第 36 卷 第 4 期 应 用 数 学 学 报 Vol. 36 No. 4
2013 年 7 月 ACTA MATHEMATICAE APPLICATAE SINICA July, 2013

一种半参数 ROC 曲面估计方法 *

万树文

(南京财经大学应用数学学院, 南京 210046)

(E-mail: wanshuwen@aliyun.com)

摘 要 ROC 曲面是诊断医学统计学里评估有多类诊断结果的诊断测试方法准确性的一个重要工具, 也是近年来的一个研究热点. 本文提出一种半参数的 ROC 曲面估计方法. 该方法可以借助于许多统计软件里的逻辑斯蒂回归程序进行计算, 所以它的实施较为方便. 相关统计模拟显示, 本文提出的方法与传统的非参数方法相比, 有效性得到了显著提高. 而与参数方法相比, 当参数模型假设是正确时仍比参数方法有略高的有效性; 而当参数模型假设不正确时, 本文提出的半参数方法明显优于参数方法.

关键词 密度函数比模型; 经验似然; 逻辑斯蒂回归模型; ROC 曲线; ROC 曲面

MR(2000) 主题分类 62G99; 62H12; 62H15

中图分类 O212.1; O212.7

1 引言

在诊断医学统计学里, ROC 曲线是一种经典的用来评价一个有两类诊断结果 (有病的和健康的) 的诊断测试方法准确性的统计分析工具. 在医学实践中, 经常会碰到需要有三类或者更多类诊断结果的诊断情形. 例如, Nakas, Yiannoutsos[1] 研究了一种连续型的神经心理成套测验用来诊断由艾滋病引起的痴呆复合症 (ADC), 该方法将被检对象诊断成三类: 完好的, ADC 阶段 0.5, 和 ADC 阶段 1–3. 这种有多个诊断结果的情形使得传统的 ROC 曲线统计分析不再适用. 近年来, ROC 曲线的研究已经进展到对 ROC 曲面的研究以处理在实践中出现的有三类或三类以上诊断结果的诊断情形. 在文献里, Nakas 和 Yiannoutsos[1] 以及 Xiong[2] 等人研究了基于一个连续型的诊断测试方法的 ROC 曲面统计分析问题; Yang, Carlin[3] 以及 Wan, Zhang[4] 研究了基于两个连续型诊断测试方法的 ROC 曲面的统计推断; 而其它研究者象 Mossman[5], Dreiseitl 等人 [6] 以及 Heckerling[7] 研究了关于两个离散型的诊断测试方法的 ROC 曲面分析. 另

本文 2011 年 10 月 25 日收到. 2012 年 2 月 3 日收到修改稿.

* 国家自然科学基金 (11001119) 资助项目.

外，在当一个连续型诊断测试方法的三类诊断结果存在一个特殊的伞状次序时，Nakas, Alonzo[8], Alonzo, Nakas[9] 研究了相应的 ROC 曲面分析问题. 本文将提出了一个半参数的统计分析方法用于对一个连续型诊断测试方法有三类诊断结果时的 ROC 曲面分析. 这也是 Nakas, Yiannoutsos[1] 以及 Xiong[2] 研究的情形，不过这两篇文献分别研究了相应的非参数和参数的方法，而本文提出了一种半参数的方法.

当诊断结果有三类时，一个连续型的诊断测试方法将产生连续型的测量数据，用以将被测对象判别成三类. 在进行 ROC 曲面分析时，一个通常的假设是第三类对象的测量数据倾向于比第二类对象的测量数据高，且相应地，第二类对象的测量数据倾向于比第一类对象的测量数据高. 记测量数据为 X, 如果任意给定两个临界值 c_1 和 c_2, 则一个自然的分类标准就是：将满足 $X \le c_1$ 的对象判别成第一类，将满足 $c_1 < X \le c_2$ 的对象判别成第二类，而将剩余的对象判别成第三类. 受 ROC 曲线定义的启发，Nakas, Yiannoutsos[1] 提出了以 TC_2 对 TC_1 和 TC_3 作图得到的曲面作为 ROC 曲面的定义，这里 TC_i 是将一个第 i 类的对象正确判别成第 i 类的概率，$i = 1, 2, 3$. 另外，Nakas, Yiannoutsos 还提出用 ROC 曲面下的体积来从整体上评价一个诊断测试方法的准确性，体积越大则相应的诊断测试方法就越准确. 可以证明，ROC 曲面下的体积在数值上就等于该诊断测试方法对于从三类中随机抽取的对象进行正确判别的概率.

在本文里，我们将提出一种半参数的方法用于 ROC 曲面的估计，其核心的思想是在一个半参数密度函数比模型下进行 ROC 曲面的构建. 在文献里，半参数密度函数比模型已成功地用在了 ROC 曲线的统计分析中. 例如，Qin, Zhang [10] 提出了密度函数比模型下的一个 ROC 曲线半参数估计量，后来 Wan, Zhang[11] 将其改进，提出了一个光滑的 ROC 曲线估计量. 另外，Wan, Zhang[12] 还研究了如何在密度函数比模型下进行相关 ROC 曲线的比较. 本文的思路是对测量数据建立适当的密度函数比模型. 在得到三类数据分布函数的半参数估计量后，ROC 曲面的估计量就可自然获得. 将密度函数比模型引入 ROC 曲面的分析中不但使我们获得了一种半参数的统计分析方法，而且这种方法实施的难度不大，可以借助于逻辑斯蒂回归程序，这在文章的第 5 部分将有解释. 接下来，本文将介绍提出的主要方法，然后将通过统计模拟来比较本文提出的半参数方法和已有的非参数和参数方法，最后我们将提出的方法用于一组实际数据的分析，并进行适当的讨论.

2　主要方法

以 $X_{k1}, \cdots, X_{kn_k}$ 表示独立同分布的从第 k 类对象得到的测量数据，$k = 1, 2, 3$. 假定从第 3 类对象测得的数据倾向于比第 2 类的数据大，而第 2 类测得的数据倾向于比第 1 类的数据大. 另令 F_1, F_2 和 F_3 分别代表 X_{11}, X_{21} 和 X_{31} 的分布函数. Nakas, Yiannoutsos[1] 提出了如下的 ROC 曲面的定义：

$$R(s_1, s_2) = F_2\big(F_3^{-1}(1 - s_2)\big) - F_2\big(F_1^{-1}(s_1)\big), \qquad 0 \le s_1, s_2 \le 1.$$

ROC 曲面下的体积 VUS 也可以衡量一个诊断测试方法的准确性，其计算为

$$\mathrm{VUS}=\int_0^1\int_0^1 R(s_1,s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2=P(X_{31}>X_{21}>X_{11}).$$

注意到， VUS 实际上就是一个诊断测试方法将从各类得到的测量值正确归类的概率. 以

$$\widehat{F}_k(x)=\frac{1}{n_k}\sum_{i=1}^{n_k}I(X_{ki}\le x)$$

代表 $F_k(x)$ 的经验分布函数，则 ROC 曲面及其 VUS 的非参数估计量为

$$\widehat{R}(s_1,s_2)=\widehat{F}_2\big(\widehat{F}_3^{-1}(1-s_2)\big)-\widehat{F}_2\big(\widehat{F}_1^{-1}(s_1)\big),\qquad \widehat{\mathrm{VUS}}=\int_0^1\int_0^1\widehat{R}(s_1,s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2.$$

在正态分布的假设下，我们可以首先分布函数的估计量 $\overline{F}_1(x)$, $\overline{F}_2(x)$ 和 $\overline{F}_3(x)$, 然后将其代人 ROC 曲面的定义，就可以得到 ROC 曲面的参数型估计量 $\overline{R}(s_1,s_2)$ 以及相应的 VUS 的估计量 $\overline{\mathrm{VUS}}=\int_0^1\int_0^1\overline{R}(s_1,s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2$.

逻辑斯蒂模型是对多类响应变量进行建模的一个常用选择. 以 $D=k$ 代表第 k 类, $k=1,2,3$. 对于给定的测试数据 $X=x$, 逻辑斯蒂回归模型为

$$\log\Big(\frac{P(D=i|X=x)}{P(D=3|X=x)}\Big)=\alpha_i^*+\beta_i^Tr(x),\qquad i=1,2,\tag{1}$$

这里， α_i^* 是一个尺度参数， β_i 是一个 $p\times1$ 的向量参数，而 $r(x)$ 是一个 $p\times1$ 的 x 的向量函数. 则 $F_k(x)=P(X_{k1}\le x|D=k)$, $k=1,2,3$. 另以 $f_k(x)$ 代表 $F_k(x)$ 的密度函数. 类似于 Qin 和 Zhang[13] 的处理，很容易通过贝叶斯公式得出模型 (1) 就等价于如下的半参数密度函数比模型

$$\begin{aligned}&X_{31},\cdots,X_{3n_3}\sim f_3(x),\\&X_{i1},\cdots,X_{in_i}\sim f_i(x)=\exp\{\alpha_i+\beta_i^Tr(x)\}f_3(x),\qquad i=1,2,\end{aligned}\tag{2}$$

这里， $\alpha_i=\alpha_i^*+\log[P(D=3)/P(D=i)]$. 在这个模型里，我们没有要求密度函数 $f_i(x)$ 是某个具体的分布，只是要求这些密度函数在形式上有一个指数化的联系. 这不是传统的参数模型，但由于在模型中有参数的存在，故也不是一个非参数的模型. 模型 (2) 是介于参数和非参数之间的一种半参数模型. 在实践中，我们常常选择 $r(x)=x$ 或者 $r(x)=(x,x^2)^T$. 当 $r(x)=x$ 时，密度函数 $f_i(x)$ 和 $f_3(x)$ 包含许多常用分布包括两个有相同方差不同均值的正态分布，以及两个有不同均值的指数分布；当 $r(x)=(x,x^2)^T$ 时，可以是两个有不同均值和不同方差的正态分布. 有时， $r(x)=\log(x)$ 也是建模的一种较好选择，此时 $f_i(x)$ 和 $f_3(x)$ 包含两个对数正态分布 $\log N(\mu_1,\sigma^2)$ 和 $\log N(\mu_2,\sigma^2)$ 且 $\mu_1\ne\mu_2$. 至于在实际应用中如何选择 $r(x)$, Kay 和 Little[14] 提供了较好的参考. 对于一个给定的 $r(x)$, 我们可以类似于 Qin 和 Zhang[13] 构造科尔莫戈罗夫 - 斯米尔诺夫型的统计量进行拟合优度检验. 当多个形式的 $r(x)$ 均可行时，根据统计学节俭的原则，应选用相对简单的模型而不是复杂的模型.

现以 $\{T_1,\cdots,T_n\}$ 记合并的样本 $\{X_{11},\cdots,X_{1n_1};X_{21},\cdots,X_{2n_2};X_{31},\cdots,X_{3n_3}\}$ 且 $n=n_1+n_2+n_3$. 另外，令 $\rho_k=n_k/n_3$，$k=1,2$. 根据观察到的数据，我们可以写出半参数经验似然函数为

$$\begin{aligned}&\mathbf{L}(\alpha_{\mathbf{1}},\alpha_{\mathbf{2}},\beta_{\mathbf{1}},\beta_{\mathbf{2}},\mathbf{F_3})\\=&\prod_{i=1}^{n_3}dF_3(X_{3i})\prod_{j=1}^{n_1}\exp\{\alpha_1+\beta_1^Tr(X_{1j})\}\,\mathrm{d}F_3(X_{1j})\prod_{k=1}^{n_2}\exp\{\alpha_2+\beta_2^Tr(X_{2k})\}\,\mathrm{d}F_3(X_{2k})\\=&\Big\{\prod_{i=1}^{n}p_i\Big\}\Big\{\prod_{j=1}^{n_1}\exp\{\alpha_1+\beta_1^Tr(X_{1j})\}\Big\}\Big\{\prod_{k=1}^{n_2}\exp\{\alpha_2+\beta_2^Tr(X_{2k})\}\Big\},\end{aligned}$$

这里 $p_i=\mathrm{d}F_3(T_i)$，$i=1,\cdots,n$ 是概率的跃变且总和为 1. 类似于 Owen[15,16], Qin, Lawless[17], 我们采用拉格朗日乘子法发现满足如下约束条件

$$\sum_{i=1}^{n}p_i=1,\qquad p_i\geq 0,\qquad \sum_{i=1}^{n}p_i\big[\exp\{\alpha_k+\beta_k^Tr(T_i)\}-1\big]=0,\qquad k=1,2,$$

下的似然函数 $\mathbf{L}$ 的最大值在

$$\widetilde{p}_i=\frac{1}{n_3}\frac{1}{1+\sum\limits_{k=1}^{2}\rho_k\exp\big\{\widetilde{\alpha}_k+\widetilde{\beta}_k^Tr(T_i)\big\}},$$

获得. 这里 $(\widetilde{\alpha}_k,\widetilde{\beta}_k)$ 是 (α_k,β_k) 的最大半参数似然估计量，其通过解如下的计分方程组而获得：

$$\begin{aligned}\frac{\partial l(\alpha_k,\beta_k)}{\partial\alpha_k}&=n_k-\sum_{i=1}^{n}\frac{\rho_k\exp\{\alpha_k+\beta_k^Tr(T_i)\}}{1+\sum\limits_{j=1}^{2}\rho_j\exp\{\alpha_j+\beta_j^Tr(T_i)\}}=0,\\\frac{\partial l(\alpha_k,\beta_k)}{\partial\beta_k}&=\sum_{m=1}^{n_k}r(X_{km})-\sum_{i=1}^{n}\frac{\rho_k\exp\{\alpha_k+\beta_k^Tr(T_i)\}r(T_i)}{1+\sum\limits_{j=1}^{2}\rho_j\exp\{\alpha_j+\beta_j^Tr(T_i)\}}=0,\end{aligned}\tag{3}$$

这里，$k=1,2$，$l(\alpha_k,\beta_k)$ 是如下的对数似然函数

$$l(\alpha_k,\beta_k;\ k=1,2)\propto\sum_{k=1}^{2}\sum_{j=1}^{n_k}\big[\alpha_k+\beta_k^Tr(X_{kj})\big]-\sum_{i=1}^{n}\log\Big\{1+\sum_{j=1}^{2}\rho_j\exp\{\alpha_j+\beta_j^Tr(T_i)\}\Big\}.$$

注意到，约束条件

$$\sum_{i=1}^{n}p_i\big[\exp\{\alpha_k+\beta_k^Tr(T_i)\}-1\big]=0,$$

等价于

$$\sum_{i=1}^{n}p_i\exp\{\alpha_k+\beta_k^Tr(T_i)\}=1,$$

其反映了 $\exp\{\alpha_k+\beta_k^T r(t)\}dF_3(t)$ 是一个分布函数的事实.

进而， $F_3(t)$ 的最大半参数似然估计量为

$$\widetilde{F}_3(t)=\sum_{i=1}^{n}\widetilde{p}_i I(T_i\le t)=\frac{1}{n_3}\sum_{i=1}^{n}\frac{I(T_i\le t)}{1+\sum\limits_{k=1}^{2}\rho_k\exp\{\widetilde{\alpha}_k+\widetilde{\beta}_k^T r(T_i)\}}.$$

相应地， $F_j(t),\ j=1,2$ 的最大半参数似然估计量为

$$\widetilde{F}_j(t)=\frac{1}{n_3}\sum_{i=1}^{n}\frac{\exp\{\widetilde{\alpha}_j+\widetilde{\beta}_j^T r(T_i)\}}{1+\sum\limits_{k=1}^{2}\rho_k\exp\{\widetilde{\alpha}_k+\widetilde{\beta}_k^T r(T_i)\}}I(T_i\le t).$$

以 $(\widetilde{F}_1,\widetilde{F}_2,\widetilde{F}_3)$ 为基础，我们提出在半参数密度函数比模型 (2) 下构建如下的 ROC 曲面 $R(s_1,s_2)$ 的估计量

$$\widetilde{R}(s_1,s_2)=\widetilde{F}_2(\widetilde{F}_3^{-1}(1-s_2))-\widetilde{F}_2(\widetilde{F}_1^{-1}(s_1)),\qquad s_1,s_2\in[0,1],$$

并且相应地，用

$$\widetilde{\mathrm{VUS}}=\int_0^1\int_0^1\widetilde{R}(s_1,s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2. \tag{4}$$

来估计 ROC 曲面下体积. 根据最大似然估计的不变性原则， $\widetilde{R}(s_1,s_2)$ 和 $\widetilde{\mathrm{VUS}}$ 分别是 $R(s_1,s_2)$ 和其下体积 VUS 在模型 (2) 下的最大半参数似然估计量.

值得注意的是在实际计算中，关于 (4) 中 $\widetilde{\mathrm{VUS}}$ 的计算不需要采用数值积分的方法就能完成. 因为半参数估计量 $\widetilde{F}_k,\ k=1,2,3$ 都是阶梯函数，经过适当的推算就可以得出下面的表达式来计算 $\widetilde{\mathrm{VUS}}$:

$$\widetilde{\mathrm{VUS}}=\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\widetilde{p}_i\widetilde{p}_j\widetilde{p}_k\exp\{\widetilde{\alpha}_1+\widetilde{\beta}_1^T r(T_i)\}\exp\{\widetilde{\alpha}_2+\widetilde{\beta}_2^T r(T_j)\}I(T_i<T_j<T_k).$$

现在我们来讨论如何对提出的 ROC 曲面半参数估计量 $\widetilde{R}(s_1,s_2)$ 进行统计推断. 不幸的是， $\mathrm{Var}(\widetilde{R}(s_1,s_2))$ 的解析表达式很难获得. 另外，根据我们以前在密度函数比模型下处理 ROC 曲线问题的经验，这样的表达式即便能获得，也是非常复杂和不便于应用的. 这里，我们提出一个简单的自助法 (bootstrap) 来构建 $\widetilde{R}(s_1,s_2)$ 的置信区间. 首先从 $\widetilde{F}_1$ 中产生一个简单随机样本 $X_{11}^*,\cdots,X_{1n_1}^*$; 以类似的方式独立地从 $\widetilde{F}_2$ 和 $\widetilde{F}_3$ 中产生样本 $X_{21}^*,\cdots,X_{2n_2}^*$ 和 $X_{31}^*,\cdots,X_{3n_3}^*$. 另外，以 $T_1^*,\cdots,T_n^*$ 代表合并的样本，并且以 $(\widetilde{\alpha}_k^*,\widetilde{\beta}_k^*;k=1,2)$ 代表在计分方程组 (3) 中将 T_i 替换成 T_i^* 得到的解，则

$$\widetilde{F}_3^*(t)=\frac{1}{n_3}\sum_{i=1}^{n}\frac{I(T_i^*\le t)}{1+\sum\limits_{k=1}^{2}\rho_k\exp\{\widetilde{\alpha}_k^*+\widetilde{\beta}_k^{*T} r(T_i^*)\}}$$

且

$$\widetilde{F}_j^*(t)=\frac{1}{n_3}\sum_{i=1}^{n}\frac{\exp\{\widetilde{\alpha}_j^*+\widetilde{\beta}_j^{*T}r(T_i^*)\}}{1+\sum\limits_{k=1}^{2}\rho_k\exp\{\widetilde{\alpha}_k^*+\widetilde{\beta}_k^{*T}r(T_i^*)\}}I(T_i^*\le t),\qquad j=1,2.$$

进而，自助法获得的 ROC 曲面估计量为

$$\widetilde{R}^*(s_1,s_2)=\widetilde{F}_2^*(\widetilde{F}_3^{*-1}(1-s_2))-\widetilde{F}_2^*(\widetilde{F}_1^{*-1}(s_1)).$$

这样很自然地，从自助法获得的 $\widetilde{R}^*(s_1,s_2)$ 的值就可以用来找到 $\widetilde{R}(s_1,s_2)$ 的临界值.

3　统计模拟实验

在这部分，我们用统计模拟实验比较了在有限样本情形下我们提出的 ROC 曲面半参数估计量 $\widetilde{R}(s_1,s_2)$ 与相应的非参数和参数估计量. 我们的模拟实验分为两种情形：一种是参数方法的模型假设是正确的；另一种是参数方法的模型假设是错误的.

在模拟的第一种情形下，我们假定 $f_i(x)$ 是 $N(\mu_i,\sigma^2)$ 的密度函数，$i=1,2,3$, 则密度函数比模型 (2) 成立，且有 $r(x)=x$ 和

$$\alpha_1=\frac{\mu_3^2-\mu_1^2}{2\sigma^2},\qquad \alpha_2=\frac{\mu_3^2-\mu_2^2}{2\sigma^2},\qquad \beta_1=\frac{\mu_1-\mu_3}{\sigma^2},\qquad \beta_2=\frac{\mu_2-\mu_3}{\sigma^2}.$$

我们设置的样本容量是 $(n_1,n_2,n_3)=(30,30,30)$ 和 $(n_1,n_2,n_3)=(60,45,30)$ 两种情形. 设置 $(n_1,n_2,n_3)=(60,45,30)$ 的原因主要是在实际研究里的数据经常看到三类对象的个数是递减的. 另外，设置 $\mu_1=0,\ \mu_2=1,\ \mu_3=2$ 和 $\sigma=1$. 这样，模型 (2) 参数的真实值为 $\alpha_1=2,\ \alpha_2=1.5,\ \beta_1=-2,\ \beta_2=-1$. 对于每种情形 (n_1,n_2,n_3), 我们从 $N(\mu_i,\sigma^2)$ $(i=1,2,3)$ 中产生 1000 组相互独立的合并样本.

对于 ROC 曲面估计量，我们考虑了

$$(s_1,s_2)\in\{(0.8,0.5),(0.8,0.2),(0.5,0.8),(0.5,0.2),(0.2,0.8),(0.2,0.5)\},$$

相应的结果总结在了表 1 中. 在表 1 中，BS 代表偏差，SE 代表标准误，而 $e(\widetilde{R},\widehat{R})$, 则代表 $\widetilde{R}(s_1,s_2)$ 对 $\widehat{R}(s_1,s_2)$ 的相对效率. 从表中可看出，$\widehat{R}(s_1,s_2)$, $\widetilde{R}(s_1,s_2)$ 和 $\overline{R}(s_1,s_2)$ 的偏差都接近于 0, 这说明这三种估计量都是渐近无偏的. 对于每一个 (s_1,s_2), 尽管 $\widetilde{R}(s_1,s_2)$ 和 $\overline{R}(s_1,s_2)$ 的标准误比较接近，但 $\widetilde{R}(s_1,s_2)$ 的标准误还是比 $\widehat{R}(s_1,s_2)$ 和 $\overline{R}(s_1,s_2)$ 都小. 因而，相对效率 $e(\widetilde{R},\widehat{R})$ 总是大于 1.7, 这说明在密度函数比模型正确的情形下，我们提出的半参数方法比非参数方法明显的好. 另外，半参数方法对参数方法的相对效率 $e(\widetilde{R},\overline{R})$ 总是比 1 稍大，这说明尽管在正态分布假设正确的情形下，我们的半参数方法也不比参数方法逊色，甚至于还稍优.

在统计模拟的另外一种情形里，我们采用了偏态的分布，这里我们举了两个例子：指数分布和对数正态分布，分别对应 $r(x)=x$ 和 $r(x)=\log x$. 在第一个例子里，我们假定 $f_i(x)$ 是指数分布 $E(\theta_i)$ 的密度函数，$i=1,2,3$. 因而这时对参数方法而言，正态分

布的假设就不再成立了. 但是, 密度函数比模型 (2) 的假设仍然成立, 且有 $r(x)=x$ 和

$$\alpha_1=\log\Big(\frac{\theta_3}{\theta_1}\Big),\qquad \alpha_2=\log\Big(\frac{\theta_3}{\theta_2}\Big),\qquad \beta_1=\frac{1}{\theta_3}-\frac{1}{\theta_1},\qquad \beta_2=\frac{1}{\theta_3}-\frac{1}{\theta_2}.$$

我们设置的样本容量是 $(n_1,n_2,n_3)=(30,30,30)$ 和 $(n_1,n_2,n_3)=(60,45,30)$ 两种情形. 另外, 设置 $\theta_1=1,\ \theta_2=2,\ \theta_3=3$. 这样, 模型 (2) 参数的真实值为 $\alpha_1=1.098,\ \alpha_2=0.405,\ \beta_1=-0.667,\ \beta_2=-0.167$. 对于每种情形 (n_1,n_2,n_3), 我们从 $E(\theta_i)\ \ (i=1,2,3)$ 中产生 1000 组相互独立的合并样本.

表 1 样本从正态分布抽取时的三种 ROC 曲面估计量的比较结果

(s_1,s_2)	$(0.8,0.5)$	$(0.8,0.2)$	$(0.5,0.8)$	$(0.5,0.2)$	$(0.2,0.8)$	$(0.2,0.5)$
$(n_1,n_2,n_3)=(30,30,30)$						
BS($\widehat{R}(s_1,s_2)$)	0.00617	0.01078	−0.01363	−0.00825	−0.01462	−0.01385
BS($\widetilde{R}(s_1,s_2)$)	−0.00283	−0.00680	0.00551	−0.01278	0.00955	−0.00476
BS($\overline{R}(s_1,s_2)$)	−0.00644	−0.00993	−0.00186	−0.01179	−0.00161	−0.00805
SE($\widehat{R}(s_1,s_2)$)	0.14346	0.13224	0.14375	0.09222	0.13892	0.09772
SE($\widetilde{R}(s_1,s_2)$)	0.10133	0.09974	0.10643	0.06323	0.10531	0.06445
SE($\overline{R}(s_1,s_2)$)	0.10508	0.10474	0.11064	0.07266	0.11141	0.07677
$e(\widetilde{R},\widehat{R})$	2.007	1.761	1.835	2.060	1.745	2.332
$e(\widetilde{R},\overline{R})$	1.078	1.108	1.078	1.302	1.110	1.427
$(n_1,n_2,n_3)=(60,45,30)$						
BS($\widehat{R}(s_1,s_2)$)	−0.00678	0.00038	−0.02035	−0.00816	−0.02280	−0.01776
BS($\widetilde{R}(s_1,s_2)$)	−0.00034	−0.00080	0.00472	−0.00545	0.00575	−0.00396
BS($\overline{R}(s_1,s_2)$)	−0.00497	−0.00532	−0.00054	−0.00842	−0.00181	−0.00935
SE($\widehat{R}(s_1,s_2)$)	0.11527	0.10272	0.12878	0.07207	0.12330	0.08504
SE($\widetilde{R}(s_1,s_2)$)	0.08261	0.07825	0.09234	0.04975	0.09318	0.05747
SE($\overline{R}(s_1,s_2)$)	0.08433	0.08046	0.09646	0.05814	0.09685	0.06632
$e(\widetilde{R},\widehat{R})$	1.954	1.723	1.989	2.101	1.804	2.274
$e(\widetilde{R},\overline{R})$	1.046	1.062	1.088	1.378	1.077	1.351

该情形下的统计模拟结果总结在了表 2 中. 从表中可看出, 除了只有一个值大于 0.01, $\widehat{R}(s_1,s_2)$ 和 $\widetilde{R}(s_1,s_2)$ 的偏差都接近于 0, 这说明此时非参数和半参数估计量都是渐近无偏的. 但这对参数估计量并不成立, 因为表中 $\overline{R}(s_1,s_2)$ 的偏差绝大多数都显著的不等于 0. 这说明在正态假设不成立时, 参数方法得到的估计量并不是渐近无偏的. 另外, 对每一个 (s_1,s_2), $\widetilde{R}(s_1,s_2)$ 的标准误比 $\widehat{R}(s_1,s_2)$ 和 $\overline{R}(s_1,s_2)$ 都小. 因此造成半参数估计量对其它两种估计量的相对效率 $e(\widetilde{R},\widehat{R})$ 和 $e(\widetilde{R},\overline{R})$ 都总是大于 1.6, 这说明半参数方法优于其它两种方法. 换句话说, 在正态假设不成立时, 半参数方法是一种理想的选择.

类似的模拟结果也出现在选择对数正态分布作模拟时. 在第 2 个例子里, 我们假定 $f_i(x)$ 是对数正态分布 $\log N(\theta_i,\sigma^2)$ 的密度函数, $i=1,2,3$. 因而这时对参数方法而言, 正态分布的假设就不再成立了. 但是, 密度函数比模型 (2) 的假设仍然成立, 且有

$r(x)=\log x$ 和

$$\alpha_1=\frac{\theta_1^2-\theta_3^2}{\sigma^2},\qquad \alpha_2=\frac{\theta_2^2-\theta_3^2}{\sigma^2},\qquad \beta_1=\frac{\theta_3-\theta_1}{\sigma^2},\qquad \beta_2=\frac{\theta_3-\theta_2}{\sigma^2}.$$

我们设置的样本容量是 $(n_1,n_2,n_3)=(30,30,30)$ 和 $(n_1,n_2,n_3)=(60,45,30)$ 两种情形. 另外，设置 $\theta_1=0$，$\theta_2=1$，$\theta_3=2$，$\sigma=1$. 这样，模型 (2) 参数的真实值为 $\alpha_1=-4$, $\alpha_2=-3$，$\beta_1=2$，$\beta_2=1$. 对于每种情形 (n_1,n_2,n_3), 我们从 $\log N(\theta_i,\sigma^2)$ $(i=1,2,3)$ 中产生 1000 组相互独立的合并样本. 该情形下的统计模拟结果总结在了表 3 中. 从表中可看出， $\widehat{R}(s_1,s_2)$ 和 $\widetilde{R}(s_1,s_2)$ 的偏差都接近于 0, 这说明此时非参数和半参数估计量都是渐近无偏的. 但这对参数估计量并不成立，因为表中 $\overline{R}(s_1,s_2)$ 的偏差绝大多数都显著的不等于 0. 这说明在正态假设不成立时，参数方法得到的估计量并不是渐近无偏的. 另外，对每一个 (s_1,s_2), $\widetilde{R}(s_1,s_2)$ 的标准误比 $\widehat{R}(s_1,s_2)$ 和 $\overline{R}(s_1,s_2)$ 都小. 因此造成半参数估计量对其它两种估计量的相对效率 $e(\widetilde{R},\widehat{R})$ 和 $e(\widetilde{R},\overline{R})$ 都总是大于 1.3, 这说明半参数方法优于其它两种方法.

表 2　样本从指数分布抽取时的三种 ROC 曲面估计量的比较结果

(s_1,s_2)	$(0.8,0.5)$	$(0.8,0.2)$	$(0.5,0.8)$	$(0.5,0.2)$	$(0.2,0.8)$	$(0.2,0.5)$
$(n_1,n_2,n_3)=(30,30,30)$						
BS$(\widehat{R}(s_1,s_2))$	−0.00473	−0.00700	−0.00487	−0.01486	−0.01209	−0.01981
BS$(\widetilde{R}(s_1,s_2))$	0.00934	−0.00622	0.01200	−0.00889	0.01307	0.00774
BS$(\overline{R}(s_1,s_2))$	0.13059	0.10055	−0.06467	0.00673	−0.11749	−0.01605
SE$(\widehat{R}(s_1,s_2))$	0.12879	0.12545	0.12084	0.12017	0.12061	0.13259
SE$(\widetilde{R}(s_1,s_2))$	0.09995	0.08604	0.07469	0.06650	0.06389	0.08697
SE$(\overline{R}(s_1,s_2))$	0.12115	0.10489	0.07575	0.08492	0.06820	0.13527
$e(\widetilde{R},\widehat{R})$	1.648	2.121	2.556	2.257	3.456	2.357
$e(\widetilde{R},\overline{R})$	3.149	2.837	1.733	1.612	4.340	2.434
$(n_1,n_2,n_3)=(60,45,30)$						
BS$(\widehat{R}(s_1,s_2))$	−0.01517	−0.01126	−0.00683	−0.01409	−0.01126	−0.02243
BS$(\widetilde{R}(s_1,s_2))$	0.00338	−0.00897	0.00843	−0.00912	0.01022	0.00502
BS$(\overline{R}(s_1,s_2))$	0.12479	0.09888	−0.06532	0.00714	−0.11740	−0.01902
SE$(\widehat{R}(s_1,s_2))$	0.12557	0.10603	0.10955	0.09902	0.11121	0.12711
SE$(\widetilde{R}(s_1,s_2))$	0.09128	0.07378	0.06826	0.06006	0.06229	0.08484
SE$(\overline{R}(s_1,s_2))$	0.11105	0.09091	0.07049	0.07826	0.06887	0.12913
$e(\widetilde{R},\widehat{R})$	1.918	2.058	2.547	2.710	3.136	2.306
$e(\widetilde{R},\overline{R})$	3.345	3.266	1.953	1.673	4.649	2.358

4　实例分析

我们将提出的方法用于一个真实数据的分析. 该数据源于 Reaven 和 Miller [18] 对糖尿病的研究. 在该数据中，有 145 个非肥胖的成年人被分成了三类：正常，糖尿病前期，糖尿病，这三类人群的数量分别是 76, 36 和 33. 我们以空腹血浆葡萄糖水平 (PLG) 为例来进行 ROC 曲面分析. 对正常人群 PLG 数据进行正态性检验，发现其不符合正态

分布，故参数方法不适用．我们将提出的半参数方法运用到该数据的分析上，发现对应于 $r(x)=x$ 的模型 (2) 有较好的拟合．模型参数的估计值为 $\widetilde{\alpha}_1=8.79384$, $\widetilde{\alpha}_2=5.39702$, $\widetilde{\beta}_1=-0.04219$, $\widetilde{\beta}_2=-0.02062$．对于 ROC 曲面的估计，我们以 $R(0.2,0.4)$ 为例．其半参数估计值为 $\widetilde{R}(0.2,0.4)=0.94757$, 相应的 95% 的置信区间为 $(0.87956,0.98100)$. 至于对 VUS 的分析，其半参数估计值为 $\widetilde{\mathrm{VUS}}=0.69084$, 相应的 95% 的置信区间为 $(0.60329,0.79527)$.

表 3 样本从对数正态分布抽取时的三种 ROC 曲面估计量的比较结果

(s_1,s_2)	$(0.8,0.5)$	$(0.8,0.2)$	$(0.5,0.8)$	$(0.5,0.2)$	$(0.2,0.8)$	$(0.2,0.5)$
$(n_1,n_2,n_3)=(30,30,30)$						
$\mathrm{BS}(\widehat{R}(s_1,s_2))$	0.00317	0.01008	−0.01583	−0.00618	−0.02049	−0.01775
$\mathrm{BS}(\widetilde{R}(s_1,s_2))$	0.00179	0.00023	0.00454	−0.00854	0.00554	−0.00598
$\mathrm{BS}(\overline{R}(s_1,s_2))$	0.07907	0.03789	−0.43243	−0.10989	−0.45573	−0.09201
$\mathrm{SE}(\widehat{R}(s_1,s_2))$	0.14385	0.13639	0.14155	0.09301	0.13100	0.09122
$\mathrm{SE}(\widetilde{R}(s_1,s_2))$	0.10066	0.10012	0.09936	0.06306	0.09762	0.06008
$\mathrm{SE}(\overline{R}(s_1,s_2))$	0.13782	0.11055	0.15266	0.07723	0.15390	0.15809
$e(\widetilde{R},\widehat{R})$	2.043	1.866	2.051	2.146	1.839	2.369
$e(\widetilde{R},\overline{R})$	2.491	1.363	21.258	4.455	24.120	9.179
$(n_1,n_2,n_3)=(60,45,30)$						
$\mathrm{BS}(\widehat{R}(s_1,s_2))$	−0.00838	0.00031	−0.01589	−0.00665	−0.01864	−0.01809
$\mathrm{BS}(\widetilde{R}(s_1,s_2))$	0.00040	−0.00248	0.00707	−0.00562	0.00884	−0.00097
$\mathrm{BS}(\overline{R}(s_1,s_2))$	0.07388	0.03265	−0.44513	−0.11235	−0.46562	−0.09162
$\mathrm{SE}(\widehat{R}(s_1,s_2))$	0.11570	0.10579	0.12551	0.07330	0.12290	0.08374
$\mathrm{SE}(\widetilde{R}(s_1,s_2))$	0.07951	0.07736	0.09000	0.04895	0.09172	0.05434
$\mathrm{SE}(\overline{R}(s_1,s_2))$	0.12032	0.09512	0.14434	0.06256	0.14682	0.14236
$e(\widetilde{R},\widehat{R})$	2.129	1.868	1.964	2.232	1.820	2.485
$e(\widetilde{R},\overline{R})$	3.153	1.688	26.866	6.812	28.076	9.704

5 讨论

在这篇文章里，我们提出了一种用于连续型诊断测试方法 ROC 曲面构建的半参数方法．统计模拟显示，无论是正态假设成立与否，本文的半参数方法均比非参数和参数方法优越．除此而外，还有一些好的性质值得我们注意．第一点值得注意的是本文提出的方法很好实现，这是因为很多统计软件里的逻辑斯蒂回归程序可以被利用．实施本文的方法中重要的一步是获得 $(\widetilde{\alpha}_i,\widetilde{\beta}_i)$ 的最大半参数似然估计值．传统的方法是采用一些象牛顿法之类解非线性方程组的数值计算方法．但在这里有一个简便的方法，那就是采用逻辑斯蒂回归程序．以 $(\widehat{\alpha}_i^*,\widehat{\beta}_i)$ 代表采用逻辑斯蒂回归得到的模型 (1) 中 (α_i^*,β_i) 的估计值．类似于 Zhang[19] 的分析，可以得出 $(\widetilde{\alpha}_i,\widetilde{\beta}_i)$ 和 $(\widehat{\alpha}_i^*,\widehat{\beta}_i)$ 有如下的联系 $\widehat{\alpha}_i^*=\widetilde{\alpha}_i+\log\rho_i$, $\widehat{\beta}_i=\widetilde{\beta}_i$. 另外一点值得注意的是，本文提出的半参数方法构建的 ROC 曲面从图形上看比非参数方法构建的曲面要光滑一些，这是因为半参数分布函数估计量 $\widetilde{F}_1(x)$, $\widetilde{F}_2(x)$ 和 $\widetilde{F}_3(x)$ 中的每一个都是以全部样本数据为支撑的．类似的现象在 Qin, Zhang[10], Wan, Zhang[12] 也出现过，即半参数 ROC 曲线或曲面估计量通常比相应的非

参数估计量来得光滑.

参 考 文 献

[1] Nakas C T, Yiannoutsos C T. Ordered Multiple-class ROC Analysis with Continuous Measurements. *Stat. Med.*, 2004, 23: 3437–3449

[2] Xiong C, Belle G, Miller J P, Morris J C. Measuring and Estimating Diagnostic Accuracy When There Are Three Ordinal Diagnostic Groups. *Stat. Med.*, 2006, 25: 1251–1273

[3] Yang H, Carlin D. ROC Surface: A Generalization of ROC Curve Analysis. *J. Biopharm. Stat.*, 2000, 10: 183–196

[4] Wan S, Zhang B. Semiparametric ROC Surfaces for Continuous Diagnostic Tests Based on Two Test Measurements. *Statist. Med.*, 2009, 28: 2370–2383

[5] Mossman D. Three-way ROCs. *Med. Decis. Making*, 1999, 19: 78–89

[6] Dreiseitl S, Ohno-machado L, Binder M. Comparing Three-class Diagnostic Tests by Three-way ROC Analysis. *Med. Decis. Making*, 2000, 20: 323–331

[7] Heckerling P S. Parametric Three-way Receiver Operating Characteristic Surface Analysis Using Mathematica. *Med. Decis. Making*, 2001, 21: 409–417

[8] Nakas C T, Alonzo T A. ROC Graphs for Assessing the Ability of a Diagnostic Marker to Detect Three Disease Classes with an Umbrella Ordering. *Biometrics*, 2007, 63: 603–609

[9] Alonzo T A, Nakas C T. Comparison of ROC Umbrella Volumes with an Application to the Assessment of Lung Cancer Diagnostic Markers. *Biometrical J.*, 2007, 49: 654–664

[10] Qin J, Zhang B. Using Logistic Regression Procedures for Estimating Receiver Operating Characteristic Curves. *Biometrika*, 2003, 90: 585–596

[11] Wan S, Zhang B. Smooth Semiparametric Receiver Operating Characteristic Curves for Continuous Diagnostic Tests. *Stat. Med.*, 2007, 26: 2565–2586

[12] Wan S, Zhang B. Comparing Correlated ROC Curves for Continuous Diagnostic Tests under Density Ratio Models. *Comput. Statist. Data Anal.*, 2008, 53: 233–245

[13] Qin J, Zhang B. A Goodness of Fit Test for Logistic Regression Models Based on Case-control Data. *Biometrika*, 1997, 84: 609–618

[14] Kay R, Little S. Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data. *Biometrika*, 1987, 74: 495–501

[15] Owen A B. Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, 1988, 75: 237–249

[16] Owen A B. Empirical Likelihood Confidence Regions. *Ann. Statist.*, 1990, 18: 90–120

[17] Qin J, Lawless J F. Empirical Likelihood and Estimating Equations. *Ann. Statist.*, 1994, 22: 300–325

[18] Reaven G M, Miller R G. An Attempt to Define the Nature of Chemical Diabetes Using a Multimensional Analysis. *Diabetologia*, 1979, 16: 17–24

[19] Zhang B. Prospective and Retrospective Analyses under Logistic Regression Models. *J. Multivariate Anal.*, 2006, 97: 211–230

A Semiparametric Method for ROC Surface Estimation

WAN Shuwen

(*Department of Applied Mathematics, Nanjing University of Finance and Economics, Nanjing* 210046)

(*E-mail: wanshuwen@aliyun.com*)

Abstract We propose a semiparametric method of estimating ROC surfaces for continuous diagnostic tests under density ratio models. Implementation of our method is easy since the usual logistic regression procedures in many statistical softwares can be employed. Simulation results show that the proposed semiparametric ROC surface estimator is more efficient than the nonparametric counterpart and the parametric counterpart whether the normality assumption of data holds or not.

Key words density ratio model; empirical likelihood; logistic regression; ROC curve; ROC surface

MR(2000) Subject Classification 62G99; 62H12; 62H15

Chinese Library Classification O212.1; O212.7