文章编号:1000-2367(2023)03-0097-08 DOI:10.16366/j.cnki.1000-2367.2023.03.012

随机缺失下阈值和灵敏性的联合估计

程伟丽[1],吴莹[2],左卫兵[1]

(1.华北水利水电大学 数学与统计学院,郑州 450046;2.云南大学 数学与统计学院,昆明 650500)

摘 要:在观察性试验研究中,某些受试者的诊断检验值可能缺失,若只用完全观测的数据可能会得到有偏的估计.考虑在诊断检验值随机缺失的情形下,基于灵敏性和特异性相等的对称点准则将逆概率加权和多重插补相结合建立起包含待估参数的非光滑估计方程,采用非参数两样本经验似然方法给出阈值和灵敏性的联合估计和置信域.在一定的正则条件下,建立了阈值和灵敏性联合估计的渐近性质.模拟研究证实所提方法的估计要优于其他方法的估计.

关键词:随机缺失;阈值;灵敏性;对称点准则;经验似然

中图分类号:O212 **文献标志码**:A

在诊断医学中,受试者测试特征曲线(简称为 ROC 曲线)是常用的衡量连续型诊断检验诊断能力的一个统计工具.随着阈值的变化,正确诊断有病体的概率—灵敏性(S_e)和正确诊断无病体的概率—特异性(S_p)会向相反方向变化,以 $1-S_p$ 为横坐标,以 S_e 为纵坐标,将这些点连起来就构成了 ROC 曲线.有关 ROC 曲线可以参看文献[1].

通过一个连续性诊断检验值来诊断受试者是有病或者是无病,需要选择一个阈值,最优阈值 τ 的选择非常重要.设有病体生物指标 X 和无病体生物指标 Y 的分布函数分别是 $F(\cdot)$ 和 $G(\cdot)$,不失一般性,假设生物指标值越大越易患病.给定阈值 τ,则灵敏性和特异性可表示为 $\theta(\tau)=\Pr(X>\tau)=1-F(r)$ 和 $\eta(\tau)=\Pr(Y\leq\tau)=G(\tau)$. 通常选择在高特异性下的阈值,一方面这会使得阈值随着特异性的变化而变化,另一方面高特异性下的灵敏性不一定很高.为选择一个最优的阈值,现有文献提出了以下几种方法.文献[2]提出基于最大化正确分类两总体概率之和的约当指标的基础上选择最优阈值,即 $\tau=\arg\max_\tau|\{\theta(\tau)+\eta(\tau)-1\}|$. 文献[3]提出用 ROC 曲线中最接近最好点(0,1)的方法选择最优阈值,即

$$\tau=\arg\max_\tau\sqrt{[1-\eta(\tau)]^2+[\theta(\tau)-1]^2}.$$

文献[4]提出最大化灵敏性和特异性乘积的方法来确定阈值 $\tau=\arg\max_\tau[\theta(\tau)\eta(\tau)]$.文献[5]提出基于两个总体被正确分类概率的基础上选择阈值,即 $\tau=\arg_\tau\{\theta(\tau)=\eta(\tau)\}$.上面所有方法选择的阈值有可能是相同的,但是在一般情况下不同的方法选择的阈值是不同的.由于对称点准则容易推广到广义对称点准则,即 $\tau=\arg_\tau\{\theta(\tau)=\eta(\tau)\}$,$\rho$ 是灵敏性和特异性的相对重视度,若 $\rho=1$,广义对称点就是标准对称点,鉴于不需要优化和实际分析中需要同等重视两总体被正确诊断的概率,本文选择基于对称点准则来确定阈值.

现有文献中,阈值与相应的灵敏性和特异性的估计有参数、半参数和非参数的方法.非参数估计由于受假定错误的影响少而备受关注.文献[6]用经验似然结合光滑化估计方程的方法在给定特异性的条件下估计灵敏性.文献[7]采用刀切经验似然结合非光滑估计方程的方法去估计给定特异性下的灵敏性.文献[8]用经验似然结合非光滑估计方程的方法在阈值、灵敏性和特异性三者中给定任意一个参数去估计剩余的两个参数.文献[9]在对称点准则下使用经验似然结合非光滑估计方程的方法去选择最优阈值和相对应的灵敏性.因此,本文也选择用两样本经验似然结合非光滑估计方程的非参方法.

收稿日期:2022-01-17;**修回日期**:2022-12-27.

基金项目:国家自然科学基金(11871419;12201550);河南省高等学校重点科研项目(22A560003).

作者简介(通信作者):程伟丽(1980—),女,河南许昌人,华北水利水电大学讲师,博士,研究方向为生物统计,E-mail:chengweili@ncwu.edu.cn.

在实际应用中,受试者有可能会由于各种各样的原因导致生物指标值的缺失,比如:研究中的退出,各种不可控因素引起的信息缺失,参看文献[10].因此近年来,在诊断检验值缺失的情形下,ROC 曲线的统计分析受到了不少的关注.文献[11-12]研究了在完全随机缺失数据下通过随机热平台插补方法得到 ROC 曲线的估计和区间估计.文献[13]研究了随机缺失下经验似然结合光滑化估计方程的方法得到高特异性下灵敏性的估计和区间估计.但光滑估计方程中窗宽的选择是一个不易解决的问题.因此,本文研究生物指标值随机缺失情形基于对称点原则下两样本经验似然结合非光滑估计方程的方法给出阈值和灵敏性的联合估计和置信域.

1 缺失数据下阈值和灵敏性的经验似然估计

1.1 符号

设有病体的观测数据是 $\{(\delta_{xi},X_i,Z_{xi})\}_{i=1}^{m}$,第 i 个有病受试者生物指标观测值是 X_i,存在缺失,δ_{xi} 是 X_i 缺失指示器,若 X_i 可观测就令 $\delta_{xi}=1$,若 X_i 缺失就令 $\delta_{xi}=0$,Z_{xi} 是完全被观测辅助协变量,跟生物指标值 X_i 有关.对于任意指标 $i\neq k$,假设 δ_{xi} 与 δ_{xk} 相互独立,并且 δ_{xi} 与 X_i 也相互独立,也即 $\Pr(\delta_{xi}=1\mid X_i,Z_{xi})=\Pr(\delta_{xi}=1\mid Z_{xi})=\pi_1(Z_{xi})$.同理,无病体的观测数据是$\{(\delta_{yj},Y_j,Z_{yj})\}_{j=1}^{n}$,$\delta_{yj}$ 是第 j 个无病受试者生物指标观测值 Y_j 的缺失指示器,若 Y_j 可观测 $\delta_{yj}=1$,否则 $\delta_{yj}=0$,Z_{yj} 是完全被观测辅助协变量且跟 Y_j 有关.同样地,对于任意 $j\neq k$,也假设 δ_{yj} 与 δ_{yk} 相互独立,δ_{yj} 与 Y_j 相互独立,即是 $\Pr(\delta_{yj}=1\mid Y_j,Z_{yj})=\Pr(\delta_{yj}=1\mid Z_{yj})=\pi_2(Z_{yj})$,其中 $\pi_1(Z_{xi})$ 和 $\pi_2(Z_{yj})$ 就是倾向得分函数.在上面的假设下,按照文献[10],本文考虑的是随机缺失机制.

1.2 两样本经验似然估计

在不存在缺失生物指标的条件下,阈值、灵敏性和特异性的两样本经验似然估计如下:基于两样本估计方程 $g_{1i}(\theta,\eta,\tau,X_i)$ 和 $g_{2j}(\theta,\eta,\tau,Y_j)$,定义参数$(\theta,\eta,\tau)$ 的两样本经验似然比函数

$$\text{Ł}(\theta,\eta,\tau)=\sup\{\prod_{i=1}^{m}(mp_i)\prod_{j=1}^{n}(nq_j)\mid p_i\geq 0,\sum_{i=1}^{m}p_i=1,\sum_{i=1}^{m}p_ig_{1i}(\theta,\eta,\tau,X_i)=0,$$

$$q_j\geq 0,\sum_{j=1}^{n}q_j=1,\sum_{j=1}^{n}q_jg_{2j}(\theta,\eta,\tau,Y_j)=0\},$$

其中 $g_{1i}(\theta,\eta,\tau,X_i)=I(X_i\leq\tau)-(1-\theta),i=1,2,\cdots,m,g_{2j}(\theta,\eta,\tau,Y_j)=I(Y_j\leq\tau)-\eta,j=1,2,\cdots,n$.在对称点 $\theta=\eta$ 的要求下,上面的两样本经验似然比函数只是关于参数(θ,τ),矩函数 $g_{1i}(\theta,\eta,\tau,X_i)$ 和 $g_{2j}(\theta,\eta,\tau,Y_j)$ 分别调整为 $g_{1i}(\theta,\tau,X_i)=I(X_i\leq\tau)-(1-\theta),i=1,2,\cdots,m$ 和$g_{2j}(\theta,\tau,Y_j)=I(Y_j\leq\tau)-\theta,j=1,2,\cdots,n$. 再如上定义两样本经验似然比函数是

$$\text{Ł}(\theta,\tau)=\sup\{\prod_{i=1}^{m}(mp_i)\prod_{j=1}^{n}(nq_j)\mid p_i\geq 0,\sum_{i=1}^{m}p_i=1,\sum_{i=1}^{m}p_ig_{1i}(\theta,\tau,X_i)=0,$$

$$q_j\geq 0,\sum_{j=1}^{n}q_j=1,\sum_{j=1}^{n}q_jg_{2j}(\theta,\tau,Y_j)=0\},$$

上面的对数经验似然比在真值点的渐近分布是自由度为 2 的标准卡方分布.这里令真值点 θ_0,τ_0 分别表示 θ,τ 的真值,且满足 $E\{[g_{1i}(\theta_0,\tau_0,X_i),g_{2j}(\theta,\tau,Y_j)]^{\mathrm{T}}\}=0$ 的唯一解.

1.3 带有缺失数据的两样本经验似然估计

当有病体或无病体观测的生物指标值有缺失时,上面所述的经验似然方法不能直接用于 (θ,τ) 的估计.虽然前面提出的经验似然估计方法可以直接用于完全观测到的数据,但是这样得到的估计可能是一个有偏的估计.为此,考虑逆概率加权和多重插补两种方法相融合的思想,构造一组插补估计方程 $g_{i1}^{A}(\theta,\tau)$ 和 $g_{2j}^{A}(\theta,\tau)$:

$$\begin{cases}g_{1i}^{A}(\theta,\tau,X_i)=\dfrac{\delta_{xi}}{\pi_1(Z_{xi})}g_{1i}(\theta,\tau,X_i)+(1-\dfrac{\delta_{xi}}{\pi_1(Z_{xi})})\dfrac{1}{\kappa}\sum_{\upsilon=1}^{\kappa}g_{1i}(\theta,\tau,\widetilde{X}_{i\upsilon}),\\ g_{2j}^{A}(\theta,\tau,Y_j)=\dfrac{\delta_{yj}}{\pi_2(Z_{yj})}g_{2j}(\theta,\tau,Y_j)+(1-\dfrac{\delta_{yj}}{\pi_2(Z_{yj})})\dfrac{1}{\kappa}\sum_{\upsilon=1}^{\kappa}g_{2j}(\theta,\tau,\widetilde{Y}_{j\upsilon}).\end{cases}$$

其中 $\widetilde{X}_{iv}$ 和 $\widetilde{Y}_{iv}$ 分别是 X_i 和 Y_j 的插补值.在样本 $\{(\delta_{xi},X_i,z_{xi})\}_{i=1}^{m}$ 和 $\{(\delta_{yj},y_i,z_{yj})\}_{j=1}^{n}$ 的基础上,有病体和无病体的条件分布函数 $F(\cdot\mid Z_{xi})$ 和 $G(\cdot\mid z_{yj})$ 的核估计分别是

$$\hat{F}(x\mid Z_{xi})=\sum_{\kappa=1}^{m}w_{1k}(Z_{xi})I(X_i\leq x),\hat{G}(y\mid Z_{yj})=\sum_{\kappa=1}^{n}w_{2k}(Z_{yj})I(Y_j\leq y),$$

其中 $w_{1\kappa}(Z_x)=\delta_{x\kappa}K_{b_1}(Z_x-Z_{x\kappa})/\{\sum_{\kappa'=1}^{m}\delta_{x\kappa'}K_{b_1}(Z_x-Z_{x\kappa'})\}$ 和 $w_{2\kappa}(Z_y)=\delta_{y\kappa}K_{b_2}(Z_y-Z_{y\kappa})/\{\sum_{\kappa'=1}^{n}\delta_{y\kappa'}K_{b_2}(Z_y-Z_{y\kappa'})\}$,$K_{b_1}(Z)=K_1(Z/b_1)$,$K_{b_2}(Z)=K_2(Z/b_2)$,$K_1(Z)$ 和 $K_2(Z)$ 可以是不同的多元核函数,b_1 和 b_2 也可以是不同的窗宽.然后,分别从 $\hat{F}(x\mid Z_{xi})$ 和 $\hat{G}(y\mid Z_{yj})$ 随机抽取缺失的 X_i 和 Y_j 的 κ 重插补值 $\{\widetilde{X}_{iv}\}_{v=1}^{\kappa}$ 和 $\{\widetilde{Y}_{jv}\}_{v=1}^{\kappa}$.这里,当 $\delta_{x\kappa}=1$ 时,插补值 $\widetilde{X}_{iv}$ 是观测值 X_κ 的概率是 $w_{1\kappa}(Z_{xi})$;当 $\delta_{y\kappa}=1$ 时,插补值 $\widetilde{Y}_{jv}$ 是观测值 Y_κ 的概率是 $w_{2k}(Z_{yj})$.

在实际应用中,$\pi_1(Z_{xi})$ 和 $\pi_2(Z_{yj})$ 通常是不知道的.为此,考虑倾向得分函数 $\pi_1(Z_{xi})$ 和 $\pi_2(Z_{yj})$ 是下面的 logistic 回归模型:

$$\text{logit}\{\pi_1(Z_{xi};\alpha_1)\}=\alpha_{1,0}+\alpha_{1,1}^{\mathrm{T}}Z_{xi},\text{logit}\{\pi_2(Z_{yj};\alpha_2)\}=\alpha_{2,0}+\alpha_{2,1}^{\mathrm{T}}Z_{yj},\tag{1}$$

其中 $\text{logit}(t)=\ln\{t/(1-t)\}$,$\pi_1(Z_{xi})=\pi_1(Z_{xi};\alpha)$,$\pi_2(z_{yj})=\pi_2(Z_{yj};\beta)$,$\alpha_1=(\alpha_{1,0},\alpha_{1,1}^{\mathrm{T}})^{\mathrm{T}}$ 和 $\alpha_2=(\alpha_{2,0},\alpha_{2,1}^{\mathrm{T}})^{\mathrm{T}}$ 是待估的未知参数.在上面所述的 logistic 回归模型中,通过最大化对数似然函数 $\ell_P(\alpha_1)=\sum_{i=1}^{m}\{\delta_{xi}(\alpha_{1,0}+\alpha_{1,1}^{\mathrm{T}}Z_{xi})-\ln(1+\mathrm{e}^{\alpha_{1,0}+\alpha_{1,1}^{\mathrm{T}}Z_{xi}})\}$ 和 $\ell_P(\alpha_2)=\sum_{j=1}^{n}\{\delta_{yj}(\alpha_{2,0}+\alpha_{2,1}^{\mathrm{T}}Z_{yj})-\ln(1+\mathrm{e}^{\alpha_{2,0}+\alpha_{2,1}^{\mathrm{T}}Z_{yj}})\}$,得到未知参数 α_1 和 α_2 的相合估计 $\hat{\alpha}_1$ 和 $\hat{\alpha}_2$,进而得到 $\pi_1(Z_{xi})$ 和 $\pi_2(Z_{yj})$ 的相合估计分别是 $\hat{\pi}_1(Z_{xi})=\pi_1(Z_{xi};\hat{\alpha}_1)$ 和 $\hat{\pi}_2(Z_{yj})=\pi_2(Z_{yj};\hat{\alpha}_2)$.最后得到估计方程组 $\hat{g}_{1i}^{A}(\theta,\tau)$ 和 $\hat{g}_{2j}^{A}(\theta,\tau)$:

$$\begin{cases}\hat{g}_{1i}^{A}(\theta,\tau)=\dfrac{\delta_{xi}}{\hat{\pi}_1(Z_{xi})}g_{1i}(\theta,\tau,X_i)+(1-\dfrac{\delta_{xi}}{\hat{\pi}_1(Z_{xi})})\dfrac{1}{\kappa}\sum_{v=1}^{\kappa}g_{1i}(\theta,\tau,\widetilde{X}_{iv}),\\ \hat{g}_{2j}^{A}(\theta,\tau)=\dfrac{\delta_{yj}}{\hat{\pi}_2(Z_{yj})}g_{2j}(\theta,\tau,Y_j)+(1-\dfrac{\delta_{yj}}{\hat{\pi}_2(Z_{yj})})\dfrac{1}{\kappa}\sum_{v=1}^{\kappa}g_{2j}(\theta,\tau,\widetilde{Y}_{jv}).\end{cases}\tag{2}$$

根据上面定义的估计方程 $\hat{g}_{1i}^{A}(\theta,\tau)$ 和 $\hat{g}_{2j}^{A}(\theta,\tau)$,参数 (θ,τ) 的经验似然比函数被定义为

$$L_A(\theta,\tau)=\sup\{\prod_{i=1}^{m}(mp_i)\prod_{j=1}^{n}(nq_j)\mid p_i\geq 0,\sum_{i=1}^{m}p_i=1,\sum_{i=1}^{m}p_i\hat{g}_{1i}^{A}(\theta,\tau)=0,$$

$$q_j\geq 0,\sum_{j=1}^{n}q_j=1,\sum_{j=1}^{n}q_j\hat{g}_{2j}^{A}(\theta,\tau)=0\}.$$

经计算后,概率质量 p_i 和 q_j 的最优解分别是 $\hat{p}_i=[m\{1+\lambda_1(\theta,\tau)\hat{g}_{1i}^{A}(\theta,\tau)\}]^{-1}$,$\hat{q}_j=[n\{1+\lambda_2(\theta,\tau)\hat{g}_{2j}^{A}(\theta,\tau)\}]^{-1}$,其中 $\lambda_1(\theta,\tau)$ 和 $\lambda_2(\theta,\tau)$ 是拉格朗日乘子,且满足

$$\begin{cases}Q_1(\lambda_1,\lambda_2\mid\theta,\tau)=\dfrac{1}{m}\sum_{i=1}^{m}\dfrac{\hat{g}_{1i}^{A}(\theta,\tau)}{1+\lambda_1(\theta,\tau)\hat{g}_{1i}^{A}(\theta,\tau)}=0,\\ Q_2(\lambda_1,\lambda_2\mid\theta,\tau)=\dfrac{1}{n}\sum_{j=1}^{n}\dfrac{\hat{g}_{2j}^{A}(\theta,\tau)}{1+\lambda_2(\theta,\tau)\hat{g}_{2j}^{A}(\theta,\tau)}=0.\end{cases}$$

因此,参数 (θ,τ) 的对数经验似然比函数是

$$\ell_A(\theta,\tau)=-2\ln L_A(\theta,\tau)=2\sum_{i=1}^{m}\ln\{1+\lambda_1(\theta,\tau)\hat{g}_{1i}^{A}(\theta,\tau)\}+2\sum_{j=1}^{n}\ln\{1+\lambda_2(\theta,\tau)\hat{g}_{2j}^{A}(\theta,\tau)\}.\tag{3}$$

而后,极大化 $-\ell_A(\theta,\tau)$ 后,可以得到 (θ,τ) 的最大经验似然估计,记为 $(\hat{\theta},\hat{\tau})$.

2　渐近理论

这一小节,将确定在一定正则条件下参数的渐近正态性和上面的对数经验似然比函数的 Wliks 定理.首先假设 $\beta=(\theta,\tau)^{\mathrm{T}}$,则可以将 $g_1(\theta,\tau,X)=I(X\leq\tau)-(1-\theta)$,$g_2(\theta,\tau,Y)=I(Y\leq\tau)-(1-\theta)$ 记为 $g_1(\beta,$

$X)$,$g_2(\beta,Y)$,且令 $\phi_1(\beta)=E[g_1(\beta,X)]$,$\phi_2(\beta)=E[g_2(\beta,Y)]$.在正则条件下,自然容易得到 $\partial\phi_1(\beta)/\partial\tau=f_x(\tau)$ 和 $\partial\phi_2(\beta)/\partial\tau=g_y(\tau)$,其中 $f_x(\cdot)$ 和 $g_y(\cdot)$ 分别是随机变量 X 和 Y 的概率密度函数,即 $f_x(x)=\partial F(x)/\partial x$ 和 $g_y(y)=\partial G(y)/\partial y$.为了符号书写简单,令 $\hat{g}_{ij}^{A}(\beta)=(\gamma_1\hat{g}_{1i}^{A}(\theta,\tau,X_i),\gamma_2\hat{g}_{2j}^{A}(\theta,\tau,Y_j))^{\mathrm{T}}$,其中 $\gamma_1=m/N$,$\gamma_2=n/N$,$N=m+n$.也令 $g(\beta)=(\gamma_1 g_1(\beta,X),\gamma_2 g_2(\beta,Y))^{\mathrm{T}}$,其期望 $\phi(\beta)=E[g(\beta)]=(\gamma_1\phi_1(\beta),\gamma_2\phi_2(\beta))^{\mathrm{T}}$.假设

$$\phi_N(\beta)=(\frac{1}{N}\sum_{i=1}^{m}\hat{g}_{1i}^{A}(\beta,X_i),\frac{1}{N}\sum_{j=1}^{n}\hat{g}_{2j}^{A}(\beta,Y_j))^{\mathrm{T}}=(\gamma_1\frac{1}{m}\sum_{i=1}^{m}\hat{g}_{1i}^{A}(\beta,X_i),\gamma_2\frac{1}{n}\sum_{j=1}^{n}\hat{g}_{2j}^{A}(\beta,Y_j))^{\mathrm{T}}.$$

假设 $\sigma_1^2=E[\pi_1^{-1}(Z_x)\mathrm{var}\{g_1(\beta,X)\mid Z_x\}+E^2\{g_1(\beta,X)\mid Z_x\}]$,$\sigma_2^2=E[\pi_2^{-1}(Z_y)\mathrm{var}\{g_2(\beta,Y)\mid Z_y\}+E^2\{g_2(\beta,Y)\mid Z_y\}]$,$V=\mathrm{diag}(\gamma_1\sigma_1^2,\gamma_2\sigma_2^2)$.且令 $\Gamma=\dfrac{\partial E[\hat{g}_{ij}^{A}(\beta)]}{\partial\beta^{\mathrm{T}}}=(\dfrac{\partial E(\hat{g}_{ij}^{A})}{\partial\theta},\dfrac{\partial E(\hat{g}_{ij}^{A})}{\partial\tau})=\begin{pmatrix}\gamma_1 & \gamma_1 f_x(\tau)\\ \gamma_2 & \gamma_2 g_y(\tau)\end{pmatrix}$.

为证明结论,需要如下条件.

(C1)当 $\min(m,n)\to\infty$,有 $m/N\to\gamma_1$,$n/N\to\gamma_2$,其中 $0<\gamma_1,\gamma_2<1$.

(C2)核函数 $K(\cdot)$是一个概率密度函数,满足:(i)有界且有紧支撑;(ii)对称且有方差 $\sigma^2=\int s^2K(s)\mathrm{d}s=1$;(iii) $\mu=\int sK(s)\mathrm{d}s=0$,在以 0 为中心的封闭区间内,$K(s)\geq d$ 某个 d,核函数 $\mathscr{W}_\kappa$,$\kappa=1,2$ 与 $\mathscr{K}_\kappa$ 条件相同.

当 $\min(m,n)\to\infty$,窗宽 b_1 和 b_2 满足 $mb_1^{b_{Z_x}}/\ln m\to\infty$,$nb_2^{d_{Z_y}}/\ln n\to\infty$,$mb_1^4\to\infty$,$mb_2^4\to\infty$,这里辅助变量 Z_x 和 Z_y 分别是 d_{Z_x} 维和 d_{Z_y} 维,窗宽 h_κ,$\kappa=1,2$ 条件与 b_κ 条件相同.

(C3)倾向得分函数 $\pi_1(Z_x)$ 和 $\pi_2(Z_y)$ 满足 $\min_i\pi_1(Z_{xi})\geq c_1$,对某个正数 $c_1>0$,$\min_j\pi_2(Z_{yj})\geq c_2$ 对某个正数 $c_2>0$.密度函数 $pz_x(Z_x)$ 在 Z_x 的支撑集上有界,关于 Z_x 至少二阶连续可导;$pz_y(Z_y)$ 在 Z_y 的支撑集上有界,关于 Z_y 至少二阶连续可导.

(C4)存在参数 $\beta_0=(\theta_0,\tau_0)\in\mathscr{B}$ 是矩函数 $\phi(\beta)=0$ 的唯一解.参数空 B 是 $\mathbf{R}^2$ 紧子集,且 $E[\sup_{\beta\in\mathscr{B}}\mid g(\beta)\mid]<\infty$ 和 $E[\sup_{\beta\in\mathscr{B}}\mid g(\beta)g^{\mathrm{T}}(\beta)\mid]$ 每个分量都有界.

(C5)函数族 $\{\hat{g}_{ij}^{A}(\beta),(\beta)\in\mathscr{B}\}$ 是 Glivenko-Cantelli.

(C6)对某个 $\epsilon>0$,$\{\hat{g}_{ij}^{A}(\beta),\beta\in\mathscr{N}_\epsilon\}$ 是 Donsker.

(C7)对于所有的 $\beta\in\mathscr{B}$ 和所有的小正数 $\epsilon=o(1)$,存在一个正数 C 和 $s\in(0,1]$,使得 $E[\sup\limits_{\beta,\beta'\in\mathscr{B}}\mid\hat{g}_{1i}^{A}(\beta)-\hat{g}_{1i}^{A}(\beta')\mid]<C\epsilon^{2s}$ 和 $E[\sup\limits_{\beta,\beta'\in\mathscr{B}}\mid\hat{g}_{2j}^{A}(\beta)-\hat{g}_{2j}^{A}(\beta')\mid]<C\epsilon^{2s}$ 成立.

(C8)当 $\kappa\to\infty$ 时,矩函数的插补部分的条件期望 $m_{g_1}(\beta,Z_x)$ 满足条件:(i) 函数族$\{m_{g_1}(\beta,Z_x),(\beta)\in\mathscr{B}\}$ 是 Glivenko-Cantelli;(ii) 对所有的 $Z_x\in Z$ 存在某个 $\epsilon_1>0$ 满足在小邻域 $\mathscr{N}_{\epsilon_1}$ 关于参数 β 有连续的偏导数 $\partial_\beta m_{g_1}(\beta,Z_x)=\partial m_{g1}(\beta,Z_x)/\partial\beta$;$E\{\sup_{\beta\in\mathscr{N}_{\epsilon_1}}\mid\partial_\beta m_{g1}(\beta,Z_x)\}$ 的每个分量都有界;(iii) 存在 $s_1\in(0,1]$ 和某个满足 $E[b(Z_x)]<\infty$ 的可测函数 $b(Z_x)$,对满足 $\sup\limits_{\beta\in\mathscr{N}_{\epsilon_1}}\|\partial_\beta\widetilde{m}_{g1}(\beta,Z_x)-\partial_\beta m_{g1}(\beta,Z_x)\|_\infty<\epsilon_1$ 的光滑函数 $\partial_\beta\widetilde{m}_{g1}(\beta,Z_x)$,有 $\mid\partial_\beta\widetilde{m}_{g1}(\beta,Z_x)-\partial_\beta\widetilde{m}_{g1}(\beta,Z_x)\mid\leq b(Z_x)\sup\limits_{\beta\in\mathscr{N}_{\epsilon_1}}\|\partial_\beta\widetilde{m}_{g1}(\beta,Z_x)-\partial_\beta m_{g1}(\beta,Z_x)\|_\infty^{s_1}$. 当 $\kappa\to\infty$ 时,另一个矩函数的插补部分的条件期望 $m_{g2}(\beta,Z_y)$ 有类似上面的要求条件.

条件(C1)是两样本的样本量平衡的条件,条件(C2)和(C3)是缺失数据中常要求满足的条件,条件(C4)~(C7)是非光滑矩函数需要满足的条件,条件(C8)是非光滑矩函数的插补部分需要满足的条件.

定理 1 假设上面的条件(C1)~(C8)成立,当 $\min(m,n)\to\infty$ 和 $\kappa\to\infty$,则有

$$N^{1/2}(\hat{\beta}-\beta_0)\xrightarrow{\mathscr{L}}\mathscr{N}(0,\sum),\tag{4}$$

其中 $\xrightarrow{\mathscr{L}}$ 表示依分布收敛,$\sum=(\Gamma^{\mathrm{T}}V^{-1}\Gamma)^{-1}$.

定理 2 假设上面的条件(C1)~(C8)成立,当 $\min(m,n)\to\infty$ 和 $\kappa\to\infty$,则有

$$\ell_A(\beta_0)\xrightarrow{\mathscr{L}}\chi_2^2,\tag{5}$$

其中 χ_2^2 是自由度为 2 的卡方分布.在灵敏性和特异性相等的条件下，阈值和灵敏性的名义水平是 $1-\alpha$ 的置信域是 $C_{1-\alpha}=\{\beta=(\theta,\tau):\ell_A(\beta)\leq\chi_2^2(1-\alpha)\}$，其中 $\chi_2^2(1-\alpha)$ 满足 $\Pr(\chi_2^2\leq\chi_2^2(1-\alpha))=1-\alpha$，这里 $\alpha\in(0,1)$.

定理 1 建立了参数向量的估计 $\hat{\beta}$ 的相合性和渐近正态性.定理 2 表明 $\ell_A(\beta_0)$ 的渐近分布是一个标准的自由度为 2 的中心卡方分布.在没有缺失数据的情形下，$\pi_1(z_x)=\pi_2(z_y)=1$，这与文献[9]所给出的结论一致.因此，本文将文献[9]的结果推广到带有随机缺失的情形下.

证明　根据条件(C1)，在 $O_p(m^{-1/2})$，$O_p(n^{-1/2})$ 和 $O_p(N^{-1/2})$ 没有必要区分.因为随机缺失是不可忽略缺失的特殊情况，所以依据文献[14]的证明思路，可以得到 $\phi_N(\beta)=(\frac{1}{N}\sum_{i=1}^{m}\tilde{g}_{1i}^{A}(\beta_0),\frac{1}{N}\sum_{j=1}^{n}\tilde{g}_{2j}^{A}(\beta_0))'+O_p(N^{-1/2})$ 和 $|\phi_N(\hat{\beta})|=O_p(N^{-1/2})$.

定理 1 和定理 2 证明的关键是建立一个经验似然比函数 $\ell_A(\beta)$ 的光滑近似函数

$$\tilde{\ell}_A(\beta)=-N[S_{21}(\beta-\beta_0)]^{\mathrm{T}}\lambda-N\phi_N(\beta_0)^{\mathrm{T}}\lambda-N\lambda^{\mathrm{T}}V\lambda.$$

可得到 $|\ell_A(\beta)-\tilde{\ell}_A(\beta)|=O_p(N^{-1})$，因此只需要考虑 $\min_{\beta\in\mathcal{B}}\sup_{\lambda\in\mathbf{R}^2}\tilde{\ell}_A(\beta)$.这需要满足 $S_{12}\tilde{\lambda}=0$，$S_{21}(\tilde{\beta}-\beta_0)+\phi_N(\beta_0)-V\tilde{\lambda}=0$，简单计算后可以有 $N^{1/2}(\tilde{\beta}-\beta_0)=(S_{12}\Gamma S_{12})^{-1}S_{12}\Gamma N^{1/2}\phi_N(\beta_0)$，再结合 $(\hat{\beta}-\tilde{\beta})=O_p(1)$，定理 1 可证.而 $\ell_A(\beta,\lambda(\beta))=N\phi_N(\beta)^{\mathrm{T}}\hat{G}^A(\beta)^{-1}\phi_N(\beta)+O_p(1)$，定理 2 可证.

3　数值模拟

在这一节，实施两个模拟研究来调查提出方法的有限样本表现.为了便于比较，考虑以下几个估计：(1) GS 估计，基于完整的数据集而不考虑缺失值计算的估计；(2)CC 估计，只用完全观测数据的估计；(3)IPW 估计，基于 logistic 倾向得分函数的逆概率加权方法的估计；(4)AIPW 估计，基于 logistic 倾向得分函数的逆概率加权和多重插补方法得到的估计.

第一个模拟，有病体和无病体分别来自于独立的正态总体 $\mathcal{N}(8.138\ 745+\sqrt{5}Z_x,\sigma^2)$ 和 $\mathcal{N}(6.5+\sqrt{3}Z_y,\sigma^2)$，其中 $\sigma^2=0.25$.协变量 Z_x 和 Z_y 相互独立，且都来自正态分布 $\mathcal{N}(0,0.5^2)$.为了调查本文提出的估计对于误差分布假设的稳健性，设置了第二个模拟，有病体是 $X=5+4.5Z_x+\varepsilon_x$，无病体是 $Y=3+4Z_y+\varepsilon_y$，这里的协变量 Z_x 和 Z_y 的设置与第一个模拟相同，误差项 ε_x 和 ε_y 相互独立，且都服从 $3\{\eta-E(\eta)\}$ 的分布，这里 $\eta\sim\beta(5,1)$.生物指标值 X_i 和 Y_j 的缺失，缺失指示器 δ_{xi} 和 δ_{yj} 相互独立，且分别来自于成功概率是 $\pi_1(Z_{xi})$ 和 $\pi_2(Z_{yj})$ 的两点分布，倾向得分函数 $\pi_1(Z_{xi})$ 和 $\pi_2(Z_{yj})$ 设置下面 3 种情形：

(a) $\mathrm{logit}\{\pi_1(Z_{xi})\}=1$，$\mathrm{logit}\{\pi_2(Z_{yj})\}=1$；

(b) $\mathrm{logit}\{\pi_1(Z_{xi})\}=1+0.4Z_{xi}$，$\mathrm{logit}\{\pi_2(Z_{yj})\}=1+0.4Z_{yj}$，其中 $\mathrm{logit}(x)=\ln\{x/(1-x)\}$；

(c) $\pi_1(Z_{xi})=\Phi(0.6+0.4Z_{xi})$，$\pi_2(Z_{yj})=\Phi(0.6+0.4Z_{yj})$，其中 $\Phi(\cdot)$ 是标准正态分布的累积分布函数.

这里(a)是本文 1.3 小节定义缺失数据机制(1)的特殊情况，即 $\alpha_{1,1}=0$ 和 $\alpha_{2,1}=0$，这对应于完全随机缺失情况；(b)满足缺失数据机制(1)给定的随机缺失数据机制下的参数模型假设；(c)是随机缺失机制，但是不满足缺失数据机制(1)的参数模型假设，这主要是对错误设定倾向得分模型的稳健分析.按上面情形产生的平均缺失率大约在 30%左右.

对于上面所述的 3 种缺失数据机制和两种误差情形，设定样本量为平衡样本 $m=n=300$ 和非平衡样本 $m=240$，$n=360$，用上面给定的方法模拟重复计算全数据集 1 000 次.对于 1 000 次模拟重复中的每一次，基于本文提出的增广逆概率加权方法和 GS，CC 方法去计算灵敏性和阈值的最大经验似然估计和名义水平 95%的覆盖率，优化采用模拟退火的方法.本文中采用高斯核函数 $K_1(u)=K_2(u)=(2\pi)^{-1/2}\exp(-u^2/2)$，窗宽 $b_1=\hat{\sigma}_{z_x}m^{-1/3}$ 和 $b_2=\hat{\sigma}_{z_y}n^{-1/3}$，其中 $\hat{\sigma}_{z_x}$ 和 $\hat{\sigma}_{z_y}$ 分别也是观测值 Z_x 和 Z_y 的样本标准差.所有的模拟结果在表 1 和表 2 中给出，其中 Bias 表示真值与 1 000 次重复的平均估计值之偏差，SD 表示 1 000 次重复得到的估计值的标准差，RMS 表示 1 000 次重复得到的估计值的均方误差的平方根，CP 表示 1 000 次重复得到

的 95%的置信区间覆盖真值的比例.

表 1 样本量 $m=n=300$ 3 种缺失设置下灵敏性和阈值的估计

Tab. 1 The estimators for sensitivity and cutoff value with $m=n=300$ under three missing settings

Case	Est	(i)							(ii)						
		Bias	SD	RMS	Bias	SD	RMS	CP/%	Bias	SD	RMS	Bias	SD	RMS	CP/%
	GS	Sensitivity			Cut-off			95.2	Sensitivity			Cut-off			93.9
		0.000	0.017	0.017	0.000	0.063	0.063		0.000	0.019	0.019	−0.001	0.118	0.118	
(a)	CC	0.000	0.020	0.020	0.000	0.073	0.073	94.8	0.000	0.022	0.022	0.004	0.136	0.136	95.3
	IPW	0.000	0.019	0.019	0.000	0.069	0.069	100.0	0.000	0.021	0.021	0.004	0.125	0.125	99.5
	AIPW	0.000	0.019	0.019	0.000	0.067	0.067	94.9	0.000	0.020	0.020	0.001	0.121	0.121	93.8
(b)	CC	0.003	0.020	0.020	0.053	0.073	0.090	90.2	0.002	0.023	0.023	0.114	0.135	0.177	88.0
	IPW	0.000	0.019	0.019	0.000	0.069	0.069	100.0	0.000	0.021	0.021	0.003	0.125	0.125	99.4
	AIPW	0.000	0.018	0.018	0.000	0.068	0.068	94.6	0.000	0.020	0.020	0.001	0.120	0.120	93.7
(c)	CC	0.005	0.020	0.021	0.090	0.073	0.116	82.3	0.004	0.023	0.023	0.194	0.135	0.237	75.0
	IPW	0.000	0.019	0.019	0.000	0.069	0.069	100.0	0.000	0.021	0.021	0.002	0.125	0.125	99.6
	AIPW	0.000	0.019	0.019	0.000	0.068	0.068	94.6	0.000	0.020	0.020	0.001	0.121	0.121	93.5

表 2 样本量 $m=240, n=360$ 3 种缺失设置下灵敏性和阈值的估计

Tab. 2 The estimators for sensitivity and cutoff value with $m=240, n=360$ under three missing settings

Case	Est	(i)							(ii)						
		Bias	SD	RMS	Bias	SD	RMS	CP/%	Bias	SD	RMS	Bias	SD	RMS	CP/%
	GS	Sensitivity			Cut-off			95.0	Sensitivity			Cut-off			94.0
		0.000	0.018	0.018	−0.002	0.066	0.066		0.000	0.019	0.019	−0.002	0.114	0.114	
(a)	CC	0.001	0.018	0.018	−0.003	0.073	0.073	95.0	0.000	0.021	0.021	0.004	0.129	0.129	96.0
	IPW	0.000	0.018	0.018	0.001	0.070	0.070	99.5	−0.001	0.019	0.019	0.000	0.120	0.120	98.0
	AIPW	0.000	0.018	0.018	−0.001	0.068	0.068	95.5	−0.001	0.019	0.019	0.001	0.118	0.118	95.0
(b)	CC	0.005	0.018	0.019	0.056	0.071	0.090	90.0	0.001	0.021	0.021	0.124	0.129	0.179	88.0
	IPW	0.000	0.017	0.017	−0.001	0.070	0.070	99.5	−0.001	0.019	0.019	−0.002	0.121	0.121	98.0
	AIPW	0.000	0.017	0.018	0.000	0.067	0.067	95.5	0.000	0.019	0.019	0.003	0.118	0.118	94.0
(c)	CC	0.006	0.020	0.021	0.086	0.073	0.113	83.0	0.003	0.021	0.022	0.190	0.134	0.232	80.0
	IPW	0.001	0.018	0.018	0.000	0.071	0.071	100.0	0.000	0.019	0.019	0.003	0.123	0.123	100.0
	AIPW	0.001	0.018	0.018	−0.001	0.070	0.070	95.0	0.000	0.019	0.019	0.004	0.122	0.122	95.0

从表 1 可以看出,误差项若是正态分布,在所设置的 3 种缺失环境下,即便是在倾向得分函数的模型假定错误的情形下,增广逆概率加权估计的所有结果都接近于没有缺失数据下基准的 GS 的结果;在完全随机缺失(a)下,只用观测到数据的 CC 估计在标准差上比 GS 估计的标准差大,不过在偏差和覆盖率上也接近于 GS 的结果,但是非随机缺失(b)和(c)下,不但标准差增大,偏差也变大,覆盖率要远小于名义水平 95%;与 GS 估计相比,逆概率加权估计的标准差虽然增大,但是偏差变化不大,覆盖率却远大于名义水平 95%,这很

可能是由于权重估计的不稳定造成的.若误差项是非正态分布，表 2 有相似的模拟结果表现.将误差项是正态分布和非正态分布情形 3 种缺失机制下阈值和灵敏性 95%的非参置信域显示(图 1)，其中上面 3 个图是正态分布误差项下 3 种缺失机制(从左到右依次是 a，b 和 c)的联合置信域，下面 3 个图是非正态分布误差项下 3 种缺失机制(从左到右依次是 a，b 和 c)的联合置信域，点图是 CC，实线是本文提出的方法.从图 1 中可以看出这两种估计是有差别的.

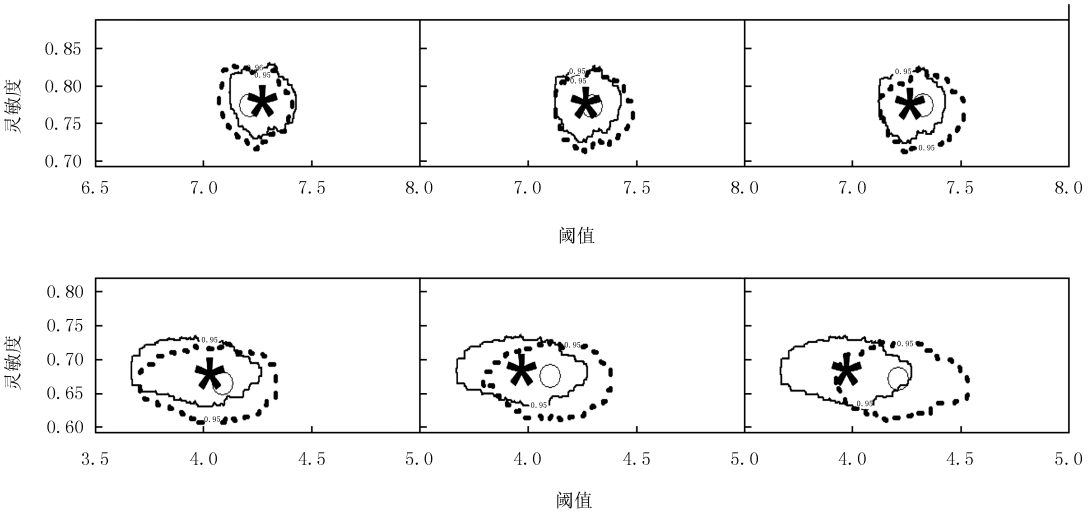

图1　阈值和灵敏性95%的联合置信域

Fig.1　The 95% joint confidence regions for cutoff value and sensitivity

参　考　文　献

[1] PEPE M S.The Statistical Evaluation of Medical Tests for Classification and Prediction[M].Oxford:Oxford University Press, 2003.

[2] YOUDEN W J.Index for Rating Diagnostic tests[J].Cancer,1950,3:32-35.

[3] PERKINS N J,SCHISTERMAN E F.The Inconsistency of "Optimal" Cutpoints Obtained Using two Criteria based on the Receiver Operating Characteristic Curve[J].American Journal of Epidemiology,2006,163:670-675.

[4] LIU X.Classification Accuracy and Cutpoint Selection[J].Statistics in Medicine,2012,31:2676-2686.

[5] JIMÉNEZ-VALVERDE A.Threshold-dependence as a Desirable Attribute for Discrimination Assessment: Implications for the Evaluation of Species Distribution Models[J].Biodiversity and Conservation,2014,23:369-385.

[6] CLAESKENS G,JING B Y,PENG L,et al.Empirical Likelihood Confidence Regions for Comparison Distributions and ROC Curves[J].Canadian Journal of Statistics,2003,31(2):173-190.

[7] GONG Y,PENG L,QI Y.Smoothed Jackknife Empirical Likelihood Method for ROC Curve[J]. Journal of Multivariate Analysis,2010,101(6):1520-1531.

[8] ADIMARI G,CHIOGNA M.Simple Nonparametric Confidence Regions for the Evaluation of Continuous-scale Diagnostic Tests[J].The International Journal of Biostatistics,2010,6(1):24.

[9] ADIMARI G,SINIGAGLIA A.Nonparametric Confidence Regions for the Symmetry Point-based Optimal Cutpoint and Associated Sensitivity of a Continuous-scale Diagnostic Test[J].Biometrical Journal,2020,62(6):1463-1475.

[10] LITTLE R J A,RUBIN D B.Statistical Analysis With Missing Data[M].London:Wiley,2002.

[11] WANG B,QIN G.Imputation-based Empirical Likelihood Inference for the Area under the ROC Curve with Missing Data[J].Statistics and Its Interface,2012,5(3):319-329.

[12] YANG H,ZHAO Y.Smoothed Jackknife Empirical Likelihood Inference for ROC Curves with Missing Data[J].Journal of Multivariate Analysis,2015,140:123-138.

[13] CHENG W,TANG N.Smoothed Empirical Likelihood Inference for ROC Curve in the Presence of Missing Biomarker Values[J].Biometrical Journal,2020,62(4):1038-1059.

[14] ZHAO P,TANG N,ZHU H.Generalized Empirical Likelihood Inferences for Nonsmooth Moment Functions with Nonignorable Missing Values[J].Statistica Sinica,2020,30(1):217-249.

The joint estimators of cutoff value and sensitivity in the absence of biomarker values

Cheng Weili[1], Wu Ying[2], Zuo Weibing[1]

(1. School of Mathematics and Statistics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China;
2. School of Mathematics and Statistics, Yunnan University, Kunming 650500, China)

Abstract: In observational studies, diagnostic test values for some subjects may be missing. Only if completely observed data are used, biased estimates may be obtained. In the presence of missing at random diagnostic test values, non-smooth estimation quations with the unknown parameters are established by combining inverse probability weighting and multiple imputation based on the symmetry point criterion with equal sensitivity and specificity. The joint nonparametric confidence regions for the optimal cutoff value and sensitivity are obtained by the two-sample empirical likelihood method. Under certain regular conditions, the asymptotic properties of the maximum empirical likelihood estimators of cutoff value and sensitivity are established. Simulation studies show that the proposed approach is better than other approaches.

Keywords: missing at random; cutoff value; sensitivity; symmetry point criterion; empirical likelihood

[责任编校 陈留院 赵晓华]