应用概率统计 第 40 卷
第 6 期 2024 年 12 月

Chinese Journal of Applied Probability and Statistics
Dec., 2024, Vol. 40, No. 6, pp. 877–890
doi: 10.12460/j.issn.1001-4268.aps.2024.2022080

生存函数中混杂因子的判断准则 *

李开灿†[1] 范凯旋[2]

[1] 湖北师范大学文理学院, 黄石, 435109

[2] 北京工商大学数学与统计学院, 北京, 102488

摘 要: 本文讨论了在响应变量和协变量都是连续型的条件下, 生存函数中判断混杂因子的准则问题. 利用分割的方法, 我们得到了协变量为一维情况下一致无关因子的充要条件, 同时得到了协变量为多维情况下条件一致无关因子向量的充分条件, 从而给出了生存函数中判断单个混杂因子和多重混杂因子的准则.

关键词: 生存函数; 混杂偏倚; 混杂因子; 一致无关因子

中图分类号: O212.4

英文引用格式: LI K C, FAN K X. Criteria for confounders in survival function [J]. *Chinese J Appl Probab Statist*, 2024, **40**(**6**): 877–890. (in Chinese)

1 引言

在流行病学的研究中, 如果某种暴露 (处理) 对某种疾病的因果作用受到协变量不同取值的影响而不同, 这种现象称为混杂, 能引起混杂且不受暴露影响的协变量称为混杂因子. 判断哪些协变量是混杂因子是因果推断中最基本的问题, 其判断准则一般分为可压缩性准则和可比较性准则这两类. 文献 [1] 给出了这两种判断准则的基本描述: 可压缩性准则根据变量间的关联测度来定义混杂因子, 而可比较性准则是基于潜在结果模型来叙述的. 由于可压缩性准则依赖于变量间的关联测度, 而关联测度的选择往往使得混杂因子的判断标准具有较强的主观性, 且有时不可压缩却对应无混杂 (文献 [2]), 因此可比较准则在流行病学的研究中深受关注. 在可比较性准则判断混杂因子方面, 文献 [3] 根据流行病学的诸多案例提出了一个判断混杂因子的准则, 文献 [4] 在潜在混杂因子集是充分控制集的假设下推广了文献 [3] 的准则, 文献 [5] 给出了混杂因子的形式化定义和离散型协变量中混杂因子的判断准则, 从文献 [6–8] 的讨论可以知道, 这些研究在流行病学中是十分有用的.

在生存分析中, 常常用一个非负随机变量 T 来表示寿命, 在研究实际问题时, 一般要考虑系统性因素对寿命的影响. 按照文献 [9] 的描述方式, 假设 X 是暴露 (处理) 变量, V 是协变量, 在一定的条件下, 考虑协变量可能使寿命和暴露之间的关系产生完全相反的结论, 即出现 Simpson 悖论. 按照流行病学的观点应该是出现了混杂, V 即为混杂因子. 文

* 国家自然科学基金项目 (批准号: 11471105) 资助.

† 通讯作者, E-mail: lkstat@163.com.

本文 2022 年 5 月 24 日收到, 2022 年 11 月 21 日收到修改稿, 2023 年 4 月 22 日录用.

献 [10] 给出了生存函数在 Aalen 加性危险模型和 Cox 比例危险模型中估计混杂偏倚大小的方法, 而估计混杂偏倚大小的前提是要判断出哪个或者哪些协变量是混杂因子, 然而他们并没有提出生存函数中判断混杂因子的准则. 文献 [5] 给出的离散型协变量是否为混杂因子的准则此时也不能用, 因为他们是在流行病学框架下讨论的判别, 而在生存函数中, 响应变量和协变量都是连续型的. 为此本文打算研究寿命 T 关于暴露 (处理) 变量 X、连续型协变量 V 在生存函数中判断混杂的准则问题, 这对生存分析中影响寿命的混杂因子的辨别具有极其重要的意义.

本文第 2 节, 我们给出了生存函数、暴露变量、响应变量、协变量和潜在生存函数等基本概念和记号. 第 3 节, 对连续型协变量的值域进行分割去定义一致无关因子, 用分割的任意性去探索协变量在暴露和不暴露时的分布差异, 建立这种差异的关联方程, 得到了生存函数中协变量为一维时一致无关因子的充要条件, 同时也得到了协变量为多维时条件一致无关因子向量的充分条件, 从而间接地给出了生存函数中混杂因子的判断准则. 第 4 节是对本文的讨论.

2 基本概念与记号

为了研究方便, 我们记 T 是寿命变量, $F(t) = \mathsf{P}(T \leqslant t)$ 是 T 的分布函数, $f(t)$ 是 T 的分布密度函数, 称 $S(t) = \mathsf{P}(T > t)$ 为 T 的生存函数. 若考虑寿命 T 受到暴露 (处理) 变量 X 和连续型协变量 V 的影响时, 则分布密度函数和生存函数分别记为 $f(t|x,v)$, $S(t|x,v) = \mathsf{P}(T > t|X = x, V = v)$, 分别称为给定 $X = x$、 $V = v$ 的条件下, T 的条件密度函数、条件生存函数, 而 $f(t|x)$, $S(t|x) = \mathsf{P}(T > t|X = x)$ 分别类似地称呼.

令 X 为二值暴露变量, 取值为 1, 0, 表示暴露与未暴露, T_x 称为潜在结果变量, 即表示假设在暴露 $X = x$ 中群体对应的寿命, 由于个体不可能同时暴露和不暴露, 所以 T_0, T_1 不可能同时观测到, 从而称为潜在变量. 用记号 $S_x(t|x) = \mathsf{P}(T_x > t|X = x)$ 表示在暴露 $X = x$ 中群体对应的潜在生存函数, $S_x(t|x') = \mathsf{P}(T_x > t|X = x')$, $x' \neq x$ 表示在暴露 $X = x'$ 中的群体如果在暴露 $X = x$ 情况下对应的潜在生存函数. 考虑连续型协变量 V 时, 则 $S_x(t|x,v) = \mathsf{P}(T_x > t|X = x, V = v)$ 和 $S_x(t|x',v) = \mathsf{P}(T_x > t|X = x', V = v)$, $x' \neq x$ 分别表示给定 $V = v$ 条件下暴露取值为 $X = x$ 的群体对应的潜在生存函数和暴露取值为 $X = x'$ 的群体如果在暴露取值为 $X = x$ 情况下对应的潜在生存函数. 从因果分析的角度来看, 假设暴露 X、协变量 V 对寿命 T 有某种影响, 我们想要研究的是连续型协变量 V 是否为混杂因子, 判断准则是什么? 截止目前, 在生存函数的因果推断中, 我们还没有发现有这方面的研究文献.

本文利用文献 [11] 的记号, $X \perp\!\!\!\perp V$ 表示随机变量 (或随机向量) X 与 V 相互独立, $T \perp\!\!\!\perp V|X$ 表示给定 X 的条件下, T 与 V 相互独立.

3　生存函数中判断混杂因子的准则

按照第一节引言关于生存函数中混杂因子的描述, 本节主要从协变量为一维连续型和多维连续型两种情况分别给出单个混杂因子和多重混杂因子的判断准则.

3.1　单个混杂因子的判断准则

文献 [5] 在暴露变量和响应变量都是二值的, 协变量是离散型时给出了一个混杂因子的判别准则, 而当寿命变量和协变量都是连续型的条件下, 我们给出判别混杂因子的准则: 一个必要条件和一个间接的充要条件. 这是本节的研究重点.

众所周知, 许多命题在离散型成立, 而将离散型变量 V 过渡到连续型去证明命题还成立往往是困难的, 例如文献 [12] 在证明分布相依的可压缩性条件是某种独立性时, 连续型变量远比离散型变量复杂. 比较发现, 他们的方法在处理判断生存函数混杂因子时是不可利用的. 本节通过对连续型协变量的值域分割去定义一致无关因子, 用分割的任意性去探索了协变量在暴露和未暴露时的分布差异, 建立了这种差异的关联方程 (见下面的引理 2、引理 3), 从而获得了针对连续型协变量是否为寿命变量的混杂因子的一种判别准则.

不失一般性, 设协变量 V 的密度函数 $f(v)$ 的支撑集为实数集 R, 我们先给出协变量为一维连续型情况下混杂偏倚和混杂因子的相关概念.

定义 1　设 $S_0(t|X=0)=\mathsf{P}(T_0>t|X=0)$ 表示在未暴露群体对应的潜在生存函数, $S_0(t|X=1)=\mathsf{P}(T_0>t|X=1)$ 表示暴露群体如果在未暴露情况下对应的潜在生存函数, 寿命 T 的潜在生存函数在暴露 X 作用下的混杂偏倚表示为

$$B=S_0(t|X=1)-S_0(t|X=0).$$

设 ω 为 R 的可测子集, $\mathsf{P}(V\in\omega)>0$, 令

$$B_\omega=\mathsf{P}(T_0>t|X=1,V\in\omega)-\mathsf{P}(T_0>t|X=0,V\in\omega),$$

称 B_ω 为潜在生存函数在暴露 X 作用下, 协变量 V 在 ω 上产生的混杂偏倚.

注记 1　(i) 如果 $\forall t>0, B=0$, 那么暴露群体和未暴露群体是可比较的, 并且是无混杂的.

(ii) 类似于定义连续型随机变量的条件分布函数, 令

$$B_v=S_0(t|X=1,v)-S_0(t|X=0,v),$$

称 B_v 是潜在生存函数在暴露 X 作用下, 协变量在 $V=v$ 处的混杂偏倚.

定义 2　对于协变量 V, 令

$$S_\delta(t|X=0)=\int S_0(t|X=0,v)\,f(v|X=1)\,\mathrm{d}v,$$

称 $S_\delta(t|X=0)$ 为按照暴露调整后的未暴露群体的潜在生存函数. 如果

$$S_\delta(t|X=0)=S_0(t|X=0),\forall t>0,$$

那么称 V 是一个无关因子.

定义 3 如果协变量 V 满足

$$|S_0(t|X=1)-S_\delta(t|X=0)|<|B|,$$

那么称 V 是一个混杂因子.

定义 3 所表述的是未暴露群体的协变量密度函数按照暴露调整后, 其潜在生存函数更接近暴露群体如果在未暴露的情况下对应的潜在生存函数值, 从而导致混杂偏倚减小, 它与文献 [3] 关于混杂因子的定义是一致的.

由定义 2 可知, 无关因子一定不是混杂因子, 所以如下定理给出了判断混杂因子的必要条件.

定理 1 如果 V 是一个混杂因子, 那么 V 满足

(i) $V \not\perp\!\!\!\perp X$, 且

(ii) $T_0 \not\perp\!\!\!\perp V|X=0$.

证明: （反证法）假设 $V \perp\!\!\!\perp X$ 或者 $T_0 \perp\!\!\!\perp V|X=0$ 成立.

令 $G=S_\delta(t|X=0)-S_0(t|X=0)$, 则

$$G=\int S_0(t|X=0,v)\left[f(v|X=1)-f(v|X=0)\right]\mathrm{d}v.$$

当条件 $V \perp\!\!\!\perp X$ 成立时, 则

$$f(v|X=1)=f(v|X=0)=f(v),$$

所以 $G=0$. 利用定义 2, V 是一个无关因子, 与协变量 V 是一个混杂因子矛盾, 故结论成立.

当条件 $T_0 \perp\!\!\!\perp V|X=0$ 成立时, 则

$$S_0(t|X=0,v)=S_0(t|X=0),$$

所以

$$\begin{aligned}G&=\int S_0(t|X=0,v)\left[f(v|X=1)-f(v|X=0)\right]\mathrm{d}v\\&=\int S_0(t|X=0)\left[f(v|X=1)-f(v|X=0)\right]\mathrm{d}v\\&=S_0(t|X=0)\int\left[f(v|X=1)-f(v|X=0)\right]\mathrm{d}v\\&=0.\end{aligned}$$

利用定义 2, V 是一个无关因子, 从而产生矛盾, 故结论成立. □

下面重点研究判断一致无关因子的准则.

定义 4　设 $\omega_1,\omega_2,\cdots,\omega_s\,(s\geqslant 2)$ 是 R 的 s 个非空可测子集, 若同时满足

(i) $\omega_i\cap\omega_j=\emptyset, i\neq j$,

(ii) $\mathsf{P}\left(V\in\omega_i\right)>0, i=1,2,\cdots,s$,

(iii) $\bigcup\limits_i\omega_i=R$,

则称 $\mathscr{P}=\{\omega_1,\omega_2,\cdots,\omega_s\}$ 为 R 关于 V 的一个分割.

定义 5　设 $\mathscr{P}=\{\omega_1,\omega_2,\cdots,\omega_s\}$ 为 R 关于 V 的一个分割, 记

$$\begin{aligned}B_{\mathscr{P}}&=\sum_{i=1}^{s}B_{\omega_i}\,\mathsf{P}\left(V\in\omega_i|X=1\right)\\&=\sum_{i=1}^{s}\left[\mathsf{P}\left(T_0>t|X=1,V\in\omega_i\right)-\mathsf{P}\left(T_0>t|X=0,V\in\omega_i\right)\right]\times\mathsf{P}\left(V\in\omega_i|X=1\right),\end{aligned}$$

称 $B_{\mathscr{P}}$ 是寿命 T 的潜在生存函数在暴露 X 作用下关于协变量 V 对分割 $\mathscr{P}$ 的混杂偏倚.

定义 6　如果关于 V 的任何分割 $\mathscr{P}$ 混杂偏倚总有

$$B_{\mathscr{P}}=B,$$

那么称协变量 V 为一个一致无关因子.

为了给出一个协变量是一致无关因子的充要条件, 我们需要如下引理.

引理 2　若 ω,ω_1,ω_2 是 R 的可测子集, 满足 $\omega_1\cap\omega_2=\emptyset$, 记

$$\begin{aligned}&m_\omega=\mathsf{P}\left(T_0>t|X=0,V\in\omega\right),\ m_R=m_0,\\&n_\omega=\mathsf{P}\left(V\in\omega|X=1\right)-\mathsf{P}\left(V\in\omega|X=0\right),\end{aligned}$$

则

(i) $m_{\omega_1+\omega_2}=\frac{m_{\omega_1}\mathsf{P}(V\in\omega_1|X=0)+m_{\omega_2}\mathsf{P}(V\in\omega_2|X=0)}{\mathsf{P}(V\in\omega_1|X=0)+\mathsf{P}(V\in\omega_2|X=0)}$;

(ii) $n_{\omega_1+\omega_2}=n_{\omega_1}+n_{\omega_2}$;

(iii) $n_\omega+n_{\bar{\omega}}=0, m_0=m_\omega\,\mathsf{P}\left(V\in\omega|X=0\right)+m_{\bar{\omega}}\left[1-\mathsf{P}\left(V\in\omega|X=0\right)\right]$.

证明:

(i)

$$\begin{aligned}m_{\omega_1+\omega_2}&=\mathsf{P}\left(T_0>t|V\in\omega_1+\omega_2,X=0\right)\\&=\frac{\mathsf{P}\left(T_0>t,V\in\omega_1+\omega_2,X=0\right)}{\mathsf{P}\left(V\in\omega_1+\omega_2,X=0\right)}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathsf{P}\left(T_0 > t, V \in \omega_1, X = 0\right) + \mathsf{P}\left(T_0 > t, V \in \omega_2, X = 0\right)}{\mathsf{P}\left(V \in \omega_1, X = 0\right) + \mathsf{P}\left(V \in \omega_2, X = 0\right)} \\
&= \frac{m_{\omega_1}\mathsf{P}\left(V \in \omega_1, X = 0\right) + m_{\omega_2}\mathsf{P}\left(V \in \omega_2, X = 0\right)}{\mathsf{P}\left(V \in \omega_1, X = 0\right) + \mathsf{P}\left(V \in \omega_2, X = 0\right)} \\
&= \frac{m_{\omega_1}\mathsf{P}\left(V \in \omega_1 | X = 0\right) + m_{\omega_2}\mathsf{P}\left(V \in \omega_2 | X = 0\right)}{\mathsf{P}\left(V \in \omega_1 | X = 0\right) + \mathsf{P}\left(V \in \omega_2 | X = 0\right)}.
\end{aligned}
$$

(ii)

$$
\begin{aligned}
n_{\omega_1+\omega_2} &= \mathsf{P}\left(V \in \omega_1 + \omega_2 | X = 1\right) - \mathsf{P}\left(V \in \omega_1 + \omega_2 | X = 0\right) \\
&= \mathsf{P}\left(V \in \omega_1 | X = 1\right) + \mathsf{P}\left(V \in \omega_2 | X = 1\right) \\
&\quad - \mathsf{P}\left(V \in \omega_1 | X = 0\right) - \mathsf{P}\left(V \in \omega_2 | X = 0\right) \\
&= n_{\omega_1} + n_{\omega_2}.
\end{aligned}
$$

(iii) 利用概率的性质和引理 2 (i) (ii) 的结果可知, (iii) 是显然的.

□

引理 3 若 V 是一个一致无关因子, 则

(i) 对于任一可测实数集 ω, 只要 $0 < \mathsf{P}\left(V \in \omega\right) < 1$, 有

$$
\left(m_\omega - m_{\bar{\omega}}\right) n_\omega = 0; \tag{1}
$$

(ii) 对于关于 V 的任一分割 $\mathscr{P} = \{\omega_1, \omega_2, \cdots, \omega_s\}, s > 1$, 恒有

$$
m_{\omega_1} = m_{\omega_2} = \cdots = m_{\omega_s} = m_0, \text{ 或者 } n_{\omega_1} = n_{\omega_2} = \cdots = n_{\omega_s} = 0.
$$

证明:

(i) 因为 V 是一个一致无关因子, 取分割 $\mathscr{P}_1 = \{\omega, \bar{\omega}\}$, 则由引理2 (iii) 可得 $n_{\bar{\omega}} = -n_\omega$, 依据 $B_{\mathscr{P}_1} = B$ 有

$$
\begin{aligned}
&\left[\mathsf{P}\left(T_0 > t | X = 1, V \in \omega\right) - \mathsf{P}\left(T_0 > t | X = 0, V \in \omega\right)\right]\mathsf{P}\left(V \in \omega | X = 1\right) \\
&+ \left[\mathsf{P}\left(T_0 > t | X = 1, V \in \bar{\omega}\right) - \mathsf{P}\left(T_0 > t | X = 0, V \in \bar{\omega}\right)\right]\mathsf{P}\left(V \in \bar{\omega} | X = 1\right) \\
&= \mathsf{P}\left(T_0 > t | X = 1\right) - \mathsf{P}\left(T_0 > t | X = 0\right),
\end{aligned}
$$

即

$$
\begin{aligned}
0 &= \mathsf{P}\left(T_0 > t | X = 0, V \in \omega\right)\left[\mathsf{P}\left(V \in \omega | X = 1\right) - \mathsf{P}\left(V \in \omega | X = 0\right)\right] \\
&\quad + \mathsf{P}\left(T_0 > t | X = 0, V \in \bar{\omega}\right)\left[\mathsf{P}\left(V \in \bar{\omega} | X = 1\right) - \mathsf{P}\left(V \in \bar{\omega} | X = 0\right)\right] \\
&= m_\omega n_\omega - m_{\bar{\omega}} n_\omega.
\end{aligned}
$$

所以式(1) 成立, 即

$$(m_\omega - m_{\bar{\omega}})\, n_\omega = 0.$$

(ii) 用反证法: 如果不然, 则必可假设 $n_{\omega_2} \neq 0$, 并且 $m_{\omega_1} \neq m_0$.

由 (i) 知 $m_{\omega_2} = m_{\bar{\omega}_2}$, 再由引理 2 (iii) 有

$$m_0 = m_{\omega_2} \mathsf{P}\left(V \in \omega_2 | X = 0\right) + m_{\bar{\omega}_2} \left[1 - \mathsf{P}\left(V \in \omega_2 | X = 0\right)\right],$$

所以 $m_{\omega_2} = m_{\bar{\omega}_2} = m_0$. 由 (i) 可知, $n_{\omega_1} = 0$, 否则 $m_{\omega_1} = m_0$, 从而 $n_{\omega_1+\omega_2} = n_{\omega_1} + n_{\omega_2} = n_{\omega_2} \neq 0$. 又由 (i) 知 $m_{\omega_1+\omega_2} = m_0$, 由引理 2 (i) 可知

$$m_{\omega_1+\omega_2} = \frac{m_{\omega_1} \mathsf{P}\left(V \in \omega_1 | X = 0\right) + m_{\omega_2} \mathsf{P}\left(V \in \omega_2 | X = 0\right)}{\mathsf{P}\left(V \in \omega_1 | X = 0\right) + \mathsf{P}\left(V \in \omega_2 | X = 0\right)},$$

所以

$$\frac{m_{\omega_1} \mathsf{P}\left(V \in \omega_1 | X = 0\right) + m_0 \mathsf{P}\left(V \in \omega_2 | X = 0\right)}{\mathsf{P}\left(V \in \omega_1 | X = 0\right) + \mathsf{P}\left(V \in \omega_2 | X = 0\right)} = m_0,$$

从而获得了 $m_{\omega_1} = m_0$, 这是矛盾的, 故引理得证.

□

定理 4　协变量 V 是潜在生存函数在暴露 X 作用下的一致无关因子的充分必要条件是

(i) $V \perp\!\!\!\perp X$, 或者

(ii) $T_0 \perp\!\!\!\perp V | X = 0$.

证明:　充分性. 如果 $V \perp\!\!\!\perp X$, 那么关于 V 的任一个分割 $\mathscr{P} = \{\omega_1, \omega_2, \cdots, \omega_s\}$,

$$\begin{aligned}
B_{\mathscr{P}} &= \sum_{i=1}^{s} B_{\omega_i} \mathsf{P}\left(V \in \omega_i | X = 1\right) \\
&= \sum_{i=1}^{s} B_{\omega_i} \mathsf{P}\left(V \in \omega_i\right) \\
&= \sum_{i=1}^{s} \left[\mathsf{P}\left(T_0 > t | X = 1, V \in \omega_i\right) - \mathsf{P}\left(T_0 > t | X = 0, V \in \omega_i\right)\right] \mathsf{P}\left(V \in \omega_i\right) \\
&= \sum_{i=1}^{s} \left[\frac{\mathsf{P}\left(T_0 > t, V \in \omega_i, X = 1\right) \mathsf{P}\left(V \in \omega_i\right)}{\mathsf{P}\left(V \in \omega_i, X = 1\right)} - \frac{\mathsf{P}\left(T_0 > t, V \in \omega_i, X = 0\right) \mathsf{P}\left(V \in \omega_i\right)}{\mathsf{P}\left(V \in \omega_i, X = 0\right)}\right] \\
&= \sum_{i=1}^{s} \left[\mathsf{P}\left(T_0 > t, V \in \omega_i | X = 1\right) - \mathsf{P}\left(T_0 > t, V \in \omega_i | X = 0\right)\right] \\
&= \mathsf{P}\left(T_0 > T | X = 1\right) - \mathsf{P}\left(T_0 > T | X = 0\right) = B.
\end{aligned}$$

故由定义 6 可知, V 是一个一致无关因子.

如果 $T_0 \perp\!\!\!\perp V|X=0$,

$$\begin{aligned}
B_{\mathscr{P}} &= \sum_{i=1}^{s} \mathsf{P}\left(T_0 > t|V \in \omega_i, X=1\right) \mathsf{P}\left(V \in \omega_i|X=1\right) \\
&\quad - \sum_{i=1}^{s} \mathsf{P}\left(T_0 > t|V \in \omega_i, X=0\right) \mathsf{P}\left(V \in \omega_i|X=1\right) \\
&= \sum_{i=1}^{s} \frac{\mathsf{P}\left(T_0 > t, V \in \omega_i, X=1\right)}{\mathsf{P}\left(V \in \omega_i, X=1\right)} \frac{\mathsf{P}\left(V \in \omega_i, X=1\right)}{\mathsf{P}\left(X=1\right)} \\
&\quad - \sum_{i=1}^{s} \frac{\mathsf{P}\left(T_0 > t, V \in \omega_i, X=0\right)}{\mathsf{P}\left(V \in \omega_i, X=0\right)} \frac{\mathsf{P}\left(V \in \omega_i, X=1\right)}{\mathsf{P}\left(X=1\right)} \\
&= \sum_{i=1}^{s} \frac{\mathsf{P}\left(T_0 > t, V \in \omega_i, X=1\right)}{\mathsf{P}\left(X=1\right)} \\
&\quad - \sum_{i=1}^{s} \frac{\mathsf{P}\left(X=0\right) \mathsf{P}\left(T_0 > t, V \in \omega_i|X=0\right)}{\mathsf{P}\left(X=0\right) \mathsf{P}\left(V \in \omega_i|X=0\right)} \frac{\mathsf{P}\left(V \in \omega_i, X=1\right)}{\mathsf{P}\left(X=1\right)} \\
&= \frac{\mathsf{P}\left(T_0 > t, X=1\right)}{\mathsf{P}\left(X=1\right)} - \sum_{i=1}^{s} \frac{\mathsf{P}\left(T_0 > t|X=0\right) \mathsf{P}\left(V \in \omega_i|X=0\right) \mathsf{P}\left(V \in \omega_i, X=1\right)}{\mathsf{P}\left(V \in \omega_i|X=0\right) \mathsf{P}\left(X=1\right)} \\
&= \mathsf{P}\left(T_0 > t|X=1\right) - \sum_{i=1}^{s} \frac{\mathsf{P}\left(T_0 > t|X=0\right) \mathsf{P}\left(V \in \omega_i, X=1\right)}{\mathsf{P}\left(X=1\right)} \\
&= \mathsf{P}\left(T_0 > t|X=1\right) - \mathsf{P}\left(T_0 > t|X=0\right) = B.
\end{aligned}$$

故由定义 6 可知, V 是一个一致无关因子.

必要性. 为了证明条件的必要性, 在逻辑上, 只要证明如果 V 是一致无关因子, 并且 $V \not\perp\!\!\!\perp X$, 必有 $T_0 \perp\!\!\!\perp V|X=0$ 即可. 根据 $T_0 \perp\!\!\!\perp V|X=0$ 的定义, 只要证明 $\forall t > 0, \forall \omega \subseteq R$ 是可测的, 总有

$$\mathsf{P}\left(T_0 > t, V \in \omega|X=0\right) = \mathsf{P}\left(T_0 > t|X=0\right) \mathsf{P}\left(V \in \omega|X=0\right) \tag{2}$$

成立, 则结论为真. 下面按这个假设予以证明.

由于 $\mathsf{P}\left(V \in \omega\right) = 0$, 式 (2) 肯定成立, 所以以下证明均假定 $\mathsf{P}\left(V \in \omega\right) > 0$.

如果 $V \not\perp\!\!\!\perp X$, 利用 X 只取 0, 1 二值的特点可知, 必存在可测集 $\omega_0 \subseteq R$, $0 < \mathsf{P}\left(V \in \omega_0\right) < 1$, 使得 $n_{\omega_0} \neq 0$, 从而利用引理 3 (ii) 有 $m_{\omega_0} = m_{\bar{\omega}_0} = m_0$. 下面分几种情形证明式 (2) 成立.

1) 当 $\omega \subseteq \bar{\omega}_0$, 且 $\mathsf{P}\left(V \in \omega\right) < \mathsf{P}\left(V \in \bar{\omega}_0\right)$ 时, 取分割 $\mathscr{P} = \{\omega_0, \omega, \overline{\omega_0 + \omega}\}$, 因为 $n_{\omega_0} \neq 0$, 那么利用引理 3 (ii) 有 $m_{\omega} = m_0$, 从 m_{ω}, m_0 的定义就是 $\mathsf{P}\left(T_0 > t|X=0, V \in \omega\right) =$

$\mathsf{P}(T_0 > t|X=0)$, 故

$$\begin{aligned}\frac{\mathsf{P}(T_0>t, X=0, V\in\omega)}{\mathsf{P}(V\in\omega, X=0)} &= \frac{\mathsf{P}(X=0)\,\mathsf{P}(T_0>t, V\in\omega|X=0)}{\mathsf{P}(X=0)\,\mathsf{P}(V\in\omega|X=0)}\\ &= \frac{\mathsf{P}(T_0>t, V\in\omega|X=0)}{\mathsf{P}(V\in\omega|X=0)}\\ &= \mathsf{P}(T_0>t|X=0). \end{aligned} \tag{3}$$

所以 $\mathsf{P}(T_0>t, V\in\omega|X=0) = \mathsf{P}(T_0>t|X=0)\,\mathsf{P}(V\in\omega|X=0)$, 故 $T_0 \perp\!\!\!\perp V|X=0$.

2) 当 $\omega\subseteq\bar{\omega}_0$, 且 $\mathsf{P}(V\in\omega) = \mathsf{P}(V\in\bar{\omega}_0)$ 时, 由于

$$\mathsf{P}(V\in\omega\cup\omega_0) = \mathsf{P}(V\in\omega) + \mathsf{P}(V\in\omega_0) - \mathsf{P}(V\in\omega\cap\omega_0) = 1,$$

所以 $\mathsf{P}(V\in\bar{\omega}_0\bar{\omega}) = 0$. 用引理 2 (ii) (iii),

$$n_{\bar{\omega}_0} = n_{\bar{\omega}_0\omega+\bar{\omega}_0\bar{\omega}} = n_{\omega+\bar{\omega}_0\bar{\omega}} = n_\omega + n_{\bar{\omega}_0\bar{\omega}} = n_\omega,$$

所以 $n_\omega = n_{\bar{\omega}_0} = -n_{\omega_0} \neq 0$, 故 $m_\omega = m_0$. 用证明式 (3) 一样的方式可知, $T_0 \perp\!\!\!\perp V|X=0$.

3) 当 $\omega\subseteq\omega_0$ 时, 分 $\omega\subseteq\omega_0$, 且 $\mathsf{P}(V\in\omega) < \mathsf{P}(V\in\omega_0)$ 和 $\omega\subseteq\omega_0$, 且 $\mathsf{P}(V\in\omega) = \mathsf{P}(V\in\omega_0)$ 两种情形. 由证明 1) 和 2) 同样的方法知, $T_0 \perp\!\!\!\perp V|X=0$.

4) 当 $\mathsf{P}(V\in\omega_0\omega) > 0$, 且 $\mathsf{P}(V\in\bar{\omega}_0\omega) > 0$ 时, 以及 $\omega_0\omega\subseteq\omega_0, \bar{\omega}_0\omega\subseteq\bar{\omega}_0$, 由证明 1) 和 2) 同样的方法可得, $m_{\omega_0\omega} = m_{\bar{\omega}_0\omega} = m_0$. 由引理 2 (i),

$$m_\omega = m_{\omega_0\omega+\bar{\omega}_0\omega} = \frac{m_{\omega_0\omega}\,\mathsf{P}(V\in\omega_0\omega|X=0) + m_{\bar{\omega}_0\omega}\,\mathsf{P}(V\in\bar{\omega}_0\omega|X=0)}{\mathsf{P}(V\in\omega_0\omega|X=0) + \mathsf{P}(V\in\bar{\omega}_0\omega|X=0)} = m_0,$$

所以用证明式 (3) 一样的方式可知, $T_0 \perp\!\!\!\perp V|X=0$.

5) 当 $\mathsf{P}(V\in\omega_0\omega) = 0$ 或者 $\mathsf{P}(V\in\bar{\omega}_0\omega) = 0$ 时, 也可以证明结论成立. 事实上, 假设 $\mathsf{P}(V\in\omega_0\omega) = 0$, 那么

$$\mathsf{P}(V\in\omega_0\bar{\omega}) = \mathsf{P}(V\in\omega_0) - \mathsf{P}(V\in\omega_0\omega) = \mathsf{P}(V\in\omega_0) > 0,$$

$$\mathsf{P}(Y\in\bar{\omega}_0\omega) = \mathsf{P}(V\in\omega) - \mathsf{P}(V\in\omega_0\omega) > 0,$$

从而 $\mathscr{P} = \{\omega_0\bar{\omega}, \omega, \bar{\omega}_0\bar{\omega}\}$ 构成分割. 另一方面, 从 $\mathsf{P}(V\in\omega_0\omega) = 0$ 和引理 2 (ii), 有

$$n_{\omega_0} = n_{\omega_0\omega+\omega_0\bar{\omega}} = n_{\omega_0\omega} + n_{\omega_0\bar{\omega}} = n_{\omega_0\bar{\omega}},$$

所以

$$n_{\omega_0\bar{\omega}} = n_{\omega_0} \neq 0.$$

再利用引理 3 (ii) 可得 $m_\omega = m_0$, 用证明式 (3) 的方式, $T_0 \perp\!\!\!\perp V|X=0$.

对于 $\mathsf{P}(V \in \bar{\omega}_0\omega) = 0$ 的情形一样讨论.

综上述 1)–5), 当 $V \not\perp\!\!\!\perp X$ 时, 对于任意的可测集 $\omega \subseteq R$, 式 (2) 恒成立, 所以 $T_0 \perp\!\!\!\perp V|X=0$. □

定理 5 如果一个协变量是一致无关因子, 则其一定是无关因子.

证明: 由定理 4 可知, 协变量 V 是潜在生存函数在暴露 X 作用下的一致无关因子可以推出 $V \perp\!\!\!\perp X$ 或者 $T_0 \perp\!\!\!\perp V|X=0$.

(i) 当 $V \perp\!\!\!\perp X$ 成立时, 则

$$f(v|X=1) = f(v|X=0),$$

所以

$$\begin{aligned} S_\delta(t|X=0) &= \int S_0(t|X=0,v)\, f(v|X=1)\,\mathrm{d}v \\ &= \int S_0(t|X=0,v)\, f(v|X=0)\,\mathrm{d}v \\ &= S_0(t|X=0). \end{aligned}$$

由定义 2 可知, 协变量 V 是一个无关因子.

(ii) 当 $T_0 \perp\!\!\!\perp V|X=0$ 成立时, 则

$$S_0(t|X=0,v) = S_0(t|X=0),$$

所以

$$\begin{aligned} S_\delta(t|X=0) &= \int S_0(t|X=0,v)\, f(v|X=1)\,\mathrm{d}v \\ &= \int S_0(t|X=0)\, f(v|X=1)\,\mathrm{d}v \\ &= S_0(t|X=0) \int f(v|X=1)\,\mathrm{d}v \\ &= S_0(t|X=0). \end{aligned}$$

由定义 2 可知, 协变量 V 是一个无关因子.

□

根据定理 1 的证明可知, 无关因子一定不是混杂因子. 又由定理 5 可得, 一致无关因子是无关因子, 如果一个协变量是一致无关因子, 则它不是混杂因子. 定理 4 给出了一个协变量是一致无关因子的充要条件, 从而间接得到了一维协变量情况下判断混杂因子的准则.

3.2　多重混杂因子的判断准则

协变量是一维连续型时, 3.1 节给出了单个混杂因子的判断准则, 但在生存函数中协变量可能不止一个, 相应地, 我们需要判别多重混杂因子. 本节主要研究协变量是多维连续型情况下多重混杂因子的判断准则.

设协变量 V 是一个随机向量, 即 $V=(V_1,V_2,\cdots,V_n)^{\top},n\geqslant 2$, 设 V 在各分量的边际密度函数 $f_{V_1}(v_1),f_{V_2}(v_2),\cdots,f_{V_n}(v_n)$ 的支撑集分别为 $R_1,R_2,\cdots,R_n$. 由于协变量中可能只有部分分量是混杂因子, 这些分量构成混杂因子向量, 而其他的分量不造成混杂, 所以我们采取分块的思想给出多重混杂因子的判断准则. 令 V' 和 V'' 表示 V 的分块, 即 $V=(V',V'')^{\top}$, 其中 V' 的维数是 q, $0<q<n$. 下面我们给出条件一致无关因子向量的相关定义.

定义 7　设 $\omega_1^{(k)},\omega_2^{(k)},\cdots,\omega_{l_k}^{(k)}$, $l_k\geqslant 2$ 是 V 的任意一个分量 V_k 密度函数的支撑集 R_k 的 l_k 个非空可测子集, $k=1,2,\cdots,n$, 若同时满足

(i) $\omega_i^{(k)}\cap\omega_j^{(k)}=\emptyset,i\neq j$,

(ii) $\mathsf{P}\left(V_k\in\omega_i^{(k)}\right)>0,i=1,2,\cdots,l_k$,

(iii) $\bigcup\limits_i\omega_i^{(k)}=R_k$,

则称 $\mathscr{P}^{(k)}=\{\omega_1^{(k)},\omega_2^{(k)},\cdots,\omega_{l_k}^{(k)}\}$ 为 R_k 关于 V_k 的一个分割, 称 $\mathscr{P}'=\mathscr{P}^{(1)}\times\mathscr{P}^{(2)}\times\cdots\times\mathscr{P}^{(q)}$ 为 $R_1\times R_2\times\cdots\times R_q$ 关于 V' 的一个分割.

定义 8　设 $\mathscr{P}'=\mathscr{P}^{(1)}\times\mathscr{P}^{(2)}\times\cdots\times\mathscr{P}^{(q)}$ 为 $R_1\times R_2\times\cdots\times R_q$ 关于 V' 的一个分割, 记

$$\begin{aligned}B_{\mathscr{P}'}=\sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q}&\Big[\mathsf{P}\left(T_0>t|X=1,V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)},V''=v''\right)\\&-\mathsf{P}\left(T_0>t|X=0,V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)},V''=v''\right)\Big]\\&\times\mathsf{P}\left(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1,V''=v''\right),\end{aligned}$$

称 $B_{\mathscr{P}'}$ 是寿命 T 的潜在生存函数在暴露 X 作用下关于协变量 V' 在给定 V'' 条件下对分割 $\mathscr{P}'$ 的混杂偏倚.

定义 9　对于 V' 的任何分割 $\mathscr{P}'$ 的混杂偏倚总有

$$B_{\mathscr{P}'}=B_{v''},$$

那么称协变量 V' 为一个给定 V'' 条件下的一致无关因子向量.

如果存在一个分割 $\mathscr{P}_0'$ 使得 $B_{\mathscr{P}_0'}\neq B_{v''}$, 那么称协变量 V' 为一个给定 V'' 条件下的混杂因子向量. 显然, 给定 V'' 条件下的协变量如果是一致无关因子向量, 则一定不是混杂因子向量. 下面给出条件一致无关因子向量的充分条件.

定理 6 令 $V=(V',V'')^{\top}$, 如果

(i) $T_0 \perp\!\!\!\perp V'|(X=0,V'')$, 或者

(ii) $V' \perp\!\!\!\perp X|V''$,

则协变量 V' 是给定 V'' 条件下的一致无关因子向量.

证明:

$$\begin{aligned}B_{\mathscr{P}'}=&\sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q}\Big[\mathsf{P}\Big(T_0>t|X=1,V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)},V''=v''\Big)\\&-\mathsf{P}\Big(T_0>t|X=0,V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)},V''=v''\Big)\Big]\\&\times\mathsf{P}\Big(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1,V''=v''\Big).\end{aligned}\tag{4}$$

当条件 (i) 成立时, 则

$$\begin{aligned}&\mathsf{P}\Big(T_0>t|X=0,V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)},V''=v''\Big)\\=&\mathsf{P}\left(T_0>t|X=0,V''=v''\right),\end{aligned}$$

代入式 (4) 得

$$\begin{aligned}B_{\mathscr{P}'}=&\sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q}\mathsf{P}\Big(T_0>t|X=1,V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)},V''=v''\Big)\\&\times\mathsf{P}\Big(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1,V''=v''\Big)\\&-\sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q}\mathsf{P}\left(T_0>t|X=0,V''=v''\right)\\&\times\mathsf{P}\Big(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1,V''=v''\Big)\\=&\mathsf{P}\left(T_0>t|X=1,V''=v''\right)-\mathsf{P}\left(T_0>t|X=0,V''=v''\right)\\&\times\sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q}\mathsf{P}\Big(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1,V''=v''\Big)\\=&S_0\left(t|X=1,v''\right)-S_0\left(t|X=0,v''\right)\\=&B_{v''}.\end{aligned}$$

故由定义 9 可知, V' 是一个给定 V'' 条件下的一致无关因子向量.

当条件 (ii) 成立时, 则

$$\begin{aligned}&\mathsf{P}\Big(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1,V''=v''\Big)\\=&\mathsf{P}\Big(V_{i_1}\in\omega_{i_1}^{(1)},V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=0,V''=v''\Big),\end{aligned}$$

代入式 (4) 得

$$
\begin{aligned}
B_{\mathscr{P}'} = & \sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q} \mathsf{P}\left(T_0 > t|X=1, V_{i_1}\in\omega_{i_1}^{(1)}, V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}, V''=v''\right)\\
& \times \mathsf{P}\left(V_{i_1}\in\omega_{i_1}^{(1)}, V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=1, V''=v''\right)\\
& - \sum_{i_1=1}^{l_1}\sum_{i_2=1}^{l_2}\cdots\sum_{i_q=1}^{l_q} \mathsf{P}\left(T_0 > t|X=0, V_{i_1}\in\omega_{i_1}^{(1)}, V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}, V''=v''\right)\\
& \times \mathsf{P}\left(V_{i_1}\in\omega_{i_1}^{(1)}, V_{i_2}\in\omega_{i_2}^{(2)},\cdots,V_{i_q}\in\omega_{i_q}^{(q)}|X=0, V''=v''\right)\\
= & S_0\left(t|X=1, v''\right) - S_0\left(t|X=0, v''\right)\\
= & B_{v''}.
\end{aligned}
$$

故由定义 9 可知, V' 是一个给定 V'' 条件下的一致无关因子向量. □

由于给定 V'' 条件下的一致无关因子向量一定不是给定相同条件下的混杂因子向量, 定理 6 给出了给定 V'' 条件下的一致无关因子向量的充分条件, 从而间接得到了多维协变量情况下判断多重混杂因子的准则.

4　讨论

本文主要讨论了生存函数中判断混杂因子的准则. 定理 1 验证了文献 [3] 的判断混杂因子的准则, 其中 $V \not\perp\!\!\!\perp X$ 表示协变量 V 在暴露与未暴露两个群体的分布是不同的, $T_0 \not\perp\!\!\!\perp V|X=0$ 表示在未暴露群体中, 协变量 V 可以用来推测寿命. 我们给出的单个混杂因子和多重混杂因子的判断准则, 均是通过判断协变量是否为不引起混杂的变量 (一致无关因子和条件一致无关因子向量) 实现的, 这是一种间接判断混杂因子的准则, 解决了文献 [10] 在生存函数中估计混杂偏倚之前, 需要先假设协变量中哪些是混杂因子的问题. 由于随机变量的条件独立性不能很好地检验, 所以要找出一致无关因子和条件一致无关因子向量并不是一件很容易的事.

本文考虑的暴露变量是二值型的, 当暴露变量是多值型或者连续型时 (比如用药剂量就是一个多值或者连续型暴露变量), 生存函数中混杂因子的判断准则又是怎样的, 这是一个十分值得我们研究的问题.

参考文献

[1] GENG Z, LIU Y, LIU C C, et al. Evaluation of causal effects and local structure learning of causal networks [J]. *Annu Rev Stat Appl*, 2019, **6**: 103–124.

[2] PEARL J. *Causality: Models, Reasoning, and Inference* [M]. Cambridge: Cambridge University Press, 2009.

[3] MIETTINEN O S, COOK E F. Confounding: Essence and detection [J]. *Am J Epidem*, 1981, **114**(**4**):

593–603.

[4] ROBINS J M. Causal inference from complex longitudinal data [M] //BERKANE M. *Latent Variable Modeling with Applications to Causality*. New York: Springer, 1997: 69–117.

[5] GENG Z, GUO J H, FUNG W K. Criteria for confounders in epidemiological studies [J]. *J R Stat Soc Ser B*, 2002, **64**(**1**): 3–15.

[6] KLEINBAUM D G, KUPPER L L, MORGENSTERN H. *Epidemiologic Research: Principles and Quantitative Methods* [M]. New York: Van Nostrand Reinhold, 1991.

[7] LASH T L, VANDERWEELE T J, HANEUSE S, et al. *Modern Epidemiology* [M]. 4th ed. Boston: Lippincott Williams & Wilkins, 2021.

[8] GREENLAND S, PEARL J, ROBINS J M. Confounding and collapsibility in causal inference [J]. *Statist Sci*, 1999, **14**(**1**): 29–46.

[9] SERIO C D, RINOTT Y, SCARSINI M. Simpson's paradox in survival models [J]. *Scand J Stat*, 2009, **36**(**3**): 463–480.

[10] MARTINUSSEN T, VANSTEELANDT S. On collapsibility and confounding bias in Cox and Aalen regression models [J]. *Lifetime Data Anal*, 2013, **19**(**3**): 279–296.

[11] DAWID A P. Conditional independence in statistical theory [J]. *J R Stat Soc Ser B*, 1979, **41**(**1**): 1–15.

[12] MA Z M, XIE X C, GENG Z. Collapsibility of distribution dependence [J]. *J R Stat Soc Ser B*, 2006, **68**(**1**): 127–133.

Criteria for Confounders in Survival Function

LI Kaican[1] FAN Kaixuan[2]

[1] College of Arts and Sciences of Hubei Normal University, Huangshi, 435109, China

[2] School of Mathematics and Statistics, Beijing Technology and Business University, Beijing, 102488, China

Abstract: In this paper, we discuss the criteria for detecting confounders in the survival function under the assumption that the response and covariates are both continuous. By the partition-based method, we propose the necessary and sufficient conditions for a uniformly irrelevant factor when a covariate is one-dimensional. Then the sufficient conditions for a conditional uniformly irrelevant factor vector are also obtained when a covariate is multidimensional. We propose the criteria for detecting a single confounder and multiple confounders in survival function.

Keywords: survival function; confounding bias; confounder; uniformly irrelevant factor

2020 Mathematics Subject Classification: 62N05