

Martingaalvahed Coxi mudeli hindamisel

Coxi mudel on lihtne: T on meid huvitav mittenegatiivne pidev juhuslik suurus, X on juhuslik vektor, mis modelleerib argumenttunnuseid. Seos nende vahel on järgmine:

$$P(T \geq t | X = x) = P(T > t | X = x) = e^{-h(x)\Lambda_0(t)}. \quad (1)$$

Siin $h(X)$ on ppsitiivne funktsioon, harilikult $h(X) = \exp[a'X]$, kus a on parameetrite vektor. Kui X on juhuslik suurus, siis $h(X) = \exp[aX]$, kus $a \in \mathbb{R}$. Mittengatiivne diferentseeruv funktsioon $\Lambda_0 : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ on baasrisk (*baseline cumulative hazard*), olgu

$$\lambda_0(t) = \Lambda_0'(t).$$

Seega tingimusel $X = x$ on T tinglik tihedus

$$f(t|x) := h(x)e^{-h(x)\Lambda_0(t)}\lambda_0(t).$$

Juhusliku suuruse T üleelamisfunktsioon ja kumulatiivne risk on seega

$$\begin{aligned} S(t) &= EP(T > t | X) = Ee^{-h(X)\Lambda_0(t)}, \\ \Lambda(t) &= -\ln S(t), \\ \lambda(t) &= \Lambda'(t) = -\frac{S'(t)}{S(t)} = \lambda_0(t) \frac{E[h(X)e^{-h(X)\Lambda_0(t)}]}{E[e^{-h(X)\Lambda_0(t)}]} = \lambda_0(t)E[h(X)|T > t], \end{aligned}$$

sest

$$E[h(X)e^{-h(X)\Lambda_0(t)}] = E[h(X)P(T > t | X)] = \int_{\{T > t\}} h(X)d\mathbf{P} = E[h(X); T > t].$$

Näide: Olgu $X \sim E(\rho)$ (eksponentjaotus), kus $\rho > 0$ ning olgu $\Lambda_0(t) = t$. Siis

$$T | X = x \sim E(h(x)).$$

Juhusliku suuruse T (tingimatu) üleelamisfunktsioon avaldub

$$S(t) = Ee^{-h(X)t} = \rho \int_0^\infty e^{-h(x)t - \rho x} dx.$$

Kui $h(x) = e^{-\rho x}$, siis ülaltoodud integraal on

$$S(t) = \rho \int_0^\infty \exp[-e^{\rho x}t - \rho x] dx = \frac{1 - e^{-t}}{t}.$$

Järelikult T jaotusfunktsioon, kumulatiivne risk ja tihedusfunktsioon on vastavalt

$$\begin{aligned} F(t) &= 1 - S(t) = \frac{t - 1 + e^{-t}}{t}, \\ \Lambda(t) &= -\ln S(t) = \ln t - \ln(1 - e^{-t}), \\ f(t) &= -S'(t) = \frac{1 - e^{-t} - te^{-t}}{t^2}. \end{aligned}$$

Näeme, et T jaotus pole üldse eksponentjaotus, kuigi tinglikult on (iga x korral) T eksponentjaotusega. Sel jaotusel puudub keskvärtus.

Juhusliku suuruse T kumulatiivne risk avaldub:

$$\Lambda(t) = \int_0^t \lambda(s) ds = \int_0^t E[h(X)|T > t] \lambda_0(s) ds = \int_0^t \bar{h}(s) \lambda_0(s) ds, \quad (2)$$

kus

$$\bar{h}(t) := E[h(X)|T > t] = \frac{E[h(X)e^{-h(X)\Lambda_0(t)}]}{E[e^{-h(X)\Lambda_0(t)}]} = \frac{Eh(X)P(T > t|X)}{S(t)} = \frac{Eh(X)P(Y(t) = 1|X)}{P(Y(t) = 1)}.$$

Siin $Y(t) := I_{\{T \geq t\}}$.

Martingaalvahe. Olgu $N(t) := I_{\{T \leq t\}}$. On teada, et

$$M(t) := N(t) - \Lambda(T \wedge t) = I_{\{T \leq t\}} - \int_0^{T \wedge t} \bar{h}(s) \lambda_0(s) ds = N(t) - \int_0^t Y(s) \bar{h}(s) \lambda_0(s) ds$$

on martingaal, siit nimi. Muuhulgas järeldub sellest, et

$$EN(t) = P(T \leq t) = E\Lambda(T \wedge t),$$

ja seda saab (ositi integreerimise kaudu) näidata ka otse.

Leiame nüüd

$$\frac{E[\Lambda(T \wedge t)|X]}{E[N(t)|X]}.$$

Nimetaja: Keskvärtust $E[N(t)|X]$ saame rehkendada mitmeti:

$$\begin{aligned} E[N(t)|X] &= P(T \leq t|X) = \int_0^t f(s|x) ds = h(X) \int_0^t e^{-h(X)\Lambda_0(s)} \lambda_0(s) ds \\ &= 1 - e^{-h(X)\Lambda_0(t)} = h(X) E[\Lambda_0(t \wedge T)|X]. \end{aligned}$$

Viimane võrdus tuleb sellest, et fikseeritud $X = x$ korral on tingliku jaotuse $T|X = x$ kumulatiivne risk $h(x)\Lambda_0(t)$. Seosest (1) saame

$$E[N(t)|X] = h(X) \int_0^t e^{-h(X)\Lambda_0(s)} \lambda_0(s) ds = h(X) \int_0^t P(Y(s) = 1) \lambda_0(s) ds.$$

Lugeja: leiame ositi integreerides (esimeses reas kasutame seost (2) ja teises reas seost (1))

$$\begin{aligned} E[\Lambda(T \wedge t)|X] &= \int_0^\infty \Lambda(t \wedge s) f(s|x) ds = -(1 - F(s|x)) \Lambda(t \wedge s)|_0^\infty + \int_0^\infty (1 - F(s|x)) \bar{h}(s) I_{s < t} \lambda_0(s) ds \\ &= \int_0^t e^{-h(X)\Lambda_0(s)} \bar{h}(s) \lambda_0(s) ds = \int_0^t P(Y(s) = 1|X) \bar{h}(s) \lambda_0(s) ds. \end{aligned}$$

Seega

$$\begin{aligned}\frac{E[\Lambda(T \wedge t)|X = x]}{E[N(t)|X = x]} &= \frac{\int_0^t \bar{h}(s) e^{-h(x)\Lambda_0(s)} \lambda_0(s) ds}{h(x) \int_0^t e^{-h(x)\Lambda_0(s)} \lambda_0(s) ds} = \frac{\int_0^t P(Y(s) = 1|X = x) \bar{h}(s) \lambda_0(s) ds}{\int_0^t P(Y(s) = 1|X = x) \lambda_0(s) ds} \\ \frac{E[\Lambda(T \wedge t)|X = x] h(x)}{E[N(t)|X = x]} &= \frac{\int_0^t e^{-h(x)\Lambda_0(s)} \bar{h}(s) \lambda_0(s) ds}{\int_0^t e^{-h(x)\Lambda_0(s)} \lambda_0(s) ds} =: \bar{h}(x, t)\end{aligned}\quad (3)$$

Seega

$$\begin{aligned}\frac{E[M(t)|X]}{E[N(t)|X]} &= 1 - \frac{E[\Lambda(T \wedge t)|X]}{E[N(t)|X]} = 1 - \frac{\bar{h}(X, t)}{h(X)} \\ 1 - \frac{E[M(t)|X]}{E[N(t)|X]} &= \frac{\bar{h}(X)}{h(X)} = \frac{E[\Lambda(T \wedge t)|X]}{E[N(t)|X]} \\ -\ln\left(1 - \frac{E[M(t)|X]}{E[N(t)|X]}\right) &= \ln(h(X)) - \ln(\bar{h}(X, t)) = -\ln\left(\frac{E[\Lambda(T \wedge t)|X]}{E[N(t)|X]}\right).\end{aligned}$$

Seega kui $x \mapsto \bar{h}(x, t)$ on ligikaudu konstant, siis

$$-\ln\left(1 - \frac{E[M(t)|X = x]}{E[N(t)|X = x]}\right) \approx \ln(h(x))$$

ja hinnates andmete põhjal $\frac{E[M(t)|X=x]}{E[N(t)|X=x]}$ saame hinnata meid huvitavat funktsiooni $h(x)$.
Et

$$\lim_{t \rightarrow \infty} E[N(t)|X] = 1, \quad \lim_{t \rightarrow \infty} E[\Lambda(T \wedge t)|X] = E[\Lambda(T)|X]$$

(tinglik domineeritud ja tinglik monotoonse koondumise teoreem), saame

$$\lim_{t \rightarrow \infty} -\ln\left(1 - \frac{E[M(t)|X]}{E[N(t)|X]}\right) = \lim_{t \rightarrow \infty} -\ln\left(\frac{E[\Lambda(T \wedge t)|X]}{E[N(t)|X]}\right) = -\ln E[\Lambda(T)|X = x] =: g(x).$$

Seega $g(x)$ peaks olema $\ln h(x)$ hinnang. Kas ikka on?

Kõik kokku: Olgu $\Lambda_0(t) = t$ ja P_X mingi tõenäosusjaotus. Vaatlusi (andmeid) denereeriv mudel:

$$\begin{aligned}X_i &\stackrel{\text{i.i.d.}}{\sim} P_X, \quad i = 1, \dots, n \\ T_i|X_i &\stackrel{\text{ind}}{\sim} E(h(X_i)) \quad i = 1, \dots, n.\end{aligned}$$

Eesmärk: vaatluste põhjal hinnata funktsiooni $h(x)$. Kui eeldame, et $h(x) = e^{aX}$, siis eesmärk on hinnata parameetrit a . Uuritav meetod käib nii. Empiirilise jaotusfunktsiooni F_n põhjal hinnatakse kumulatiivset riski

$$\hat{\Lambda}(s) = -\ln(1 - F_n(s)).$$

Siis, fikseeritud ajahetke t korral leotakse martingaaljäägid

$$M_i(t) := I_{T_i \leq t} - \hat{\Lambda}(T_i \wedge t), \quad i = 1, \dots, n.$$

Nende põhjal tuleb hinnata tinglikke keskväärtusi:

$$E[\widehat{M(t)}|X = x], \quad EP[\widehat{T \leq t}|X = x], \quad \forall x$$

ja nii saame hinnangu

$$g(t, x) = -\ln \left(1 - \frac{E[\widehat{M(t)}|X = x]}{EP[\widehat{T \leq t}|X = x]} \right).$$

Kas $x \mapsto g(t, x)$ on mõistlik hinnang funktsioonile $\ln h(x)$?