Teil 2: Kaplan-Meier Schätzer

Nicola Valenti

22.05.2006

1 Herleitung des Kaplan-Meier Schätzer

Die Beziehung zwischen Λ (die kumulierte Risikofunktion) und S (die Ueberlebensfunktion) liefert eine Methode, um die Ueberlebensfunktion zu Schätzen. Durch die Definition (F(t) Verteilungsfunktion von einer nichtnegativen Zufallsvariable)

$$\Lambda(t) = \int_0^t [1 - F(s-)]^{-1}\, dF(s)$$

so

$$d\Lambda(s) = \frac{dF(s)}{1 - F(s-)}$$

Wenn $F(0) = 0$ dann folgt:

$$F(t) - \underbrace{F(0)}_{=0} = \int_0^t dF(s) = \int_0^t [1 - F(s-)]\, d\Lambda(s) \tag{1}$$

so ist F von Λ eindeutig bestimmt.
Um S zu Schätzen, benutzt man den *Nelson Schätzer*:

$$\widehat{\Lambda}(t) = \int_0^t [\overline{Y}(s)]^{-1}\, d\overline{N}(s)$$

für Λ in der Gleichung (1), und man definiert $\widehat{S}$ (der Schätzer von der Ueberlebensfunktion) rekursiv:

$$\widehat{S}(t) = 1 - \int_0^t \widehat{S}(s-)\, d\widehat{\Lambda}(s) \tag{2}$$

Dann

$$\begin{aligned} \widehat{S}(t-) - \widehat{S}(t) &= -\Delta\widehat{S}(t) \\ &= \widehat{S}(t-)\tfrac{\Delta\overline{N}(t)}{\overline{Y}(t)} \end{aligned}$$

$$\widehat{S}(t) \;=\; \widehat{S}(t-)\left[1-\tfrac{\Delta\overline{N}(t)}{\overline{Y}(t)}\right]$$

und

$$\boxed{\widehat{S}(t)=\prod_{s\le t}\left[1-\tfrac{\Delta\overline{N}(s)}{\overline{Y}(s)}\right]}$$

$\widehat{S}(t)$ ist der *Kaplan-Meier Schätzer*, der im Kapitel 0 eingeführt worden ist.

2 Systematicher Schätzfehler/Bias

Jetzt benutzen wir das nächste Theorem, um die Eigenschaft von $\widehat{S}$ zu untersuchen.

Theorem 2.1 *Sei $S(t) > 0$, dann gilt*

$$\frac{\widehat{S}(t)}{S(t)} = 1 - \int_0^t \frac{\widehat{S}(s-)}{S(s)}\left(\frac{d\overline{N}(s)}{\overline{Y}(s)} - d\Lambda(s)\right) \tag{3}$$

□

Aus Theorem 2.1 folgt:

Korollar 2.1 *Für jedes t so dass $S(t) > 0$, gilt*

$$\widehat{S}(t) - S(t) = -S(t)\int_0^t \frac{\widehat{S}(s-)}{S(s)}\frac{I_{\{\overline{Y}(s)>0\}}}{\overline{Y}(s)}\,dM(s) + B(t) \tag{4}$$

wo

$$B(t) = I_{\{T<t\}}\frac{\widehat{S}(T)[S(T)-S(t)]}{S(T)}$$

und $\quad T = inf\{s : \overline{Y}(s) = 0\}$

□

Die Gleichung (4) gibt uns direkte Informationen über der Bias von Kaplan-Meier Schätzer.

Lemma 2.1 *Wenn $S(t) > 0$ folgt:*

1.

$$\begin{aligned} E[\widehat{S}(t) - S(t)] &= E[B(t)] \\ &= E\left[I_{\{T<t\}} \frac{\widehat{S}(T)[S(T)-S(t)]}{S(T)}\right] \\ &\geq 0 \end{aligned}$$

und

2. Wenn $\pi_j(t) = \pi(t)$ für jedes j,

$$E[\widehat{S}(t) - S(t)] \leq [1 - S(t)][1 - \pi(t)]^n$$

□

The Kaplan-Meier Schätzer hat so einen nichtnegativen Bias, der mit exponentialer Rate für $n \to \infty$ gegen 0 konvergiert.

Durch Lemma 2.1, hat $\widehat{S}(t)$ einen Bias (aber es ist nicht so gross), nur wenn es eine positive Wahrscheinlichkeit gibt, so dass $\widehat{S}(T) > 0$ und $S(t) < S(T)$ sind. In dem zufallszensierten Modell für unzensierten Daten (also für $P[U_j \geq T_j] = 1$), $\widehat{S}(t) \equiv 0$ für $t > T$, hat $\widehat{S}(t)$ keinen Bias (also ist $\widehat{S}(t)$ erwartungstreu für S(t)) für jedes t. In unzensierten Daten, reduziert Kaplan-Meier Schätzer zu der empirischen kumulierten Verteilungsfunktion, und hat keinen Bias. Für eine Anwendung, wenn die Zensierung für ein Intervall $[a; b]$ benötig wird, in dem S konstant ist; dann hat $\widehat{S}(t)$ keinen Bias für alle $t \leq b$.

Beispiel 2.1 *Die Einschreibungszeit zu einer Stichprobe für einen klinischen Versuch ist gegeben.*

Der Versuch beginnt, wenn die Anzahl von Teilnehmern eine festgesetzte Grösse n in der Zeit a (Einheitszeit) erreicht wird. Nach einer Zeit $b - a$ sind die Daten des Versuchs analysiert.

Die Daten sind von der sorgfältigen Wahl von der Anzahl der Teilnehmern und von der Zeitanalyse bedingt. Man kann annehmen, dass die Eingangszeit E_i von dem i-tem Subjekt eine wahrscheinliche Dichte $f_E(t)$ in $[0; a]$ hat und die zensierte Ueberlebensfunktion für jedes Subjekt folgende ist

$$C(t) = \begin{cases} 1 & 0 \leq t \leq b - a \\ \int_t^b f_E(b - s)\, ds & b - a < t \leq b \\ 0 & b < t \end{cases}$$

Wenn $T = max\{X_j : j = 1, \ldots, n\}$, dann hat die Version von Kaplan-Meier, die hier definiert ist folgender Bias:

$$E\left[I_{\{T<t\}}\frac{\widehat{S}(T)[S(T)-S(t)]}{S(T)}\right]=0$$

für $t<b-a$, *weil* $\widehat{S}(T)=0$ *für* $T<b-a$ *ist. Der Kaplan-Meier Schätzer hat keinen Bias für die Werte von* t, *so dass die Zensierung unmöglich ist. Für* $t>b-a$, *ist die Aussage für Bias kompliziert, so wie für kleine Stichproben. Eine einfache, aber nützliche obere Schranke für den Bias ist mit Lemma 2.1 Punkt* 2 *gegeben. Wenn die Verteilungsfunktion der Eingangszeit Uniform in* $[0;a]$ *ist, dann*

$$C(t)=\begin{cases}1 & 0<t\leq b-a\\ a^{-1}(b-t) & b-a<t\leq b\\ 0 & b<t<\infty.\end{cases}$$

$\widehat{S}$ *wird keinen Bias in* $[0;b-a]$ *haben, und man wird ein Bias von*

$$[1-S(t)]\left[1-S(t)\left(\frac{b-t}{a}\right)\right]^n$$

in $b-a<t\leq b$ *beschränken. Wenn man keine Daten über* $t=b$ *hat, kann man den Schätzer für* $t>b$ *nicht berechnen, und so ist das Problem des Bias irrelevant. Der relative Bias in* $[b-a;b]$ *wird kleiner oder gleich als*

$$\left[\frac{1-S(t)}{S(t)}\right]\left[1-S(t)\left(\frac{b-t}{a}\right)\right]^n$$

sein.

Nehmen wir an, dass der Versuch für eine vierjährige Registrierung offen ist, mit 5 Patienten pro Jahr, die für zwei Jahren analysiert werden. Setzen wir voraus, dass ein Schlusspunkt die Zeit von der Registrierung bis zur Entwicklung von einer Krankheit ist.
Wenn die Rate von Entwiklung für drei Jahre 35% *ist, dann erschliesst man, dass die Schätzung des relativen Bias auf* $S(3)$ *von*

$$\left(\frac{0.65}{0.35}\right)\left[1-(0.35)\frac{3}{4}\right]^{20}=4.21\times10^{-3}$$

begrenzt ist.

□

3 Varianz

Eine Abschätzung der Varianz von $\widehat{S}(t)$ kann man mit dem gleichen Argument erhalten wie die Varianz von Nelson Schätzer.
Noch einmal, nimmt man $\pi_j(t) = \pi(t) > 0$ für jedes j an. Dann ist

$$\text{var } \widehat{S}(t) = E\left[\{\widehat{S}(t) - [S(t) + E[B(t)]]\}^2\right] \tag{5}$$

und

$$E\left[\{\widehat{S}(t) - [S(t) + E[B(t)]]\}^2\right] - E\left[\{\widehat{S}(t) - S(t) - B(t)\}^2\right] \to 0$$

exponential schnell für $n \to 0$, wo die Konvergenz von Lemma 2.1 und $0 \leq B(t) < 1$ folgt. Also

$$\begin{aligned} n^{-1}V(t) &\equiv E\left[\{\widehat{S}(t) - S(t) - B(t)\}^2\right] \\ &= E\left[\{S(t)\int_0^t \frac{\widehat{S}(s-)}{S(s)} \frac{I_{\{\overline{Y}(s)>0\}}}{\overline{Y}(s)}\, dM(s)\}^2\right] \\ &= S^2(t)\int_0^t E\left[\frac{\widehat{S}^2(s-)}{S^2(s)} \frac{I_{\{\overline{Y}(s)>0\}}}{\overline{Y}(s)}\right][1 - \Delta\Lambda(s)]\, d\Lambda(s) \end{aligned} \tag{6}$$

Mit (5), kann man die endliche-Stichprobe Varianz von dem Kaplan-Meier Schätzer schätzen und man erhält die folgende Schätzung $n^{-1}\widehat{V}(t)$ von $n^{-1}V(t)$. Ersetzt man S, $\Delta\Lambda$ und $d\Lambda$ in (6) respektiv mit $\widehat{S}$, $\frac{\Delta\overline{N}}{\overline{Y}}$ und $\frac{d\overline{N}}{\overline{Y}}$, so folgt dass:

$$\begin{aligned} n^{-1}\widehat{V}(t) &\equiv \widehat{S}^2(t)\int_0^t \left\{\frac{\prod_{v<s}\left[1-\frac{\Delta\overline{N}(v)}{\overline{Y}(v)}\right]}{\prod_{v\leq s}\left[1-\frac{\Delta\overline{N}(v)}{\overline{Y}(v)}\right]}\right\}^2 \frac{I_{\{\overline{Y}(s)>0\}}}{\overline{Y}(s)}\left[1 - \frac{\Delta\overline{N}(s)}{\overline{Y}(s)}\right]\frac{d\overline{N}(s)}{\overline{Y}(s)} \\ &= \widehat{S}^2(t)\int_0^t \left[1 - \frac{\Delta\overline{N}(s)}{\overline{Y}(s)}\right]^{-2}\left[1 - \frac{\Delta\overline{N}(s)}{\overline{Y}(s)}\right]\frac{d\overline{N}(s)}{\overline{Y}^2(s)} \\ &= \widehat{S}^2(t)\int_0^t \frac{d\overline{N}(s)}{[\overline{Y}(s) - \Delta\overline{N}(s)]\overline{Y}(s)} \end{aligned}$$

Bemerkung 3.1 $n^{-1}\widehat{V}(t) = 0$ *für* t^* *so dass* $\overline{Y}(t^*) = \Delta\overline{N}(t^*)$ *wegen* $\widehat{S}(t^*) = 0$.

□

Benutzt man eine andere Annährung, verwendet man $\widehat{V}(t)$ um $\text{var}(\sqrt{n}\widehat{S}(t))$ zu schätzen.

Die Aussage (6) für die approximative Varianz von $\widehat{S}(t)$ liefert gefühlsmässig die asymptotische Verteilung von $\widehat{S}(t)$. Man hat

$$V(t) = S^2(t)\int_0^t E\left[\frac{\widehat{S}^2(s-)}{S^2(s)}\frac{n}{\overline{Y}(s)}I_{\{\overline{Y}(s)>0\}}\right][1 - \Delta\Lambda(s)]\, d\Lambda(s)$$

Da $V(t) - \text{var}(\sqrt{n}\widehat{S}(t)) \to 0$ exponential schnell ist, errcheint es sinnvoll, dass die Verteilung von $\sqrt{n}(\widehat{S}(t) - S(t))$, für grösse n, approximativ $\mathcal{N}(0, \sigma^2)$ (Normalverteilung) ist, wo

$$\sigma^2(t) = S^2(t) \int_0^t \frac{S^2(s-)}{S^2(s)} [\pi(s)]^{-1} [1 - \Delta\Lambda(s)] \, d\Lambda(s)$$

Die Schätzung $\widehat{S}$ ist tatsächlich asymptotisch Normal verteilt, und die Aussage für $\sigma^2(t)$ ist sehr einfach:
wegen der Beziehung zwischen Λ und F,

$$\begin{aligned} 1 - \Delta\Lambda(s) &= 1 - \frac{F(s) - F(s-)}{1 - F(s-)} \\ &= \frac{1 - F(s)}{1 - F(s-)} \end{aligned}$$

und

$$d\Lambda(s) = \frac{dF(s)}{1 - F(s-)}$$

so

$$\begin{aligned} \sigma^2(t) &= -S^2(t) \int_0^t \frac{S^2(s-)}{S^2(s)} [\pi(s)]^{-1} \frac{S(s)}{S^2(s-)} \, dS(s) \\ &= -S^2(t) \int_0^t \frac{dS(s)}{\pi(s) S(s)} \end{aligned}$$

Das ist das Ergebnis von Breslow und Growley (1974).
Dieses Ergebnis ist keine gute Idee, um $\widehat{S}$ über die letzte beobachtete Zeit zu erweitern. Das hier abgeleitete Ergebnis, benutzt die günstige Konvention $\widehat{S}(t) = \widehat{S}(T)$ für $t \geq T$. Im Fall von mässig oder stark zensierten Daten, ist diese Konvention in der Praxis schwierig zu verteidigen und ein Datenanalytiker betrachtet, dass $\widehat{S}$ unbestimmt über den (Zufalls-) Intervall $(T; \infty]$ ist, oder höchstens unbeobachtbar mit einem Wert $0 \leq \widehat{S}(t) \leq \widehat{S}(T)$ für $t > T$.

Die Zahlprozesse und die martingale Annäherung ergeben eine Struktur, um ein Schätzer $\widehat{S}$ von S vorzuschlagen, um die Formel zu erhalten, die von der exakten Varianz von $n^{1/2}\widehat{S}$ abzuleiten ist, um eine Annährung (V) für diese Varianze und einen Schätzer($\widehat{V}$) zu haben und um die asymptotische Varianz (σ^2) von $n^{1/2}\widehat{S}$ abzuleiten.