCSIMs)是一类重要的半参数模型,它不仅保留了非参数模型的特点,还能避免“维数灾祸”问题,因此是统计分析的重要工具,尤其在处理高维数据时十分有优势,模型的结构如下:

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

可以看出,该模型具有一般性,它不仅兼有变系数模型和单指标模型的特点,还包含了许多其他重要的半参数模型作为特例:对于变系数部分,函数 $\boldsymbol{\theta}(\cdot)$ 表示的是 Z 和 U 的相互作用,如果假设 $\boldsymbol{\theta}(\cdot)$ 是常数向量,那么模型就可以看作是部分线性单指标模型,若进一步令系数函数向量 $\boldsymbol{\theta}(\cdot)$ 的维数 q 等于 1,那么就可以得到单指标模型。对于单指标部分,若令联系函数 $g(\boldsymbol{\beta}^{\mathrm{T}}X)\equiv\boldsymbol{\beta}^{\mathrm{T}}X$,那么模型就变成了部分线性变系数模型,进一步取参数 $\boldsymbol{\beta}$ 的维数 p 为 1,模型就退化成了变系数模型。

近年来,关于模型式(1)的研究成果已经十分丰富。为了研究化学污染物水平与每天因呼吸系统

收稿日期:2021-01-01;**修回日期**:2021-02-17.

作者简介:孙兴(1995—),男,安徽肥西人,硕士研究生,从事非参数与半参数统计.

疾病住院的总人数的关系,以及温度和相对湿度对入院人数的影响,Wong 等[2]首次提出 PVCSIMs,他们结合二元局部线性方法、平均方法和一步回拟技术(One-step Back-fitting Technique)得到函数和参数的有效估计。基于他们的研究,Huang 和 Zhang[3]进一步利用广义似然法(GLR)解决了模型中变系数部分的检验问题。Li 和 Zhang[4]通过将系数函数和联系函数样条化,提出了模型的惩罚样条估计方法,该方法可以同时得到未知参数和函数的估计值。Wang 和 Xue[5]的研究指出,Wong 等[2]用二元局部线性光滑进行估计可能会导致估计量不相合,因此,提出一种较为稳定的逐步估计法对模型进行估计,其基本思想是:假设参数已知,将模型转换成变系数模型,并利用 Nadaraya-Watson 核估计方法和局部线性回归方法逐步得到系数函数和联系函数的初步估计,然后根据这些初始估计量计算未知参数的估计值,文章还讨论了估计量的渐近性质,并且建立了逐点置信区间和置信域。Huang[6]通过经验似然方法研究了单指标参数的极大似然估计,并且利用截面经验似然方法构造了各参数分量的置信区间。Huang[7]等结合 SCAD(Smoothly Clipped Absolute Deviation)惩罚和逐步估计法研究了指标参数 β 的变量选择问题,在一定的正则化条件下,还构建了估计量的大样本性质。最近,受到图像数据分析的启发,Li 等[8]讨论了函数型数据下 PVCSIMs 的估计问题,利用局部线性方法逐步迭代得到了系数函数、联系函数、指标参数以及方差函数的估计值,并且证明提出的方法相较于 Wong 等[2]和 Wang 和 Xue[5]中的方法更加稳定。

虽然半参数模型可以用来解决大部分回归拟合问题,然而在实际应用中,由于操作人员的失误和测量工具不精确等问题,收集到数据中往往带有测量误差,因此进一步研究半参数误差模型是很有必要的。扭曲测量误差(Distorted Measurement Error)作为误差的常见形式,是近年来学者们研究的热点问题。Sentürk 和 Müller[9]提出协变量调整回归(Covariate-adjusted Regression, CAR)模型,即假设响应变量和协变量都含有扭曲测量误差,通过将线性误差模型转换成变系数模型,并结合分箱法给出了对未知参数的估计。Cui 等[10]提出一种非参数误差回归模型的一般估计方法,该方法通过核估计得到误差函数的估计量并以此计算受污染变量的校正值,然后根据校正后的变量对目标参数进行估计。Delaigle[11]等进一步讨论了在不同假设条件下非参数误差回归模型的估计问题,在弱化对未知变量或扭曲函数的假设后,提出了更一般的估计方法,并且建立了相应估计量的渐近性质。此外,Qian 和 Huang[12]研究了含扭曲测量误差的部分非线性变系数模型的统计推断问题。Dai 和 Huang[13]将协变量含扭曲测量误差的情形推广到部分非线性变系数模型。

对于糖尿病数据集,根据数据本身的特点,利用单指标部分含扭曲测量误差的 PVCSIMs 模型进行拟合。该模型具有复杂的结构,可以灵活地拟合变量之间的关系。进一步考虑了测量误差的存在,使得模型更加符合实际情形,可以更好地挖掘变量之间潜在的联系。模型的估计方法主要参考 Cui 等[10]和 Huang[6]中的思想。

1 模型建立

含有扭曲测量误差的部分变系数单指标模型具有如下形式:

$$\begin{cases} Y_i=\boldsymbol{\theta}^{\mathrm{T}}(U_i)Z_i+g(\boldsymbol{\beta}_0^{\mathrm{T}}X_i)+\varepsilon_i \\ \tilde{Y}_i=\psi(V_i)Y_i \\ \tilde{X}_{ir}=\varphi_r(V_i)X_{ir} \end{cases} \tag{2}$$

其中,$i=1,2,\cdots,n$,$r=1,2,\cdots,p$,Y_i 是响应变量,$X_i\in R^p$,$Z_i\in R^q$,$U_i\in R^1$ 是协变量,$\boldsymbol{\beta}_0$ 是未知的目标参数,$\boldsymbol{\theta}(\cdot)$ 是未知的系数函数向量,$\boldsymbol{g}(\cdot)$ 是未知的联系函数,ε_i 是期望为 0 方差为 $\sigma(U_i)$ 的模型误差,且与变量(X_i,Z_i)独立。$\tilde{Y}_i$和$\tilde{X}_i$分别是真实值Y_i和X_i的直接观测值,V 是一维的混淆变量且与变量(Y,X)独立。为了保证模型的可识别性,对模型(2)做出如下假设:

(1) $\|\beta_0\|=1$,其中 $\|\cdot\|$ 表示 Euclid 模;

(2) $E|\psi(U)|=1$,$E|\varphi_r(U)|=1$;

(3) 模型误差 ε 的方差是有限的。

其中,$r=1,2\cdots,p$。假设条件(1)保证了单指标参数的唯一性,假设条件(2)保证了误差函数的可识别性,假设条件(3)保证了估计量的渐近性质。

接下来,介绍模型(2)的估计方法。首先估计

误差函数,根据 Cui 等[10]和 Delaigle 等[11]中提出的方法,结合假设条件(2)可以得到:

$$\psi(V)=\frac{E(|\tilde{Y}||V)}{E(|Y|)},\varphi_r(V)=\frac{E(|\tilde{X}_r||V)}{E(|X_r|)}$$

因此可以用 N-W 核估计方法可以得到误差函数的估计式分别为

$$\widehat{\varphi}(v)=\frac{\sum_{i=1}^{n}K_{h_1}(v-V_i)\,|\tilde{Y}_i|}{\sum_{i=1}^{n}K_{h_1}(v-V_i)}\times\frac{1}{\bar{|\tilde{Y}|}}$$

$$\widehat{\varphi}_r(v)=\frac{\sum_{i=1}^{n}K_{h_1}(v-V_i)\,|\tilde{X}_{ir}|}{\sum_{i=1}^{n}K_{h_1}(v-V_i)}\times\frac{1}{\bar{|\tilde{X}_r|}}$$

其中,$\bar{|\tilde{Y}|}=\frac{1}{n}\sum_{i=1}^{n}|\tilde{Y}_i|$,$\bar{|\tilde{X}_r|}=\frac{1}{n}\sum_{i=1}^{n}|\tilde{X}_{ir}|$,$K_{h_1}(\cdot)=K(\cdot/h_1)/h_1$,$K(\cdot)$ 和h_1 分别表示核函数和带宽。这里采用变量的绝对值可以避免估计方法失效的问题,因为若 $E(|Y|)=0$ 或 $E(|X_r|)=0$,那么变量 Y 或 X 是恒为 0 的变量,然而这样的变量是没有研究价值的。接着,可以计算出变量(X,Y)的校正估计值:

$$\widehat{Y}_i=\frac{\tilde{Y}_i}{\widehat{\psi}(V_i)},\widehat{X}_{ir}=\frac{\tilde{X}_{ir}}{\widehat{\varphi}(V_i)}$$

根据校正后的变量,可以得到模型(1)的一个近似形式

$$\widehat{Y}_i\approx\boldsymbol{\theta}^{\mathrm{T}}(U_i)Z_i+g(\boldsymbol{\beta}_0^{\mathrm{T}}\widehat{X}_i)+\varepsilon_i \tag{3}$$

模型(3)的估计过程主要参考文献[6]中的思想和方法,简要表述过程如下:

令 $B=\{\boldsymbol{\beta}\in R^p:\|\boldsymbol{\beta}\|=1\}$,可以推出$\boldsymbol{\beta}_0\in B$。由于假设条件(1)的存在,目标函数

$$Q(\boldsymbol{\beta})=\sum_{i=1}^{n}[\widehat{Y}_i-\boldsymbol{\theta}^{\mathrm{T}}(U_i)Z_i-g(\boldsymbol{\beta}^{\mathrm{T}}\widehat{X}_i)]^2$$

$\boldsymbol{\beta}\in B$,在$\boldsymbol{\beta}_0$处的一阶导数不存在,因此考虑使用 Zhu 和 Xue[14]中的“去一分量”法得到$\boldsymbol{\beta}_0$的有效估计。不妨假设$\boldsymbol{\beta}$ 的第 r 个分量$\beta_r>0$,定义

$$\boldsymbol{\beta}^{(r)}=(\beta_1,\beta_2,\cdots,\beta_{r-1},\beta_{r+1},\cdots\beta_p)^{\mathrm{T}}$$

则有

$$\boldsymbol{\beta}=\boldsymbol{\beta}(\boldsymbol{\beta}^{(r)})=$$

$$(\beta_1,\beta_2,\cdots,\beta_{r-1},(1-\|\beta^{(r)}\|^2)^{\frac{1}{2}},\beta_{r+1},\cdots,\beta_p)^{\mathrm{T}}$$

接着,可以计算出$\boldsymbol{\beta}$ 关于$\boldsymbol{\beta}^{(r)}$的 Jacobian 矩阵为

$$\boldsymbol{J}_{\beta^{(r)}}=\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\beta}^{(r)}}=(\gamma_1,\gamma_2,\cdots,\gamma_p)^{\mathrm{T}}$$

其中$\gamma_s(1\leqslant s\leqslant p,s\neq r)$是第 s 个分量为 1 的 $p-1$ 维单位向量,$\gamma_r=-(1-\|\boldsymbol{\beta}^{(r)}\|^2)^{1/2}\cdot\boldsymbol{\beta}^{(r)}$。注意到,$Q(\boldsymbol{\beta})=Q(\boldsymbol{\beta}(\boldsymbol{\beta}^{(r)})\equiv Q(\boldsymbol{\beta}^{(r)})$,可以通过计算$\boldsymbol{\beta}^{(r)}$的估计值,然后经过简单地变换得到参数$\boldsymbol{\beta}_0$的估计值$\widehat{\beta}$。引入辅助随机变量

$$\eta_i(\boldsymbol{\beta}^{(r)})=[\widehat{Y}_i-\boldsymbol{\theta}^{\mathrm{T}}(U_i)Z_i-g(\boldsymbol{\beta}^{\mathrm{T}}\widehat{X}_i)]g'(\boldsymbol{\beta}^{\mathrm{T}}\widehat{X}_i)\boldsymbol{J}_{\beta^{(r)}}^{\mathrm{T}}\widehat{X}_i$$

不难推出,$\{\eta_i(\boldsymbol{\beta}^{(r)}),i=1,2,\cdots,n\}$是相互独立的且$E(\eta_i(\boldsymbol{\beta}^{(r)}))=0$。因此$\boldsymbol{\beta}^{(r)}$的经验对数似然比函数可以定义为

$$\widehat{P}_1(\boldsymbol{\beta}^{(r)})=-2\max\{\sum_{i=1}^{n}\log(np_i):$$

$$p_i\geqslant 0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_i\widehat{\eta}_i(\boldsymbol{\beta}^{(r)})=0\}$$

其中$\widehat{\eta}_i(\boldsymbol{\beta}^{(r)})$ 表示将式$\eta_i(\boldsymbol{\beta}^{(r)})$ 中的函数 $\boldsymbol{\theta}(U_i)$,$g(\boldsymbol{\beta}^{\mathrm{T}}X_i)$ 和$g'(\boldsymbol{\beta}^{\mathrm{T}}X)$ 分别替换成它们的函数初始估计值$\breve{\boldsymbol{\theta}}^{\mathrm{T}}(U_i;\boldsymbol{\beta})$,$\breve{g}(\boldsymbol{\beta}^{\mathrm{T}}\widehat{X}_i;\boldsymbol{\beta})$ 和$\breve{g}'(\boldsymbol{\beta}^{\mathrm{T}}\widehat{X}_i;\boldsymbol{\beta})$。通过最小化 $-\widehat{P}_1(\boldsymbol{\beta}^{(r)})$ 即可得到参数$\boldsymbol{\beta}_0^{(r)}$ 的估计值$\widehat{\boldsymbol{\beta}}^{(r)}$,可以证明,这相当于求解方程组

$$\begin{cases}\sum_{i=1}^{n}\widehat{\eta}_i(\boldsymbol{\beta}^{(r)})=0\\ \|\boldsymbol{\beta}\|-1=0\end{cases} \tag{4}$$

接下来,只要求解函数的初始估计值即可。对于函数的初始估计,考虑利用局部线性光滑方法分步求解系数函数和联系函数,具体的估计步骤如下:首先假设参数$\boldsymbol{\beta}_0$已知,对模型(3)的两边求条件期望,经过简单的计算可以得到:

$$\breve{Y}_i\approx\boldsymbol{\theta}^{\mathrm{T}}(U_i)\breve{Z}_i+\epsilon_i,$$

其中

$$\breve{\widehat{Y}}_i=\widehat{Y}_i-\omega_1(\boldsymbol{\beta}_0^{\mathrm{T}}\widehat{X}_i,U_i)$$

$$\breve{Z}_i=Z_i-\omega_2(\boldsymbol{\beta}_0^{\mathrm{T}}\widehat{X}_i,U_i)$$

且有

$$\omega_1(t,u)=E(\widehat{Y}|\boldsymbol{\beta}_0^{\mathrm{T}}\widehat{X}_i=t,U_i=u)$$

$$\omega_2(t,u)=E(Z|\boldsymbol{\beta}_0^{\mathrm{T}}\widehat{X}_i=t,U_i=u)$$

参考 Einmahl 和 Mason[15]的方法,应用 N-W 核估计求解两个二元函数的估计表达式。选择核函数$K_1(t,u)=K(t)\cdot K(u)$和带宽$h_2(n)\to 0$,$\omega_j(\cdot,\cdot)$ $(j=1,2)$的估计量分别为

$$\hat{\omega}_1(t,u;\boldsymbol{\beta}_0)=\frac{\sum_{i=1}^{n}\hat{Y}_i K_1\left(\frac{\boldsymbol{\beta}_0^{\mathrm{T}}X_i-t}{h_2^{1/2}},\frac{U_i-u}{h_2^{1/2}}\right)}{\sum_{i=1}^{n}K_1\left(\frac{\boldsymbol{\beta}_0^{\mathrm{T}}X_i-t}{h_2^{1/2}},\frac{U_i-u}{h_2^{1/2}}\right)}$$

$$\hat{\omega}_2(t,u;\boldsymbol{\beta}_0)=\frac{\sum_{i=1}^{n}Z_i K_1\left(\frac{\boldsymbol{\beta}_0^{\mathrm{T}}X_i-t}{h_2^{1/2}},\frac{U_i-u}{h_2^{1/2}}\right)}{\sum_{i=1}^{n}K_1\left(\frac{\boldsymbol{\beta}_0^{\mathrm{T}}X_i-t}{h_2^{1/2}},\frac{U_i-u}{h_2^{1/2}}\right)}$$

接着,令 $\boldsymbol{a}=(a_1,\cdots,a_q)^{\mathrm{T}}$ 和 $\boldsymbol{b}=(b_1,\cdots,b_q)^{\mathrm{T}}$,在 u 的邻域内有

$$\boldsymbol{\theta}(u)\approx\theta_j(u)+\theta_j'(u)(U-u)\equiv a_j+b_j(U-u)$$

通过最小化加权平方和

$$\sum_{i=1}^{n}[\hat{Y}_i-[a_j+b_j(U_i-u)]\check{Z}_i]^2K_{h_2}(U_i-u)$$

可以得到系数函数 $\theta(\cdot)$ 及其一阶导数 $\theta'(\cdot)$ 的初始估计量为

$$(\hat{\boldsymbol{a}}^{\mathrm{T}},h_2\hat{\boldsymbol{b}}^{\mathrm{T}})=(\check{\boldsymbol{\theta}}(u;\boldsymbol{\beta}_0)^{\mathrm{T}},h_2\check{\boldsymbol{\theta}}'(u;\boldsymbol{\beta}_0)^{\mathrm{T}})=$$

$$[\boldsymbol{\Gamma}(u)^{\mathrm{T}}\boldsymbol{\kappa}(u)\boldsymbol{\Gamma}(u)]^{-1}[\boldsymbol{\Gamma}(u)^{\mathrm{T}}\boldsymbol{\kappa}(u)\hat{Y}_i]$$

其中 $\boldsymbol{\kappa}(u)=\mathrm{diag}(K_{h_2}(U_1-u),K_{h_2}(U_2-u),\cdots,K_{h_2}(U_n-n))$ 以及

$$\boldsymbol{\Gamma}(u)=\begin{pmatrix}\check{\boldsymbol{Z}}_1^{\mathrm{T}} & h_2^{-1}(U_1-u)\check{\boldsymbol{Z}}_1^{\mathrm{T}}\\ \cdots & \cdots\\ \check{\boldsymbol{Z}}_n^{\mathrm{T}} & h_2^{-1}(U_n-u)\check{\boldsymbol{Z}}_n^{\mathrm{T}}\end{pmatrix}$$

然后,带入系数函数估计量到式(3)中,通过局部线性方法,类似地,可以得到联系函数 $g(\cdot)$ 及其一阶导数 $g'(\cdot)$ 的初始估计方程估计分别为

$$\check{g}(t;\boldsymbol{\beta}_0)=\frac{\sum_{i=1}^{n}W_{ni}(t,\boldsymbol{\beta}_0)(\hat{Y}_i-\check{\boldsymbol{\theta}}^{\mathrm{T}}(U_i;\boldsymbol{\beta}_0)Z_i)}{\sum_{i=1}^{n}W_{ni}(t,\boldsymbol{\beta}_0)}\#$$

$$\check{g}'(t;\boldsymbol{\beta}_0)=\frac{\sum_{i=1}^{n}\widetilde{W}_{ni}(t,\boldsymbol{\beta}_0)(\hat{Y}_i-\check{\boldsymbol{\theta}}^{\mathrm{T}}(U_i;\boldsymbol{\beta}_0)Z_i)}{\sum_{i=1}^{n}\widetilde{W}_{ni}(t,\boldsymbol{\beta}_0)}\#$$

令 $\hat{x}_i(t)=\boldsymbol{\beta}_0^{\mathrm{T}}\hat{X}_i-t$,那么

$$W_{ni}(t;\boldsymbol{\beta}_0)=K_{h_3}(\hat{x}_i(t))[S_{n,2}(t;\boldsymbol{\beta}_0)-\hat{x}_i(t)S_{n,1}(t;\boldsymbol{\beta}_0)]$$

$$\widetilde{W}_{ni}(t;\boldsymbol{\beta}_0)=K_{h_3}(\hat{x}_i(t))[\hat{x}_i(t)S_{n,0}(t;\boldsymbol{\beta}_0)-S_{n,1}(t;\boldsymbol{\beta}_0)]$$

以及

$$S_{n,l}(t;\boldsymbol{\beta}_0)=\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i(t))^l K_{h_3}(\hat{x}_i(t))$$

其中 $l=0,1,2$。到此,就完成了未知函数的初始估计。下一步,求解方程组(4)即可得到估计值 $\hat{\beta}^{(r)}$,根据 $\boldsymbol{\beta}$ 与 $\boldsymbol{\beta}^{(r)}$ 之间的关系,进一步可以计算出参数估计值 $\hat{\beta}$。将 $\hat{\beta}$ 分别代入系数函数向量和联系函数的初始估计量中,就可以得到函数的最终估计 $\hat{\theta}(u;\hat{\beta})$ 和 $\hat{g}(t;\hat{\beta})$。

2 模型拟合

糖尿病数据集共包含 442 个样本观测值和 11 个变量,数据集来自 Efron 等[16]。Zhang 等[17]利用含测量误差的部分线性单指标模型研究了该数据集,并且证明变量 *Sex* 和 6 种血清的化验数据与变量 *BMI* 可能存在非线性关系,受到该实验结果的启发,考虑用部分变系数单指标模型重新分析该数据集,观察变量 *Sex* 和 6 种血清的化验数据变量的系数是否受到 *BMI* 的影响。此外,考虑到病人的 *BMI* 可能与他的年龄和血压存在关系,因此假设变量 *Age* 和 *BP* 受到 *BMI* 的污染,具体的变量意义见表 1。实验前,对各个变量进行了标准化处理。

表 1 糖尿病数据集变量及其含义

Table 1 The variables and meanings of diabetes data set

变量名称	变量意义
Y	患者定量测量数据
X_1	年龄(*Age*)
X_2	平均血压(*BP*)
Z_1-Z_6	血清的化验数据
Z_7	性别(*Sex*)
U,V	体质指数(*BMI*)

在非参数回归方法中,带宽对估计精度较大,因此使用合适的带宽选择方法十分重要。由于对于误差函数的估计方法比较简单,因此可以采用基于经验的拇指规则(Rule of Thumb)来选择带宽。选取 $h_1=n^{-1/3}SE(V)$,其中

$$SE(V)=\left[\frac{1}{n}\sum_{i=1}^{n}(V_i-\bar{V})^2\right]^{\frac{1}{2}}$$

$$\bar{V}=\frac{1}{n}\sum_{i=1}^{n}V_i$$

而对于系数函数向量 $\boldsymbol{\theta}(\cdot)$ 和联系函数 $g(\cdot)$ 的估计比较复杂,因此使用交叉验证法(Cross-validation)选择最优带宽 h_{cv},并令 $h_2=h_{cv}$,$h_3=h_{cv}(n\log n)^{-1/20}$。

根据第 2 节的模型估计方法,本次实验的具体算法流程如下:

Step1　利用绝对值核估计法估计误差函数,并计算校准后的变量$(\widehat{Y}_i,\widehat{X}_i)$;

Step2　选择符合假设(1)的初始参数β_{ini};

Step3　代入参数β,分别计算系数函数向量和联系函数的初始估计值$\breve{\theta}(u;\boldsymbol{\beta})$,$\breve{g}(t;\boldsymbol{\beta})$;

Step4　代入 Step3 中得到估计值$\breve{\theta}(u;\boldsymbol{\beta})$和$\breve{g}(t;\boldsymbol{\beta})$,求解方程组(4)得到估计值$\boldsymbol{\beta}_k$;

Step5　判断估计值β_k是否收敛,若是,则代入到$\breve{\theta}(u;\boldsymbol{\beta})$和$\breve{g}(t;\boldsymbol{\beta})$得到函数的最终估计值,否则返回 Step3 直到结果收敛。

实验的思路十分清晰:首先使用 N-W 核估计法得到误差函数的估计值,并据此计算出变量(Y_i,X_i)的校正估计值;然后,通过局部线性估计法和非线性最小二乘方法计算未知函数的初始估计值;接着,利用函数的估计值,结合"去一分量"法和最小二乘方法得到目标参数的估计值$\widehat{\beta}_k$;最后,若$\widehat{\beta}_k$收敛,则将估计值代入到函数的初始估计量中,得到它们的最终的估计值。

使用模型(2)对糖尿病数据集进行拟合,变量选择如表 1 所示。首先得到的是误差函数$\psi(\cdot)$和$\varphi_r(\cdot)(r=1,2)$的估计曲线,见图 1 和图 2,其中图 2(a)和图 2(b)分别代表函数$\varphi_1(\cdot)$和$\varphi_2(\cdot)$的估计曲线。从估计曲线的趋势可以看出,各条估计曲线都不是水平的,这说明变量(Y,X_1,X_2)与混淆因子 V 存在非线性的关系,这验证了之前的实验假设:糖尿病患者定量测量数据,*Age* 和 *BP* 受到了混淆因子 *BMI* 的污染。

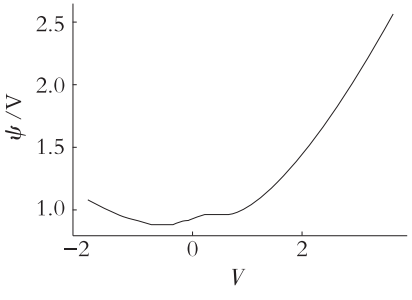
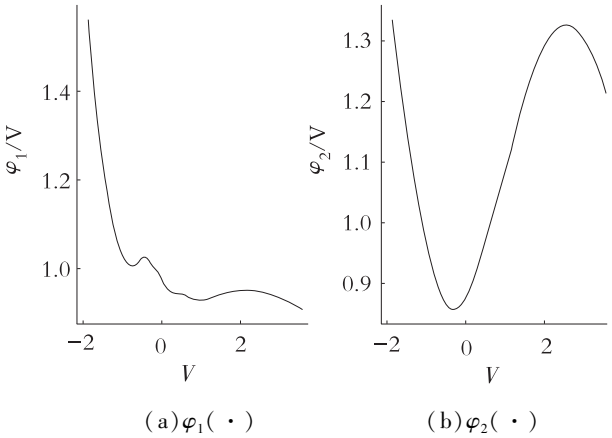


图 1　$\psi(\cdot)$的估计值

Fig. 1　The estimation of $\psi(\cdot)$



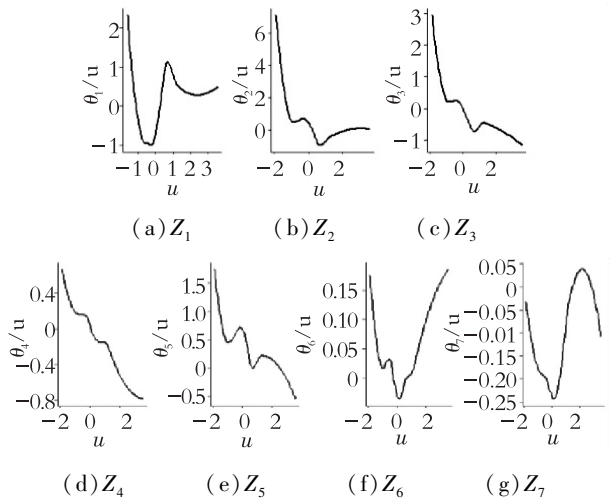
(a)$\varphi_1(\cdot)$　　(b)$\varphi_2(\cdot)$

图 2　$\varphi_1(\cdot)$和$\varphi_2(\cdot)$的估计值

Fig. 2　The estimations of $\varphi_1(\cdot)$ and $\varphi_2(\cdot)$

为了更好地说明扭曲测量误差在模型估计上的影响,实验将比较考虑误差时得到的估计结果与未考虑测量误差时得到的实验结果。在考虑误差的情形下,参数β_0的最终估计值为(0.387 0,0.922 1);当不考虑误差时,得到参数β_0的估计值为(0.329 9,0.944 0)。可以看出,两组估计值存在较大的差距,当考虑测量误差时,*BP* 对指标量$\boldsymbol{\beta}_0^{\mathrm{T}}X$ 的影响要小于未考虑测量误差时的结果,而 *Age* 的影响在考虑误差时反而增大了,这说明测量误差可能会对参数的估计结果有较大的影响。

图 3 展示的是考虑测量误差时得到的系数函数估计图,其中图 3(a)–图 3(g)分别表示变量Z_1–Z_7的系数函数估计曲线。可以看到,函数的估计曲线不是十分平滑,这可能是窗宽 h 较小导致的,其次,还可以观察到,各条估计曲线都不是水平的,函数值均随变量 U 的变化而变化,这说明变量 Z 的系数受到 U 的影响,即变量 *Sex* 和 6 种血清的化验数据变量的系数是否受到 *BMI* 的影响。图 4 反映的是联系函数的估计曲线在不同情形下的比较结果,其中实线表示考虑测量误差时的函数的估计曲线,虚线表示未考虑测量误差时函数的估计曲线。总体来看,考虑测量误差的函数估计要小于未考虑测量误差的估计值,且当变量$\boldsymbol{\beta}_0^{\mathrm{T}}X$ 较小时,两条曲线的增长趋势较为接近。但是,当变量$\boldsymbol{\beta}_0^{\mathrm{T}}X$ 较大时,两条曲线存在较大的差距,并且呈现出相反的增长趋势:实线是先减小后增大而虚线则是先增大后减小。这说明,引入测量误差可能会影响联系函数 $g(\cdot)$ 的估计,且在值$\boldsymbol{\beta}_0^{\mathrm{T}}X$ 较大时,这种影响更加显著。



(a)Z_1 (b)Z_2 (c)Z_3

(d)Z_4 (e)Z_5 (f)Z_6 (g)Z_7

图 3 系数函数向量 $\boldsymbol{\theta}(\cdot)$ 的估计曲线

Fig. 3 The estimation curve of coefficient function vector $\boldsymbol{\theta}(\cdot)$

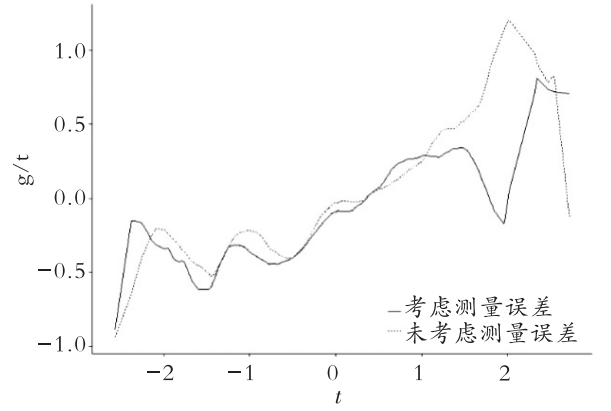


图 4 联系函数 $g(\cdot)$ 的估计曲线

Fig. 4 The estimation curve of contact function $g(\cdot)$

3 结束语

对于高维数据分析,相较于传统的参数模型和非参数模型,半参数模型不仅可以较好地拟合数据,还可以避免“维数灾祸”问题,因此广泛地应用医药和经济等领域。此外,在实际应用中,由于外界因素的干扰,很难避免测量误差的产生,如果忽略误差的影响,就有可能导致模型拟合产生偏差,因此研究带有测量误差的模型应用问题是比较重要的。这里,针对糖尿病数据集,应用部分变系数单指标模型进行拟合,并假设变量 Y 和 X 受到混淆因子 *BMI* 的乘积污染。观察实验结果发现,6 种血清测量数据和变量 *Sex* 的系数会随着 *BMI* 的变化而变化,并且对比带有测量误差和不含测量误差两种情形的结果发现,糖尿病人定量测量值、*Age* 和 *BP* 均受到 *BMI* 的污染。这些结果说明选择单指标部分带有测量误差的部分变系数单指标模型对该数据集进行拟合是合理的,并且相较于不含测量误差的半参数模型,可以更好地挖掘数据中的信息。

参考文献(References):

[1] CHO N H, SHAW J E, KARURANGA S, et al. IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045[J]. Diabetes Research and Clinical Practice, 2018,138(1):271—281.

[2] WONG H, IP W C, ZHANG R Q. Varying-coefficient single-index model[J]. Computational Statistics & Data Analysis, 2008, 52(3):1458—1476.

[3] HUANG Z S, ZHANG R Q. Tests for varying-coefficient parts on varying-coefficient single-index model [J]. Journal of the Korean Mathmatical Society, 2010, 47(2):385—407.

[4] LI J B, ZHANG R Q. Penalized spline varying-coefficient single-index model [J]. Communications in Statistics-Simulation and Computation, 2010, 39(2):221—239.

[5] WANG Q H, XUE L G. Statistical inference in partially-varying-coefficient single-index model [J]. Journal of Multivariate Analysis, 2011, 102(1):1—19.

[6] HUANG Z S. Efficient inferences on the varying-coefficient single-index model with empirical likelihood [J]. Computational Statistics and Data Analysis, 2012, 56(12):4413—4420.

[7] HUANG Z S, LIN B Q, FENG F, et al. Efficient penalized estimating method in the partially varying-coefficient single-index model[J]. Journal of Multivariate Analysis, 2013, 114(1): 189—200.

[8] LI J L, HUANG C, ZHU H T, et al. A functional varying-coefficient single-index model for functional response data[J]. Journal of the American Statistical Association, 2017, 112(519):1169—1181.

[9] ŞENTÜRK D, MÜLLER H G. Covariate-adjusted regression[J]. Biometrika, 2005, 92(1):75—89.

[10] CUI X, GUO W S, LIN L, et al. Covariate-adjusted nonlinear regression[J]. The Annals of Statistics, 2009, 37(4):1839—1870.

[11] DELAIGLE A, HALL P, ZHUO W X. Nonparametric covariate-adjusted regression[J]. The Annals of Statistics, 2016, 44(5):2190—2220.

[12] QIAN Y Y, HUANG Z S. Statistical inference for a varying-coefficient partially nonlinear model with measurement errors[J]. Statistical Methodology, 2016, 32(1):122—130.

[13] DAI S, HUANG Z S. Estimation for varying coefficient partially nonlinear models with distorted measurement errors[J]. Journal of the Korean Statistical Society, 2019, 48(1):117—133.

[14] ZHU L X, XUE L G. Empirical likelihood confidence regions in a partially linear single-index model[J]. Journal of the Royal Statistical Society, Series B, 2006, 68(3):549—570.

[15] EINMAHL U, MASON D M. Uniform in bandwidth consistency of kernel-type function estimators[J]. The Annals of Statistics, 2005, 33(3):1380—1403.

[16] EFRON B, HASTIE T, JOHNSTONE I, et al. Least angle regression[J]. The Annals of Statistics, 2004, 32(2):407—499.

[17] ZHANG J, YU Y, ZHU L X, LIANG H. Partial linear single index models with distortion measurement errors[J]. Annals of the Institute of Statistical Mathematics, 2013, 65(2):237—267.

Research on Diabetes Data Based on Semi-parametric Model with Measurement Error

SUN Xing, HUANG Zhen-sheng

(School of Science, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: For diabetes data, the partially varying-coefficient single-index model with distorted measurement error was used for fitting. Due to the large dimension of experimental data and compared with the traditional parametric model and non-parametric model, the application of semi-parametric model can not only fit the data better, but also avoid the problem of "curse of dimensionality". In addition, if the influence of error is ignored during fitting, it may lead to deviation in model estimation. Therefore, body mass index (BMI) was further selected as a potential confounding factor, and it was assumed that both the response variable and the single-index parameter were contaminated by the BMI. The observation of the experimental results showed that the coefficient of the measurement data of the six serum indicators and sex would vary with BMI, and comparing the results in two different situations, it can be found that the quantitative measurement value, age and average blood pressure of diabetic patients were all polluted by BMI. These results indicate that it is reasonable to select the partially varying-coefficient single-index model with measurement error for the fitting of this data set, and compared with the semi-parameter model without measurement error, this semi-parametric model can better mine the information in the data.

Key words: partially varying-coefficient single-index model; product distortion measurement error; diabetes data

责任编辑:罗姗姗

引用本文/Cite this paper:

孙兴,黄振生. 基于含测量误差半参数模型的糖尿病数据研究[J]. 重庆工商大学学报(自然科学版),2022,39(1):85—91.

SUN Xing, HUANG Zhen-sheng. Research on Diabetes Data Based on Semi-parametric Model with Measurement Error[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2022,39(1):85—91.