随机效应半参数二值响应模型的估计研究①

孙 燕[1,2]

(1. 上海财经大学经济学院，上海 200433; 2. 上海财经大学数理经济学重点实验室，上海 200433)

摘要: 为了研究诸如在收入差距和健康关系中某些解释变量的影响效应可能依赖于其他解释变量的状况，降低可能存在的模型设定和遗漏变量偏误，本文提出了随机效应半参数二值响应模型，其中非参数的设定还可用于数据的初探性分析. 随后本文探讨了模型非参数和参数部分的估计. 由于似然函数中随机效应和非参数的存在加大了估计难度，为此本文采用 B 样条方法逼近非参数部分，同时将随机效应视为缺失数据，利用 EM 算法和 MCMC 方法建立了模型参数的估计，并研究了其一致性. 模拟研究结果表明估计量在有限样本下的表现良好，最后将模型运用于收入差距与健康关系的实例研究中，结果表明数据支持收入差距弱假说.

关键词: 随机效应半参数二值响应模型; B 样条估计; EM 算法; MCMC 算法; 收入差距假说

中图分类号: F224.0 **文献标识码**: A **文章编号**: 1007 - 9807(2015) 11 - 0059 - 11

0 引 言

众所周知，虽然经济理论告知变量之间可能存在的关系但并没有告知确切的函数形式. 自 20 世纪 70 年代以来，关于收入、收入差距与健康状况关系的研究受到了越来越多经济学家和社会学家的关注. 早期的研究指出收入对健康有正向影响[1]，然而自 90 年代以来有研究表明相对收入或收入差距同样是健康的决定因素[2]，并逐步形成了收入差距强假说和收入差距弱假说[3,4]. 概括而言收入差距强假说表明收入差距对健康的影响效应不依赖于收入，即无论个体本身的收入状况如何，收入差距对每个个体的影响是相同的; 而收入差距弱假说则表明收入差距对健康的不利影响随个体收入的增加而减少，即收入差距对健康的影响效应依赖于收入.

中国在过去的 20 余年间收入差距迅速扩大，由此带来的问题是日益扩大的收入差距是否会对健康的改善带来不利的影响. 现有文献在探讨中国的收入差距对健康的影响时，均采用分层抽样获得的微观数据，基于不同的模型对收入差距假说进行了检验，如 Li 和 Zhu[5] 在线性 probit 模型中分别引入收入和收入的二次形式来验证收入差距强假说，又通过引入收入与收入差距的交叉项来验证收入差距弱假说. 封进和余央央[6] 也用类似的方法检验了中国农村地区收入差距与健康的关系.

事实上，为了避免模型设定偏误，检验收入差距对健康状况的影响是否依赖于收入，本文考虑把感兴趣的收入差距影响效应设定为收入的未知函数. 同时考虑到可得数据的特点，将提出一类随机效应半参数二值响应模型，并将建立模型中未知部分的估计，讨论其一致性，同时将借助蒙特卡洛模拟方法研究估计量在有限样本下的表现，最后将模型应用于我国收入差距与健康关系的研究中.

① 收稿日期: 2013 - 03 - 12; 修订日期: 2014 - 04 - 15.
基金项目: 教育部“新世纪优秀人才支持计划”资助项目(NECT - 10 - 0562); 国家自然基金资助项目(11271242); 上海财经大学“讲席教授”资助项目.
作者简介: 孙 燕(1977—)，女，上海人，博士，教授. Email: sunyan@ mail. shufe. edu. cn

1 模型及文献综述

本文提出的模型具有如下形式

$$\Pr(Y_{ij}=1\mid U_{ij},Z_{ij},X_{ij},e_j)=F\left(\alpha_{00}(U_{ij})+\sum_{l=1}^{p}Z_{ijl}\alpha_{0l}(U_{ij})+X_{ij}^{T}\beta_0+e_j\right)$$
$$j=1,2,\cdots,m;\ i=1,2,\cdots,n_j \qquad (1)$$

其中 Y_{ij} 表示第 j 个组中第 i 个观测值取 0 或 1 的二值响应变量; $U_{ij}\in R$, $Z_{ij}=(Z_{ij1},\cdots,Z_{ijp})^{T}\in R^{p}$, $X_{ij}\in R^{q}$ 为解释变量; $\alpha_{0l}(\cdot)\in R, l=0,\cdots,p$ 为未知真实光滑函数,称为函数系数,表示解释变量的影响效应可能依赖于其他解释变量, $\beta_0\in R^{q}$ 为真实回归系数; $e_j\in R$ 为随机效应,表示不可观测或没有观测到的组特征遗漏变量; $F(\cdot)$ 为 logstic 或标准正态分布函数.

若将模型(1)应用于我国收入差距与健康关系研究中,则 Y_{ij} 表示第 j 个社区中第 i 个个体的健康状况是否良好; $U_{ij}\in R$ 表示个体收入; 未知函数 $\alpha_{00}(\cdot)$ 体现了收入和健康之间可能存在的非线性关系[7]; $Z_{ij}\equiv Z_j\in R$ 表示社区收入差距,其系数 $\alpha_{01}(\cdot)$ 取为收入的未知函数; $X_{ij}\equiv(Q_{ij}^{T},W_j^{T})^{T}$,其中 $Q_{ij}\in R^{p_1}$ 表示其他个体特征解释变量,包括年龄、教育等; $W_j\in R^{p_2}$ 表示其他社区特征解释变量,如社区位于农村还是城市等; 随机效应 $e_j\in R$ 表示其他不可观测或没有观测到的社区变量,如该社区的环境状况等. 显见,当样本估计结果有证据显示 $\alpha_{01}(u)$ 为常数时,支持收入差距强假说. 而当 $\alpha_{01}(u)$ 为非常数函数时,支持收入差距弱假说. 特别是如果 $\alpha_{01}(u)=a+bu$ 为线性函数时,表明在模型中加入交叉项的做法不会引起模型设定偏误,是合理的. 可见,本文提出的模型可用于初步探索实证分析中参数模型设定的合理性,提供实证模型设定是否准确的数据证据,这种非参数的灵活设定在变量间的关系尚存争议时非常有用.

本文称模型(1)为随机效应半参数二值响应模型,具体地,当 $F(\cdot)$ 取 logistic 分布时,也称为随机效应半参数 logit 模型,而当 $F(\cdot)$ 为标准正态分布时,称为随机效应半参数 probit 模型,是笔者在文献中没有见到过的. 该模型中既含有参数分量又含有非参数分量,它把与响应变量有明确关系的这部分解释变量设为参数形式,而把与响应变量关系不够明确的或者是研究感兴趣的这部分解释变量设为非参数形式,结合了参数和非参数模型的优点,既充分利用了数据中的信息,又能防止模型设定偏误可能导致的感兴趣变量关系估计的非一致性. 而随机效应的引入能刻画同组数据间的相关性、减少遗漏变量偏差等,该处理方法常见于面板数据建模中[8,9]. 因此,模型(1)适用于分群数据或面板数据相关问题的分析,具有广泛的适用性.

显见,当模型(1)中取 $\alpha_{00}(u)=a_0+b_0u$, $\alpha_{0l}(u)=a_l, l=1,\cdots,p$,且不存在随机效应 e_j 时,模型(1)就化为了实证研究中常见的线性 logit 或 probit 计量模型[10,11],该类模型的估计一般采用最大似然方法. 而当随机效应 e_j 存在时,即使 $\alpha_{00}(\cdot),\alpha_{0l}(\cdot)$ $(l=1,\cdots,p)$ 仍为参数形式,也很难获得模型参数的最大似然估计. 这是由于随机效应 e_j 的存在,导致参数的最大似然估计需要最大化下面的积分

$$L(\theta)=\int f(y\mid e)\,p_e(e)\,\mathrm{d}e \qquad (2)$$

其中 $f(y\mid e)$ 为给定解释变量和随机效应 e 时 y 的概率密度函数, $p_e(e)$ 为给定解释变量时随机效应的概率密度函数, θ 为分布中的未知参数. 此时,式(2)中的积分没有解析式,故参数估计首先就是要处理该棘手的积分,如 Anderson 和 Aitkin[12] 利用 Gauss-Hermite 积分法得到了似然函数的近似值,然后通过最大化该近似似然函数导出了参数的最大似然估计; Wei 和 Tanner[13] 则采用了 Monto Carlo 积分法; 而 Raudenbush 等[14] 采用了高阶 Laplace 近似方法; 等.

本文将研究模型(1)中未知函数系数 $\alpha_{0l}(u), l=0,1,\cdots,p$ 和参数 β_0 的估计. 由于模型未知部分的似然估计也必然涉及类似式(2)中的积分,且本文的模型中含有非参数部分,因而将给模型的估计带来更大的挑战.

2 随机效应半参数二值响应模型的估计

本文假定模型(1)中的随机效应 e_j 是在给定

解释变量条件下从总体 e 中抽取的独立同分布样本,服从 $N(0,\sigma_e^2)$,其中 $\sigma_e^2>0$ 未知; 且给定解释变量和随机效应 e_j 条件下,Y_{ij} 相互独立,显然同一组中的 Y_{ij} 之间由于随机效应的存在是非独立的; 进一步本文假定不同组之间是独立的.

2.1　非参数的 B 样条逼近方法

关于非参数估计方法,通常有样条估计和局部多项式估计方法[15],由于 B 样条方法具有良好的局部性质和稳定的数值解[16],本文将基于 B 样条方法估计模型(1)中的非参数部分. 即,对于定义在闭区间 $[a,b]$ 上的未知光滑函数 $\alpha_{0l}(u)$ $(l=0,\cdots,p)$,本文将采用 B 样条函数进行逼近. 设 $\Re_l$ 是由 B 样条基函数$\{B_{ls}(u),s=1,\cdots,K_l\}$ 张成的线性空间,其中 K_l 是该线性空间的维数,它可用于调节待估函数的光滑程度. 假定对每个 $\alpha_{0l}(u)$ $(l=0,\cdots,p)$,存在 $\alpha_l(u)=\sum_{s=1}^{K_l}B_{ls}(u)\gamma_{ls}^*\in\Re_l$,使得

$$\alpha_{0l}(u)\approx\sum_{s=1}^{K_l}B_{ls}(u)\gamma_{ls}^*,u\in[a,b] \tag{3}$$

令 $\boldsymbol{B}(u)=\begin{pmatrix}B_{01}(u)&\cdots&B_{0K_0}(u)&0&\cdots&0&0&\cdots&0\\ &\vdots&&&\vdots&&&\vdots&\\ 0&\cdots&0&0&\cdots&0&B_{p1}(u)&\cdots&B_{pK_p}(u)\end{pmatrix}$, $\tilde{Z}_{ij}=(1,Z_{ij}^{\mathrm{T}})^{\mathrm{T}}$

$R_{ij}=(\tilde{Z}_{ij}^{\mathrm{T}}B(U_{ij}),X_{ij}^{\mathrm{T}})^{\mathrm{T}}$,$\gamma^*=(\gamma_0^{*\mathrm{T}},\cdots,\gamma_p^{*\mathrm{T}})^{\mathrm{T}}$,$\gamma_l^*=(\gamma_{l1}^*,\cdots,\gamma_{lK_l}^*)^{\mathrm{T}}$,$\theta_0=(\gamma^{*\mathrm{T}},\beta_0^{\mathrm{T}})^{\mathrm{T}}$. 于是基于样本数据$\{Y_{ij}\}$ 的似然函数为

$$\prod_{j=1}^{m}\left\{\int_{e_j}\left[\prod_{i=1}^{n_j}f(Y_{ij}\mid e_j;\theta)\right]p_e(e_j;\sigma_e^2)\,\mathrm{d}e_j\right\} \tag{5}$$

其中 $f(y\mid e)$ 为给定解释变量和随机效应时 y 的概率密度函数(当 $F(\cdot)$ 为 logistic 分布函数时为 logsitic 密度函数,如果 $F(\cdot)$ 为标准正态分布时为标准正态密度函数),$p_e(\cdot)$ 为给定解释变量时随机效应的正态概率密度函数. 正如前文指出的,随机效应的存在导致似然函数(5)没有解析式,必须借助其它近似计算方法,且式(5)涉及多维积分,尤其是当 $F(\cdot)$ 取为标准正态时难以处理,因而难以直接获得参数的最大似然估计.

本文将采用间接方法求解式(5)的对数形式的最优化问题,即将随机效应视为缺失数据,利用 EM 算法求解该最大化问题,其介绍可参见 Fahrmeir 和 Tutz[17] 的附录. 下记 $Y_e=(e_1,\cdots,e_m)^{\mathrm{T}}$,$Y=(Y_1^{\mathrm{T}},\cdots,Y_m^{\mathrm{T}})^{\mathrm{T}}$,$Y_j=(Y_{1j},\cdots,Y_{n_jj})^{\mathrm{T}}$,EM 算法考虑如下基于完全数据$(Y,Y_e)$ 的完全对数似然函数

$$l(\theta,\sigma_e^2)\equiv\sum_{j=1}^{m}\{\ln f(Y_j\mid e_j;\theta)+\ln p_e(e_j;\sigma_e^2)\} \tag{6}$$

具体算法如下:

E-step

$$Q(\theta,\sigma_e^2\mid\theta^{(k)},\sigma_e^{2(k)})=E[l(\theta,\sigma_e^2)\mid Y;\theta^{(k)},\sigma_e^{2(k)}]$$

其中 $\theta^{(k)}$ 和 $\sigma_e^{2(k)}$ 为第 k 次迭代的估计. 直接计算并利用概率密度函数的归一化性质可得

$$E[l(\theta,\sigma_e^2)\mid Y;\theta^{(k)},\sigma_e^{2(k)}]=\sum_{j=1}^{m}\int_{e_j}\{\ln f(Y_j\mid e_j;\theta)\}p(e_j\mid Y_j;\theta^{(k)},\sigma_e^{2(k)})\,\mathrm{d}e_j+\sum_{j=1}^{m}\int_{e_j}\{\ln p_e(e_j;\sigma_e^2)\}p(e_j\mid Y_j;\theta^{(k)},\sigma_e^{2(k)})\,\mathrm{d}e_j \tag{7}$$

其中 $p(\cdot\mid\cdot)$ 为给定解释变量时随机效应的后验

其中对不同的 $\alpha_{0l}(u)$ 可取不同的 K_l,表明允许不同的未知函数有不同的光滑程度. 而 K_l 可利用交叉验证法进行选取,由于其只取整数值,故大大降低了交叉验证法的计算量,具体讨论见后. 于是对任意的 $l=0,\cdots,p$,先获取 γ_{ls}^*,$s=1,\cdots,K_l$ 的估计 $\hat{\gamma}_{ls}$,然后代入式(3)就能得到 $\alpha_{0l}(u)$ 的估计. 这里需要说明的是,不同的样条基函数可张成相同的空间 $\Re_l$,从而对应不同的 $\hat{\gamma}_{ls}$,但得到的 $\alpha_{0l}(u)$ 估计是相同的,即由样条基函数张成的线性空间 $\Re_l$ 唯一确定了 $\alpha_{0l}(u)$ $(l=0,\cdots,p)$ 的估计.

将式(3)代入模型(1)可得

$$\Pr(Y_{ij}=1\mid U_{ij},Z_{ij},X_{ij},e_j)\approx F\Big(\sum_{s=1}^{K_0}B_{0s}(U_{ij})\gamma_{0s}^*+\sum_{l=1}^{p}\sum_{s=1}^{K_l}Z_{ijl}B_{ls}(U_{ij})\gamma_{ls}^*+X_{ij}^T\beta_0+e_j\Big)$$
$$j=1,2,\cdots,m;\ i=1,2,\cdots,n_j \tag{4}$$

2.2　基于 EM-MCMC 算法的估计方法

分布. 由该式可见,第一项只与 θ 有关而与 σ_e^2 无关,记为 $Q_1(\theta \mid \theta^{(k)},\sigma_e^{2(k)})$,第二项只与 σ_e^2 有关而与 θ 无关,记为 $Q_2(\sigma_e^2 \mid \theta^{(k)},\sigma_e^{2(k)})$. 因此 M-step 归结为:

M-step

$$\theta^{(k+1)} = \arg\max_{\theta} Q_1(\theta \mid \theta^{(k)},\sigma_e^{2(k)})$$

$$\sigma_e^{2(k+1)} = \arg\max_{\sigma_e^2} Q_2(\sigma_e^2 \mid \theta^{(k)},\sigma_e^{2(k)})$$

重复上述 E-step 和 M-step 直至收敛就能得到 θ 和 σ_e^2 的估计,其中在 M-step 中,本文只需寻找某 $\theta^{(k+1)}$ 满足

$$Q_1(\theta^{(k+1)} \mid \theta^{(k)},\sigma_e^{2(k)}) > Q_1(\theta^{(k)} \mid \theta^{(k)},\sigma_e^{2(k)})$$

即可[18]. 为此,关于 θ,本文采用 Newton-Raphson 一步迭代算法

$$I(\theta^{(\text{old})} \mid \theta^{(k)},\sigma_e^{2(k)})(\theta^{(\text{new})} - \theta^{(\text{old})}) = \frac{\partial Q_1(\theta \mid \theta^{(k)},\sigma_e^{2(k)})}{\partial \theta}\Big|_{\theta=\theta^{(\text{old})}} \tag{8}$$

其中

$$I(\theta \mid \theta^{(k)},\sigma_e^{2(k)}) = \sum_{j=1}^{m}\int_{e_j} f_1(e_j;\theta)\, p(e_j \mid Y_j;\theta^{(k)},\sigma_e^{2(k)})\,\mathrm{d}e_j,$$

$$f_1(e_j;\theta) = -E\left\{\frac{\partial^2 \ln f(Y_j \mid e_j;\theta)}{\partial\theta\partial\theta^T}\Big| \{U_{ij},Z_{ij},X_{ij},e_j\}\right\}$$

将 $\theta^{(\text{old})}$ 取为 $\theta^{(k)}$,基于式(4)和直接求导可得

$$\theta^{(k+1)} = \theta^{(k)} + \left\{\sum_{j=1}^{m} E_{e_j|Y_j}(\boldsymbol{R}_j^{\mathrm{T}}\boldsymbol{\omega}_j\boldsymbol{R}_j)\right\}^{-1} \times \left\{\sum_{j=1}^{m} E_{e_j|Y_j}\left[\boldsymbol{R}_j^{\mathrm{T}}\boldsymbol{\omega}_j\dot{\boldsymbol{\eta}}_j(\boldsymbol{Y}_j-\boldsymbol{\mu}_j)\right]\right\}\Big|_{\theta=\theta^{(k)},\sigma_e^2=\sigma_e^{2(k)}} \tag{9}$$

其中 $\boldsymbol{R}_j = (R_{1j},\cdots,R_{n_j j})^{\mathrm{T}}$, $\boldsymbol{\mu}_j = (\mu_{1j},\cdots,\mu_{n_j j})^{\mathrm{T}}$, $\mu_{ij} = F(\eta_{ij})$, $\eta_{ij} = R_{ij}^{\mathrm{T}}\theta + e_j$, $\dot{\boldsymbol{\eta}}_j$ 为对角元素分别为 $\dot{\eta}_{1j},\cdots,\dot{\eta}_{n_j j}$ 的对角阵, $\dot{\eta}_{ij} = \partial\eta_{ij}/\partial\mu_{ij}$, $\boldsymbol{\omega}_j$ 为对角元素分别为 $\omega_{1j},\cdots,\omega_{n_j j}$ 的对角阵, $\omega_{ij} = (\partial\mu_{ij}/\partial\eta_{ij})^2/\mathrm{var}(Y_{ij} \mid e_j)$, $\mathrm{var}(Y_{ij} \mid e_j)$ 为给定解释变量和随机效应下的方差, $E_{e_j|Y_j}(\cdot)$ 为给定解释变量条件下随机效应后验分布的期望,文中关于后验分布期望的计算,采用常用的 Metropolis-Hastings 算法的 MCMC 方法[19].

关于 M-step 中 σ_e^2 的部分,最大化 $Q_2(\sigma_e^2 \mid \theta^{(k)},\sigma_e^{2(k)})$ 可得

$$\sigma_e^{2(k+1)} = \frac{1}{m}\sum_{j=1}^{m} E(e_j^2 \mid Y_j;\theta^{(k)},\sigma_e^{2(k)}) = \frac{1}{m}\sum_{j=1}^{m} E_{e_j|Y_j}(e_j^2) \tag{10}$$

2.3 估计算法概括

步骤 1 选取 θ,σ_e^2 的初始值 $\theta^{(0)}$ 和 $\sigma_e^{2(0)}$,如可取 $\theta^{(0)}$ 为全零向量, $\sigma_e^{2(0)} = 0.1$(非零值);

令 $k = 1$

步骤 2 用 Metropolis-Hastings 算法的 MCMC 抽样方法得到 $p_{e_j|Y_j}(e \mid Y;\theta^{(k)},\sigma_e^{2(k)})$ 的 N 个随机样本 $e_{j1},\cdots,e_{jN}$;

步骤 3 利用步骤 2 中的这 N 个样本基于样本均值近似计算式(9)中的条件期望可得 $\theta^{(k+1)}$;同理近似计算式(10)中的条件期望可得 $\sigma_e^{2(k+1)} = \frac{1}{mN}\sum_{j=1}^{m}\sum_{t=1}^{N} e_{jt}^2$;

步骤 4 如果迭代收敛,则 $\theta^{(k)},\sigma_e^{2(k)}$ 即为最终估计,分别记为 $\hat{\theta},\hat{\sigma}_e^2$,将 $\hat{\theta}$ 中对应的分量代入式(3)即得 $\alpha_{0l}(u)$ $(l = 0,1,\cdots,p)$ 的估计 $\hat{\alpha}_l(u)$;否则令 $k = k+1$,回到步骤 2 继续迭代.

上述算法中迭代收敛的准则取为:对事先给定的任意小的数 $\varepsilon_0 > 0$,当 $k+1$ 步的估计值 $\hat{\vartheta}^{(k+1)}$ 和 k 步的估计值 $\hat{\vartheta}^{(k)}$ 满足

$$\|\hat{\vartheta}^{(k+1)} - \hat{\vartheta}^{(k)}\| / \|\hat{\vartheta}^{(k)}\| < \varepsilon_0 \tag{11}$$

时迭代终止,这里的 $\|\cdot\|$ 表示向量的 2 范数.

2.4 光滑参数 $K_0,\cdots,K_p$ 的选取

对于未知函数 $\alpha_{0l}(u)$,在其定义区间 $[a,b]$ 上取 J_l 个内节点: $a < u_1 < \cdots < u_{J_l} < b$,则 $K_l = J_l + m + 1$, m 为样条基函数的次数(degree), J_l 为节点个数,可随样本容量的增大而增大. 用 B 样条函数来逼近未知函数时,首要的是节点的选择,包括节点位置和节点个数两个方面. 关于节点的位置,常见的有两种方法:均匀节点和非均匀节点(主要是等间隔的样本分位点). 比如设回归变量 U 在 $[0,1]$ 上取值,若选择 3 个均匀节点,则 3 个内节点分别是 0.25,0.50,0.75;若选择等间隔样本分位数的非均匀节点时,则节点为 U 的第 25,50,75 分位点,这样可确保 $[0,1]$ 中每个子区间上

的样本点个数基本相同.

光滑参数的选取,常用方法有交叉验证法(CV)、广义交叉验证法(GCV)、AIC、BIC 等. 本文将基于 O'Sullivan 等[20] 定义的 GCV 选取光滑参数

$$\mathrm{GCV}(K_0,\cdots,K_p) = \min_{K_0,\cdots,K_p} \frac{1}{n} \times \sum_{j=1}^{m}\sum_{i=1}^{n_j} \left\{ \frac{[Y_{ij} - E(\hat{\mu}_{ij})] / \sqrt{\mathrm{var}(\hat{Y}_{ij})}}{1 - (K_0 + \cdots + K_p + q)/n} \right\}^2 \quad (12)$$

其中期望和方差都是给定解释变量条件下的,$\mathrm{var}(\hat{Y}_{ij}) = E(\hat{\mu}_{ij})(1 - E(\hat{\mu}_{ij}))$,期望本文将利用 1 000 次的 Monte Carlo 方法进行计算.

3　估计量的一致性及其证明

为了度量函数系数估计的表现,令 $g(u) = (g_0(u),\cdots,g_p(u))^{\mathrm{T}}$,其中 $g_l(u)(l = 0,\cdots,p)$ 为区间 $[a,b]$ 上的平方可积实函数,定义 $\| g(u) \|_{L_2} = \left\{ \sum_{l=0}^{p} \int_a^b [g_l(u)]^2 \mathrm{d}u \right\}^{1/2}$ 为向量函数 $g(u)$ 在区间 $[a,b]$ 上的 L_2 范数. 为简化证明且不失一般性,设 $\{B_{ls}(u), s = 1,\cdots,K_l\}$ 为线性空间 $\Re_l$ 中的一组标准正交基. 并记 $K_n = \max_{0\leqslant l\leqslant p} K_l$. 下面将基于以下条件给出函数系数 $\alpha_{0l}(u)$ $(l = 0,\cdots,p)$ 的 B 样条估计和参数 β_0 估计的一致性.

条件 1　$\alpha_{0l}(u), l = 0,\cdots,p$ 的二阶导数 $\alpha''_{0l}(u)$ 在区间 $[a,b]$ 上连续;

条件 2　U 在 $[a,b]$ 上取值,其密度函数 $p_U(\cdot)$ 非零有界;

条件 3　对不同的 i, j, (Z_{ij}, U_{ij}, X_{ij}) 独立同分布来自于总体 (Z,U,X);且随机效应 e_j 是从总体 e 中取出的独立同分布样本;$E(VV^{\mathrm{T}} \mid U = u)$ 可逆且其所有特征根非零有界,其中 $V = (1,Z^{\mathrm{T}},X^{\mathrm{T}})^{\mathrm{T}}$.

注: 条件 1 确保一定能够找到在 $[a,b]$ 上一致收敛于待估函数 $\alpha_{0l}(\cdot)$ 的 B 样条函数. 即对区间内的一切 x,存在 B 样条函数逼近于该待估函数且误差以一定速度趋于零,参见 Schumaker[21]. 而条件 2 中 U 的所有可能取值对应该待估函数的定义域,简单起见,同多数文献(Wang 等[22],等),本文设条件 1 和 2 成立. 若条件 2 改为 U 在某 Γ 上取值,则条件 1 中要改为在 Γ 上存在某一 B 样条函数一致收敛于待估函数,但该条件在实际应用中不易验证.

定理 1　若条件 1 - 3 满足,$0 < \max_j n_j < \infty$, $K_n/n \to 0$,则当 $n \to \infty$ 时有

$$\| \hat{\theta} - \theta_0 \| = \{ \| \hat{\gamma} - \gamma^* \|^2 + \| \hat{\beta} - \beta_0 \|^2 \}^{1/2} = O_p(\{K_n/n\}^{1/2}).$$

证明　由 EM 算法可见参数 θ 的估计是以下似然方程的解

$$\frac{\partial E\left(\left[\sum_{j=1}^{m} \ln f(Y_j \mid e_j;\theta) \right] \mid Y \right)}{\partial \theta} = 0$$

记 $\ell_n(\theta) = \frac{1}{n} \sum_{j=1}^{m} E_{e_j \mid Y_j} \{ \ln f(Y_j \mid e_j;\theta) \}$,其中 $E_{e_j \mid Y_j}(\cdot)$ 同前,表示给定 Y_j 和解释变量时随机效应 e_j 的条件期望,$S_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta}$, $H_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}$. 利用积分型余项的 Taylor 展开式可得

$$\ell_n(\theta) = \ell_n(\theta_0) + (\theta - \theta_0)^{\mathrm{T}} S_n(\theta_0) + (\theta - \theta_0)^{\mathrm{T}} \times \left\{ \int_0^1 (1-t) H_n(\theta_0 + t(\theta - \theta_0)) \mathrm{d}t \right\} (\theta - \theta_0)$$

可以证明(见后)(i) 对任意常数 $M_1 > 0$,有 $\lim_{c\to\infty} \lim_{n\to\infty} P(\| S_n(\theta_0) \| \geqslant M_1 c (K_n/n)^{1/2}) = 0$;(ii) 存在概率等于 1 的集合 Ω_0 和某一常数 $M_2 > 0$,当 $n \to \infty$ 时在 Ω_0 上几乎处处成立

$$(\theta - \theta_0)^{\mathrm{T}} \left\{ \int_0^1 (1-t) H_n(\theta_0 + t(\theta - \theta_0)) \mathrm{d}t \right\} \times (\theta - \theta_0) \leqslant -M_2 \| \theta - \theta_0 \|^2, \text{a. s.}$$

于是由(i) 可得,对任意的 $\theta \in R^{K_0+K_1\cdots+K_p+q}$ 且 $\| \theta - \theta_0 \| = c(K_n/n)^{1/2}$,取(ii) 中的常数 $M_2 > 0$,当 n 充分大时有, $\mathrm{P}\{ | (\theta - \theta_0)^{\mathrm{T}} S_n(\theta_0) | < M_2 c^2 (K_n/n) \} = 1$. 结论(ii) 和 Taylor 展开式可得 n 充分大时有 $\mathrm{P}\{ l_n(\theta) < l_n(\theta_0)$,对任意的 $\theta \in R^{K_0+K_1\cdots+K_p+q}$ 且 $\| \theta - \theta_0 \| = c(K_n/n)^{1/2} \} = 1$. (*)

下证 $H_n(\theta)$ 为负定矩阵. 显见 $H_n(\theta)$ 为独立和的形式,故基于独立和的大数定律可知

$$H_n(\theta) \xrightarrow{P} H(\theta)$$

其中 $H(\theta) = -E\{ [\tilde{Z}^{\mathrm{T}} B(U), X^{\mathrm{T}}]^{\mathrm{T}} [\tilde{Z}^{\mathrm{T}} B(U), X^{\mathrm{T}}]^{\mathrm{T}} g(\eta) \}$, $\eta = \tilde{Z}^{\mathrm{T}} B(U) \gamma + X^{\mathrm{T}} \beta + e$, $\tilde{Z} = (1, Z^{\mathrm{T}})^{\mathrm{T}}$,

$g(\eta) = f^2(\eta)/\mathrm{var}(Y \mid e) = f^2(\eta)/\{F(\eta)(1-F(\eta))\}$，显然 $g(\cdot)$ 非负有界.

取 $\tilde{\boldsymbol{B}}(U) = \begin{pmatrix} B(U) & 0 \\ 0 & I_q \end{pmatrix}$，其中 I_q 表示 q 阶单位矩阵，则对任意的非零列向量 $\tilde{\gamma} = (\gamma_1^{\mathrm{T}}, \gamma_2^{\mathrm{T}})^{\mathrm{T}}$，由标准正交基的性质和条件 2 - 3 可知，存在正常数 M_3，有

$$\tilde{\gamma}^{\mathrm{T}} E(\tilde{B}^{\mathrm{T}}(U) VV^{\mathrm{T}} \tilde{\boldsymbol{B}}(U))\tilde{\gamma} = \int \tilde{\gamma}^{\mathrm{T}} \tilde{\boldsymbol{B}}^{\mathrm{T}}(u) E(VV^{\mathrm{T}} \mid U = u) \tilde{\boldsymbol{B}}(u) \tilde{\gamma} p_U(u)\,\mathrm{d}u \geqslant M_3 \{\int \gamma_1^{\mathrm{T}} B^{\mathrm{T}}(u) B(u) \gamma_1 \mathrm{d}u + \gamma_2^{\mathrm{T}}\gamma_2\} = M_3 \| \tilde{\gamma} \|^2$$

故可得 $-H(\theta)$ 为正定矩阵，于是 $H_n(\theta)$ 为负定矩阵，因此 $l_n(\theta)$ 为凹函数，故 $\hat{\theta}$ 存在且由上述(*)式知，当 n 充分大有 $P\{\| \hat{\theta} - \theta_0 \| \leqslant c(K_n/n)^{1/2}\} = 1$，取 c 充分大即可得 $\| \hat{\theta} - \theta_0 \| = O_p(\{K_n/n\}^{1/2})$.

(i) 的证明　由不同 j 之间的独立性、Cauchy-Schwartz 不等式、Jensen 不等式、条件期望的性质、密度函数与条件方差的有界性、标准正交样条基的性质和条件 2 - 3 可得，存在正常数 M_4

$$E(\| S_n(\theta_0) \|^2) = \sum_{l=0}^{K_0+\cdots+K_p+q} E\left\{\frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n_j} E_{e_j|Y_j}[R_{ijl}(Y_{ij} - F(R_{ij}^{\mathrm{T}}\theta_0 + e_j)) f(R_{ij}^{\mathrm{T}}\theta_0 + e_j)/\mathrm{var}(Y_{ij} \mid e_j)]\right\}^2$$
$$\leqslant \frac{1}{n^2}\sum_{l=0}^{K_0+\cdots+K_p+q}\sum_{j=1}^{m} n_j \sum_{i=1}^{n_j} E[R_{ijl}(Y_{ij} - F(R_{ij}^{\mathrm{T}}\theta_0 + e_j)) f(R_{ij}^{\mathrm{T}}\theta_0 + e_j)/\mathrm{var}(Y_{ij} \mid e_j)]^2$$
$$\leqslant \frac{M_4 K_n}{n^2}\sum_{j=1}^{m}\sum_{i=1}^{n_j} E(R_{ij}^{\mathrm{T}} R_{ij}) = O(\{K_n/n\})$$

于是由 Chebyshev 不等式即得式(i).

(ii) 的证明　由独立和的强大数定律知存在概率等于 1 的集合 Ω_0，当 $n \to \infty$ 时，在 Ω_0 上几乎处处成立 $H_n(\theta) = H(\theta)$. 前面已证 $\boldsymbol{A} \equiv -H(\theta)$ 为正定矩阵. 记 $\lambda_{\min}(\boldsymbol{A})$ 为矩阵 $\boldsymbol{A}$ 的最小特征根，则存在常数 $M_5 > 0$ 满足 $\lambda_{\min}(\boldsymbol{A}) \geqslant M_5$，于是可得对任意与 $\boldsymbol{A}$ 维数相同的非零列向量 δ，成立 $\delta^{\mathrm{T}}\boldsymbol{A}\delta \geqslant \lambda_{\min}(\boldsymbol{A})\delta^{\mathrm{T}}\delta \geqslant M_5\delta^{\mathrm{T}}\delta$，因此当 $n \to \infty$ 时，在 Ω_0 上几乎处处有

$$(\theta-\theta_0)^{\mathrm{T}}\left\{\int_0^1 (1-t) H_n(\theta_0 + t(\theta-\theta_0))\,\mathrm{d}t\right\}(\theta - \theta_0) \leqslant -M_5\int_0^1 (1-t)\,\mathrm{d}t \| \theta-\theta_0 \|^2 \equiv -M_2 \| \theta-\theta_0 \|^2$$

即(ii)成立. 定理证毕.

定理 2　若条件 1 - 3 满足，$0 < \max_j n_j < \infty$，$K_n \to \infty$，$K_n/n \to 0$，则当 $n \to \infty$ 时，对任意给定的 $u \in [a,b]$ 有

$$\| \hat{\alpha}(u) - \alpha(u) \|_{L_2} \xrightarrow{P} 0, \text{且 } \hat{\beta} \xrightarrow{P} \beta_0$$

证明　由 Schumaker[21] 和条件 1 可得，

$$\sup_{u\in[a,b]} | \alpha_{0l}(u) - \sum_{s=1}^{K_l} B_{ls}(u)\gamma_{ls}^* | = O(K_l^{-2}).$$

基此并由标准正交基的性质和定理 1 得

$$\| \hat{\alpha}(u) - \alpha_0(u) \|_{L_2} \leqslant \| B(u)\hat{\gamma} - B(u)\gamma^* \|_{L_2} + \| B(u)\gamma^* - \alpha_0(u) \|_{L_2} \leqslant \| \hat{\gamma} - \gamma^* \| + \left\{\sum_{l=0}^{p}[\sup_{u\in[a,b]} | \sum_{s=1}^{K_l} B_{ls}(u)\gamma_{ls} - \alpha_{0l}(u) |]^2\right\}^{1/2} = O_P(\{K_n/n\}^{1/2} + K_n^{-2}).$$

结论 $\hat{\beta} \xrightarrow{P} \beta_0$ 由定理 1 直接可得.

注：由定理 2 可见本文提出的随机效应半参数二值响应模型中 $\alpha_{0l}(\cdot)$ 的 B 样条估计和 β_0 的估计满足一致性.

4　蒙特卡洛模拟实验研究

数值模拟例子

$$\Pr(Y_{ij} = 1 \mid U_{ij}, Z_j, X_{ij}, W_j, e_j) = F(\alpha_1(U_{ij}) + Z_j\alpha_2(U_{ij}) + \beta_1 X_{ij} + \beta_2 W_j + e_j) \tag{13}$$

其中 $\alpha_1(u) = 1 + 2u(1-u)$，$\alpha_2(u) = 0.1\exp(-0.5 + 3u)$，$\beta_1 = -2$，$\beta_2 = -1$，$U_{ij}$ 取自 $(0,1)$ 均匀分布，$Z_j \sim N(0.5, 0.6^2)$，$X_{ij} \sim N(1, 0.5^2)$，$W_j \sim B(1, 0.4)$，随机效应 $e_j \sim N(0, 0.5^2)$. 在这个例子中，分别从该模型中独立抽取样本容量

$m = 50, 100, n_j = 4$ 的样本,在每种样本容量组合下重复抽取100次,生成样本容量相同的100个模拟数据集. 基于文中的算法步骤,采用均匀节点,基于 GCV 值,本文取二次 B 样条函数. 在 MCMC 方法中,取建议分布为随机效应的先验正态分布,当 F 取 logistic 分布函数时非参数部分的估计结果见图 1,F 为 probit 分布时的结果类似. 参数估计的结果见表 1,总体上本文提出的估计在有限样本下表现较好,且随着样本容量的增大,估计结果更精确.

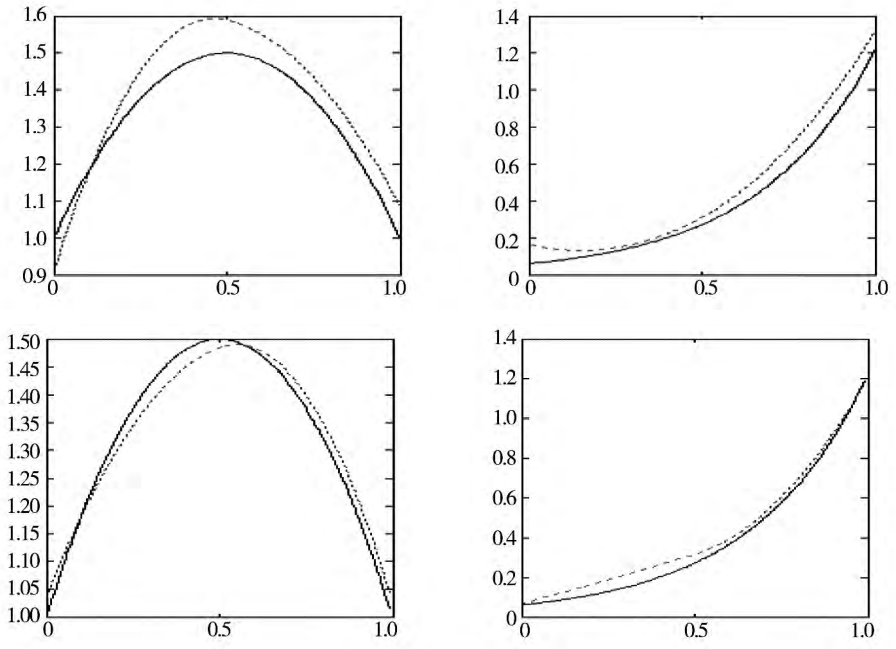

图 1　左右图分别为 $\alpha_1(u)$ 和 $\alpha_2(u)$ 的 B 样条估计(虚线)及其真值(实线)
(上半部分对应 $m = 50$,下半部分对应 $m = 100$)

Fig. 1 B spline estimation (dashed line) of $\alpha_1(u)$ (left) and $\alpha_2(u)$ (right) and their true values (real line)
(the top half part with $m = 50$, the lower half part with $m = 100$)

表 1　模拟结果

Table 1 Simulation results

样本容量	$m = 50, n_j = 4$		$m = 100, n_j = 4$	
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
F 取 logistic 分布	-2.153 (0.422)	-1.057 (0.447)	-2.021 (0.287)	-1.016 (0.313)
F 取标准正态分布	-1.928 (0.351)	-0.970 (0.304)	-1.940 (0.251)	-1.034 (0.230)

注: 表中参数估计值为 100 次估计的平均值,括号内为 100 次估计的标准差. 结果由 Matlab 编程运算得到.

5　在收入差距与健康关系中的应用

5.1　样本数据和变量说明

本文使用的数据来自于由北卡罗来纳大学人口研究中心和中国疾病控制与预防中心合作开展的中国健康与营养调查(CHNS)②. 现阶段该调查覆盖了辽宁、黑龙江、江苏、山东、河南、湖北、湖南、广西和贵州 9 个省. 每省的样本家庭是采用多阶段随机分层抽样获得的,具体方案是每省分别随机抽取2个城市和4个农村,其中每个城市又随机抽取4个街区,而每个农村随机抽取4或5个

② 参见 http://www.cpc.unc.edu/projects/China.

村庄，本文将街区和村庄统一定义为社区，最后从每个社区中随机抽取约 20 户家庭进行健康与营养调查. 本文的样本来自 2006 年的调查数据.

这里采用自评健康作为衡量健康的标准，虽然自评健康具有很强的主观性，却是常用的指标之一，它能有力地反应个人的健康状况[2,3]. 本文令自评健康状况非常好和好时健康指标取 1，一般和差时取 0. 每个社区的收入差距采用按户计算的 Gini 系数度量，即按户总收入排序，以调查户数倒数为权重计算. 其他个体和社区控制变量基本同 Li 和 Zhu[5]，具体描述见表 2.

限于相关个体的自评健康和人口特征变量（年龄、性别、婚姻状况、教育、家庭收入、家庭环境、家庭人数等）数据没有缺失的年龄在 16 岁 – 65 岁的有效个人样本. 另外，由于本文要计算收入差距指标，因此排除家庭收入非正的且社区中只含有一个收入为正的家庭的样本，最后得到的有效样本容量为 6 044 个个体，覆盖 217 个社区.

表 2 其他控制变量的定义

Table 2 Definitions of the other control variables

变量	定义	变量	定义
收入	家庭人均收入，单位: 元	教育	接受正规教育的年数
年龄	样本只限于 16 岁 – 65 岁的成人，参考了劳动力年龄标准	家庭规模	家庭人口数
性别指示变量	男性取 1，女性取 0	家庭环境指示变量	房屋周围没有排泄物取 1，其余情形取 0
婚姻指示变量	在婚取 1，未婚、离异和丧偶取 0	农村指示变量	社区是农村则取 1，否则取 0

5.2 估计结果及分析

基于模型(1) 且不防取 $F(\cdot)$ 为 logsitic 分布函数，采用文中的算法步骤，其中考虑到收入分布的非均匀性，取等间隔的样本分位数作节点并用 GCV 选取节点个数，利用二次 B 样条函数，且在 MCMC 方法中，取建议分布为随机效应的先验正态分布，B 样条估计结果见图 2. 由图 2 中的右图可见收入差距对健康影响效应 $\alpha_{01}(u)$ 的估计呈现出非线性的形式，该估计结果支持收入差距弱假说，即收入差距对穷人和富人而言其影响是不同的，并且随着收入差距的扩大，收入增加更加有助于健康的改善，收入下降则更加不利于健康的改善. 这可能是由于收入差距的扩大会使低收入人群更不易对自身进行医疗卫生与教育投资. 此外，收入差距的扩大会增加低收入人群的挫败感及压力，引起吸烟和酗酒等不良习惯，从而恶化其健康水平. 收入函数 $\alpha_{00}(u)$ 的估计表明收入与健康之间基本呈类似倒 U 的非线性关系，其严格的统计检验还需作进一步研究. 其余控制变量对健康的影响效应如预期，比如教育、良好的环境有利于健康，见表 3.

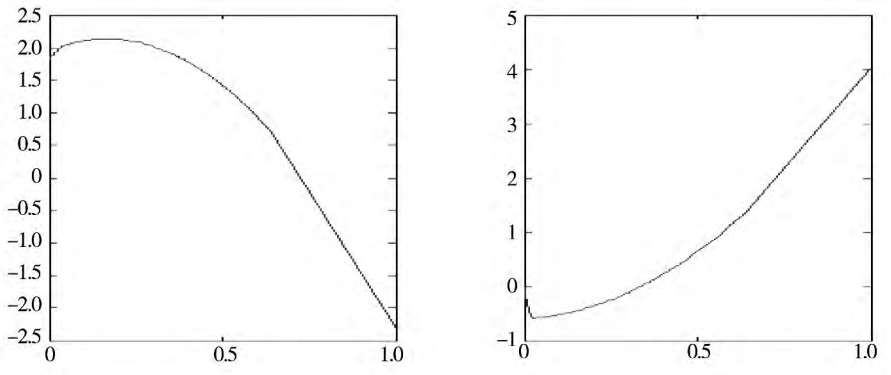

图 2 左图为 $\alpha_{00}(u)$ 的 B 样条估计，右图为 $\alpha_{01}(u)$ 的 B 样条估计③

Fig. 2 B spline estimation of $\alpha_{00}(u)$ (left) and $\alpha_{01}(u)$ (right)

③ 非参数估计时本文根据收入的最大最小值将其化为了有限区间 $[0,1]$ 内的取值，利于 B 样条的估计.

表 3　随机效应半参数 logit 模型的参数估计结果

Table 3 Parameter estimations of the semi-parametric logit model with random effects

变量		变量		变量	
年龄	-0.049 9	性别指示变量	0.307 8	婚姻指示变量	0.314 1
教育	0.023 7	家庭规模	-0.017 0	家庭环境指示变量	0.075 1
农村指示变量	0.307 0	随机效应标准差		0.665 8	

6　结束语

为了研究某些尚存争议的变量之间的关系，本文提出了一类随机效应半参数二值响应模型，既充分挖掘了数据中的信息，又能防止模型设定偏误可能导致的感兴趣变量关系估计的非一致性，具有很大的灵活性，并且随机效应的引入能够反应不可观测或没有观测到的社区特征变量对响应变量的影响，刻画同组数据间的相关性，降低可能存在的遗漏变量偏误等.

本文随后探讨了模型的估计问题，基于 B 样条的非参数估计方法，在由于随机效应的存在造成似然函数中积分难以求解的困境时，将随机效应视为缺失数据，采用 EM 算法，结合 Newton-Raphson 和 MCMC 方法得到了模型函数系数的 B 样条估计和回归系数的估计，并证明了其一致性. 模拟研究结果表明估计方法在有限样本下表现良好，最后将模型运用于我国收入差距和健康关系的研究中，结果表明样本数据支持收入差距弱假说，即收入差距对穷人和富人而言其影响是不同的，当收入差距的扩大致使贫者愈贫、富者愈富时，其产生的收入效应加剧了健康不平等的现象. 另外非参数的设定也可用于数据的初探性分析，如在收入差距和健康实例中，本文发现数据很可能支持 U 型函数的设定，关于其严格的统计检验还需要作进一步深入的研究.

虽然模型(1) 是在研究收入差距与健康关系中引入的，但事实上该模型为许多实际问题的研究提供了一个很自然的分析框架. 如在劳动经济学中，许多学者研究发现教育的边际收益随受教育程度的不同而变化[24]. 如果工作经验也是雇主考虑的一个重要因素，那么教育的边际收益也应该随着经验的不同而变化[25]. 因此，为避免模型设定偏误，在研究教育与个体收入高低状况的关系时，可以假定教育的边际收益是经验的未知函数; 又比如在个体失业状况的研究中，在发现农民工“用工荒”现象的同时又观察到了大学生毕业即失业的现象，因此某些影响因素，如教育对失业的影响效应可能随工作类别的不同而有所差别. 因此，本文提出的随机效应半参数二值响应模型具有广泛的应用前景.

最后，需要说明的是，若模型(1) 中的组效应与个体解释变量相关，则同 Chamberlain[26] 和 Mundlak[27]，可将效应假定为解释变量组均值的线性回归函数: $e_j = (\overline{U}_j, \overline{X}_j^{\mathrm{T}})\pi + v_j$，其中 $\overline{U}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} U_{ij}, \overline{X}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$，$\pi$ 为回归系数，v_j 与所有解释变量独立. 显然，本文提出的估计方法可直接应用于该情形. 故虽然本文仅考虑了随机效应模型，但可推广到类似 Chamberlain[26] 和 Mundlak[27] 的固定效应模型.

参 考 文 献:

[1] Grossman Michael. On the concept of health capital and the demand for health [J]. Journal of Political Economy, 1972, 80(2): 223-255.

[2] Waggstaff A, Doorslaer E V. Income inequality and health: What does the literature tell us? [J]. Annual Review of Public Health, 2000, 21: 543-567.

[3] Wilkinson Richard G. Unhealthy Societies: The Affliction of Inequality [M]. London: Routledge Press, 1996.

[4] Wilkinson Richard G. Health inequalities: Relative or absolute material standards? [J]. British Medical Journal, 1997, 314: 591 – 595.

[5] Li Hongbin, Zhu Yi. Income, income inequality, and health: Evidence from China [J]. Journal of Comparative Economics, 2006, 34(4): 668 – 693.

[6] 封 进, 余央央. 中国农村的收入差距与健康 [J]. 经济研究, 2007, 1: 79 – 88.
Feng Jin, Yu Yangyang. Income inequality and health in rural China [J]. Economic Research Journal, 2007, 1: 79 – 88. (in Chinese)

[7] Feinstein Jonathan S. The relationship between socioeconomic status and health: A review of the literature [J]. The Milbank Quarterly, 1993, 71: 279 – 322.

[8] Wooldridge Jeffrey. 计量经济学导论(第三版)英文版 [M]. 北京: 清华大学出版社, 2007.
Wooldridge Jeffrey. Introductory Econometrics (Third Edition) English Version [M]. Beijing: Tsinghua University Press, 2007. (in Chinese)

[9] 刘莉亚, 丁剑平, 覃 筱, 等. 面板数据计量模型适应性的比较研究 [J]. 管理科学学报, 2011, 14(2): 86 – 96.
Liu Liya, Ding Jianping, Qin Xiao, et al. Simulation comparison on adaptability of econometrical models based on panel data [J]. Journal of Management Sciences in China, 2011, 14(2): 86 – 96. (in Chinese)

[10] 姜旭平, 王 鑫. 影响搜索引擎营销效果的关键因素分析 [J]. 管理科学学报, 2011, 14(9): 37 – 45.
Jiang Xuping, Wang Xin. Study on the core factors impacting search engine marketing [J]. Journal of Management Sciences in China, 2011, 14(9): 37 – 45. (in Chinese)

[11] 肖作平. 公司治理影响债务期限结构类型吗—来自中国上市公司的经验证据 [J]. 管理工程学报, 2010, 24(1): 110 – 123.
Xiao Zuoping. Does corporate governance affect the type of debt maturity structure? Empirical evidence from Chinese listed companies [J]. Journal of Industrial Engineering and Engineering Management, 2010, 24(1): 110 – 123. (in Chinese)

[12] Anderson D A, Aitkin M. Variance component models with binary response: Interviewer variability [J]. Journal of the Royal Statistical Society Series B, 1985, 47: 204 – 210.

[13] Wei G C G, Tanner M A. A Monte Carlo implementation of the EM algorithm and the poor man's augmentation algorithms [J]. Journal of the American Statistical Association, 1990, 86: 669 – 795.

[14] Raudenbush Stephen W, Yang Meng-Li, Yosef Matheos. Maximum likelihood for generalized linear models with nested random effects via high-order multivariate Laplace approximation [J]. Journal of Computational and Graphical Statistics, 2000, 9(1): 141 – 157.

[15] 叶阿忠. 非参数计量经济学 [M]. 天津: 南开大学出版社, 2007.
Ye Azhong. Nonparametric Econometrics [M]. Tianjin: Nankai University Press, 2007. (in Chinese)

[16] Boor C D E. A Practical Guide to Splines [M]. New York: Springer Press, 1978.

[17] Fahrmeir Ludwig, Tutz Gerhard. Multivariate Statistical Modeling Based on Generalized Linear Models [M]. New York: Springer-Verlage, 1994.

[18] Bickel P J. One-step huber estimates in the linear models [J]. Journal of the American Statistical Association, 1975, 350(70): 428 – 433.

[19] 朱新玲. 马尔科夫链蒙特卡洛方法研究综述 [J]. 统计与决策, 2009, 21: 151 – 153.
Zhu Xinling. Research summary on MCMC methods [J]. Statistics and Decision, 2009, 21: 151 – 153. (in Chinese)

[20] O'Sullivan F, Yandell B, Raynor W. Automatic smoothing of regression functions in generalized linear models [J]. Journal of the American Statistical Association, 1986, 81: 96 – 103.

[21] Schumaker L L. Spline Functions [M]. New York: Wiley, 1981.

[22] Wang Li, Liu Xiang, Liang Hua, et al. Estimation and variable selection for generalized additive partial linear models [J]. 2011, 39: 1827 – 1851.

[23] Gerdtham U G, Johannesson M, Lundberg L, et al. The demand for health: Result from new measures of health capital

[J]. European Journal of Political Economy, 1999, 15: 501 – 521.

[24] Schultz T P. Human capital, schooling and health, IUSSP, XXIII [C]. General Population Conference, Yale University, 1997.

[25] Card D. Estimating the return to schooling: Progress on some persistent ecnometric problems [J]. Econometrica, 2001, 69: 1127 – 1160.

[26] Chamberlain G. Analysis of covariance with qualitative data [J]. Review of Economic Studies, 1980, 47: 225 – 238.

[27] Mundlak Yair. On the pooling of time series and cross section data [J]. Econometrica, 1978, 46: 69 – 85.

Estimation for semi-parametric binary response model with random effects

SUN Yan[1,2]

1. School of Economics, Shanghai University of Finance and Economics, Shanghai 200433, China;
2. Key Laboratory of Mathematical Economics, Shanghai University of Finance and Economics, Shanghai 200433, China

Abstract: In order to study that the marginal effect of some explanatory variables may depend on other covariates such as in the problem of income inequality on health, we propose a semi-parametric binary response model with random effects which can reduce both model specification and omitted variable bias. Its nonparametric specification can also be used to the exploratory data analysis. Then the estimation procedures are established, where the difficulties lies in the evaluation of the intractable integral caused by random effects, and the existence of nonparametric parts makes it even tougher. We solve it by using EM and MCMC algorithm together with non-parametric B spline method. The estimators are shown to be consistent. Simulation studies show the good properties of the estimators. A practical application of our model supports the weak version of income inequality hypothesis.

Key words: semi-parametric binary response model with random effects; B spline estimation; EM algorithm; MCMC algorithm; income inequality hypothesis