

基于缺失数据的 B 样条单指标模型估计 *

李建波*

(江苏师范大学数学与统计学院, 徐州, 221116)

孙晶

(澳门大学科技学院, 澳门)

摘要: 本文主要研究基于响应变量随机缺失的单指标模型的逆概率加权估计问题. 首先通过 B 样条逼近未知单指标函数, 然后构建逆概率加权最小二乘损失函数, 接着通过两阶段牛顿迭代算法获得指标函数和指标系数的估计, 最后通过大量模拟例子和实例分析说明了我们所提估计方法的有效性和合理性.

关键词: 缺失数据; B 样条; 单指标模型

中图分类号: O212.7

英文引用格式: LI J B, SUN J. B-spline estimation of single index models with missing data [J]. Chinese J Appl Probab Statist, 2019, 35(5): 525–534. (in Chinese)

§1. 引言

单指标模型通过未知连接函数非常灵活并更加接近实际的方式刻画了所关心的响应变量及其相应的解释变量之间的数量关系, 是一类半参数回归模型, 在金融经济和生物医学等领域具有广泛的应用. 假设 $Y \in R$ 为研究对象的响应变量, $X \in R^p$ 为相应的解释变量, 单指标模型具有如下形式:

$$Y = g(X'\beta) + \varepsilon, \quad (1)$$

其中 $g(\cdot)$ 是一元未知函数, 一般称之为指标函数; $\beta \in R^p$ 为回归系数向量, 一般称之为指标参数向量, ε 为随机误差项且满足条件 $E(\varepsilon | X) = 0$, $Var(\varepsilon | X) = \sigma^2 > 0$. 为了模型的可识别性, 通常还需要假设 $\|\beta\| = 1$, 且 β 的第一个非零元素为正数, 这里 $\|\cdot\|$ 表示 L_2 范数.

相对线性模型而言, 模型 (1) 数据建模更灵活且保持了模型的易解释性; 相对完全非参数模型而言, 模型 (1) 避免了所谓的“维数祸根”问题, 起到了模型降维的目的, 因此该类模型是一类更广泛、更灵活的半参数统计模型. 关于模型 (1) 的研究已经很多, 例如, Ichimura^[1] 提出了一类最小二乘和加权最小二乘估计方法; Zhu 等^[2] 研究了一类具有单调指标函数的单指标模型. Kamil^[3] 使用单指标模型进行投资组合研究分析. 薛留根^[4] 综述

*国家自然科学基金面上项目(批准号: 11571148)、江苏省统计学优势学科资助项目、江苏省“六大人才高峰”高层次人才项目(批准号: RJFW-038)、江苏省“青蓝工程”中青年学术带头人支持项目和江苏师范大学本科教育教学教研课题(批准号: JYKTZ201907)资助.

*通讯作者, E-mail: ljianb66@163.com.

本文 2018 年 1 月 28 日收到, 2018 年 10 月 12 日收到修改稿.

了单指标模型的估计、模型检验、变量选择等问题的研究现状. Kong 和 Xia^[5] 使用分离交叉核实法提出了一类有效的单指标模型中变量选择方法. Lin 和 Kulasekera^[6] 基于纵向数据研究了线性单指标模型的统计分析问题. Lai 等^[7] 将单指标模型 (1) 推广到异方差的情形, 提出了一类有效的估计方程方法并研究了异方差线性单指标模型的变量选择问题.

在大量实际问题中, 由于记录者的疏忽、复杂条件下数据不可观测性、存储过程错误操作、存储设备的损坏等原因, 数据缺失现象普遍存在, 例如本文实例分析中的 HIV 问题. 一般情况下, 缺失数据包含一定的有效信息, 如果直接删除缺失数据而仅仅使用完全数据进行统计建模与分析, 有可能会造成信息的浪费和统计分析结果较低的可靠性. 根据数据缺失机制不同, 数据缺失机制分为完全随机缺失、随机缺失机制和不可忽略缺失机制三种类型. 不依赖于任何其他变量的数据缺失机制为完全随机缺失 (MCAR); 数据缺失仅依赖于观测数据的数据缺失机制为随机缺失 (MAR); 依赖观测数据缺失部分的数据缺失机制称为不可忽略或非随机缺失. 为了充分挖掘缺失数据在统计分析中的有效信息, 很多统计学者提出了大量缺失数据补全策略, 例如逆概率加权方法、回归借补方法、逆概率加权借补方法等. Zhou 等^[8] 使用回归借补方法研究了缺失数据处理问题. 王启华等^[9] 综述了缺失数据处理方法以及缺失数据下参数与半参数模型的统计推断. 李志强和薛留根^[10] 研究了协变量随机缺失的广义半参数模型的加权拟似然估计方法. 赵培信^[11] 研究了响应变量缺失情形下变系数部分线性模型的逆概率加权经验似然方法. Lai 和 Wang^[12] 在响应变量随机缺失的情形下, 研究了异方差部分线性模型的半参数估计问题. Tan^[13] 研究了缺失数据下条件均值模型的有效限制估计. Liang 等^[14] 探讨了在响应变量缺失以及协变量出现错误的情形下部分线性模型的参数估计问题. Zhao 等^[15] 在协变量缺失的情形下, 研究了估计方程回归分析方法. Wang 等^[16] 研究了协变量缺失情形下回归分析中的加权半参数估计问题. 赵洋等^[17] 在缺失响应下, 通过比较参数方法和非参数方法对选择概率建模的优缺点, 提出一类利用单指标模型对选择概率建模的半参数估计方法.

为了简化缺失数据统计分析, 大部分学者假设数据缺失机制为随机缺失, 这一般也符合众多问题实际情况, 即给定观测变量, 数据缺失与缺失变量相互独立. 关于单指标模型 (1) 中响应变量缺失情形下, 假设 δ 为缺失变量, 即当 Y 缺失时 $\delta = 0$, 否则 $\delta = 1$. 随机缺失蕴含

$$\mathbb{P}(\delta = 1 | Y, X) = \mathbb{P}(\delta = 1 | X) = \pi(X).$$

关于缺失概率 $\pi(X)$ 的统计建模有参数与非参数两类方法, 赵洋等^[17] 比较了参数方法和非参数方法对选择概率建模的优缺点. 当影响因素对缺失机制影响未知时, 往往采用非参数建模方法, 一般使用 Nadaraya-Watson 核来估计缺失概率, 即

$$\pi(X_i) = \frac{\sum_{j=1}^n W_{h_p}(X_j - X_i) \delta_j}{\sum_{j=1}^n W_{h_p}(X_j - X_i)},$$

其中 h_p 为核估计窗宽, $W_{h_p} = W(\cdot/h_p)/h_p^p$, $W(\cdot)$ 为 p 维核函数, $\{(X_i, \delta_i)\}_{i=1}^n$ 为 (X, δ) 的一组样本容量为 n 的样本. 显然, 非参数方法有“维数祸根”问题, 特别当 X 维数过高时, 非参数方法精度不高, 而参数方法简单易操作, 故常常被用来对缺失概率进行建模. 参数方法较为常用的统计模型是 logistics 回归模型, 即

$$\text{logit}(\pi(X)) = X'\alpha, \quad (2)$$

其中 $\text{logit}(p) = \ln[p/(1-p)]$, $\alpha \in R^p$ 为模型回归系数.

目前, 对基于缺失数据的单指标模型统计推断研究较少. 本文将采用逆概率加权最小二乘方法来估计模型 (1) 中的未知参数和未知函数. 首先, 使用模型 (2) 来估计响应变量 Y 的缺失概率, 然后通过 B 样条逼近技术逼近模型 (1) 中未知指标函数并基于估计的缺失概率构建逆概率加权最小二乘损失函数, 从而获得模型未知参数和函数的估计, 最后通过数值模拟和实例分析来说明我们所研究方法的有限样本性质. B 样条的重要优势之一是, 给定指标系数 β , 模型 (1) 就转变为线性模型, 因此, 基于 B 样条逼近的模型 (1) 估计将非常容易操作.

本文接下来, 在第 2 节介绍 B 样条逼近缺失数据单指标模型的估计方法; 在第 3 节给出一个模拟例子和一个实例分析说明我们所研究方法有限样本性质, 最后在第 4 节对本文做出总结并展望本文的研究问题的未来拓展.

§2. 估计方法

假设 (X_i, Y_i, δ_i) , $i = 1, 2, \dots, n$ 是来自总体 (X, Y, δ) 的独立样本, 且 (X, Y) 可通过单指标模型 (1) 来刻画. 给定协变量向量 X_i , 记 $\pi_i = \pi(X_i, \alpha)$ 为响应变量 Y_i 的缺失概率, 其中 α 表明该缺失概率通过模型 (2) 依赖于协变量 X_i . 为了使用 B 样条逼近模型 (1) 中的未知函数 $g(\cdot)$, 一般还假设存在实数 a, b 使得 $P(a \leq \|X\| \leq b) = 1$, 因此考虑到 $\|\beta\| = 1$, $X'\beta$ 也具有有限支撑, 不妨假设为 $[a_0, b_0]$.

给定样条节点集合 $\{\xi_i\}_{i=-(l-1)}^{k+l}$ 且满足 $\xi_{-(l-1)} = \xi_{-(l-2)} = \dots = \xi_0 = a_0 < \xi_1 < \xi_2 < \dots < \xi_k < b_0 = \xi_{k+1} = \xi_{k+2} = \dots = \xi_{k+l}$, 假设 l 阶 B 样条基为 $B_1(t), B_2(t), \dots, B_{k+l}(t)$, 并且对任意 $t \in [a_0, b_0]$ 满足 $B_i(t) > 0$, $\sum_{i=1}^{k+l} B_i(t) = 1$, 其中 k 为 B 样条内节点个数. 于是对任意的 $t \in [a_0, b_0]$ 均有

$$g(t) = \gamma_1 B_1(t) + \gamma_2 B_2(t) + \dots + \gamma_{k+l} B_{k+l}(t) = B(t)'\gamma,$$

其中 $B(t) = (B_1(t), B_2(t), \dots, B_{k+l}(t))'$ 为 B 样条基函数向量, $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{k+l})'$ 为 B 样条系数. 进而模型 (1) 可逼近为

$$Y = B(X'\beta)'\gamma + \varepsilon.$$

于是 B 样条逼近的逆概率加权最小二乘损失函数为

$$l(\beta, \gamma) = \sum_{i=1}^n \frac{\delta_i}{\pi_i} [Y_i - B(X'_i \beta)' \gamma]^2. \quad (3)$$

注意到 $\|\beta\| = 1$, $\beta_1 > 0$, 记 $\beta^{(-1)} = (\beta_2, \beta_3, \dots, \beta_p)'$, 则

$$\|\beta^{(-1)}\| < 1, \quad \beta_1 = \sqrt{1 - \|\beta^{(-1)}\|^2},$$

从而 $\beta = (\sqrt{1 - \|\beta^{(-1)}\|^2}, \beta^{(-1)'})'$. 于是损失函数 (3) 又可表达为

$$\ell(\beta^{(-1)}, \gamma) = \sum_{i=1}^n \frac{\delta_i}{\pi_i} [Y_i - B(X'_i \beta)' \gamma]^2.$$

基于缺失数据模型 (2), 并通过极大化其似然函数

$$\tau(\alpha) = \prod_{i=1}^n \pi^{\delta_i}(X_i, \alpha) [1 - \pi(X_i, \alpha)]^{1-\delta_i}$$

获得 α 的估计, 记为 $\hat{\alpha}$, 其中 $\pi(X_i, \alpha) = [1 + \exp(-X'_i \alpha)]^{-1}$, 进而可得到 π_i 的估计, $\hat{\pi}_i = \pi(X_i, \hat{\alpha})$, 最终我们可得估计的损失函数

$$\hat{\ell}(\beta^{(-1)}, \gamma) = \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} [Y_i - B(X'_i \beta)' \gamma]^2. \quad (4)$$

注记 1 由广义线性模型基本理论可知, 这里的 $\hat{\pi}_i$ 是 π_i 的一致估计, 在模型 (2) 假设下, 该估计与参数 $(\beta^{(-1)}, \gamma)$ 无关, 因此可把损失函数 (4) 中的 $\hat{\pi}_i$ 看成常量, 而把 (4) 仅看作关于 $(\beta^{(-1)}, \gamma)$ 的函数.

注记 2 由于 $E[\hat{\ell}(\beta^{(-1)}, \gamma)/n] = E\{[Y - g(X'\beta)]^2\}$ 是总体损失 $E\{[Y - g(X'\beta)]^2\}$ 的无偏估计, 因此极小化 (4) 而得到的参数估计是有效估计.

注记 3 损失函数 (4) 关于 $(\beta^{(-1)}, \gamma)$ 是非线性, 因此迭代算法是优化该函数的选择之一. 注意到, 给定 $\beta^{(-1)}$, 该损失函数关于参数 γ 对应线性模型的加权最小二乘损失, 从而有显示解, 我们可设计一套牛顿迭代算法对损失函数进行求解.

控制 $\|\beta^{(-1)}\| < 1$, 关于 $(\beta^{(-1)}, \gamma)$ 极小化损失函数 4, 我们便得到 $(\beta^{(-1)}, \gamma)$ 的逆概率加权最小二乘估计, 记为 $(\hat{\beta}^{(-1)}, \hat{\gamma})$. 从而指标回归系数估计为 $\hat{\beta} = (\sqrt{1 - \|\hat{\beta}^{(-1)}\|^2}, \hat{\beta}^{(-1)'})'$. 对于 $\forall t \in [a_0, b_0]$, 指标函数可估计为 $\hat{g}(t) = B(t)' \hat{\gamma}$. 下边我们给出两阶段牛顿迭代法对估计的损失函数 (4) 进行极小化.

Step 1 设定参数 $(\beta^{(-1)}, \gamma)$ 初始估计 $(\beta_0^{(-1)}, \gamma_0)$;

Step 2 给定参数 γ 的估计 γ_0 , 通过下式更新 β 的估计,

$$\hat{\beta}^{(-1)} = \beta_0^{(-1)} - \left[\frac{\partial^2 \hat{\ell}(\beta^{(-1)}, \gamma_0)}{\partial \beta^{(-1)} \partial \beta^{(-1)'}} \Big|_{\beta^{(-1)}=\beta_0^{(-1)}} \right]^{-1} \frac{\partial \hat{\ell}(\beta^{(-1)}, \gamma_0)}{\partial \beta^{(-1)}} \Big|_{\beta^{(-1)}=\beta_0^{(-1)}},$$

其中

$$\begin{aligned}\frac{\partial \hat{\ell}(\beta^{(-1)}, \gamma)}{\partial \beta^{(-1)}} &= -2J \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} [Y_i - B(X'_i \beta)' \gamma] [\dot{B}(X'_i \beta)' \gamma] X_i, \\ \frac{\partial^2 \hat{\ell}(\beta^{(-1)}, \gamma)}{\partial \beta^{(-1)} \partial \beta^{(-1)'} } &= -2 \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} \{[Y_i - B(X'_i \beta)' \gamma] [\ddot{B}(X'_i \beta)' \gamma] - [\dot{B}(X'_i \beta)' \gamma]^2\} J X_i X'_i J', \\ J &= [-\beta_0^{(-1)} / \sqrt{1 - \|\beta_0^{(-1)}\|^2}, I_{p-1}],\end{aligned}$$

I_{p-1} 为 $p-1$ 维单位矩阵, $\dot{B}(t) = [\dot{B}_1(t), \dot{B}_2(t), \dots, \dot{B}_d(t)]'$, $\ddot{B}(t) = [\ddot{B}_1(t), \ddot{B}_2(t), \dots, \ddot{B}_d(t)]'$, $\dot{B}_i(t)$, $\ddot{B}_i(t)$ 分别为 $B_i(t)$ 的一阶和二阶导数. 如果 $\|\hat{\beta}^{(-1)}\| > 1$, 取 $\hat{\beta}^{(-1)} = \hat{\beta}^{(-1)} / (\|\hat{\beta}^{(-1)}\| + 1/n)$.

Step 3 给定 $\beta^{(-1)}$ 的估计 $\hat{\beta}^{(-1)}$, 即 β 的估计 $\hat{\beta} = (\sqrt{1 - \|\hat{\beta}^{(-1)}\|^2}, \hat{\beta}^{(-1)'})'$, 通过线性模型 $Y = B(X' \hat{\beta})' \gamma + \varepsilon$ 的加权最小二乘估计方法更新 γ 的估计为

$$\hat{\gamma} = [\mathcal{B}(\hat{\beta})' W \mathcal{B}(\hat{\beta})]^{-1} \mathcal{B}(\hat{\beta})' W Y,$$

其中

$$\mathcal{B}(\beta) = [B(X'_1 \beta), B(X'_2 \beta), \dots, B(X'_n \beta)]', \quad W = \text{diag} \left\{ \frac{\delta_1}{\hat{\pi}_1}, \frac{\delta_2}{\hat{\pi}_2}, \dots, \frac{\delta_n}{\hat{\pi}_n} \right\}.$$

Step 4 令 $\beta_0^{(-1)} = \hat{\beta}^{(-1)}$, $\gamma_0 = \hat{\gamma}$, 重复迭代 Step 2-3 直到满足算法收敛条件停止迭代.

在本文中, 如果 $\|\hat{\beta}^{(-1)} - \beta_0^{(-1)}\| < 10^{-4}$, 我们就停止迭代过程. 从以上算法可看出, B 样条逼近技术的重要优势之一在于给定指标回归系数的估计, 单指标模型 (1) 就转化为线性模型的结构, 从而提高了模型未知量估计的效率和速度. 同时, 我们所提出的方法是全局最优估计方法.

§3. 数值研究

3.1 模拟研究

我们从以下半参数单指标模型模拟产生样本量为 $n = 60, 200, 500$ 的缺失样本数据 $\{(Y_i, X_i, \delta_i)\}_{i=1}^n$,

$$Y = \sin \left(\frac{X' \beta \pi}{2} \right) + \varepsilon, \quad \text{logit}(\pi) = \mu + X' \alpha,$$

其中 $\beta = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})'$, $X \sim N_3(0, I_3)$, $\varepsilon \sim N(0, 1)$, $\alpha = (0.2, -0.5, 0.3)'$, μ 是截矩项. 通过调节 μ 的取值, 考虑平均缺失概率三种情形: sp = 15%, 30%, 50%. 为了进一步说明本文提出的估计方法有限样本表现效果, 本文还基于忽略缺失数据的样本的非线性最小

二乘法对模型 (1) 中的未知量进行估计, 两种方法均做 2000 次数值模拟. 我们使用均方误差 (MSE) 来说明指标回归系数, 均方误差平方根 (RMSE) 来说明指标函数的拟合效果, 即

$$\text{MSE} = \mathbb{E}(\hat{\beta}^{(-1)} - \beta^{(-1)})^2, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n [\hat{g}(t_i) - g(t_i)]^2}.$$

从以上两式我们很容易看出, MSE 越小表明指标参数的拟合效果越好, 而 RMSE 越小表明指标函数的拟合效果越好. 一般而言, B 样条的阶数对指标函数逼近精度影响较小. 我们可以使用如下 BIC 准则来选择 B 样条节点个数,

$$\text{BIC} = \hat{\ell}(\hat{\beta}^{(-1)}, \hat{\gamma}) + (p + k + d - 1) \ln(n),$$

其中 $\hat{\beta}^{(-1)}, \hat{\gamma}$ 为内节点个数为 k 设置下的 $\beta^{(-1)}, \gamma$ 的逆概率加权最小二乘估计, d 为 B 样条阶数. 为了简化计算复杂度, 我们均选择 $\lceil n^{0.2} \rceil$ 个内节点来构造指标函数的四阶 B 样条逼近. 而当样条节点个数确定后, 我们采用 $\{X'_1 \hat{\beta}, X'_2 \hat{\beta}, \dots, X'_n \hat{\beta}\}$ 的等距样本分位数作为 B 样条逼近多项式节点, 其中 $\lceil \cdot \rceil$ 表示向上取整算子.

表 1 β 和 $g(\cdot)$ 估计的 MSE 和 RMSE

sp	n	缺失数据		完全数据	
		MSE	RMSE	MSE	RMSE
15%	60	0.0016	0.0232	0.0018	0.0143
	200	0.0014	0.0122	0.0016	0.0126
	500	0.0004	0.0105	0.0012	0.0121
30%	60	0.0392	0.0424	0.0653	0.0467
	200	0.0017	0.0215	0.0027	0.0223
	500	0.0005	0.0112	0.0016	0.0218
50%	60	0.0611	0.0541	0.0724	0.0643
	200	0.0239	0.0224	0.0167	0.0336
	500	0.0007	0.0118	0.0019	0.0221

表 1 展示了基于缺失数据和删除缺失数据的“完全数据”两种情形下指标参数估计的 MSE 和指标函数估计的 RMSE 的模拟结果. 一方面, 基于缺失数据的 MSE 和 RMSE 均比较小, 并且随着样本量的增加或者缺失率的降低, MSE 和 RMSE 均呈现下降的趋势, 这表明基于缺失数据的逆概率加权最小二乘估计方法, 当缺失率一定时, 随着样本量的增大, 指标参数和指标函数的逆概率加权最小二乘估计效果表现越来越好; 而当样本容量一定, 估计效果随着缺失率的降低而表现越来越好. 另一方面, 直接使用删除缺失数据后的“完全数据”进行模型估计, 我们发现指标参数和指标函数的最小二乘估计的 MSE 和 RMSE 均比相应的基于缺失数据的逆概率加权最小二乘估计偏大, 这说明基于缺失数据的逆概率加权最小二乘估计效果要优于基于“完全数据”的最小二乘估计.

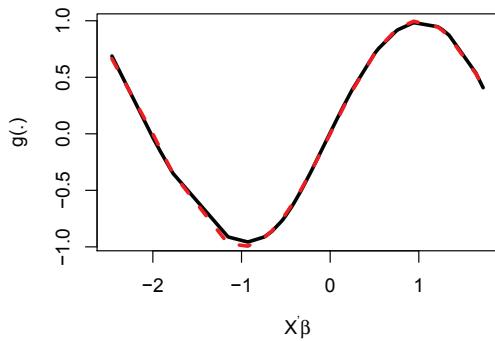
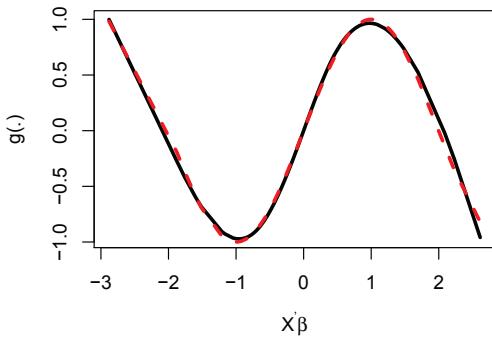
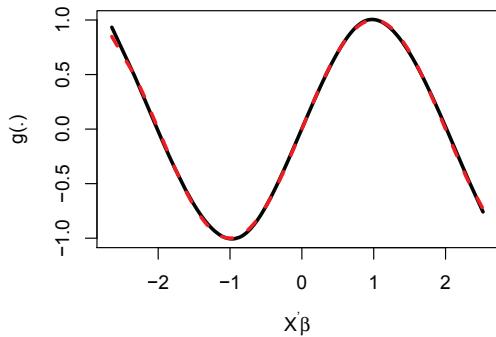
图 1 $n = 60$ 指标函数曲线图 2 $n = 200$ 时指标函数曲线图 3 $n = 500$ 指标函数曲线

图 1–3 分别刻画了样本容量为 60, 200, 500 典型样本的指标函数的拟合曲线, 其中实线为真实曲线, 虚线为拟合曲线. 这里的典型样本是其 RMSE 为 2000 个指标函数估计的 RMSE 的中位数. 三张图表明指标函数逆概率加权最小二乘估计曲线均非常接近相应的真值曲线, 并且样本量较大接近程度越好.

总之, 我们所提出的基于 B 样条逼近的逆概率加权最小二乘估计方法的有限样本表现很好, 并且优于删除缺失数据的最小二乘估计方法.

3.2 实例分析

我们用本文所研究的逆概率加权最小二乘估计方法来分析 AIDS 病例临床试验数据 (ACTG175). 该组数据曾被许多统计学者进行过研究, 例如文献 [12].

CD4 细胞是一种重要的免疫细胞, 是 HIV 病毒受体, CD4 细胞数量能够直接反映人体免疫功能, 是刻画 HIV 病毒感染患者免疫系统受损最明确的指标之一. 该数据分析的目标是通过 CD4 细胞的数量检测结果来刻画某种 HIV 治疗手段的医治效果以及其他生理因素对治疗效果的影响的因素分析. 该临床试验包括四种不同逆转录病毒治疗方案, 逆转录病毒可以降低患有中期艾滋病病毒和无症状的人的风险. 本试验中有 532 名受试者, CD4

细胞正常数量在每立方毫米血液 200 到 500 之间, 响应变量是 96 ± 5 周内的 CD4 细胞浓度 (cell/mm^3), 即 CD496 (Y). CD8 细胞是另一种 T 淋巴细胞, 它可消灭受 HIV 病毒感染的 CD4 细胞. 因此 CD4 细胞数目和 CD8 细胞数目是倒置的情况, 也就是说 CD8 细胞数目对 CD4 细胞起负作用. 本数据分析选取年龄 (X_1)、体重 (X_2)、CD4 细胞初始数目 (CD40 , X_3)、在 20 ± 5 周内 CD4 细胞数目 (CD420 , X_4)、CD8 细胞初始数目 (CD80 , X_5)、在 20 ± 5 周内 CD8 细胞数目 (CD820 , X_6) 作为解释变量并构建单指标模型 (1), 其中 $X = (X_1, X_2, \dots, X_6)'$. 由于调查人员疏忽或其他原因导致其中一些响应变量缺失, 此时假设数据缺失机制是 MAR, 响应变量的缺失率为 0.419.

我们仍使用三次 B 样条逼近指标函数, 通过 BIC 准则在 $\{2, 3, \dots, 8\}$ 中选择内节点个数为 4. 通过基于 B 样条逼近的逆概率加权最小二乘估计方法获得模型参数估计和指标函数拟合曲线, 分别见表 2 和图 4.

表 2 HIV 病例模型指标参数 β 估计结果

	估计	标准差	t 值
β_1	0.5891	0.7706	0.573
β_2	-0.0474	0.4144	-0.343
β_3	0.3474	0.0829	4.933
β_4	0.7263	0.0768	7.211
β_5	-0.0003	0.0251	-1.306
β_6	-0.0502	0.0281	-1.049

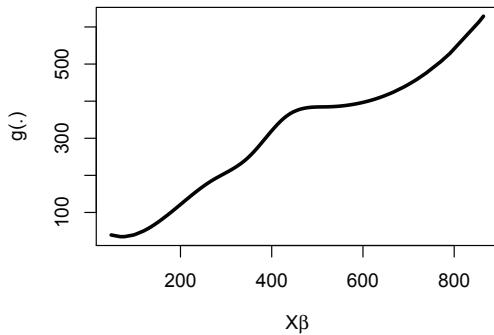


图 4 HIV 病例模型指标函数拟合曲线

由表 2 可看出 96 ± 5 周内的 CD4 细胞浓度仅与 CD4 的初始浓度、 20 ± 5 周内的 CD4 细胞数目有显著性关系, 而与患者年龄、性别和 CD8 各阶段的细胞数目之间无显著性关系. 另外, 结合图 4 中指标函数的非线性递增趋势, 两阶段的 CD4 细胞数对 CD496 的浓度具有正影响, 而两阶段 CD8 细胞数目对 CD496 浓度具有负影响, 因此为了控制 HIV 病毒的进一步扩展, 重点应放在控制 CD4 细胞在初期和前期的数目, 否则后期会加速患者病毒

感染.

§4. 结 论

本文基于样条技术逼近单指标模型, 研究了单指标缺失数据逆概率加权最小二乘估计问题, 通过若干数值研究说明了我们所研究的方法的有效性和合理性.

本论文在讨论缺失数据问题时, 仅考虑了响应变量缺失的情况, 我们还可以考虑协变量缺失或者协变量与响应变量同时缺失的情况. 另外, 本论文假设响应变量随机缺失, 我们还可以进一步探讨缺失机制为完全随机缺失和非随机缺失的情形. 同时还可以把该估计方法推广到部分线性单指标模型、部分变系数单指标模型等, 所有这些问题将是我们后续研究的重要内容.

参 考 文 献

- [1] ICHIMURA H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models [J]. *J Econometrics*, 1993, **58(1-2)**: 71–120.
- [2] ZHUL P, YANG X Y, YU Z, et al. An analysis of single-index model with monotonic link function [J]. *Appl Math J Chinese Univ Ser B*, 2008, **23(1)**: 107–112.
- [3] KAMIL A A. Portfolio analysis using single index model [J]. *WSEAS Trans Math*, 2003, **2(1)**: 83–91.
- [4] 薛留根. 单指标模型的统计推断 [J]. 数理统计与管理, 2012, **31(1)**: 55–78; **31(2)**: 226–246.
- [5] KONG E F, XIA Y C. Variable selection for the single-index model [J]. *Biometrika*, 2007, **94(1)**: 217–299.
- [6] LIN W, KULASEKERA K B. Testing the equality of linear single-index models [J]. *J Multivariate Anal*, 2010, **101(5)**: 1156–1167.
- [7] LAI P, WANG Q H, ZHOU X H. Variable selection and semiparametric efficient estimation for the heteroscedastic partially linear single-index model [J]. *Comput Statist Data Anal*, 2014, **70**: 241–256.
- [8] ZHOU Y, WAN A T K, WANG X J. Estimating equations inference with missing data [J]. *J Amer Statist Assoc*, 2008, **103(483)**: 1187–1199.
- [9] 王启华, 史宁中, 耿直. 现代统计研究基础 [M]. 北京: 科学出版社, 2010.
- [10] 李志强, 薛留根. 协变量随机缺失的广义半参数模型 [J]. 北京工业大学学报, 2007, **33(7)**: 761–765.
- [11] 赵培信. 半参数变系数部分线性模型的统计推断 [D]. 北京: 北京工业大学, 2010.
- [12] LAI P, WANG Q H. Semiparametric efficient estimation for partially linear single-index models with responses missing at random [J]. *J Multivariate Anal*, 2014, **128**: 33–50.
- [13] TAN Z. Efficient restricted estimators for conditional mean models with missing data [J]. *Biometrika*, 2011, **98(3)**: 663–684.
- [14] LIANG H, WANG S J, CARROLL R J. Partially linear models with missing response variables and error-prone covariates [J]. *Biometrika*, 2007, **94(1)**: 185–198.
- [15] ZHAO L P, LIPSITZ S, LEW D. Regression analysis with missing covariate data using estimating equations [J]. *Biometrics*, 1996, **52(4)**: 1165–1182.

- [16] WANG C Y, WANG S J, ZHAO L P, et al. Weighted semiparametric estimation in regression analysis with missing covariate data [J]. *J Amer Statist Assoc*, 1997, **92**(438): 512–525.
- [17] 赵洋, 薛留根, 胡玉琴. 缺失数据下线性模型的半参数估计 [J]. 数学的实践与认识, 2016, **46**(16): 191–197.
- [18] 何晓群, 刘文卿. 应用回归分析 [M]. 3 版. 北京: 中国人民大学出版社, 2011.
- [19] WANG L, YANG L J. Spline estimation of single-index models [J]. *Statist Sinica*, 2009, **19**(2): 765–783.
- [20] 王斌会. 多元统计分析及 R 语言建模 [M]. 广州: 暨南大学出版社, 2010.

B-Spline Estimation of Single Index Models with Missing Data

LI Jianbo

(School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou, 221116, China)

SUN Jing

(Faculty of Science and Technology, University of Macau, Macau, China)

Abstract: In this paper, we studied the inverse probability weighted least squares estimation of single-index model with response variable missing at random. Firstly, the B-spline technique is used to approximate the unknown single-index function, and then the objective function is established based on the inverse probability weighted least squares method. By the two-stage Newton iterative algorithm, the estimation of index parameters and the B-spline coefficients can be obtained. Finally, through many simulation examples and a real data application, it can be concluded that the method proposed in this paper performs very well for moderate sample.

Keywords: missing data; B-spline; single-index model

2010 Mathematics Subject Classification: 62G05