

多指标可加模型及在医疗费用预测中的应用*

潘 青 赵晓兵*

(浙江财经大学数据科学学院, 杭州, 310018)

摘 要: 对医疗费用的建模分析与合理预测是医疗保险费用厘定的基础与根本. 医疗费用中的高维附加信息在长期预测中具有重要作用. 然而, 传统的统计建模方法不适用于处理高维纵向数据下的医疗费用. 本文提出部分线性多指标可加模型, 对具有高维特征的纵向医疗费用数据进行拟合与预测, 并且使用两种不同的降维估计方法进行模型估计, 并将该模型应用于一组含高维协变量的纵向医疗费用数据中进行实例分析. 结果表明该模型以及两种不同的降维方法均对纵向医疗费用进行了很好的拟合.

关键词: 部分线性多指标可加模型; 纵向医疗费用; 部分充分降维; 最小平均方差估计

中图分类号: O212.1

英文引用格式: PAN Q, ZHAO X B. Multi-index additive model and its application in medical cost forecast [J]. Chinese J Appl Probab Statist, 2022, 38(1): 43-52. (in Chinese)

§1. 引 言

医疗费用的分析一直是卫生健康领域研究的热点, 医疗费用的评估不仅是民生问题研究的热点之一, 同时也影响着医疗体系的构建, 医疗保险的发展, 医疗健康等产业链的形成. 当医疗费用数据是纵向数据时, 产生的医疗费用为纵向医疗费用. 分析这类数据的文章在国外早已展开, 如: Zhou 和 Liang^[1] 采用单指标模型分析纵向医疗数据. Castelli 等^[2] 提出了马尔科夫转换模型. Liu 等^[3] 提到, 在对纵向医疗费用建模时, 如果不考虑终止事件的影响, 会造成估计带有偏差. Sun 等^[4] 基于复发事件与终止事件提出了一个联合模型, 处理纵向医疗费用与复发事件或终止事件的关系. Zhao 等^[5] 在分析慢性心力衰竭患者的纵向医疗费用数据时, 使用了边际模型对纵向医疗费用进行了数据分析. Chen 等^[6] 提出用广义部分线性单指标模型分析纵向医疗数据, 之后, Chen 等^[7] 又提出了广义部分线性模型.

但是, 由于收集到的纵向医疗费用数据往往含有高维附加信息, 导致数据存在异方差、偏斜、“维数祸根”等问题. 此时, 传统的统计建模方法不再适用, 需要提出一种新的研究方法, 对含有高维协变量的纵向医疗费用数据进行拟合分析, 找出影响医疗费用的关键因素, 以此做出正确决策. 本文基于 Zhao 等^[5] 提出的模型, 将该模型延伸至可含高维协变量的

*国家自然科学基金项目(批准号: 11271317)、浙江省自然科学基金项目(批准号: LY16A010007)和浙江省一流学科A类(浙江财经大学统计学)(项目编号: Z0111120010/010)资助.

*通讯作者, E-mail: maxbzhao@126.com.

本文 2019 年 10 月 30 日收到, 2020 年 7 月 6 日收到修改稿.

部分线性多指标可加模型, 然后使用了两种不同的估计方法, 更精确的评估医疗费用. 该模型有两个特点: 一是模型中可含高维协变量, 二是连接函数可以是未知的. 最后, 通过实例分析说明本文所提出的新模型.

§2. 模 型

考虑一个样本容量为 n 的纵向数据的随机样本, n 个个体之间相互独立. 对于个体 $i = 1, 2, \dots, n$, 令 $Y_i(t)$ 为响应变量, X_i 表示为 p 维协变量, $N_i^*(t)$ 是个体 i 在 $(0, t)$ 时间段内复发事件发生的次数, 由于在实际的实验中, 通常是在有限时间段内观察个体, 因此 $N_i^*(t)$ 是不可能完全观察到的. 记个体 i 总的观测次数为 K_i , $Y_i(t)$ 是在不同的时间点 $T_{K_i,1} < T_{K_i,2} < \dots < T_{K_i,K_i}$ 观测得到. 观察到的个体 i 复发事件的数量由 $N_i(t) = \sum_{j=1}^{K_i} I(T_{K_i,j} \leq t) = N_i^*(\min(t, C_i))$ 表示, 其中 $I(\cdot)$ 为示性函数, C_i 为个体 i 复发事件的删失时间, 有 $K_i = N_i^*(C_i)$.

对于纵向的医疗费用数据, Zhao 等^[5] 提出了如下模型:

$$E[Y_i(t) | X_i, W_i, N_i(t-)] = \mu(t) + \beta' X_i + \alpha' N_i(t-) W_i, \quad (1)$$

其中 W_i 可认为是协变量 X_i 中的一部分. $\mu(t)$ 是时间 t 的光滑函数, β 是 p 维的回归参数, α 是 q 维的回归系数. $N_i(t-)$ 为在个体 i 的历史观测时间 t 之前的观测总数. 然而, 该模型存在着如下局限性: 一是将协变量定义为低维, 忽略了高维协变量. 二是该模型的联系函数被固定化, 模型不具一般性. 为了避免“维数灾难”, 首要任务是使用降维方法对协变量 X_i 进行降维.

为了提高模型的灵活性, 并且充分考虑到收集到的高维协变量信息, 本文在文献 [5] 模型的基础上, 提出了部分线性多指标可加模型,

$$Y_i(t) = \mu(t) + \phi(X_i' \beta_1, X_i' \beta_2, \dots, X_i' \beta_d) + \alpha' N_i(t-) W_i + \varepsilon_i(t), \quad (2)$$

其中 $Y_i(t)$ 为响应变量, $X_i = (X_{i,1}, X_{i,2}, \dots, X_{i,p})^T$ 是 p 维协变量, $d \ll p$. 在该模型中, 假定连接函数 $\phi(\cdot)$ 为未知的, 进一步提高了模型的灵活性与适用性, 并且允许高维协变量的存在. 显然模型 (2) 是由部分线性可加模型 (1) 衍生出的部分线性多指标可加模型. 为了模型可识别, 需要假定 $E[\mu(t)] = 0$, $E[\phi(X^T B)] = 0$, 以及指标模型的可识别条件. 而模型 (1) 的方法无法再适用于模型 (2), 因此, 必须寻找适合模型 (2) 的估计方法.

对于模型 (2) 的估计, 本文使用了部分充分降维 (partial sufficient dimension reduction) 与最小平均方差估计 (minimum average variance estimation, MAVE) 两种不同的降维方法进行模型估计.

§3. 模型估计

1) 基于部分充分降维的核估计

为了避免“维数灾难”, 第一任务是使用部分充分降维对协变量 X_i 进行降维. 部分充分降维是找出协变量 X_i 的 d 个线性组合 $X_i'\beta_1, X_i'\beta_2, \dots, X_i'\beta_d$. 当 $d \ll p$ 时, 就将高维协变量降至低维了.

第一步将利用文献 [8] 的部分充分降维方法——PDEE (partial discretization-expectation estimation), 对多指标部分进行降维. 这种方法不需要规范任何模型, 同时保留完全的回归信息. 具体过程如下:

Y 为响应变量, X 为 $p \times 1$ 维协变量, 是首要感兴趣的预测变量. $W, N(t-), t$ 则是第二感兴趣的预测变量, 记 $\Delta = (N(t-), W, t)$. 部分充分降维的方法就是要在 \mathbb{R}^p 上找出一个所有子空间 S 的交集 $S_{Y|X}^{(\Delta)}$, 且满足: $Y \perp X | (P_s X, \Delta)$, 其中 \perp 表示 Y 与 X 条件独立, P_s 表示内积的投影算子. 满足这个条件的子空间称为部分中心子空间 $S_{Y|X}^{(\Delta)}$, 部分中心子空间的结构维数 $d = \dim(S_{Y|X}^{(\Delta)})$. 分离连续的变量 $\Delta = (N(t-), W, t)$, 使其成为二元变量. 其中 $N(t-) = \{N_1(t-), N_2(t-), \dots, N_n(t-)\}$, $W = (W_1, W_2, \dots, W_n)$, $t = (t_1, t_2, \dots, t_n)$. 定义 $r = (N_i(t-), W_i, t_i)$ 是在个体 i 的某个历史观测时间 t 的任意一组观测值. 于是定义 $\Delta(r) = (I_{\{N(t-) \leq N_i(t-)\}}, I_{W \leq W_i}, I_{t \leq t_i})$, 把样本容量划分为: 当 $\{Y_i, X_i\}_{\Delta(r)=\{1,1,1\}}$ 时, 定义样本容量为 n_1 . 由此可知, 样本容量划分为 $n = \sum_{i=1}^8 n_i$. 接下来, 估计出部分中心子空间 $S_{Y|X}^{(\Delta)}$, 进而确定结构维数 d . 对于任意给定的 r , 定义 $M(r)$ 是 $p \times p$ 的半正定矩阵, R 是 Δ 的独立副本, M 是部分中心降维子空间, 其中 $M = E[M(R)]$, $\text{Span}\{M(r)\} = S_{Y|X}^{(\Delta(r))}$, 那么 $\text{Span}\{M\} = S_{Y|X}^{(\Delta)}$. 则需估计出 M 和 $M(r)$, 对于任意给定的 $r \in \Delta$ 可以得到对 $M(r)$ 的 \sqrt{n} 相合估计 $M_n(r)$. 对于 $M_n(r)$ 的估计有许多不同的方法可以采用. 例如, Chiaromonte 等 [9] 提出的部分切片逆回归 (SIR), Li 和 Wang [10] 提出的部分方向回归 (DR), Zhu 等 [11] 提出的累积均值估计 (CUME) 等. 本文采用 Zhu 等 [11] 提出的累积均值估计方法, 得到 $M_n(r_i)$ 的估计.

对已划分的 8 块样本使用累积均值估计 (CUME) 方法, 从而估计出基 B_1, B_2, \dots, B_8 , 其中 $B_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{id_i})$, $i = 1, 2, \dots, 8$. 于是有矩阵 $M_n(r_i) = \hat{B}_{r_i} \hat{B}_{r_i}^T$, 其中 $\hat{B} = (\hat{B}_1, \hat{B}_2, \dots, \hat{B}_8)$. M 的估计则通过求 $M(r)$ 在所有随机向量 R 上的期望得到, 即 $M = E[M(R)]$. 最后, 从 R 中选取 l_n 个元素为 $\{r_1, r_2, \dots, r_{l_n}\}$, 于是 M_n 可以被下式所估计:

$$\lim_{l_n \rightarrow \infty} \frac{1}{l_n} \sum_{i=1}^{l_n} M(r_i) = E[M(R)] = \lim_{l_n \rightarrow \infty} M_n,$$

M_n 作为 M 的估计, 对矩阵 M_n 进行谱分解. 利用矩阵 M_n 最大的特征值所对应的特征向量去估计部分中心子空间 $S_{Y|X}^{(\Delta)}$.

第二步, 采用文献 [8] 中提出的修正的 BIC 方法估计结构维数 d .

$$\hat{d} = \arg \max_{k \in \{1, 2, \dots, p\}} \frac{n \sum_{m=1}^k [\ln(\hat{\lambda}_m + 1) - \hat{\lambda}_m]}{2 \sum_{m=1}^k [\ln(\hat{\lambda}_m + 1) - \hat{\lambda}_m]} - 2C_n \times \frac{k(k+1)/2}{p},$$

其中 $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_p$ 为矩阵 $M_{n,n}$ 的特征向量, C_n 是一个罚常数, $k(k+1)/2$ 为自由参数. 在本文的实例分析中, 我们取 $C_n = n^{1/3}p^{2/3}$.

通过部分充分降维, 得到协变量 X_i 的线性组合: $X_i'\beta_1, X_i'\beta_2, \dots, X_i'\beta_d$, $d < p$, 从而模型 (2) 转化为标准的部分线性可加模型.

接下来, 利用 Manzan 和 Zerom^[12] 提出的核估计方法估计模型 (2) 中的回归系数 α 和联系函数 $\phi(\cdot)$, 光滑函数 $\mu(\cdot)$. 为表示方便, 令 $\mu(t_i) = m_1(z_{1i})$, $\phi(X_i'\beta_1, X_i'\beta_2, \dots, X_i'\beta_d) = m_2(z_{2i})$, 其中 $z_{1i} = t_i$, $z_{2i} = X_i'\beta$, 令 w_j 表示除了 z_j 以外的所有 z 变量的元素. 定义

$$\gamma(z_j, w_j) = \frac{p_z(z_j)p_w(w_j)}{p(z_j, w_j)}, \quad j = 1, 2, \dots, q.$$

其中 $p_z(\cdot)$ 和 $p_w(\cdot)$ 分别是 z_j 和 w_j 的概率密度函数, $p(\cdot)$ 是 $z = (z_j, w_j)$ 的联合概率函数. 令

$$h_j^{(Y)}(z_{ji}) \equiv E[\gamma(z_{ji}, w_{ji})Y_i | z_{ji}],$$

则 $h_j^{(Y)}(z_j) = \int Y p_w(w_j) dw_j$, 从而得出可加模型 $h^{(Y)}(z) = \sum_{j=1}^q h_j^{(Y)}(z_j)$ 是 Y 的渐近估计.

同理, $h^{(W)}(z) = \sum_{j=1}^q h_j^{(W)}(z_j)$ 是 W 的渐近估计. 则

$$h_j^{(Y)}(z_{ji}) = m_j(z_{ji}) + \alpha' N_i(t_i -) [h_j^{(W)}(z_{ji})], \quad (3)$$

其中,

$$h_j^{(Y)}(z_{ji}) = \frac{1}{(n-1)h} \sum_{l \neq i}^n K\left(\frac{z_{jl} - z_{ji}}{h}\right) \frac{\hat{p}_w(w_{jl})}{\hat{p}(z_{jl}, w_{jl})} Y_l,$$

这里 $K(\cdot)$ 为核函数, h 为窗宽, $\hat{p}(\cdot)$ 和 $\hat{p}_w(\cdot)$ 分别是 z 和 w_j 的概率密度函数的估计值, 因此, 可解得 $h^{(Y)}(z) = \sum_{j=1}^q h_j^{(Y)}(z_j)$, 同理可求得 $h^{(W)}(z)$.

接下来用 (2) 式减去 (3) 式的 q 维相加可得:

$$Y_i - h^{(Y)}(z_{ji}) = \alpha' N_i(t_i -) [W_i - h^{(W)}(z_{ji})],$$

于是可用 OLS 解得 $\hat{\alpha}$, 代入 (3) 式可得:

$$\begin{aligned} \hat{\mu}(t) &= m_1(z_{1i}) = h_1^{(Y)}(z_{1i}) - \hat{\alpha}' N_i(t_i -) h_1^{(W)}(z_{1i}), \\ \hat{\phi}(\cdot) &= m_2(z_{2i}) = h_2^{(Y)}(z_{2i}) - \hat{\alpha}' N_i(t_i -) h_2^{(W)}(z_{2i}). \end{aligned}$$

2) 基于最小平均方差估计 (MAVE)

使用 MAVE 方法对模型 (2) 进行估计, 需要同时对回归函数 $E(Y | X)$ 即 $E(Y | X^T B)$ 和均值中心降维子空间 $S_{E(Y|X)}$ 进行估计, 即 B 的解要满足:

$$\min_B \{E[Y - E(Y | X^T B)]^2\}. \quad (4)$$

对于任意的标准正交矩阵 $B = (\beta_1, \beta_2, \dots, \beta_d)$, $X^T B$ 的条件方差

$$\sigma_B^2(X^T B) = E\{[y - E(y | X^T B)]^2 | X^T B\}, \quad (5)$$

根据重期望法则, 于是有

$$E[y - E(y | X^T B)]^2 = E[\sigma_B^2(X^T B)].$$

那么对于 B , 最小化式 (4) 等同于最小化:

$$E[\sigma_B^2(X^T B)], \quad \text{其中 } B^T B = I. \quad (6)$$

此时, 令 $\psi_B(z_1, z_2, \dots, z_d) = \psi(X^T \beta_1, X^T \beta_2, \dots, X^T \beta_d)$, 对于任意给定的 X_0 , $E(Y_i | X_i^T B)$ 在 X_0 处的局部线性展开式为

$$E(Y_i | X_i^T B) \approx \alpha' N_i(t-) W_i + a + b^T B^T (X_i - X_0) + \mu(t), \quad (7)$$

其中, $a = \psi(X_0^T B)$, $b^T = (b_{(1)}, b_{(2)}, \dots, b_{(d)})$,

$$b_{(k)} = \left. \frac{\partial \psi(z_1, z_2, \dots, z_d)}{\partial z_k} \right|_{z_1=X_0^T \beta_1, z_2=X_0^T \beta_2, \dots, z_d=X_0^T \beta_d}, \quad k = 1, 2, \dots, d.$$

由式 (5) 和 (7), 按照局部线性光滑的思想, $\sigma_B^2(X_0^T B)$ 的估计可以利用近似:

$$\sum_{i=1}^n [Y_i - E(Y_i | X_i^T B)] w_{i0} \approx \sum_{i=1}^n \{Y_i - [a^* + b^T B^T (X_i - X_0) + \hat{\mu}(t)]\}^2 w_{i0}, \quad (8)$$

其中 $\hat{\mu}(t)$ 为 $\mu(t)$ 的估计, $a^* = \alpha' N_i(t-) W_i + a$, $w_{i0} \geq 0$, $\sum_{i=1}^n w_{i0} = 1$. 文献 [13] 中提到, 在寻找有效的降维空间时, 权重 w_{i0} 的选取至关重要, 一般情况下

$$w_{i0} = K_h(B^T (X_i - X_0)) / \sum_{l=1}^n K_h(B^T (X_l - X_0)),$$

其中, $K_h(\cdot) = h^d K(\cdot/h)$, d 是 $K(\cdot)$ 的维数, 为了便于阐述, $K(\cdot)$ 表示为在不同情况下的不同核函数.

由此, 通过式 (8) 可以得到 a 和 b 的估计. 所以, σ_B^2 在 $X_0^T B$ 的估计为

$$\hat{\sigma}_B^2(X_0^T B) = \min_{a, b} \left\{ \sum_{i=1}^n \{Y_i - [a^* + b^T B^T (X_i - X_0) + \hat{\mu}(t)]\}^2 w_{i0} \right\}. \quad (9)$$

Xia 等^[13]指出, 在满足一些温和的条件下, 有 $\hat{\sigma}_B^2(X_0^\top B) - \sigma_B^2(X_0^\top B) = o_p(1)$. 基于式 (4)、(6)、(9), 可以得到 B 的估计, 即需满足下式:

$$\begin{aligned} & \min_{B: B^\top B = I} \left\{ \sum_{j=1}^n \hat{\sigma}_B^2(X_j^\top B) \right\} \\ &= \min_{\substack{B: B^\top B = I \\ a_j, b_j, j=1, 2, \dots, n}} \left\{ \sum_{j=1}^n \sum_{i=1}^n \{Y_i - [a_j^* + b_j^\top B^\top (X_i - X_j) + \hat{\mu}(t)]\}^2 \right\} w_{ij}, \end{aligned} \quad (10)$$

其中 $b_j^\top = (b_{j1}, b_{j2}, \dots, b_{jd})$.

对于一个给定 B_0 , 首先估计出 $\mu(t)$. $\mu(t)$ 、 α 的估计, 方法与部分充分降维中的连接函数估计相同, 这里不再出. 得到 $\hat{\mu}(t)$ 的估计后, 再利用式 (9) 进行两步估计, 进而求得收敛的 B .

在文献 [13] 中提到, 权重 w_{ij} 的选取若与 B 有关, 那么式 (10) 的解是非平凡的, 且式 (10) 中权重的本质是 X_i 与 X_j 距离的函数. 本文根据 Xia 等^[13]提出的两种方法选取权重, 求解出权重的精确估计, 然后代入式 (10), 并通过最小化式 (10) 对 B 再次估计, 重复上述过程, 直至 B 收敛.

在得到了中心降维子空间的基方向后, 接下来需要对中心降维子空间的结构维数进行估计. 本文采用 Xia 等^[13]提出的改进的交叉验证法估计中心降维子空间的结构维数. 假定 d 为中心降维子空间的结构维数, 定义

$$\hat{a}_{d0,j} = \frac{\sum_{i=1, i \neq j}^n K_{h_d}^{(i,j)} [y_i - \hat{\mu}(t)]}{\sum_{i=1, i \neq j}^n K_{h_d}^{(i,j)}},$$

其中 $K_{h_d}^{(i,j)} = \{\hat{\beta}_1^\top (X_i - X_j), \hat{\beta}_2^\top (X_i - X_j), \dots, \hat{\beta}_d^\top (X_i - X_j)\}$. 并定义

$$CV(d) = n^{-1} \sum_{j=1}^n [y_j - \hat{\mu}(t) - \hat{a}_{d0,j}]^2, \quad d = 1, 2, \dots, p$$

为 y_j 与 y_j 的核估计的残差. Xia 等^[13]指出, 最佳窗宽 $h_d \sim n^{-1/(d+4)}$. 因此, d 的估计 \hat{d} 可以表示为:

$$\hat{d} = \arg \min_{0 \leq d \leq p} \{CV(d)\}.$$

§4. 实例分析

为了说明模型 (2) 的适用程度, 本文选取了慢性心力衰竭患者的医疗费用数据, 该数据来自于弗吉尼亚大学卫生系统的临床数据存储库. 该项研究包括 1475 名年龄在 60 岁到 89 岁的患者, 他们在 2004 年第一次被检查出患有心力衰竭并接受治疗. 对于每一个病人, 所观察到的信息包括临床就诊次数, 每次就诊的医疗费用和一些协变量. 在 Virginia 部门统计的死亡证明数据中发现, 病人的最终入院记录是 2006 年 7 月 31 日. 初步研究表明, 随

着患者访问医院的次数增多, 往往患者的医疗费用也随即增多, 同时这些患者也有更高的死亡率. 利用模型 (2) 分析该数据, 寻找协变量的线性组合从而进行降维. 先简单描述数据结构:

1. 个体 i 每次观测到的时间 t_i (单位: 月)
2. 个体 i 在时间 t_i 之前的观测总数 $N_i(t_i-)$ ($N_i(t_i-) > 0$)
3. 个体 i 每次观测到的医疗费用 $Y_i(t_i)$
4. 性别 X_{1i} (0 表示女性, 1 表示男性)
5. 种族 X_{2i} (0 表示非白种人, 1 表示白种人)
6. 年龄 X_{3i} (单位: 岁)
7. 住院服务 X_{4i} (0 表示没有, 1 表示有)
8. 随访期间死亡 X_{5i} (0 表示死亡, 1 表示存活)
9. 随访时间 X_{6i}

其中, 医疗费用作为因变量, 性别、种族、年龄、住院服务、随访期间死亡以及随访时间作为可能影响医疗费用的变量, 除此之外, 我们还选取了 $N_i(t_i) * X_{1i}$ 作为可能影响医疗费用的交互项, 并且取 $N_i(t_i) = [N_i(t_i-) - 10]/8$, $W_i = X_{1i}$, 取年龄的平方项代替年龄这项因素.

1) 基于部分充分降维的核估计

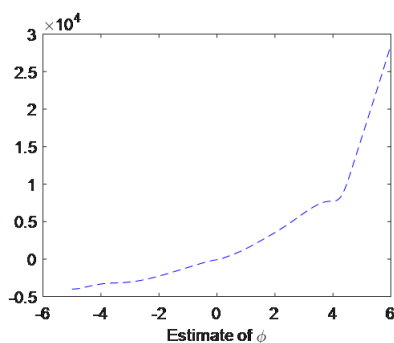
对数据使用部分充分降维方法. 首先标准化所有协变量, 便于消除变量间量纲的差异. 接着使用部分充分降维的方法使协变量降至低维, 并得到了估计的部分中心子空间的结构维数 $d = 1$ 和基方向 $\hat{\beta}$, 进而找出了协变量的线性组合 $X^T \beta$. 最后, 使用核估计方法估计联系数函数 $\phi(\cdot)$, 并得到回归系数 $\hat{\alpha}$.

表 1 基方向 $\hat{\beta}$ 与回归系数 $\hat{\alpha}$ 的值

β_1	β_2	β_3	β_4	β_5	β_6	α
0.0023	-0.0693	-0.0044	0.9680	0.2410	0.0070	0.1560

由表 1 和图 1, 我们可以得到以下结论:

1. 在降维得到的线性组合中, β_1 、 β_2 、 β_3 分别表示性别、种族、年龄的平方的系数, 分别为 0.0023, -0.0693, -0.0044, 结合图 1, 多指标的联系函数为近似单调递增函数, 可以得出性别正相关于因变量, 种族和年龄负相关于因变量. 这里 0 表示女性, 1 表示男性, 表明男性的平均医疗花费多于女性. 0 表示非白种人, 1 表示白种人, 表明非白种人的平均医疗花费多于白种人, 且医疗费用随着年龄平方的增长而下降. 这点与文献 [5] 的结论相同.

图 1 ϕ 的估计曲线

2. 在降维得到的线性组合中, β_4 、 β_5 、 β_6 分别表示住院服务、随访期间死亡、随访时间的系数, 分别为 0.9680, 0.2410, 0.0070. 结合图 1, 得出住院服务, 随访期间死亡, 随访时间皆正相关与因变量. 这里住院服务 0 表示没有, 1 表示有, 表明有住院服务的患者医疗花费比没有住院服务的患者高. 随访期间死亡 0 表示死亡, 1 表示存活, 表明在随访这些患者医疗费用的过程中, 存活的患者的医疗花费比死亡的患者高, 且随访时间越久, 患者花费越多. 在文献 [5] 中, 没有选取这三个变量.

3. 回归系数 α 表示历史观测总数与性别的交互项的系数, 为 0.1560. 表明此交互项与医疗费用呈正相关, 得出患者访问医院的次数越多, 医疗花费越高. 本文选取了文献 [5] 其中一个交互项, 两者结论吻合.

2) 基于最小平均方差估计 (MAVE)

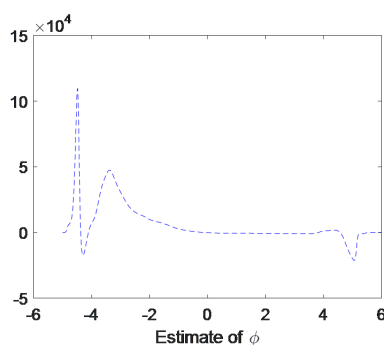
对数据使用 MAVE 方法. 首先标准化所有协变量, 便于消除变量间量纲的差异. 通过降维, 根据 $CV(d)$ 最小, 选出结构维数 $d = 1$. 从而得到部分中心子空间的基 β 与回归系数 α 的值.

表 2 基方向 β 与回归系数 α 的值

β_1	β_2	β_3	β_4	β_5	β_6	α
0.1515	0.2223	0.7911	-0.4373	-0.1896	-0.2731	0.1348

由表 2 和图 2, 可以得到以下结论:

1. 各协变量系数的估计值如表 2 所示, β_1 、 β_2 、 β_3 分别表示性别、种族、年龄的平方的系数, 分别为 0.1515, 0.2223, 0.7911. 结合图 2, 发现多指标函数可以分为 3 段分析. 首先第一段变化: 单调递增快速达到第一个峰值即图像最高点后迅速单调递减, 在达到极小值点后继续单调递增达到第二个小峰值, 这一段的函数图像波动变化非常明显. 第二段变化: 在函数图像达到第二个小峰值后, 函数图像持续单调递减并且达到图像最低点. 第三段变

图 2 ϕ 的估计曲线

化: 在达到最小值后函数图像又小幅上升. 在性别、种族、年龄的平方的系数三项中, 得出年龄这一变量对医疗费用影响最大, 其次是种族与性别, 这两项变量的影响效果要弱于年龄这一变量.

2. 在表 2 中, β_4 、 β_5 、 β_6 分别表示住院服务、随访期间死亡、随访时间的系数, 分别为 -0.4373, -0.1896, -0.2731. 在这三项变量中, 能够得出住院服务对医疗费用的影响最大, 其次是随访时间, 最后是随访期间死亡.

3. 在表 2 中, α 表示历史观测总数与性别的交互项的系数, 为 0.1348. 表明医疗费用也与患者访问医院的次数有关, 且患者访问医院的次数对其医疗费用有一定的影响.

§5. 总 结

在本篇文章中, 将文献 [5] 中的模型扩展至可含高维协变量的医疗费用评估模型, 使得联系函数非参数化, 模型从而更具灵活性, 并且将高维协变量纳入考虑之中. 在模型参数的估计上, 使用了两种不同的估计方法, 部分充分降维与 MAVE. 可以看出, 在实例分析中, 两种估计方法均对数据做了很好的分析, 找出了主要影响医疗费用的变量. 针对本文采用的高维纵向数据, Kong 和 Xia^[14] 提出了一种分位数回归的降维方法. 可以在部分线性多指标可加模型的基础上使用分位数回归方法, 作为日后的研究方向.

参 考 文 献

- [1] ZHOU X H, LIANG H. Semi-parametric single-index two-part regression models [J]. *Comput Statist Data Anal*, 2006, **50**(5): 1378–1390.
- [2] CASTELLI C, COMBESURE C, FOUCHER Y, et al. Cost-effectiveness analysis in colorectal cancer using a semi-Markov model [J]. *Stat Med*, 2007, **26**(30): 5557–5571.
- [3] LIU L, WOLFE R A, KALBFLEISCH J D. A shared random effects model for censored medical costs and mortality [J]. *Stat Med*, 2007, **26**(1): 139–155.

- [4] SUN L Q, SONG X Y, ZHOU J, et al. Joint analysis of longitudinal data with informative observation times and a dependent terminal event [J]. *J Amer Statist Assoc*, 2012, **107(498)**: 688–700.
- [5] ZHAO X Q, DENG S R, LIU L, et al. Sieve estimation in semiparametric modeling of longitudinal data with informative observation times [J]. *Biostatistics*, 2014, **15(1)**: 140–153.
- [6] CHEN J S, KIM I, TERRELL G R, et al. Generalised partial linear single-index mixed models for repeated measures data [J]. *J Nonparametr Stat*, 2014, **26(2)**: 291–303.
- [7] CHEN J S, LIU L, SHIH Y C T, et al. A flexible model for correlated medical costs, with application to medical expenditure panel survey data [J]. *Stat Med*, 2016, **35(6)**: 883–894.
- [8] FENG Z H, WEN X R M, YU Z, et al. On partial sufficient dimension reduction with applications to partially linear multi-index models [J]. *J Amer Statist Assoc*, 2013, **108(501)**: 237–246.
- [9] CHIAROMONTE F, COOK R D, LI B. Sufficient dimensions reduction in regressions with categorical predictors [J]. *Ann Statist*, 2002, **30(2)**: 475–497.
- [10] LI B, WANG S L. On directional regression for dimension reduction [J]. *J Amer Statist Assoc*, 2007, **102(479)**: 997–1008.
- [11] ZHU L P, ZHU L X, FENG Z H. Dimension reduction in regressions through cumulative slicing estimation [J]. *J Amer Statist Assoc*, 2010, **105(492)**: 1455–1466.
- [12] MANZAN S, ZEROM D. Kernel estimation of a partially linear additive model [J]. *Statist Probab Lett*, 2005, **72(4)**: 313–322.
- [13] XIA Y C, TONG H, LI W K, et al. An adaptive estimation of dimension reduction space [J]. *J R Stat Soc Ser B Stat Methodol*, 2002, **64(3)**: 363–410.
- [14] KONG E F, XIA Y C. An adaptive composite quantile approach to dimension reduction [J]. *Ann Statist*, 2014, **42(4)**: 1657–1688.

Multi-index Additive Model and Its Application in Medical Cost Forecast

PAN Qing ZHAO Xiaobing

(School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, 310018, China)

Abstract: Modeling analysis and reasonable prediction of medical costs are the basis and foundation for the determination of medical insurance costs. High-dimensional additional information in medical costs plays an important role in long-term prediction. This paper proposes a partial linear multi-indicator additive model to fit and predict longitudinal medical cost data with high-dimensional features and uses two different dimensionality reduction estimation methods to estimate the model and applies the model to a set of high-dimensional dimensions. The longitudinal medical cost data of the variable is used for case analysis.

Keywords: partial linear multi-indicator additive model; longitudinal medical costs; partial sufficient dimension reduction; minimum average variance estimation

2020 Mathematics Subject Classification: 62N02; 65G05