应用概率统计　第二十七卷
第二期　2011年4月

Chinese Journal of Applied Probability
and Statistics Vol.27 No.2 Apr. 2011

数据非随机缺失时单变量的分布函数估计

陆　福　忠

(嘉兴学院数学与信息工程学院, 嘉兴, 314001)

摘　要

本文研究数据非随机缺失下的分布函数估计问题. 在确定缺失数据是否属于某些指定区间的前提下, 对一维随机变量Y的分布函数$F(y)$作出了估计. 此时, 假定数据缺失机制形式已知, 但包含某未知多维参数θ. 本文证明了未知参数θ的估计量$\widehat{\theta}$的相合性和渐近正态性, 也证明了分布函数$F(y)$的估计量$\widehat{F}(y)$的相合性和渐近正态性.

关键词: 数据缺失机制, 分布函数, 区间数据.

学科分类号: O212.7.

§1. 引　　言

数据缺失下进行分布函数估计是数理统计学最基本的问题之一, 文献中往往运用非参数极大似然估计法估计其分布函数, 如Hu & Lawless (1998), Leign (1988)等等. 上述文献只是考虑到了简单形式的数据缺失机制, 都是假定数据缺失机制已知或是形式已知但包含一个未知参数. 本文研究数据缺失机制需要用多维参数来刻画的情形. 问题如下: 我们需要估计某一维随机变量Y的分布函数$F(y)$, 于是做了样本量为N的随机抽样, 但我们只得到n个数据, 其余$N-n$个数据缺失. 考虑到可识别性, 我们设计了新的调查方案使得我们能够获得足够的额外信息. 本文允许数据缺失机制包含未知参数, 目标仍然是对随机变量Y的分布函数作估计.

用随机变量M来指示随机变量Y是否缺失. 如果y_i被观察到, 也就是说y_i出现在样本中, 即给出了具体数据为y_i, 则令$M_i=0$; 如果y_i没有被观察到, 即缺失, 则记$M_i=1$. 数据y_i缺失时, 得到的是区间形式的数据, 即数据y_i在的一个已知区间. 本文的数据缺失机制是非随机的, 所用到的区间是事先设计好的. 本文只需要利用这些区间各自包含有多少个缺失的值. 由于是非随机缺失, 要一致地估计出分布函数$F(y)$, 必须要考虑数据缺失机制的影响.

我们得到的数据如下:

本文2008年8月11日收到, 2010年12月22日收到修改稿.

$$\left(\begin{array}{cc} Y_1 & M_1=0 \\ \cdots & \cdots \\ Y_n & M_{n_o}=0 \\ (a_1,b_1) & M_{n_o+1}=1 \\ \cdots & \cdots \\ (a_1,b_1) & M_{n_o+n_1}=1 \\ (a_2,b_2) & M_{n_o+n_1+1}=1 \\ \cdots & \cdots \\ (a_2,b_2) & M_{n_o+n_1+n_2}=1 \\ \cdots & \cdots \\ (a_k,b_k) & M_{n_o+n_1+\cdots+n_{k-1}+1}=1 \\ \cdots & \cdots \\ (a_k,b_k) & M_N=M_{n_o+n_1+n_2+\cdots+n_k}=1 \end{array}\right), \tag{1.1}$$

各次观察相互独立, 其中$(a_1,b_1),\cdots,(a_k,b_k)$是我们自行设计的区间, 如对全空间的某个互不相交的划分. 我们将在第§4节进一步讨论区间的来源以及其它相关的问题.

假定数据缺失机制如下:

$$\mathsf{P}(M_i=0|Y_i=y_i,\theta)=g(y_i,\theta)>0, \tag{1.2}$$

其中函数g的形式已知, θ为多维未知参数, 函数g关于变量θ连续. 更详细的假定条件在下文给出. 这一数据缺失机制显然是非随机的数据缺失, 现在的目标仍然是估计随机变量Y的分布函数$F(y)$, 当然还包括未知参数θ, 也即要估计$(\theta,F(y))$.

§2. 对未知参数θ的估计

首先, 给出本文的基本条件.

A1 Y为取值于区间$\mathrm{u}\subseteq\mathbf{R}$的一维随机变量, 其分布函数记为$F(y)$, 区间$(a_1,b_1),\cdots,(a_k,b_k)$是对区间u的互不相交的划分. 即

$$\bigcup_j((a_j,b_j)|j\in\{1,\cdots,k\})=\mathrm{u},$$
$$(a_i,b_i)\cap(a_j,b_j)=\emptyset,\quad \forall\, i\neq j.$$

A2 未知参数θ所在的参数空间Θ为$\mathbf{R}^k$中的开集, 参数θ的真值记为θ_0.

A3 数据缺失机制中的函数$g(y,\theta)$对于变量y连续, 对于θ二次连续可微, 在其定义域内满足

$$0<g(y,\theta)\leq 1, \tag{2.1}$$

且存在真值θ_0的某领域$\mathbf{B}(\theta_0,\delta_0)$, 使得

$$g(y,\theta)\geq C_0>0, \tag{2.2}$$

其中C_0是某一确定常数. 邻域$\mathbf{B}(\theta_0,\delta_0)\triangleq(\theta\in\mathbf{R}^k:\|\theta-\theta_0\|\leq\delta_0)$, 其中, 记号$\|\cdot\|$是条件A4中的范数.

A4 本文中出现的范数记号$\|\cdot\|$均指欧氏空间$\mathbf{R}^n$的基本范数, 即: 对任意的$x=(x_1,\cdots,x_n)\in\mathbf{R}^n$, 令$\|x\|=\sqrt{x_1^2+x_2^2+\cdots+x_n^2}$.

A5 存在真值θ_0的某领域$\mathbf{B}(\theta_0,\delta)$, 使得函数$g(y,\theta)$在$(\mathrm{u},\mathbf{B}(\theta_0,\delta))$内, 对于任意的$i\in\{1,\cdots,k\}$, $j\in\{1,\cdots,k\}$, 满足条件$|\dot{g}_{\theta_i}(y,\theta)|\leq h_1(y)$, 和$|\ddot{g}_{\theta_i,\theta_j}(y,\theta)|\leq h_2(y)$. 其中, $\dot{g}_{\theta_i}$表示对θ_i求导数, $\ddot{g}_{\theta_i,\theta_j}$表示分别对$\theta_i$和$\theta_j$求导数. 要求函数$h_1(y)$, $h_2(y)$满足

$$0<\int h_1^2(y)\mathrm{d}F(y)<\infty,$$
$$0<\int h_2(y)\mathrm{d}F(y)<\infty.$$

A6 对于条件A5中指定的δ, 对任意的小正数$\delta'\leq\delta$, 存在$\varepsilon_{\delta'}>0$, 使得

$$\lim_{N\to\infty}\Big(\sup_{\|\theta-\theta_0\|=\delta'}\{(\theta-\theta_0)^T\times\dot{C}_N(\theta)\times(\theta-\theta_0)\}\Big)\overset{\text{a.s.}}{<}-\varepsilon_{\delta'}, \tag{2.3}$$

其中记号$\cdot$表示对参数θ求导, $C_N(\theta)$的具体表达式见(2.6)式.

A7 $k\times k$矩阵$\mathsf{E}(\dot{C}_N(\theta_0))$可逆. $\mathsf{E}(\dot{C}_N(\theta_0))$的具体表达式为

$$\begin{pmatrix}
-\mathsf{E}\Big(\frac{\dot{g}_{\theta_1}*I(Y\in(a_1,b_1))}{g(Y,\theta)}\Big) & -\mathsf{E}\Big(\frac{\dot{g}_{\theta_2}*I(Y\in(a_1,b_1))}{g(Y,\theta)}\Big) & \cdots & -\mathsf{E}\Big(\frac{\dot{g}_{\theta_k}*I(Y\in(a_1,b_1))}{g(Y,\theta)}\Big)\\
-\mathsf{E}\Big(\frac{\dot{g}_{\theta_1}*I(Y\in(a_2,b_2))}{g(Y,\theta)}\Big) & -\mathsf{E}\Big(\frac{\dot{g}_{\theta_2}*I(Y\in(a_2,b_2))}{g(Y,\theta)}\Big) & \cdots & -\mathsf{E}\Big(\frac{\dot{g}_{\theta_k}*I(Y\in(a_2,b_2))}{g(Y,\theta)}\Big)\\
\cdots & \cdots & \cdots & \cdots\\
-\mathsf{E}\Big(\frac{\dot{g}_{\theta_1}*I(Y\in(a_k,b_k))}{g(Y,\theta)}\Big) & -\mathsf{E}\Big(\frac{\dot{g}_{\theta_2}*I(Y\in(a_k,b_k))}{g(Y,\theta)}\Big) & \cdots & -\mathsf{E}\Big(\frac{\dot{g}_{\theta_k}*I(Y\in(a_k,b_k))}{g(Y,\theta)}\Big)
\end{pmatrix}. \tag{2.4}$$

本文构建估计方程来估计未知参数θ. 不难证明, 对于参数空间Θ中的任何θ和所有可能的分布$F(y)$, 如下方程成立.

命题 2.1 各种记号如上, 在条件A1-A3下, 如下方程成立

$$\begin{aligned}
&\mathsf{E}\Big(\frac{I(M=0)I(Y\in(a_1,b_1))}{g(Y,\theta)}-I(Y\in(a_1,b_1))\Big)=0,\\
&\mathsf{E}\Big(\frac{I(M=0)I(Y\in(a_2,b_2))}{g(Y,\theta)}-I(Y\in(a_2,b_2))\Big)=0,\\
&\qquad\vdots\\
&\mathsf{E}\Big(\frac{I(M=0)I(Y\in(a_k,b_k))}{g(Y,\theta)}-I(Y\in(a_k,b_k))\Big)=0.
\end{aligned} \tag{2.5}$$

证明: 直接运用条件期望公式, 条件A1-A3保证了(2.5)式中的各个期望有限. □

根据上述命题可以构造θ的估计量$\widehat{\theta}$. 记

$$\begin{aligned} C_j(\theta) &= \frac{\sum_{i=1}^{N} C_{ji}}{N} \\ &\triangleq \frac{\sum_{i=1}^{N} I(Y_i \in (a_j, b_j)) * \left(\frac{I(M_i = 0)}{g(Y_i, \theta)} - 1\right)}{N}, \qquad j \in \{1, \cdots, k\}, \\ C_N(\theta) &= (C_1(\theta), \cdots, C_k(\theta)). \end{aligned} \tag{2.6}$$

则方程(2.5)可以表示为$\mathsf{E}(C_N(\theta)) = 0$, 这里的期望$\mathsf{E}$指对于随机向量$(Y, M)$的分布求期望. 所以方程$C_N(\theta) = 0$为一无偏估计方程. 则满足$C_N(\theta) = 0$的估计量$\widehat{\theta}$是未知参数$\theta$的一个合理的估计量. 即

$$\begin{aligned} C_1(\theta) &= 0, \\ &\vdots \\ C_k(\theta) &= 0. \end{aligned} \tag{2.7}$$

上式的解记为$\widehat{\theta}$. 即为未知参数θ的估计.

一般地, 估计方程(2.7)可能会有多个解, 但是本文的条件A1-A7可以保证估计方程(2.7)的解唯一.

引理 2.1 在条件A1-A5下, 对$\forall\, \varepsilon > 0$, 存在$\delta > 0$, 使得

$$\lim_{N \to \infty} \left(\sup_{\theta, \theta' \in B(\theta_0, \delta)} (\|\dot{C}_N(\theta) - \dot{C}_N(\theta')\|) \right) \overset{\text{a.s.}}{<} \varepsilon. \tag{2.8}$$

证明: 应用条件A1-A5, 细致的证明可见陆福忠(2007)第三章. □

下面的定理是关于估计量$\widehat{\theta}$的存在唯一性质的.

定理 2.1 在条件A1-A7下, 估计方程(2.7)在条件A5指定的邻域内有解$\widehat{\theta}$, 且在该邻域内解唯一, 并且$\widehat{\theta}$是未知参数θ的强相合估计, 即$\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0$. 若方程(2.7)在条件A5指定的邻域之外还有解, 则此解不会是相合的.

证明: 由于$C(\theta_0) \triangleq \lim\limits_{N \to \infty} C_N(\theta_0) = \mathbf{0}$, 而$\widehat{\theta}$是估计方程$C_N(\theta) = \mathbf{0}$的解. 主要的证明过程是反函数定理的运用. 条件A7保证函数$C_N(\theta)$在真值θ_0的某邻域内渐近可逆, 引理2.1保证了函数$\dot{C}_N(\theta)$在条件A5指定的邻域内关于θ一致收敛到$\mathsf{E}(\dot{C}_N(\theta))$, 条件A6相当于似然函数估计中Hessian矩阵为负定. 对于此定理的细致证明, 可参照经典的估计方程理论和参数似然函数理论, 如Heyde (1997)和Foutz (1977). 本文的条件A6是与Heyde (1997)第十二章中(12.4)式等价的但是更加直观的条件, 也类似于Foutz (1977)文中条件(B). □

现在证明估计量$\widehat{\theta}$的渐近正态性质, 由于$\widehat{\theta}$是估计方程(2.7)的解, 为此, 我们首先证明$(C_N(\theta_0))^T=(C_1(\theta_0),\cdots,C_k(\theta_0))^T$的联合渐近正态性. 即须证明如下引理.

引理 2.2 在条件A1-A4下, $\sqrt{N}\times(C_1(\theta_0),\cdots,C_k(\theta_0))^T$渐近服从一联合正态分布, 即

$$\sqrt{N}\times(C_1(\theta_0),\cdots,C_k(\theta_0))^T\xrightarrow{\mathfrak{D}}\mathrm{MVN}(\mathbf{0},\Sigma_C),\tag{2.9}$$

其中协方差Σ_C的具体表达式在下文中给出.

证明: 由于$\{C_j(\theta_0),j\in\{1,\cdots,k\}\}$各项都是某独立同分布随机变量和的平均值, 因此可以直接应用多变量中心极限定理得到引理的结论, 即$C_N(\theta_0)$渐近服从某联合正态分布. 其协方差矩阵Σ_C可以直接计算得出, Σ_C为一对角矩阵, 其具体表达式为

$$\Sigma_C=\begin{pmatrix}\mathsf{E}(A_1)&0&\cdots&0\\0&\mathsf{E}(A_2)&\cdots&0\\\cdots&\cdots&\cdots&\cdots\\0&0&\cdots&\mathsf{E}(A_k)\end{pmatrix},\tag{2.10}$$

$$A_i=\frac{I(Y\in(a_i,b_i))*(1-g(Y,\theta))}{g(Y,\theta)},\quad i=1,\cdots,k.$$

由条件A1-A4不难得到, Σ_C的表达式中出现的各个期望项均存在有限. □

于是,有如下定理.

定理 2.2 在条件A1-A7下, 估计量$\widehat{\theta}$服从渐近正态分布, 即

$$\sqrt{N}\times(\widehat{\theta}-\theta_0)\xrightarrow{\mathfrak{D}}\mathrm{MVN}(\mathbf{0},\Sigma),$$

其中协方差矩阵Σ为

$$\Sigma=(\mathsf{E}(\dot{C}_N(\theta_0)))^{-1}\times\Sigma_C\times((\mathsf{E}(\dot{C}_N(\theta_0)))^{-1})^T,\tag{2.11}$$

其中$\mathsf{E}(\dot{C}_N(\theta_0))$见(2.4)式, 记号$T$表示矩阵的转置.

证明: 定理2.1已经证明了估计量$\widehat{\theta}$的强相合性, $\widehat{\theta}\overset{\text{a.s.}}{\to}\theta_0$, 从而可以将函数$C_N(\widehat{\theta})=(C_1(\widehat{\theta}),\cdots,C_k(\widehat{\theta}))$在$\theta_0$处作泰勒展开得

$$\mathbf{0}=C_N(\widehat{\theta})=C_N(\theta_0)+\dot{C}_N(\theta_0)(\widehat{\theta}-\theta_0)+\frac{1}{2}(\widehat{\theta}-\theta_0)^T\ddot{C}_N(\widetilde{\theta})(\widehat{\theta}-\theta_0),\tag{2.12}$$

其中$\widetilde{\theta}$位于线段$\widehat{\theta}\theta_0$之间. 第一项$C_N(\theta_0)$已由引理2.2得到其服从渐近正态分布, 第二项中的$\dot{C}_N(\theta_0)$也是独立同分布变量和的样本均值, 由大数定律, $\dot{C}_N(\theta_0)\overset{\text{a.s.}}{\to}\mathsf{E}(\dot{C}_N(\theta_0))$, 第三项中的$\ddot{C}_N(\widetilde{\theta})$是一个$k\times(k\times k)$数组, 它的第$(i,\cdot,\cdot)$, $i=1,\cdots,k$项都是$k\times k$矩阵. 由条件A5并经过代数计算后得知, $\ddot{C}_N(\widetilde{\theta})$项的每个元素都可以被不包含$\theta$的函数

$h_1(Y_1,\cdots,Y_n)$和$h_2(Y_1,\cdots,Y_n)$的某线性组合如$f(h_1(Y_1,\cdots,Y_n),h_2(Y_1,\cdots,Y_n))$所控制, 此时, $f(h_1(Y_1,\cdots,Y_n),h_2(Y_1,\cdots,Y_n))$服从大数定律, 即可以知道$\ddot{C}_N(\widetilde{\theta})$项依概率有界. 因此, 整理表达式(2.12)得

$$\begin{aligned} -C_N(\theta_0) &= \Big(\mathsf{E}(\dot{C}_N(\theta_0))+o_p(1)+\frac{1}{2}(\widehat{\theta}-\theta_0)*O_p(1)\Big)(\widehat{\theta}-\theta_0) \\ &= (\mathsf{E}(\dot{C}_N(\theta_0))+o_p(1))(\widehat{\theta}-\theta_0). \end{aligned}$$

上式成立是因为$\widehat{\theta}\xrightarrow{\text{a.s.}}\theta_0$, 从而$(\widehat{\theta}-\theta_0)*O_p(1)=o_p(1)*O_p(1)=o_p(1)$, 条件A7假定矩阵$\mathsf{E}(\dot{C}_N(\theta_0))$可逆, 所以, 矩阵$(\mathsf{E}(\dot{C}_N(\theta_0))+o_p(1))$依概率可逆. 于是有

$$\begin{aligned} \sqrt{N}\times(\widehat{\theta}-\theta_0) &= -(\mathsf{E}(\dot{C}_N(\theta_0))+o_p(1))^{-1}\times\sqrt{N}\times C_N(\theta_0) \\ &= -(\mathsf{E}(\dot{C}_N(\theta_0)))^{-1}\times\sqrt{N}\times C_N(\theta_0)+o_p(1). \end{aligned} \tag{2.13}$$

上式的成立是因为$\sqrt{N}\times C_N(\theta_0)$依概率有界. 由引理2.2的结论即可完成$\sqrt{N}\times(\widehat{\theta}-\theta_0)$的渐近正态性证明. 渐近方差$\Sigma$经计算得到, 即为表达式(2.11). □

§3. 对分布函数$F(x,y)$的估计

参数θ已经估计出, 从而$g(y,\theta)$已经估计出, 就可以进行随机变量Y的分布函数$F(y)$的估计. 同样用函数项$g(y,\theta)$对观察到的数据进行合适加权, 构造估计量$\widehat{F}(y)$如下

$$\begin{aligned} \widehat{F}(y,\widehat{\theta}) &= \sum_{i=1}^{N}\frac{\dfrac{I(Y_i\le y)}{g(Y_i,\widehat{\theta})}}{N} \\ &= \sum_{i=1}^{N}\frac{\dfrac{I(M_i=0)I(Y_i\le y)}{g(Y_i,\widehat{\theta})}}{N} \\ &= \sum_{i=1}^{N}\frac{I(M_i=0)I(Y_i\le y)}{N*g(Y_i,\widehat{\theta})}. \end{aligned} \tag{3.1}$$

现在证明$\widehat{F}(y,\widehat{\theta})$的渐近性质.

定理 3.1 在条件A1-A7下, $\widehat{F}(y,\widehat{\theta})$是分布函数$F(y)$的强相合估计量, 且是渐近正态的, 即

$$\widehat{F}(y,\widehat{\theta})\to F(y)\qquad \text{a.s.}$$

且

$$\sqrt{N}\times(\widehat{F}(y,\widehat{\theta})-F(y))\xrightarrow{\mathfrak{D}}N(0,\sigma^2),$$

其中渐近方差σ^2的具体表达式在下文中给出.

在证明定理3.1之前, 不妨先证明以下几个引理.

引理 3.1 在条件A1-A7下,

$$\widehat{F}(y,\widehat{\theta})-\widehat{F}(y,\theta_0)\stackrel{\text{a.s.}}{\to}0.$$

证明: 定理2.1已经证明了$\widehat{\theta}$的强相合性, 注意到恰当应用条件A1-A7, 有

$$\begin{aligned}
|\widehat{F}(y,\widehat{\theta})-\widehat{F}(y,\theta_0)| &= \left|\frac{\sum_{i=1}^{N}\frac{I(M_i=0)I(Y_i\le y)}{g(Y_i,\widehat{\theta})}}{N}-\frac{\sum_{i=1}^{N}\frac{I(M_i=0)I(Y_i\le y)}{g(Y_i,\theta_0)}}{N}\right|\\
&\le \left|\frac{\sum_{i=1}^{N}\frac{1}{g(Y_i,\widehat{\theta})}}{N}-\frac{\sum_{i=1}^{N}\frac{1}{g(Y_i,\theta_0)}}{N}\right|\\
&\le \text{Const}\times\left|\frac{\sum_{i=1}^{N}g(Y_i,\widehat{\theta})-g(Y_i,\theta_0)}{N}\right|\\
&\le \text{Const}\times\frac{\sum_{i=1}^{N}|g(Y_i,\widehat{\theta})-g(Y_i,\theta_0)|}{N}\\
&\le \text{Const}\times\frac{\sum_{i=1}^{N}\|\dot{g}(Y_i,\widetilde{\theta})\|}{N}\times\|\widehat{\theta}-\theta_0\|\\
&\le \text{Const}'\times\|\widehat{\theta}-\theta_0\|\\
&\stackrel{\text{a.s.}}{\to} 0.
\end{aligned}$$

引理得证. □

引理 3.2 在条件A1-A7下, 估计量

$$\widehat{F}(y,\theta_0)\stackrel{\text{a.s.}}{\to}F(y).$$

证明: 直接应用大数定律. □

引理 3.3 在条件A1-A4下, 对任意固定的$y\in\mathbf{u}$, $(C_1(\theta_0),\cdots,C_k(\theta_0),\widehat{F}(y,\theta_0)$–$F(y))^T$渐近服从联合正态分布, 即

$$\sqrt{N}\times(C_1(\theta_0),\cdots,C_k(\theta_0),\widehat{F}(y,\theta_0)-F(y))^T\stackrel{\mathfrak{D}}{\longrightarrow}\text{MVN}(\mathbf{0},\Sigma_{FC}).\tag{3.2}$$

方差Σ_{FC}的表达式在下文给出.

证明: 由于$\{C_j(\theta_0), j \in \{1,\cdots,k\}\}$和$\widehat{F}(y,\theta_0)-F(y)$都是某独立同分布随机变量和的平均值, 证明类似于引理2.2的证明. 其协方差矩阵Σ_{FC}可以直接计算得出, 其具体表达式为

$$\Sigma_{FC}=\begin{pmatrix}\Sigma_C & (\mathrm{VEC})^T\\ \mathrm{VEC} & \mathsf{E}\Big(\dfrac{I(Y\le y)}{g(Y,\theta_0)}\Big)-(F(y))^2\end{pmatrix}$$

其中VEC是一个$1\times k$维向量, 其第个j元素为

$$\mathrm{VEC}_j=\int_{a_j}^{b_j}\frac{I(a\le y)*(1-g(a,\theta_0))}{g(a,\theta_0)}\mathrm{d}F(a),\qquad j=1,\cdots,k. \tag{3.3}$$

注意到引理中的条件A1-A4的应用, 经过计算后不难得知, $\mathrm{VEC}_j<\infty$, $j=1,\cdots,k$, 且$\mathsf{E}(I(Y\le y)/g(Y,\theta_0))\ll\infty$, 从而协方差矩阵$\Sigma_{FC}$的每个元素均为有限值的. □

现在证明定理3.1.

证明: 相合性: 根据引理3.1和3.2得

$$\begin{aligned}\widehat{F}(y,\widehat{\theta})-F(y) &= \widehat{F}(y,\widehat{\theta})-\widehat{F}(y,\theta_0)+\widehat{F}(y,\theta_0)-F(y)\\ &\overset{\text{a.s.}}{\to} 0.\end{aligned}$$

渐近正态性: 将估计量

$$\widehat{F}(y,\widehat{\theta})=\sum_{i=1}^{N}\frac{I(M_i=0)I(Y_i\le y)}{N*g(Y_i,\widehat{\theta})}$$

在θ_0处作一阶泰勒展开, 并应用引理3.1和大数定律, 得到

$$\begin{aligned}\widehat{F}(y,\widehat{\theta})-F(y) &= \widehat{F}(y,\theta_0)-F(y)+\dot{\widehat{F}}(y,\widetilde{\theta})(\widehat{\theta}-\theta_0)\\ &= \widehat{F}(y,\theta_0)-F(y)+(\dot{\widehat{F}}(y,\theta_0)+o_p(1))(\widehat{\theta}-\theta_0)\\ &= \widehat{F}(y,\theta_0)-F(y)+(\mathsf{E}(\dot{\widehat{F}}(y,\theta_0))+o_p(1))(\widehat{\theta}-\theta_0).\end{aligned}$$

其中$\widetilde{\theta}$位于线段$\widehat{\theta}\theta_0$之间. 又根据表达式(2.13), 得到

$$\begin{aligned}\sqrt{N}\times(\widehat{F}(y,\widehat{\theta})-F(y)) &= \sqrt{N}\times(\widehat{F}(y,\theta_0)-F(y))-(\dot{\widehat{F}}(y,\theta_0)+o_p(1))\\ &\quad\times((\mathsf{E}(\dot{C}_N(\theta_0)))^{-1}\times\sqrt{N}\times C_N(\theta_0)+o_p(1))\\ &= \sqrt{N}\times(\widehat{F}(y,\theta_0)-F(y))-\mathsf{E}(\dot{\widehat{F}}(y,\theta_0))(\mathsf{E}(\dot{C}_N(\theta_0)))^{-1}\\ &\quad\times\sqrt{N}\times C_N(\theta_0)+o_p(1).\end{aligned} \tag{3.4}$$

注意到$\widehat{F}(y,\theta_0)$为独立同分布随机变量和的平均值, 而且根据引理3.3知道, $\sqrt{N}\times(\widehat{F}(y,\theta_0),C_N(\theta_0))$渐近服从联合正态分布, 因此, 表达式中(3.4)的$\sqrt{N}\times(\widehat{F}(y,\widehat{\theta})-F(y))$为某联合正

态分布的线性组合也服从联合正态分布, 即定理3.1成立. 直接计算渐近方差σ^2为

$$\begin{aligned}\sigma^2 &= (\mathsf{E}(\dot{\hat{F}})(\mathsf{E}(\dot{C}_N))^{-1}\Sigma_C - \mathrm{VEC}) \times (\mathsf{E}(\dot{\hat{F}})(\mathsf{E}(\dot{C}_N))^{-1})^T \\ &\quad - \mathsf{E}(\dot{\hat{F}})(\mathsf{E}(\dot{C}_N))^{-1}(\mathrm{VEC})^T + \int \frac{I(a \le y)}{g(a,\theta_0)}\mathrm{d}F(a) - F^2(y), \end{aligned} \tag{3.5}$$

其中$\mathsf{E}(\dot{\hat{F}})$项是一个$1 \times k$维向量, 其第个$j$元素为

$$\mathsf{E}(\dot{\hat{F}})_j = -\int \Big(\frac{I(a \le y)\dot{g}_{\theta_{0_j}}}{g(a,\theta_0)}\Big)\mathrm{d}F(a), \qquad j = 1, \cdots, k,$$

其中$\mathsf{E}(\dot{C}_N)$见(2.4)式, Σ_C见(2.10)式, VEC见(3.3)式. 应用条件A1-A7, 不难得到, $\mathsf{E}(\dot{\hat{F}})$的各个元素均有限, 上文中已经说明了其它项也是有限的, 所以方差σ^2为有限值. □

注记 1 (1) 如果没有任何数据缺失, 即函数$g(y,\theta) \equiv 1$, 则

$$\dot{g}(y,\theta) = \frac{\partial g(y,\theta)}{\partial \theta} \equiv \mathbf{0},$$

从而$\mathsf{E}(\dot{\hat{F}}) = \mathbf{0}$及$\mathrm{VEC} = \mathbf{0}$. 所以此时方差$\sigma^2 = F(y) - F^2(y)$, 即此时由于可以得到完全数据, 则本文的分布函数估计就与经验分布函数估计一致了.

(2) 如果θ已知, 即θ不需要估计, 则$\dot{g}(y,\theta) \equiv \mathbf{0}$, 从而$\mathsf{E}(\dot{\hat{F}}) = \mathbf{0}$, 所以此时方差

$$\sigma^2 = \iint \frac{I(a \le x, b \le y)}{g(a,b,\theta_0)}\mathrm{d}F(a,b) - F^2(x,y),$$

即此时方差大于经验分布函数估计的方差仅仅是因为有些数据的缺失, 且此时区间提供的信息没有起到作用, 方差没有因为有区间的信息而减少.

(3) 如果只有$l < k$个互不相交的区间, 则考察(2.4)式后可知, 此时矩阵$\mathsf{E}(\dot{C}_N)$的行列式为零, 则矩阵$\mathsf{E}(\dot{C}_N)$不可逆, 或认为矩阵$\mathsf{E}(\dot{C}_N)^{-1}$的行列式$|\mathsf{E}(\dot{C}_N)|^{-1}$为无穷大, 则方差$\sigma^2$为无穷大. 也意味此时分布函数不能够用本文的方法估计, 事实上, 本文要求的k个互不相交的区间是可以识别未知参数θ和分布函数$F(y)$的一个充分条件.

如果需要估计σ^2, 可以逐个估计(3.5)式中的各项. 可用

$$-\sum_{i=1}^{N} \frac{I(M_i = 0, Y_i \le y) * \dot{g}_{\widehat{\theta}_j}}{N * g^2(Y_i, \widehat{\theta})}$$

来估计$\mathsf{E}(\dot{\hat{F}})_j$, $j = 1, \cdots, k$, 用

$$\sum_{i=1}^{N} \frac{I(M_i = 0, Y_i \le y, Y_i \in (a_j, b_j)) * (1 - g(Y_i, \widehat{\theta}))}{N * g^2(Y_i, \widehat{\theta})}$$

来估计VEC_j, $j = 1, \cdots, k$, 用

$$\sum_{i=1}^{N} \frac{I(M_i = 0, Y_i \le y)}{N * g^2(Y_i, \widehat{\theta})}$$

来估计$\int [I(a \le y)/g(a,\theta_0)]\mathrm{d}F(a)$, 用$\widehat{F}(y,\widehat{\theta})$估计$F(y)$.

$\mathsf{E}(\dot{C}_N)$的估计为

$$-\begin{pmatrix} \sum\limits_{i=1}^{N}\dfrac{B_1*\dot{g}_{\widehat{\theta}_1}}{N*g^2(Y_i,\widehat{\theta})} & \sum\limits_{i=1}^{N}\dfrac{B_1*\dot{g}_{\widehat{\theta}_2}}{N*g^2(Y_i,\widehat{\theta})} & \cdots & \sum\limits_{i=1}^{N}\dfrac{B_1*\dot{g}_{\widehat{\theta}_k}}{N*g^2(Y_i,\widehat{\theta})} \\ \sum\limits_{i=1}^{N}\dfrac{B_2*\dot{g}_{\widehat{\theta}_1}}{N*g^2(Y_i,\widehat{\theta})} & \sum\limits_{i=1}^{N}\dfrac{B_2*\dot{g}_{\widehat{\theta}_2}}{N*g^2(Y_i,\widehat{\theta})} & \cdots & \sum\limits_{i=1}^{N}\dfrac{B_2*\dot{g}_{\widehat{\theta}_k}}{N*g^2(Y_i,\widehat{\theta})} \\ \cdots & \cdots & \cdots & \cdots \\ \sum\limits_{i=1}^{N}\dfrac{B_k*\dot{g}_{\widehat{\theta}_1}}{N*g^2(Y_i,\widehat{\theta})} & \sum\limits_{i=1}^{N}\dfrac{B_k*\dot{g}_{\widehat{\theta}_2}}{N*g^2(Y_i,\widehat{\theta})} & \cdots & \sum\limits_{i=1}^{N}\dfrac{B_k*\dot{g}_{\widehat{\theta}_k}}{N*g^2(Y_i,\widehat{\theta})} \end{pmatrix},$$

$$B_j = I(M_i = 0, Y_i \in (a_j, b_j)), \qquad j = 1, \cdots, k.$$

Σ_C的估计也是一个$k \times k$对角矩阵, 其对角线元素为

$$\sum_{i=1}^{N}\frac{I(M_i = 0, Y_i \in (a_j, b_j)) * (1 - g(Y_i, \widehat{\theta}))}{N * g^2(Y_i, \widehat{\theta})}, \qquad j = 1, \cdots, k.$$

由于已经证明$\widehat{\theta} \overset{\text{a.s.}}{\to} \theta_0$, 易知上述各个估计量均是相合的, 从而方差项$\sigma^2$也可以相合地估计出.

§4. 讨　　论

首先, 我们指出, 本文的观察数据中有一些区间形式的数据, 这些数据是用来识别未知多维参数θ的辅助信息. 如果未知参数θ是一维的, 就可以不需要这些辅助信息而能够识别出未知参数θ, Leigh (1988)已经做过这一工作, 不过, 该文献的目标是未知参数θ本身, 并没有讨论分布函数的估计的渐近性质. 而当未知参数θ是多维时, 没有一定的辅助信息是不能识别出θ的.

其次, 本文的方法与区间数据的分布函数估计方法完全不同. 区间数据的分布函数估计方法, 参阅郑祖康2004年的文章[6], 丁邦俊2002年文[7]和邓文丽2004年文[8]中的方法, 都要求得到的区间是随机区间, 区间的端点是活动的, 从而大量的观察可能得到逐渐变窄的区间. 而在本文中, 观察到的区间$\{(a_1, b_1), \cdots, (a_k, b_k)\}$是固定的, 因此, 对于本文的数据, 用区间数据的分布函数估计方法无法得到分布函数$F(y)$的估计.

最后, 如果区间是活动的, 即区间是被调查者自行报道的, 本文的方法仍然可行, 只是在构造估计方程前必须对区间作些预处理. 由于此时区间是被调查者随机报告的, 区间会有很多, 而我们估计k维参数θ只需要$k-1$个互不相交的区间(全空间也可以构造一个估计方程, 本文用k个区间只是为了有形式上的对称性), 不妨设除了观察到n个确定的值Y_i之外, 还观察到了$N-n$个区间$\{(c_j, d_j), j = 1, \cdots, N-n\}$, 缺失值$Y_{\text{mis}}$就分别隐藏在这些区间

中. 要想得到构造估计方程所需要的$k-1$个互不相交的区间, 可以采取如下的原则: 每个新的区间$\{(c_l', d_l'), l = 1, \cdots, k-1\}$都是某些原始区间$\{(c_j, d_j), j = 1, \cdots, N-n\}$的并, 且

$$\bigcup_l ((c_l', d_l')|l = 1, \cdots, k-1) = \bigcup_j ((c_j, d_j)|j = 1, \cdots, N-n).$$

为了提高估计效率, 每个新的区间$\{(c_l', d_l'), l = 1, \cdots, k-1\}$应该尽可能均匀地包含那些确定的值$Y_i$. 即没有哪个新区间包含的确定值$Y_i$的数量特别少. 这里, 只有一个互不相交的条件不太容易满足, 但是如果假设区间$\{(c_j, d_j), j = 1, \cdots, N-n\}$的宽度有一定的限制, 如每个区间$\{(c_j, d_j), j = 1, \cdots, N-n\}$的宽度都小于某个常数的条件下, 则不难得到互不相交的区间$\{(c_l', d_l'), l = 1, \cdots, k-1\}$. 得到了区间$\{(c_l', d_l'), l = 1, \cdots, k-1\}$后,用本文的方法就能够估计出$k$维参数$\theta$, 然后进行分布函数的估计. 此时, 若采用区间数据估计方法(该方法不需要n个确定的值Y_i), 原则上是可行的, 但是该方法必须考虑用来作区间截断的随机变量的诸多性质, 且对于分布函数$F(y)$本身有很强的限制, 如对其支撑集的限制, 对其连续性的限制等等. 而本文的方法不用考虑被调查者报告的区间$\{(c_j, d_j), j = 1, \cdots, N-n\}$是如何随机生成的, 只要求不要报告诸如$(0, \infty)$之类太宽的区间即可, 而且对分布函数$F(y)$几乎没有任何限制.

§5. 数值模拟

取随机变量Y为标准一元正态随机变量, 其分布函数$F(y)$为一元正态分布, 即$Y \sim N(0, 1)$. 数据缺失机制为

$$g(y, \theta_1, \theta_2) \triangleq \mathsf{P}(M = 0|y, \theta_1, \theta_2) = \Phi(\theta_1 + y * \theta_2),$$

其中$(\theta_1, \theta_2) = (0.1, 0.1)$, Φ为标准一元正态分布的分布函数. 先根据分布函数$F(y)$随机产生N个随机样本, 数据缺失机制中的函数$g(y, \theta_1, \theta_2)$是一个概率值, 它以这个概率值决定这个数据是否能够被观察到. 如果观察不到具体值, 则得到一个这个数据所在的一个区间. 所以, 最终观察到的是$(y_i, i \in \{1, 2, \cdots, n\})$和各缺失数据所在的区间, 这里只需要两个区间, 不妨设置为$(-\infty, 0]$和$(0, \infty)$. 我们只需要知道各缺失数据是否大于零. 记下区间$(-\infty, 0]$和$(0, \infty)$包含缺失数据的个数.

根据本文的方法, 可以得到(θ_1, θ_2)的估计量$(\widehat{\theta}_1, \widehat{\theta}_2)$和分布函数$F(y)$的估计量$\widehat{F}(y)$, 具体结果见表1.

对于分布函数$F(y)$的估计量$\widehat{F}(y)$的渐近性, 可以考察其与真实分布$F(y)$的差异, 即$\mathrm{err}(y) \triangleq \widehat{F}(y) - F(y)$. 且不妨只需考虑该项在所有观察项的范围, 即最大值和最小值, 结果见表2.

可以看出, 缺失机制函数的真值ϕ约为0.5394, 当N逐渐变大时, 其估计值越来越接近于真值. 当$N = 400$, 作300次模拟时, 平均每次约观察到216个样本值, 其它约184个数据观

察不到, 只知道它们所在的区间, 是大于零还是小于零. 两个待估参数的真值为$\theta_1 = 0.1$, $\theta_2 = 0.1$, 其估计值也与真值越来越接近. 分布函数的估计也与真值有相合的趋势, 300次模拟平均而言, 最大差距的绝对值也逐渐缩小.

表1 参数θ的估计量$\widehat{\theta}$(300次模拟)

真值	缺失概率 $\phi = 0.5394$		$\theta_1 = 0.1$	$\theta_2 = 0.1$	
N	$\widehat{\phi}$	$\widehat{\mathsf{Var}}(\widehat{\phi})$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\mathsf{Cov}}(\widehat{\theta}_1, \widehat{\theta}_2)$
50	0.5467	0.00483022	0.15245210	0.09611547	$\begin{pmatrix} 0.02650288 & 0.00290207 \\ 0.00290207 & 0.04424340 \end{pmatrix}$
100	0.5417	0.00230100	0.11511474	0.11376222	$\begin{pmatrix} 0.01454585 & -0.001111 \\ -0.001111 & 0.02510317 \end{pmatrix}$
200	0.5398	0.00115822	0.11022953	0.05742143	$\begin{pmatrix} 0.00743417 & 0.00011734 \\ 0.00011734 & 0.01157424 \end{pmatrix}$
400	0.5388	0.00059248	0.10004180	0.10372044	$\begin{pmatrix} 0.00377749 & 0.00016259 \\ 0.00016259 & 0.00618457 \end{pmatrix}$

表2 $\widehat{F}(y) - F(y)$

N	$\max(\widehat{F}(y) - F(y))$	$\min(\widehat{F}(y) - F(y))$
50	0.10545107	-0.08433090
100	0.07928897	-0.06588293
200	0.05742143	-0.04603773
400	0.04236434	-0.03497579

参 考 文 献

[1] Hu, X.J. and Lawless, J.F., Nonparametric estimation of a lifetime distribution when censoring times are missing, *Technometrics*, **40**(1998), 3–12.

[2] Leign, G.M., A comparison of estimates of natural mortality from fish tagging experiments, *Biometrika*, **75**(1988), 347–353.

[3] 陆福忠, 数据缺失下的分布函数估计问题, 上海复旦大学管理学院博士论文, 2007.

[4] Heyde, C.C., *Quasi-Likelihood and Its Application: a General Approach to Optimal Parameter Estimation*, Springer-Verlag, 1997.

[5] Foutz, R.V., On the unique consistent solution to the likelihood equations, *J.A.S.A.*, **72**(1977), 147–148.

[6] 郑祖康, 丁邦俊, 关于区间数据的分布函数估计问题, 应用概率统计, **20**(2004), 119–125.
[7] 丁邦俊, 区间数据的分布函数估计问题, 上海复旦大学管理学院博士论文, 2002.
[8] 邓文丽, 区间数据的若干问题研究, 上海复旦大学管理学院博士论文, 2004.

Estimating a Distribution Function with Missing Data

Lu Fuzhong

(*College of Mathematics and Information Engineering, Jiaxing University, Jiaxing, 314001*)

In this paper, if some extra interval data are available, that is, we assume we can know for sure whether the missing data belong to some prescribed intervals or not, and the missing data mechanism is known up to a k-dimension unknown parameter θ, then, the estimators of the underlying distribution function and the unknown parameter can be derived, and their asymptotical properties are studied.

Keywords: Missing data mechanism, distribution function, interval data.
AMS Subject Classification: 62G10.