Bayesovska analiza doživljenja

Malović, Irena

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Irena Malović

BAYESOVSKA ANALIZA DOŽIVLJENJA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, srpanj 2014.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Tema ovog rada je bayesovski pristup u analizi doživljenja. Zašto baš bayesovski pristup? Općenito, opće je poznato da je modele doživljenja poprilično teško modelirati, pogotovo kada je dio podataka cenzuriran. Kroz ovaj diplomski rad ćemo probati predložiti neke od poluparametarskih modela koji će uz bayesovki pristup nastojati olakšati modeliranje problema.
Prva dva poglavlja ovog diplomskog rada su uvodna poglavlja. Prvo je uvod u analizu doživljenja gdje ćemo se bolje upoznati s pojmovima kao što su funkcija doživljenja, funkcija hazarda, funkcija vjerodostojnosti i funkcija parcijalne vjerodostojnosti. Također ćemo se upoznati sa tipovima cenzuriranja podataka, te ćemo neke od navedenih stvari probati prikazati na primjerima.
Drugo poglavlje je uvod u Bayesovu statistiku. U njemu ćemo objasniti pojmove apriorne i aposteriorne distribucije, informativnim i neinoformativnim apriornim distribucijama, te ćemo na primjerima pokazati kako preko bayesovskog pristupa doći do tih distribucija.
Nakon toga u trećem poglavlju se bavimo poluparametarskim modelima. Objasnit ćemo četiri različita modela, te se na kraju u zadnjem poglavlju orijentirati na generalizaciju Coxovog modela. Taj model ćemo probati i pokazati na primjeru jedne studije.
Izračuni i grafovi na u tom primjeru su napravljeni u programskom jeziku R.

Poglavlje 1

Uvod u analizu doživljenja

1.1 Definicije

Analiza doživljenja je grana statistike koja se bavi analizom perioda vremena do jednog ili više događaja. Primjeri nekih takvih događaja su vrijeme do smrti, vrijeme do nastupanja AIDS-a kod HIV pozitivnih osoba ili vrijeme do kvara na nekom stroju. Sama analiza doživljenja nam pokušava odgovoriti, na primjer, na sljedeća pitanja. Koliki dio populacije će preživjeti određeno vrijeme? Kako određene okolnosti ili svojstva povećavaju ili smanjuju vjerojatnost preživljenja?
Neka je T neprekidna, nenegativna slučajna varijabla koja predstavlja vrijeme doživljenja neke jedinke u populaciji. Nadalje, ukoliko se ne naglasi drugačije, podrazumijevamo da su sve funkcije definirane na intervalu [0,+∞). Označimo funkciju gustoće slučajne varijable T sa $f(t)$, te sa $F(t)$ njenu funkciju distribucije:

$$F(t) = P(T \leq t) = \int_0^t f(u)du. \tag{1.1}$$

Funkciju doživljenja $S(t)$ definiramo kao vjerojatnost pojedinca da doživi vrijeme t i ona je dana formulom

$$S(t) = 1 - F(t) = P(T > t). \tag{1.2}$$

Funkcija $S(t)$ ima sljedeća svojstva:

- $S(t)$ je monotono padajuća funkcija
- $S(0) = 1$, tj. vjerojatnost doživljenja trenutka $t = 0$ je jednaka 1
- $S(\infty) = \lim_{t\to\infty} S(t) = 0$, tj. kada $t \to \infty$, funkcija doživljenja teži ka 0.

Funkcija hazarda $h(t)$ se definira kao trenutna stopa neuspjeha u trenutku t i dana je formulom

$$h(t) = \lim_{\triangle t \to 0} \frac{P(t < T \leq t + \triangle t | T > t)}{\triangle t} = \frac{f(t)}{S(t)} \tag{1.3}$$

Funkcije $f(t)$, $F(t)$, $S(t)$ i $h(t)$ se lako mogu dobiti jedna iz druge. Kako je $f(t) = -\frac{d}{dt}S(t)$ iz toga slijedi da je

$$h(t) = -\frac{d}{dt}\log(S(t)). \tag{1.4}$$

Iz te dvije formule dalje slijedi da se funkcija doživljenja može izraziti preko funkcije hazarda narednom formulom:

$$S(t) = \exp\left(-\int_0^t h(u)du\right). \tag{1.5}$$

Nadalje, definiramo i funkciju kumulativnog hazarda $H(t)$ sa $H(t) = \int_0^t h(u)du$, koja je sa funkcijom doživljenja povezana formulom $S(t) = \exp(-H(t))$. Kako je
$S(\infty) = 0$, slijedi da je $H(\infty) = \lim_{t\to\infty} H(t) = \infty$. Iz toga zaključujemo da funkcija hazarda ima iduća svojstva:

- nenegativnost: $h(t) \geq 0$,
- nepostojanje gornje granice: $\int_0^\infty h(t)dt = \infty$.

Iz (1.3) i (1.5) slijedi da funkciju gustoće slučajne varijable T možemo zapisati formulom:

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right). \tag{1.6}$$

1.2 Modeli proporcionalnih hazarda

Modeli proporcionalnih hazarda su klasa statističkih modela doživljenja. Općenito, funkcija hazarda ovisi i o vremenu i o skupu kovarijata, od kojih neke mogu biti ovisne o vremenu. Modeli proporcionalnih hazarda odvajaju te dvije komponente te je hazard u trenutku t za pojedinca, čiji je vektor kovarijable $\mathbf{x} = (x_1, x_2, ..., x_n)$, dan formulom:

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t)G(\mathbf{x}, \beta) \\ &= h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n), \end{aligned} \tag{1.7}$$

gdje $h_0(t)$ zovemo *funkcija osnovnog hazarda*, a $\beta = (\beta_1, \beta_2, ..., \beta_n)$ je vektor regresijskih koeficijenata. Funkcija osnovnog hazarda $h_0(t)$ se dobiva kada sve kovarijate izjednačimo sa 0, tj. $x_1 = x_2 = ... = x_n = 0$, iz čega uvrštavanjem u formulu (1.7) dobivamo:

$$\begin{aligned} h_0(t|\mathbf{x}) &= h_0(t)\exp(\beta_1 \cdot 0 + \beta_2 \cdot 0 + ... + \beta_n \cdot 0) \\ &= h_0(t). \end{aligned}$$

S obzirom da dio formule $G(\mathbf{x}, \beta)$ mora bit pozitivan koristi se u eksponencijalnoj formi. Takav model implicira da je omjer hazarda kod dva pojedinca iz populacije konstantan tijekom danog vremena i da se kovarijate ne mijenjaju tijekom vremena. Učinak povećanja udjela u kovarijatima multiplikativan je s obzirom na stopu hazarda

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}'\beta). \tag{1.8}$$

$\eta = \mathbf{x}'\beta$ nazivamo linearni prediktor. Omjer hazarda dva pojedinca ovisi o razlici između njihovih linearnih prediktora. Iz jednadžbe (1.8) znamo da je funkcija hazarda za i-tog pojedinca jednaka

$$h(t|\mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i'\beta), \tag{1.9}$$

te iz (1.9) dobivamo funkciju osnovnog hazarda kao funkciju hazarda i-tog pojedinca za kojeg vrijedi $\mathbf{x}_i = 0$. Ako podijelimo (1.9) sa $h(t|\mathbf{x}_j)$ dobivamo jednadžbu

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_j)} = \frac{\exp(\mathbf{x}_i'\beta)}{\exp(\mathbf{x}_j'\beta)},$$

koja ne ovisi o vremenu. Iz nje vidimo odakle je došao termin 'proporcionalan'.

1.3 Cenzuriranje podataka

Dosta česta pojava u analizi doživljenja je cenzuriranje podataka. Cenzuriranje je prisutno kada imamo neke informacije o vremenu promatranog događaja za nekog pojedinca, no ne znamo točno vrijeme tog događaja. Postoje tri tipa cenzuriranja podataka i to su:

1. Desno cenzuriranje: kažemo da je opservacija desno cenzurirana u c ako točna vrijednost opservacije nije poznata, nego je veća ili jednaka c.
   Primjer: pojedinac koji sudjeluje u istraživanju se povuče iz istog, te za njega neznamo kada i da li je nastupio promatrani događaj.

2. Lijevo cenzuriranje: kažemo da je opservacija lijevo cenzurirana u c, ako znamo samo da je manja ili jednaka c.
   Primjer: starost djeteta u kojem nauče neki zadatak. Ako je dijete već znalo to napraviti na početku istraživanja taj podataka će biti lijevo cenzuriran.

3. Intervalno cenzuriranje: znamo jedino da je opservacija u nekom intervalu (c_1, c_2).
   Primjer: da bi otkrili povratak raka debelog crijeva nakon operacije, bolesnik se prati svaka tri mjeseca nakon resekcije primarnog tumora.

Tri najčešća razloga zbog kojih dolazi do cenzuriranja su:

- pojedinac ne doživi događaj prije kraja istraživanja,
- tijekom promatranog razdoblja su podaci o pojedincu izgubljeni (kompjuterskom ili ljudskom greškom),
- pojedinac se povuče iz istraživanja prije njegovog kraja.

Do kraja diplomskog rada ćemo se susretati samo sa desno cenzuriranim podacima, te ćemo pretpostaviti da je cenzuriranje neinformativno, što znači da im je vrijeme cenzuriranja nezavisno od vremena opaženih događaja. Konstruirajmo sada funkciju vjerodostojnosti za model proporcionalnog hazarda sa desno cenzuriranim podacima. Pretpostavimo da imamo n pojedinaca u istraživanju. Sa t_i ćemo označiti vrijeme doživljenja $i-$tog pojedinca, a sa c_i fiksno cenzurirano vrijeme $i-$tog pojedinca. Također, pretpostavimo da su vremena opažanja događaja t_i nezavisna i jednako distribuirana sa funkcijom gustoće $f(t)$ i funkcijom doživljenja $S(t)$. Vrijeme doživljenja t_i ćemo opažati samo u slučaju ako je $t_i \leq c_i$. Podatke sada možemo zapisati u obliku uređenog para (y_i, ν_i), gdje je

$$y_i = \min(t_i, c_i) \tag{1.10}$$

i

$$\nu_i = \begin{cases} 1, & \text{ako } t_i \leq c_i, \\ 0, & \text{ako } t_i > c_i. \end{cases} \tag{1.11}$$

Tada je funkcija vjerodostojnosti za $(\beta, h_0(.))$ za desno cenzurirane podatke o n pojedinaca dana sa

$$\begin{aligned} L(\beta, h_0(\cdot)|D) &\propto \prod_{i=1}^{n} [h_0(y_i)\exp(\eta_i)]^{\nu_i} S(y_i|\eta_i) \\ &= \prod_{i=1}^{n} [h_0(y_i)\exp(\eta_i)]^{\nu_i} S_0(y_i)^{\exp\eta_i} \\ &= \prod_{i=1}^{n} [h_0(y_i)\exp(\eta_i)]^{\nu_i} \exp\left(-\sum_{i=1}^{n} \exp(\eta_i) H_0(y_i)\right) \\ &= \prod_{i=1}^{n} h_0(y_i)^{\nu_i} \exp\left[\sum_{i=1}^{n} \nu_i \eta_i - \sum_{i=1}^{n} H_0(y_i)\exp(\eta_i)\right], \end{aligned} \tag{1.12}$$

gdje su $D = (n, \mathbf{y}, \mathbf{X}, \nu)$, $\mathbf{y} = (y_1, y_2, ..., y_n)'$, $\nu = (\nu_1, \nu_2, ..., \nu_n)'$, $\eta_i = \mathbf{x}_i'\beta$. η_i je linerani prediktor za pojedinca i, $\mathbf{x}_i$ je $p \times 1$ vektor kovarijabli za pojedinca i, $\mathbf{X}$ je $n \times p$ matrica kovarijabli sa $i-$tim retkom $\mathbf{x}_i'$, i $S_0(t)$ je osnovna funkcija doživljenja, koja je povezana sa $h_0(.)$ preko formule $S_0(t) = \exp\left(-\int_0^t h_0(u)du\right) = \exp(-H_0(t))$.

1.4 Primjeri

Izvest ćemo funkcije vjerodostojnosti za neke parametarske modele koji će nam trebati kasnije u diplomskom radu.

1.4.1 Eksponencijalni model

Eksponencijalni model je osnovni parametarski model u analizi doživljenja. Pretpostavimo da imamo nezavisna, jednako distribuirana vremena doživljenja $\mathbf{y} = (y_1, ..., y_n)'$. Svako to vrijeme je distribuirano eksponencijalno s parametrom λ. Pretpostavimo također da je $\nu = (\nu_1, ..., \nu_n)'$ indikatorska funkcija cenzuriranja zadana kao u (1.11). Funkcija gustoće varijable y_i je zadana sa $f(y_i|\lambda) = \lambda\exp(-\lambda y_i)$, a funkcija doživljenja sa $S(y_i|\lambda) = \exp(-\lambda y_i)$. Sa $D = (n, \mathbf{y}, \nu)$ označimo skup opservacija. Tada je funkcija vjerodostojnosti za λ jednaka

$$\begin{aligned} L(\lambda|D) &= \prod_{i=1}^{n} f(y_i|\lambda)^{\nu_i} S(y_i|\lambda)^{(1-\nu_i)} \\ &= \prod_{i=1}^{n} (\lambda\exp(-\lambda y_i))^{\nu_i} (\exp(-\lambda y_i))^{1-\nu_i} \\ &= \lambda^d \exp\left(-\lambda \sum_{i=1}^{n} y_i\right), \end{aligned} \tag{1.13}$$

gdje je $d = \sum_{i=1}^{n} \nu_i$. To je broj necenzuriranih podataka.

1.4.2 Log-normalni model

Pretpostavka ovog modela je da su logaritmi vremena doživljenja $\mathbf{y} = (y_1, ..., y_n)'$ nezavisni i normalno distribuirani sa parametrima (μ, σ), u oznaci $y_i \sim \mathcal{LN}(\mu, \sigma^2)$. Funkcija gustoće varijable y_i je tada dana sa

$$f(y_i|\mu, \sigma) = (2\pi)^{\frac{-1}{2}} (y_i\sigma)^{-1} \exp\left\{-\frac{1}{2\sigma^2}(log(y_i) - \mu)^2\right\},$$

a funkcija doživljenja sa

$$S(y_i|\mu, \sigma) = 1 - \Phi\left(\frac{log(y_i) - \mu}{\sigma}\right).$$

Funkcija vjerodostojnosti od (μ, σ) je tada jednaka

$$
\begin{aligned}
L(\mu,\sigma|D) &= \prod_{i=1}^{n} f(y_i|\mu,\sigma)^{\nu_i} S(y_i|\mu,\sigma)^{(1-\nu_i)} \\
&= \prod_{i=1}^{n} \left((2\pi)^{\frac{-1}{2}} (y_i\sigma)^{-1} \exp\left\{ -\frac{1}{2\sigma^2}(log(y_i)-\mu)^2 \right\} \right)^{\nu_i} \left(1 - \Phi\left(\frac{log(y_i)-\mu}{\sigma} \right) \right)^{1-\nu_i} \\
&= (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \nu_i (log(y_i)-\mu)^2 \right\} \times \prod_{i=1}^{n} y_i^{-\nu_i} \left(1 - \Phi\left(\frac{log(y_i)-\mu}{\sigma} \right) \right)^{1-\nu_i}.
\end{aligned}
\tag{1.14}
$$

1.4.3 Gama model

Gama model je generalizacija eksponencijalnog modela. Pretpostavimo da su nam vremena doživljenja $\mathbf{y} = (y_1, ..., y_n)'$ jednako distribuirana po gama distribuciji sa parametrima (α, λ), u oznaci $y_i \sim \mathcal{G}(\alpha, \lambda)$. Funkcija gustoće je tada dana sa

$$f(y_i|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)} y_i^{\alpha-1} \exp(\alpha\lambda - y_i \exp(\lambda)),$$

a funkcija doživljenja

$$S(y_i|\alpha,\lambda) = 1 - IG(\alpha, y_i \exp(\lambda)),$$

gdje je

$$IG(\alpha, y_i\exp(\lambda)) = \frac{1}{\Gamma(\alpha)} \int_0^{y_i\exp(\lambda)} u^{\alpha-1}\exp(-u)du$$

nepotpuna gama funkcija. Funkcija vjerodostojnosti je tada funkcija od (α, λ)

$$
\begin{aligned}
L(\alpha,\lambda) &= \prod_{i=1}^{n} f(y_i|\alpha,\lambda)^{\nu_i} S(y_i|\alpha,\lambda)^{(1-\nu_i)} \\
&= \prod_{i=1}^{n} \left(\frac{1}{\Gamma(\alpha)} y_i^{\alpha-1} \exp(\alpha\lambda - y_i\exp(\lambda)) \right)^{\nu_i} (1 - IG(\alpha, y_i\exp(\lambda)))^{1-\nu_i} \\
&= \frac{1}{(\Gamma(\alpha))^d} \exp\left\{ d\alpha\lambda + \sum_{i=1}^{n} \nu_i (\alpha log(y_i) - y_i\exp(\lambda)) \right\} \\
&\times \prod_{i=1}^{n} y_i^{-\nu_i} (1 - IG(\alpha, y_i\exp(\lambda)))^{1-\nu_i}.
\end{aligned}
\tag{1.15}
$$

1.5 Funkcija parcijalne vjerodostojnosti

U ovom diplomskom radu ćemo se baviti Coxovom verzijom modela proporcionalnog hazarda. Model je poluparametarski u smislu da osnovna funkcija hazarda $h_0(t)$ nije modelirana kao parametarska funkcija od t. Funkcija $h_0(t)$ može poprimati proizvoljne vrijednosti, s obzirom na to da ona ne ulazi u jednadžbe procjene parametara. U nastavku navedimo Coxovu funkciju parcijalne vjerodostojnosti.
Pretpostavimo da imamo n opservacija, te da od tih n opservacija postoji d različitih opaženih vremena događaja i $n - d$ desno cenzuriranih vremena doživljenja. Pretpostavimo također da samo jedan pojedinac umire u određenom trenutku vremena, tako da nema ponavljanja u podacima. Uređena vremena doživljenja, koja su međusobno različita, označimo sa $y_{(1)}$,$y_{(2)}$,...,$y_{(d)}$, tako da je $y_{(j)}$ $j-$to po redu vrijeme doživljenja. Sa $\mathcal{R}_j$ označimo skup pojedinaca koji su pod rizikom od nastupanja događaja i opažamo ih neposredno prije trenutka $y_{(j)}$. Nazivamo ga još i skup rizika. U svakom trenutku smrti $y_{(j)}$, doprinos funkciji vjerodostojnosti je:

$$\begin{aligned} L_j(\beta) &= \mathbb{P}(\text{za pojedinca } j \text{ je nastupio događaj} \mid \text{događaj je nastupio za nekog pojedinca iz } \mathcal{R}_j) \\ &= \frac{\mathbb{P}(\text{za pojedinca } j \text{ je nastupio događaj} \mid \text{rizičan u trenutku } y_{(j)})}{\sum_{l \in \mathcal{R}_j} \mathbb{P}(\text{za pojedinca } l \text{ je nastupio događaj} \mid \text{rizičan u trenutku } y_{(j)})} \\ &= \frac{h(y_{(j)}|\mathbf{x}_{(j)})}{\sum_{l \in \mathcal{R}_j} h(y_{(j)}|\mathbf{x}_l)} \end{aligned}$$

Iz gornje jednadžbe za vjerodostojnost i formule (1.8) dobivamo da je Coxova funkcija parcijalne vjerodostojnosti za β dana sa:

$$PL(\beta|D) = \prod_{j=1}^{d} \frac{h_0(y_{(j)})\exp(\mathbf{x}'_{(j)}\beta)}{\sum_{l \in \mathcal{R}_j} h_0(y_{(j)})\exp(\mathbf{x}'_l\beta)} = \prod_{j=1}^{d} \frac{\exp(\mathbf{x}'_{(j)}\beta)}{\sum_{l \in \mathcal{R}_j} \exp(\mathbf{x}'_l\beta)}. \tag{1.16}$$

Nazivnik u formuli (1.16) je suma svih vrijednosti $\exp(\mathbf{x}'_i\beta)$ pojedinaca i koji su pod rizikom u vremenu neposredno do $y_{(j)}$. Funkcija parcijalne vjerodostojnosti može se zapisati i na sljedeći način:

$$PL(\beta|D) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}'_{(i)}\beta)}{\sum_{l \in \mathcal{R}_i} \exp(\mathbf{x}'_l\beta)} \right)^{\nu_i},$$

gdje je ν_i definiran kao u (1.11). Procjena maksimalne parcijalne vjerodostojnosti od β može se dobiti maksimizacijom izraza (1.16) po β. Za taj postupak se koriste numeričke metode kao što je Newton-Raphsonova metoda.

Poglavlje 2

Uvod u Bayesovu statistiku

2.1 Apriorna i aposteriorna distribucija i Bayesov teorem

Neka je $\mathbf{X} = (X_1, X_2..., X_n)$ slučajni uzorak iz parametrizirane vjerojatnosne distribucije. Sa $f_i(x|\theta)$ označimo funkciju gustoće varijable X_i, koja nam je poznata. Pretpostavljamo da su parametri $\theta = (\theta_1, ..., \theta_n)$ nepoznati i svi dolaze iz prostora Θ. Baysovski pristup statističkom modelu pretpostavlja da mi iz opservacije $\mathbf{X} = \mathbf{X}(\omega)$ možemo zaključiti nešto o nepoznatom parametru θ, te pomoću te informacije doći do vjerojatnosne distribucije za njega.
Apriorna distribucija je distribucija parametra θ koja je određena bez znanja podataka iz slučajnog uzorka $\mathbf{X}$. Njenu funkciju gustoće označavamo sa $\pi(\theta)$. Ona se ponajviše bazira na subjektivnosti statističara.

Za daljnji raspis Bayesovog pristupa statističkim modelima potreban nam je Bayesov teorem.

Teorem 2.1.1. *(Bayesova formula) Neka je $(H_i, i = 1, 2, ..)$ potpun sistem događaja u vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ i $A \in \mathcal{F}$ takav da je $\mathbb{P}(A) > 0$. Tada za svako i vrijedi*

$$\mathbb{P}(H_i|A) = \frac{\mathbb{P}(H_i)\mathbb{P}(A|H_i)}{\sum_j \mathbb{P}(H_j)\mathbb{P}(A|H_j)}.$$

Kada tu Bayesovsku formulu primjenimo na funkcije gustoće nekih slučajnih varijabli X i Y dobivamo formulu čiji dokaz možemo naći u [6].

$$f_{Y|X}(y|x) = \frac{f_Y(y)f_{X|Y}(x|y)}{f_X(x)} = \frac{f_Y(y)f_{X|Y}(x|y)}{\int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)dy}, \tag{2.1}$$

gdje je $f_{X|Y}(x|y)$ uvjetna gustoća od X uz dano Y, a $f_Y(y)$ je (marginalna) funkcija gustoće slučajne varijable Y.

Aposteriorna distribucija je distribucija koja se dobiva iz apriorne gustoće $\pi(\theta)$ uz uvjetovanje na slučajni uzorak $\mathbf{X}$. Njenu funkciju gustoće označavamo sa $\pi(\theta|\mathbf{x})$. Dobivena je preko Bayesovog teorema po formuli

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x}|\xi)\pi(\xi)d\xi}. \tag{2.2}$$

Kada promatramo slučajan događaj koji ovisi o parametru θ, statističke metode nam dozvoljavaju da zaključimo nešto o tom parametru θ, dok vjerojatnosni modeli karakteriziraju buduće ponašanje opservacije uvjetno na θ. Tim načinom dolazimo do funkcije vjerodostojnosti $L(\theta|\mathbf{x})$, koja je jednaka

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta),$$

te se sada aposteriorna gustoća može napisati

$$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi(\theta)}{\int_\Theta L(\xi|\mathbf{x})\pi(\xi)d\xi}. \tag{2.3}$$

Iz (2.3) vidimo da je $\pi(\theta|\mathbf{x})$ proporcionalna funkciji vjerodostojnosti pomnoženoj sa apriornom distribucijom s gustoćom, $\pi(\theta|\mathbf{x}) \propto L(\theta|\mathbf{x})\pi(\theta)$. Označimo sa $m(\mathbf{x}) = \int_\Theta L(\xi|\mathbf{x})\pi(\xi)d\xi$. Tada $m(\mathbf{x})$ zovemo normalizirajuća konstanta od $\pi(\theta|\mathbf{x})$ ili marginalna gustoća distribucije od $\mathbf{X}$.

U većini modela, $m(\mathbf{x})$ nema analitički zatvorenu formu, tj. ne može se izraziti analitički pomoću konačnog broja "poznatih" funkcija, pa onda ni $\pi(\theta|\mathbf{x})$ nema zatvorenu formu. Prirodno pitanje koje proizlazi iz tog problema je kako uzeti uzorak iz multivarijatne distribucije sa gustoćom $\pi(\theta|\mathbf{x})$ kada nemamo zatvorene forme za nju? Jedna od najpopularnijih metoda koja daje rješenje tog problema je Gibbsova metoda uzorkovanja koja omogućuje da uzimamo uzorak iz $\pi(\theta|\mathbf{x})$ bez toga da znamo normalizirajuću konstantu. Uz nju koristi se i MCMC metoda (metoda Monte Carlo-Markovljev lanac). Sada kada smo naveli sve bitne definicije i formule definirajmo točno Bayesov statistički model:

Definicija 2.1.2. *Bayesovski statistički model je model sastavljen od parametarskog statističkog modela $\mathcal{P} = \{f(\cdot|\theta) : \theta \in \Theta\}$ i apriorne distribucije za parametre reprezentirane gustoćom $\pi(\theta)$.*

2.2 Informativna i neinformativna apriorna distribucija i konjugirani prior

Informativna apriorna distribucija izražava specifičnu određenu informaciju o varijabli. Primjerice apriorna distribucija za temperaturu sutra u podne. Razuman pristup za taj

problem bi bio da za apriornu distribuciju uzmemo normalnu distribuciju sa očekivanom vrijednosti koja je jednaka današnjoj temperaturi u podne, te homogenom dnevnom varijancom.

Neinformativna apriorna distribucija izražava nejasnu ili generalnu informaciju o varijabli. Naziva se još i objektivna apriorna distribucija jer ona izražava "objektivne" informacije kao što su: varijabla je pozitivna, varijabla je manja od neke vrijednosti x. Za neinformativnu apriornu distribuciju najčešće se uzima uniformna distribucija jer ona dodjeljuje svakoj vrijednosti istu vjerojatnost (iznos gustoće).

Iako neinformativne apriorne distribucije mogu biti korisnije i lakše za odrediti kod nekih problema, njih ne možemo koristiti za sve modele. One mogu izazvati probleme u aposteriornim procjenama i tako dovesti do problema s konvergencijama u Gibbsovoj metodi. Povrh toga, neinformativne apriorne distribucije ne koriste pravu apriornu informaciju koju možemo imati o nekom problemu. Korisne su u istraživanjima gdje provoditelj istraživanja ima pristup prijašnjim podacima i rezultatima sličnih istraživanja. Primjerice, u mnogim kliničkim istraživanjima o raku ili AIDS-u, trenutna istraživanja često koriste postupke koji su vrlo slični ili malo modificirani nekim starijim istraživanjima. Tada je prirodno uklopiti podatke iz starih sličnih u novo trenutno istraživanje, kvantificirajući ih sa prikladnom apriornom distribucijom na parametrima modela.

Konjugiranim priorom nazivamo apriornu distribuciju koja zadovoljava svojstvo da su aposteriorna $\pi(\theta|\mathbf{x})$ i apriorna $\pi(\theta)$ vjerojatnosna distribucija iz iste familije. Tada apriornu i aposteriornu distribuciju nazivamo konjugiranim distribucijama.

Primjer 2.2.1. *Neka je X slučajna varijabla s Bernoulijevom distribucijom $B(n, \theta)$, $\theta \in (0, 1)$. Njena funkcija vjerojatnosti je tada*

$$p(x) = \binom{n}{x} \theta^x (1-\theta)^{(n-x)}.$$

Ako izrazimo to kao funkciju od θ dobivamo:

$$\pi(\theta) \propto \theta^a (1-\theta)^b,$$

gdje su a i b neke konstante. Uobičajeno, ta forma će još imati neki multiplikativni faktor, tj. normalizirajuću konstantu, koja nikad neće ovisiti o θ, već će u većini slučajeva biti funkcija od a i b. Za taj faktor uzmemo funkciju beta, te je konjugirani prior tada beta distribucija sa parametrima α i β, te je njegova funkcija vjerojatnosti dana sa

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}, \theta \in (0, 1).$$

Parametre α i β odaberemo iz podataka, te ih nazivamo hiperparametrima (kako bi ih razlikovali od parametara početnog modela). Sada računamo aposteriornu distribuciju. Neka je $d = \sum_{i=1}^{n} \nu_i$ broj necenzuriranih podataka (uspjeha) i $f = n - \sum_{i=1}^{n} \nu_i$ broj cenzuriranih podataka (neuspjeha)

$$p(f, d|\theta) = \binom{n}{d}\theta^d(1-\theta)^f$$

$$\begin{aligned}\pi(\theta|f, d) &= \frac{p(f, d|\theta)\pi(\theta)}{\int p(f, d|\theta)\pi(\theta)d\theta} \\ &= \frac{\theta^{(d+\alpha-1)}(1-\theta)^{(f+\beta-1)}}{B(d+\alpha, f+\beta)},\end{aligned}$$

To je opet beta distribucija, samo sa parametrima $d + \alpha$ i $f + \beta$.

2.3 Primjeri

Izvedimo sada konjugirane priore za primjere modela iz poglavlja 1.

2.3.1 Eksponencijalni model

Funkcija vjerodostojnosti eksponencijalnog modela je dana formulom (1.13). Konjugirani prior za λ je tada gama apriorna distribucija sa parametrima (α_0, λ_0) i gustoćom danom formulom $\pi(\lambda|\alpha_0, \lambda_0) \propto \lambda^{(\alpha_0-1)}\exp(-\lambda_0\lambda)$. Iz toga slijedi da je onda aposteriorna distribucija dana sa

$$\begin{aligned}\pi(\lambda|D) &\propto L(\lambda|D)\pi(\lambda|\alpha_0, \lambda_0) \\ &\propto \left(\lambda^{\sum_{i=1}^{n}\nu_i}\exp\left\{-\lambda\sum_{i=1}^{n} y_i\right\}\right)(\lambda^{\alpha_0-1}\exp(-\lambda_0\lambda)) \\ &= \lambda^{\alpha_0+d-1}\exp\left\{-\lambda(\lambda_0 + \sum_{i=1}^{n} y_i)\right\}\end{aligned}$$

Dakle aposteriorna distribucija ima gama distribuciju sa parametrima $(\alpha_0 + d, \lambda_0 + \sum_{i=1}^{n} y_i)$. Za izradu regersijskog modela, kovarijate uvrštavamo preko λ i pišemo $\lambda_i = \varphi(\mathbf{x}_i'\beta)$, gdje je $\mathbf{x}_i$ $p \times 1$ vektor kovarijata, β je $p \times 1$ vektor regresijskih koeficijenata, a $\varphi(.)$ je poznata funkcija. Najčešće korišteni oblik za funkciju φ je $\varphi(\mathbf{x}_i'\beta) = \exp(\mathbf{x}_i'\beta)$. Koristeći taj φ

funkcija vjerodostojnosti glasi:

$$\begin{aligned} L(\beta|D) &= \prod_{i=1}^{n} f(y_i|\lambda_i)^{\nu_i} S(y_i|\lambda_i)^{(1-\nu_i)} \\ &= \prod_{i=1}^{n} \left[\exp(\mathbf{x}_i'\beta) \cdot \exp(-y_i \exp(\mathbf{x}_i'\beta))\right]^{\nu_i} \left[\exp(-y_i \cdot \exp(\mathbf{x}_i'\beta))\right]^{(1-\nu_i)} \\ &= \exp\left\{\sum_{i=1}^{n} \nu_i \mathbf{x}_i'\beta\right\} \exp\left\{-\sum_{i=1}^{n} y_i \exp(\mathbf{x}_i'\beta)\right\}. \end{aligned}$$

Skup D je kao i do sada $D = (n, \mathbf{y}, \mathbf{X}, \nu)$, gdje je $\mathbf{X}$ $n \times p$ matrica kovarijabli sa i-tim retkom $\mathbf{x}_i'$. Ako promatramo opservacije bez kovarijata tada je $D = (n, \mathbf{y}, \nu)$. Najčešće korištene apriorne distribucije za β su uniformna takva da $\pi(\beta) \propto 1$ i normalna.

2.3.2 Log-normalni model

Funkcija vjerodostojnosti log-normalnog modela dana je formulom (1.14). Neka je $\tau = \frac{1}{\sigma^2}$. Ako pretpostavimo da su oba parametra (μ, τ) nepoznati, ne postoji konjugirani prior za njih. U tom slučaju uzima se $\mu|\tau \sim N\left(\mu_0, \frac{1}{\tau\tau_0}\right)$ i $\tau \sim \mathcal{G}(\frac{\alpha_0}{2}, \frac{\lambda_0}{2})$. Tada je zajednička aposetriorna distribucija od (μ, τ) dana formulom

$$\begin{aligned} \pi(\mu, \tau|D) &\propto L(\mu, \sigma|D)\pi(\mu, \tau|\mu_0, \tau_0, \alpha_0, \lambda_0) \\ &\propto \tau^{\frac{\alpha_0+d}{2}-1} \exp\left\{-\frac{\tau}{2}\left[\sum_{i=1}^{n} \nu_i(\log(y_i) - \mu)^2 + \tau_0(\mu - \mu_0)^2 + \lambda_0\right]\right\} \\ &\times \prod_{i=1}^{n} y_i^{-\nu_i} \left(1 - \Phi\left(\tau^{\frac{1}{2}}(\log(y_i) - \mu)\right)\right)^{(1-\nu_i)}. \end{aligned} \tag{2.4}$$

Zajednička aposteriorna distribucija nema zatvorenu formu. Kako bi modelirali regresijski model uvodimo kovarijate preko μ i pišemo $\mu_i = \mathbf{x}_i'\beta$. Zajednička apriorna distribucija za β uključuje uniformnu apriornu distribuciju $\pi(\beta) \propto 1$ ili normalnu apriornu distribuciju $\beta|\tau \sim N_p(\mu_0, \tau^{-1}\Sigma_0)$. U normalnom slučaju, aposteriorna distribucija za (β, τ) je dana formulom

$$\begin{aligned} \pi(\beta, \tau|D) &\propto \tau^{\frac{\alpha_0+d}{2}-1} \exp\left\{-\frac{\tau}{2}\left[\sum_{i=1}^{n} \nu_i(\log(y_i) - \mathbf{x}_i'\beta)^2 + (\beta - \mu_0)'\Sigma_0^{-1}(\beta - \mu_0) + \lambda_0\right]\right\} \\ &\times \prod_{i=1}^{n} y_i^{-\nu_i} \left(1 - \Phi\left(\tau^{\frac{1}{2}}(\log(y_i) - \mathbf{x}_i'\beta)\right)\right)^{(1-\nu_i)}. \end{aligned} \tag{2.5}$$

2.3.3 Gama model

Funkcija vjerodostojnosti gama modela dana je formulom (1.15). Ako pretpostavimo da su oba koeficijenta (α, λ) u formuli (1.15) nepoznata, ne možemo zadanoj funkciji pridružiti niti jedan konjugirani prior. Uzmimo da su α i λ nezavisni, te $\alpha \sim \mathcal{G}(\alpha_0, \kappa_0)$ i $\lambda \sim N(\mu_0, \sigma_0^2)$. Uz tu pretpostavku zajednička aposteriorna distribucija za (α, λ) je dana sa

$$\begin{aligned}\pi(\alpha, \lambda|D) &\propto L(\alpha, \lambda|D)\pi(\alpha, \lambda|\alpha_0, \kappa_0, \mu_0, \sigma_0) \\ &\propto \frac{\alpha^{\alpha_0-1}}{(\Gamma(\alpha))^d}\exp\left\{d\alpha\lambda + \sum_{i=1}^{n} \nu_i(\alpha log(y_i) - y_i\exp(\lambda))\right\} \\ &\times \prod_{i=1}^{n} y_i^{-\nu_i}(1 - IG(\alpha, y_i\exp(\lambda)))^{1-\nu_i} \\ &\times \exp\left(-\kappa_0\alpha - \frac{1}{2\sigma_0^2}(\lambda - \mu_0)^2\right).\end{aligned}$$

Kako bi modelirali regresijski model uvodimo kovarijate preko parametra λ i pišemo $\lambda_i = \mathbf{x}_i'\beta$. Zajednička apriorna distribucija za β uključuje uniformnu apriornu distribuciju $\pi(\beta) \propto 1$ ili normalnu apriornu distribuciju $\beta \sim N_p(\mu_0, \Sigma_0)$. U normalnom slučaju, aposteriorna distribucija za (β, α) je dana formulom

$$\begin{aligned}\pi(\beta, \alpha|D) \propto \frac{\alpha^{\alpha_0-1}}{(\Gamma(\alpha))^d}\exp&\left\{\sum_{i=1}^{n} \nu_i[\alpha(\mathbf{x}_i'\beta + \log(y_i)) - y_i\exp(\mathbf{x}_i'\beta)]\right\} \\ &\times \prod_{i=1}^{n} y_i^{-\nu_i}\left(1 - \mathrm{IG}\left(\alpha, y_i\exp(\mathbf{x}_i'\beta)\right)\right)^{(1-\nu_i)} \\ &\times \exp\left(-\kappa_0\alpha - \frac{1}{2}(\beta - \mu_0)'\Sigma_0^{-1}(\beta - \mu_0)\right).\end{aligned} \tag{2.6}$$

2.4 Prednosti i nedostaci Bayesovog pristupa

Bayesova metoda, isto kao i sve klasične metode ima i prednosti i nedostataka. Navedimo neke od njih.

Prednosti Bayesovog pristupa:

- Na prirodan način se kombiniraju apriorne informacije o parametrima sa podacima. Možemo modelirati prošle informacije o parametrima i tako formirati apriornu distribuciju za buduće analize. Kada nove opservacije postanu dostupne, staru aposteriornu distribuciju možemo koristiti kao apriornu.

- Dobiveni zaključci su egzaktni i dobiveni uvjetno na podatke, bez oslanjanja na asimptotsku aproksimaciju. Prema tome, do zaključaka na malom uzorku se dolazi isto kao i do zaključaka na velikim uzorcima.
- Poštuje princip vjerodostojnosti. Ako dva različita načina uzimanja uzoraka ili opažanja rezultiraju sa proporcionalnim funkcijama vjerodostojnosti za θ, tada su svi zaključci o θ za ta dva uzorka isti.
- Prirodan je koncept za mnoge modele, kao što su hijerarhijski modeli, te modeli s problemima nedostajećih vrijednosti. Metoda MCMC, uz numeričke metode, omogućava računanje procjena parametara izvodljivim za gotovo sve parametarske modele.

Nedostaci Bayesovog pristupa:

- Sama metoda ne govori o tome kako odabrati apriornu distribuciju, te ne postoji točan način kako ju odabrati. Bayesovski zaključci zahtjevaju vještine za prevođenje subjektivne apriorne distribucije u matematički formuliranu apriornu distribuciju. Ako se ona ne odabere pažljivo, može dovesti do krivih rezultata.
- Neki puta je teško odrediti koju apriornu distribuciju uzeti. Pa se može dogoditi da se uz ”loše” određenu apriornu distribuciju, dobije kriva aposteriorna distribucija.
- Ukoliko model ima veliki broj parametara dolazi do opterećenja računalnih sustava.

Poglavlje 3

Poluparametarski modeli

Od sada na dalje u radu pod pojmom "za pojedinca i je nastupio događaj" podrazumijevat ćemo da se odnosi na "pojedinac i je umro", tj. događaj će nam biti smrt pojedinca i.

3.1 Model po dijelovima konstantnog hazarda

Ovaj model je jedan od najpopularnijih poluparametarskih modela u bayesovskoj analizi doživljenja. Pretpostavimo da je $0 < s_1 < s_2 < ... < s_J$ konačna particija varijable vremena, gdje je $s_J > y_i$ za svaki $i = 1, 2, ..., n$. Tada imamo J intervala $(s_j, s_{j+1}]$, $j = 0, 1, 2, ..., J-1$, uz $s_0 = 0$. Pretpostavimo da je za $j-$ti interval funkcija osnovnog hazarda konstantna i jednaka je $h_0(y) = \lambda_j$, za $y \in I_j = (s_{j-1}, s_j]$. Označimo sa $D = (n, \mathbf{y}, \mathbf{X}, \nu)$ skup opservacija, gdje su $\mathbf{y} = (y_1, ..., y_n)'$, $\nu = (\nu_1, ..., \nu_n)'$ zadane formulama (1.10) i (1.11), a $\mathbf{X}$ je $n \times p$ matrica kovarijata sa i-tim retkom $\mathbf{x}_i'$. Neka je $\lambda = (\lambda_1, ..., \lambda_J)'$. Tada je funkcija vjerodostojnosti za (β, λ) jednaka

$$L(\beta, \lambda|D) = \prod_{i=1}^{n} \prod_{j=1}^{J} (\lambda_j \exp(\mathbf{x}_i'\beta))^{\delta_{ij}\nu_i} \exp\left\{-\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1} \lambda_g(s_g - s_{g-1})\right]\exp(\mathbf{x}_i'\beta)\right\}, \tag{3.1}$$

gdje je $\delta_{ij} = 1$, ako je vrijeme i-tog pojedinca u j-tom intervalu cenzurirano ili ako je on umro u j-tom intervalu, a 0 inače. $\mathbf{x}_i' = (x_{i1}, x_{i2}, ..., x_{ip})$ je $p \times 1$ vektor kovarijabli za i-tog pojedinca i $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ je pripadni vektor regresijskih koeficijenata. Ako u ovom modelu stavimo $J = 1$ dobivamo parametarski eksponencijalni model sa parametrom $\lambda = \lambda_1$, pa se zbog toga gore opisani model naziva još i "po dijelovima eksponencijalni model".

Najčešće korištena apriorna distribucija za funkciju osnovnog hazarda λ je nezavisna gama distribucija, $\lambda_j \sim \mathcal{G}(\alpha_{0j}, \lambda_{0j})$, $j = 1, 2, \ldots, J$. Parametri α_{0j} i λ_{0j} mogu se dobiti iz apriorne distribucije preko matematičkog očekivanja i varijance od λ_j.

U slučaju diskretnog modela, gornja formula funkcije vjerodostojnosti glasi:

$$L(\beta, \lambda|D) \propto \prod_{j=1}^{J} G_j*, \tag{3.2}$$

gdje je

$$G_j* = \exp\left\{-\lambda_j \triangle_j \sum_{k\in\mathcal{R}_j-\mathcal{D}_j} \exp(\mathbf{x}_k'\beta)\right\} \times \prod_{l\in\mathcal{D}_j}[1 - \exp\{-\lambda_j \triangle_j \exp(\mathbf{x}_l'\beta)\}]. \tag{3.3}$$

U formuli (3.3) $\mathcal{R}_j$, kao i prije, označava skup pojedinaca koji su pod rizikom od smrti neposredno prije trenutka $y_{(j)}$, a $\mathcal{D}_j$ je skup pojedinaca koji su umrli u j-tom intervalu $\triangle_j = s_j - s_{j-1}$.

3.2 Modeli koji koriste gama procese

Gama proces je najkorišteniji neparametarski apriorni proces za Coxov model o kojem ćemo više kasnije pisati. Neka je $\mathcal{G}(\alpha, \lambda)$ gama distribucija sa parametrima $\alpha > 0$ i $\lambda > 0$, te neka je $\alpha(t)$, $t \geq 0$ neprekidna slijeva, rastuća funkcija takva da je $\alpha(0) = 0$ i $Z(t)$, $t \geq 0$ stohastički proces sa svojstvima:

- $Z(0) = 0$
- $Z(t)$ ima nezavisne priraste
- za $t > s$, $Z(t) - Z(s) \sim \mathcal{G}(c(\alpha(t) - \alpha(s)), c)$.

Takav proces $\{Z(t) : t \geq 0\}$ nazivamo gama proces i označavamo ga sa $Z(t) \sim \mathcal{GP}(\alpha(t), c)$. $\alpha(t)$ je tada matematičko očekivanje slučajne varijable $Z(t)$, a c je težina. Prirasti tog gama procesa su gotovo sigurno rastuće funkcije. Takav proces je posebna vrsta Levyevih procesa čija je karakteristična funkcija dana sa

$$\varphi_{(Z(t)-Z(s))}(y) = \mathbb{E}[\exp\{iy(Z(t) - Z(s))\}] = (\phi(y))^{\alpha(t)-\alpha(s)},$$

gdje je ϕ karakteristična funkcija beskonačno djeljive funkcije distribucije sa očekivanjem 1. Gama proces je samo specijalni slučaj $\phi(y) = \{\frac{c}{c-iy}\}^c$.

3.2.1 Gama procesi sa kumulativnim hazardom

Kod Coxovog modela, zajednička vjerojatnost preživljenja n pojedinaca uz matricu dizajna $\mathbf{X}$ dana je sa:

$$P(\mathbf{Y} > \mathbf{y}|\beta, \mathbf{X}, H_0) = \exp\left\{-\sum_{j=1}^{n} \exp(\mathbf{x}_j'\beta) H_0(y_j)\right\}. \tag{3.4}$$

Gama proces $\mathcal{GP}(c_0 H^*, c_0)$ koristimo kao apriornu distribuciju za kumulativnu funkciju osnovnog hazarda $H_0(y)$, gdje je $H^*(y)$ neprekidna slijeva rastuća funkcija za koju vrijedi $H^*(0) = 0$. Često pretpostavljamo da je H^* poznata parametarska funkcija sa vektorom hiperparametara γ_0. Marginalna funkcija doživljenja dana je sa (vidjeti [2])

$$P(\mathbf{Y} > \mathbf{y}|\beta, \mathbf{X}, \gamma_0, c_0) = \prod_{j=1}^{n} [\phi(iV_j)]^{c_0(H^*(y_{(j)}) - H^*(y_{(j-1)}))}, \tag{3.5}$$

gdje je $V_j = \sum_{l \in \mathcal{R}_j} \exp(\mathbf{x}_l'\beta)$, $y_{(1)} < y_{(2)} < ... < y_{(n)}$ su različita vremena smrti koja su uređena, a $\mathcal{R}_j$ je skup pojedinaca koji su pod rizikom od smrti neposredno prije vremena $y_{(j)}$.

3.2.2 Gama procesi sa vjerodostojnosti u odnosu na grupirane podatke

Kao u modelu po dijelovima konstantnog hazarda, podijelimo varijablu vremena na konačne particije tako da imamo J disjunktnih intervala. Označimo te intervale sa $I_j = (s_{j-1}, s_j]$. Pretpostavimo i da je $s_J > y_i$ za sve $i = 1, ..., n$. Označimo sa D skup opservacija. On je grupiran po intervalima tako da je $D = (\mathbf{X}, \mathcal{R}_j, \mathcal{D}_j : j = 1, 2, ..., J)$, gdje je $\mathcal{R}_j$ skup pojedinaca koji su pod rizikom od smrti, a $\mathcal{D}_j$ skup pojedinaca koji su umrli u intervalu I_j. Sa h_j označimo prirast kumulativne funkcije osnovnog hazarda u j-tom intervalu

$$h_j = H_0(s_j) - H_0(s_{j-1}), j = 1, 2, ..., J.$$

Gama proces iz poglavlja prije implicira da su h_j nezavisni i

$$h_j \sim \mathcal{G}(c_0(\alpha_{0,j} - \alpha_{0,j-1}), c_0), \tag{3.6}$$

gdje je $\alpha_{0,j} = c_0 H^*(s_j)$, a H^* i c_0 su definirani kao u prijašnjem podpoglavlju. Hiperparametri (H^*, c_0) za h_j se sastoje od određene parametarske kumulativne funkcije hazarda $H^*(y)$, koja je procijenjena na rubovima intervala, i od skalara $c_0 > 0$. Dakle, H_0 je gama proces $H_0 \sim \mathcal{GP}(c_0 H^*, c_0)$, iz čega vidimo da svaki prirast u H_0 ima apriornu distribuciju zadanu sa (3.6). Tada se grupirani prikaz podataka može dobiti kao

$$P(y_i \in I_j|\mathbf{h}) = \exp\left\{-\exp(\mathbf{x}_i'\beta) \sum_{k=1}^{j-1} h_k\right\} [1 - \exp\{-h_j \exp(\mathbf{x}_i'\beta)\}], \tag{3.7}$$

gdje je $\mathbf{h} = (h_1, h_2, ..., h_J)'$. Iz toga dobivamo funkciju vjerodostojnosti grupiranih podataka

$$L(\beta, \mathbf{h}|D) \propto \prod_{j=1}^{J} G_j, \tag{3.8}$$

gdje je

$$G_j = \exp\left\{-h_j \sum_{k \in \mathcal{R}_j - \mathcal{D}_j} \exp(\mathbf{x}_k'\beta)\right\} \prod_{l \in \mathcal{D}_j} \left[1 - \exp\{-h_j \exp(\mathbf{x}_l'\beta)\}\right]. \tag{3.9}$$

Primijetimo kako izraz za funkciju vjerodosotojnosti za grupirane podatke nije ograničen samo na slučaj kada je h_j realizacija gama procesa na H_0. Kako se kumulativna osnovna funkcija hazarda H_0 u formuli (3.9) pojavljuje samo preko h_j, naši parametri u funkciji vjerodostojnosti su $(\beta, \mathbf{h})$, tako da nam treba jedino zajednička apriorna distribucija za $(\beta, \mathbf{h})$. Kada za h_j uzmemo $h_j = \triangle_j \lambda_j$, gdje je $\triangle_j = s_j - s_{j-1}$ kao što su zadani u modelu po dijelovima konstantnog hazarda, vidimo veliku sličnost između funkcija vjerodostojnosti (3.3) i (3.9). U slučaju kada nemamo kovarijata, (3.9) se svodi na

$$G_j = \exp\{-h_j(r_j - d_j)\}\{1 - \exp(-h_j)\}^{d_j},$$

gdje su r_j i d_j brojevi pojedinaca u skupovima $\mathcal{R}_j$, odnosno $\mathcal{D}_j$. Često korištena apriorna distribucija za β je $N_p(\mu_0, \Sigma_0)$ distribucija. Tada je zajednička aposteriorna distribucija od $(\beta, \mathbf{h})$ dana formulom

$$\pi(\beta, \mathbf{h}|D) \propto \prod_{j=1}^{J} \left[G_j h_j^{(\alpha_{0j} - \alpha_{o,j-1})-1} \exp(-c_0 h_j)\right] \times \exp\left\{-\frac{1}{2}(\beta - \mu_0)\Sigma_0^{-1}(\beta - \mu_0)\right\}.$$

3.2.3 Odnos sa parcijalnom funkcijom vjerodostojnosti

U ovom podpoglavlju ćemo pokazati kako se parcijalna funkcija vjerodostojnosti, definirana od strane Coxa 1975. godine, može dobiti kao poseban slučaj marginalne aposteriorne distribucije od β u Coxovom modelu u kojem koristimo gama proces za kumulativnu funkciju osnovnog hazarda. Pretpostavimo da je varijabla vremena podijeljena na intervale $(0, s_1]$, $(s_1, s_2]$,...,$(s_{J-1}, s_J]$ i pretpostavimo da je $H_0 \sim \mathcal{GP}(c_0 H^*, c_0)$. Neka su $y_{(1)} < y_{(2)} < ... < y_{(n)}$ uređena vremena smrti ili cenzuriranja. Stoga, ako je $h_j = H_0(y_{(j)}) - H_0(y_{(j-1)})$, onda $h_j \sim \mathcal{G}(c_0 h_{0j}, c_0)$, gdje je $h_{0j} = H^*(y_{(j)}) - H^*(y_{(j-1)})$. Neka je $A_j = \sum_{l \in \mathcal{R}_j} \exp(\mathbf{x}_l'\beta)$. Označimo sa E_{GP} očekivanje za apriornu distribuciju gama procesa. Tada je (iz (3.4))

$$\begin{aligned} P(\mathbf{Y} > \mathbf{y}|\mathbf{X}, \beta, H_0) &= \exp\left\{-\sum_{j=1}^{n} \exp(\mathbf{x}_j'\beta) H_0(y_{(j)})\right\} \\ &= \exp\left\{-\sum_{j=1}^{n} h_j \sum_{l \in \mathcal{R}_j} \exp(\mathbf{x}_l'\beta)\right\} \end{aligned}$$

i

$$E_{GP}\left[P(\mathbf{Y} > \mathbf{y}|\mathbf{X}, \beta, H_0)|H^*\right] = \prod_{j=1}^{n}\left(\frac{c_0}{c_0 + A_j}\right)^{c_0 h_{0j}}$$
$$= \prod_{j=1}^{n} \exp\left\{c_0 H^*(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)\right\}.$$

Označimo sa $\theta = (\beta', h_0, c_0)'$, gdje je $h_0 = \frac{d}{dy}H^*(y)$. Tada funkciju vjerodostojnosti možemo pisati kao

$$L(\theta|D) = \prod_{j=1}^{n} \exp\left\{c_0 H^*(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)\right\}$$
$$\times \left\{-c_0 \frac{dH^*(y_{(j)})}{dy}\left(\log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)\right)\right\}^{\nu_i}$$
$$= \prod_{j=1}^{n} \exp\left\{H^*(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)^{c_0}\right\}$$
$$\times \left\{-c_0 h_0(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)\right\}^{\nu_i}.$$

Nadalje, neka je $d = \sum_{i=1}^{n} \nu_i$ i $h^* = \prod_{j=1}^{n}[h_0(y_{(j)})]^{\nu_i}$. Tada imamo

$$\lim_{c_0 \to 0} \exp\left\{H^*(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)^{c_0}\right\} = 1$$

za $j = 1, 2, ..., n$, i

$$\lim_{c_0 \to 0} \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right) = \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{A_j}\right) \approx -\frac{\exp(\mathbf{x}'_j\beta)}{A_j}$$

za $j = 1, 2, ..., n - 1$. Iz toga nam slijedi

$$\lim_{c_0 \to 0} \frac{L(\theta|D)}{c_0^d \log(c_0) h^*} \approx \prod_{j=1}^{n}\left[\frac{\exp(\mathbf{x}'_j\beta)}{A_j}\right]^{\nu_i}. \tag{3.10}$$

Iz formule (3.10) vidimo da je desna strana u njoj jednaka Coxovoj funkciji parcijalne vjerodostojnosti. Sada ako puštamo $c_0 \to \infty$, dobivamo funkciju vjerodostojnosti baziranu na (β, h_0).

$$\lim_{c_0 \to \infty}\left[\exp\left\{H^*(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)^{c_0}\right\} \times \left\{-c_0 h_0(y_{(j)}) \log\left(1 - \frac{\exp(\mathbf{x}'_j\beta)}{c_0 + A_j}\right)\right\}^{\nu_j}\right]$$
$$= \exp\left\{-H^*(y_{(j)})\exp(\mathbf{x}'_j\beta)\right\}\left\{h_0(y_{(j)})\exp(\mathbf{x}'_j\beta)\right\}^{\nu_j},$$

pa iz toga slijedi

$$\lim_{c_0\to\infty} L(\beta, c_0, h_0|D) = \prod_{j=1}^{n} \exp\left(-H^*(y_{(j)})\exp(\mathbf{x}'_j\beta)\right)\left(h_0(y_{(j)})\exp(\mathbf{x}'_j\beta)\right)^{\nu_j}. \tag{3.11}$$

Prema tome, vidimo da je (3.11) funkcija vjerodostojnosti od (β, h_0) bazirana na modelu proporcionalnog hazarda.

3.2.4 Gama procesi na osnovnom modelu hazarda

Sada razmatramo diskretnu aproksimaciju gama procesa. Modeliramo funkciju vjerodostojnosti koristeći model po dijelovima konstantnog osnovnog hazarda i informacije o tome u koji interval "upadnu" vremena smrti. Neka je $0 = s_0 < s_1 < ... < s_J$ konačna particija varijable vremena. Označimo sa $\delta_j = h_0(s_j) - h_0(s_{j-1})$ prirast u funkciji osnovnog hazarda na intervalu $(s_{j-1}, s_j]$, $j = 1, 2, ..., J$, gdje je $\delta = (\delta_1, \delta_2, ..., \delta_n)'$. Za proizvoljnog pojedinca iz populacije, funkcija doživljenja prema Coxovom modelu u trenutku y je dana sa

$$\begin{aligned} S(y|x) &= \exp\left\{-\eta \int_0^y h_0(u)du\right\} \\ &\approx \exp\left\{-\eta\left(\sum_{i=1}^{J} \delta_i (y - s_{i-1})^+\right)\right\}, \end{aligned} \tag{3.12}$$

gdje je $h_0(0) = 0$, $(u)^+ = u$ ako je $u > 0$, 0 inače, i $\eta = \exp(\mathbf{x}'\beta)$. Prva aproksimacija proizlazi iz definicije od δ. Sama δ ne određuje cijeli prirast hazarda, već to određuje δ_j. U svrhu aproksimacije, uzmimo da se povećanje prirasta hazarda δ_j dogodi neposredno nakon s_{j-1}. Sa p_j označimo vjerojatnost smrti u intervalu $(s_{j-1}, s_j]$, $j = 1, 2, ..., J$. Koristeći formulu (3.12), za funkciju doživljenja dobivamo

$$p_j = S(s_{j-1}) - S(s_j) \approx \exp\left\{-\eta \sum_{l=1}^{j-1} \delta_l (s_{j-1} - s_{l-1})\right\}\left[1 - \exp\left\{-\eta(s_j - s_{j-1}) \sum_{l=1}^{j} \delta_l\right\}\right].$$

Prema tome, doprinos funkciji vjerodostojnosti za smrt pojedinca u j-tom intervalu je jednak p_j ili $S(s_j)$ ako je opservacija desno cenzurirana. Označimo sada sa d_j broj umrlih, sa $\mathcal{D}_j$ skup pojedinaca koji su umrli, sa c_j broj desno cenzuriranih opservacija, te sa C_j skup pojedinaca koji su cenzurirani i to sve u j-tom intervalu. Neka je $D = (n, \mathbf{y}, \mathbf{X}, \nu)$ skup opservacija. Vjerodostojnost u odnosu na grupirane podatke je tada dana sa

$$L(\beta, \delta|D) = \prod_{j=1}^{J} \left\{\exp\{-\delta_j(a_j + b_j)\} \prod_{k\in\mathcal{D}_j} [1 - \exp\{-\eta_k T_j\}]\right\}, \tag{3.13}$$

gdje je $\eta_k = \exp(\mathbf{x}_k'\beta)$, i

$$a_j = \sum_{l=j+1}^{J} \sum_{k \in \mathcal{D}_l} \eta_k (s_{l-1} - s_{j-1}), \tag{3.14}$$

$$b_j = \sum_{l=j}^{J} \sum_{k \in C_l} \eta_k (s_l - s_{j-1}), \tag{3.15}$$

$$T_j = (s_j - s_{j-1}) \sum_{l=1}^{j} \delta_l. \tag{3.16}$$

U formuli (3.13) umjesto da uvjetujemo na točna vremena kad se dogodila smrt, mi uvjetujemo na skup događaja (smrti pojedinaca) i skup desno cenzuriranih podataka na svakom intervalu. Na taj način, aproksimiramo neprekidne, desno cenzurirane podatke po grupama.

3.3 Metoda traženja apriorne distribucije na temelju starih istraživanja

U ovoj metodi ćemo se susretati sa pojmom *potencirane apriorne distribucije* (*eng.*power prior). Neka je $a_0 \in [0, 1]$ neki broj. Apriorne distribucije koje se temelje na ideji potenciranja funkcije vjerodostojnosti podataka D_0 na potenciju a_0 se nazivaju potencirane apriorne distribucije. Sa $D_0 = (n_0, \mathbf{y}_0, \mathbf{X}_0, \nu_0)$ smo označili skup podataka iz nekog prijašnjeg istraživanja, gdje je n_0 veličina tog uzorka (broj opservacija), $\mathbf{y}_0$ je vektor desno cenzuriranih vremena doživljenja sa indikatorskom funkcijom cenzuriranja ν_0, i $\mathbf{X}_0$ je $n \times p$ matrica kovarijata. Ova metoda nastoji iskoristiti slične primjere problema u svrhu pravilnog odabira apriorne distribucije za budući problem. U pravilu, nema pravilnog načina odabira same apriorne distribucije, pa se najčešće pokušava uspostaviti neka ravnoteža između teoretskog, praktičnog i kompjuterski izvedivog odabira distribucije. Problem kako upotrijebiti D_0 u nekom novom istraživanju modela doživljenja ponajviše ovisi o tome koliko su ta istraživanja slična. Na različitosti istraživanja može utjecati mnogo faktora. Primjerice, provođenje istraživanja u različitim institucijama ili na različitim geografskim lokacijama, korištenje različitih mjernih instrumenata itd. Zbog tih različitosti, analiza koja kombinira staro i novo istraživanje može biti prikladna nakon utežavanja podataka.
Označimo sa $\pi_0(\beta, \delta)$ apriornu distribuciju za (β, δ). Potencirana apriorna distribucija za (β, δ) je dana sa

$$\pi(\beta, \delta | D_0, a_0) \propto \{L(\beta, \delta | D_0)\}^{a_0} \pi_0(\beta, \delta), \tag{3.17}$$

gdje je $L(\beta, \delta | D_0)$ funkcija vjerodostojnosti od (β, δ) bazirana na podacima D_0. Ta funkcija vjerodostojnosti je dana u formuli (3.13) uz uvjet da je na svim mjestima D zamijenjen sa

$D_0 = (n_0, \mathbf{y}_0, \mathbf{X}_0, \nu_0)$. Kako bi pojednostavili apriornu distribuciju stavimo da je

$$\pi_0(\beta, \delta) = \pi_0(\beta|c_0)\pi_0(\delta|\theta_0),$$

gdje su c_0 i θ_0 fiksirani hiperparametri. Za $\pi_0(\beta|c_0)$ uzimamo da je to gustoća p- dimenzionalne multivarijatne normalne distribucije $N_p(0, c_0 W_0)$, sa očekivanjem 0 i matricom kovarijanci $c_0 W_0$, gdje je c_0 skalar, a W_0 $p \times p$ dijagonalna matrica. Za $\pi_0(\delta|\theta_0)$ uzmemo da je gustoća produkta M nezavisnih gama distribucija, od kojih svaka ima očekivanje $\frac{f_{0i}}{g_{0i}}$ i varijancu $\frac{f_{0i}}{g_{0i}^2}$, $i = 1, 2, ..., M$. Dakle dobivamo,

$$\pi_0(\delta|\theta_0) \propto \prod_{i=1}^{M} \delta_i^{f_{0i}-1} \exp\{-\delta_i g_{0i}\},$$

gdje je $\theta_0 = (f_{01}, g_{01}, ..., f_{0M}, g_{0M})'$. Određivanje apriorne distribucije je gotovo nakon što odaberemo apriornu distribuciju za a_0. Za nju uzmimo beta distribuciju $a_0 \sim \mathcal{B}(\alpha_0, \lambda_0)$. Tada je

$$\pi(a_0|\alpha_0, \lambda_0) \propto a_0^{\alpha_0-1}(1 - a_0)^{\lambda_0-1},$$

iz čega dobivamo zajedničku apriornu distribuciju

$$\pi(\beta, \delta, a_0|D_0) \propto L(\beta, \delta|D_0)^{a_0}\pi_0(\beta|c_0)\pi_0(\delta|\theta_0)\pi(a_0|\alpha_0, \lambda_0). \tag{3.18}$$

3.3.1 Aproksimacija apriorne distribucije

Jedna od mana formule (3.18) je da nema zatvorenu formu i da može biti teško kompjuterski izvediva. Zbog toga, treba je aproksimirati na prikladan način.
Pretpostavimo da su prirasti funkcije osnovnog hazarda i regresijski koeficijenti međusobno nezavisni. Tada je zajednička gustoća apriorne distribucije dana sa

$$\pi(\beta, \delta|D_0) = \pi(\beta|D_0)\pi(\delta|D_0). \tag{3.19}$$

Kako je pretpostavka o nezavisnosti β i δ dosta osjetljiva, uzimamo u obzir potpunu p-dimenzionalnu multivarijatnu normalnu apriornu distribuciju za β danu sa

$$(\beta|a_0, \mu_0, T_0) \sim N_p(\mu_0, a_0^{-1}T_0^{-1}), \tag{3.20}$$

gdje je μ_0 očekivanje apriorne distribucije, $a_0 T_0$ kovarijacijska matrica, a $a_0 \in [0, 1]$. Uzimamo da je μ_0 rješenje Coxove funkcije parcijalne vjerodostojnosti za β koristeći podatke D_0. Pretpostavimo da je d_0 broj smrti, a $n_0 - d_0$ broj desno cenzuriranih podataka u $\mathbf{y}_0$ iz skupa podataka D_0. Coxova funkcija parcijalne vjerodostojnosti je tada dana sa

$$PL(\beta|D_0) = \prod_{i=1}^{d_0} \left(\frac{\exp(\mathbf{x}'_{0i}\beta)}{\sum_{l \in \mathcal{R}_{y_{0i}}} \exp(\mathbf{x}'_{0l}\beta)} \right), \tag{3.21}$$

gdje je $\mathbf{x}'_{0i}$ i-ti redak matrice $\mathbf{X}_0$ iz skupa podataka D_0, $(y_{01}, ..., y_{0d_0})$ su uređena vremena smrti, a $\mathcal{R}_{(y_{0i})}$ je skup pojedinaca koji su pod rizikom od smrti neposredno prije trenutka y_{0i}. Sada uzmemo da je μ_0 rješenje za

$$\frac{\partial \log(PL(\beta|D_0))}{\partial \beta} = 0.$$

Za matricu T_0 uzmemo da je ona matrica Fisherovih informacija od β bazirana na parcijalnoj funkciji vjerodostojnosti (3.21). Tada je

$$T_0 = \left[\frac{-\partial^2}{\partial\beta\partial\beta'} \log(PL(\beta|D_0))\right]\Bigg|_{\beta=\mu_0}. \tag{3.22}$$

Ove apriorne distribucije predstavljaju sažetak podataka D_0 preko (μ_0, T_0), koji su dobiveni iz Coxove funkcije parcijalne vjerodostojnosti. Naš glavni cilj je bio doći do zaključaka o β, te kako je δ gledan kao pomoćni parametar (*eng*. nuisance parameter), pa je prema tome neinformativna apriorna distribucija $\pi(\delta|a_0, D_0)$ primjerena za odabir. Za $\pi(\delta|a_0, D_0)$ možemo uzeti i produkt M gama distribucija sa očekivanjem ϕ_i i varijancom $a_0^{-1}\gamma_i$, za $i = 1, 2, ..., M$. Parametri ϕ_i i γ_i mogu biti procijenjeni iz podataka D_0 na način, $\phi_i = \hat{\delta}_{0i}$ i $\widehat{\gamma_i} = \widehat{\text{Var}}(\hat{\delta}_{0i})$, gdje je $\hat{\delta}_{0i}$ procjenitelj dobiven metodom maksimalne vjerodostojnosti i $\widehat{\text{Var}}(\hat{\delta}_{0i})$ je odgovarajuća procijenjena varijanca od δ_i izračunata iz $L_c(\delta|D_0, \beta = \mu_0)$.
Neke od prednosti korištenja apriorne distribucije dane formulama (3.19) i (3.20):

- ima zatvorenu formu,
- (μ_0, T_0) i a_0 u potpunosti opisuju apriornu distribuciju za β,
- relativno su jednostavne i kompjuterski izvedive,
- model se dodatno pojednostavljuje zbog pretpostavke nezavisnosti između β i δ.

3.3.2 Izbor hiperparametara

Izbor parametara igra bitnu ulogu u Bayesovoj analizi. Recimo prvo nešto o izboru parametara (α_0, λ_0). U svrhu određivanja apriorne distribucije lakše je raditi sa parametrima $\mu_{a_0} = \frac{\alpha_0}{\alpha_0+\lambda_0}$ i $\sigma^2_{a_0} = \frac{\mu_{a_0}(1-\mu_{a_0})}{\alpha_0+\lambda_0+1}$. Uniformna apriorna distribucija ($\alpha_0 = \lambda_0 = 1$), koja odgovara $(\mu_{a_0}, \sigma^2_{a_0}) = (\frac{1}{2}, \frac{1}{2})$ može biti prikladna kao neinformativna apriorna distribucija za početak, te kasnije olakšati analizu za ostale izbore distribucija. Parametar μ_{a_0} odabiremo da bude mali ($\mu_{a_0} \leq 0.1$), ili veliki ($\mu_{a_0} \geq 0.5$) u ovisnosti o tome da li želimo podacima u D_0 dati malu ili veliku težinu. Preporuča se uzeti više različitih odabira za $(\mu_{a_0}, \sigma^2_{a_0})$, te provesti analizu sa više njih.

Razumno je uzeti neinformativnu apriornu distribuciju za $\pi_0(\beta|c_0)$, jer ta apriorna distribucija za β odgovara nekom drugom istraživanju i ne sadržava informacije o podacima D_0. Skalar $c_0 > 0$ je skalar varijance koji služi za kontrolu utjecaja $\pi_0(\beta|c_0)$ na $\pi(\beta, \delta, a_0|D_0)$. Da bi $\pi_0(\beta|c_0)$ bila neinformativna, uzimamo velike vrijednosti za skalar c_0 zato da je $\pi_0(\beta|c_0)$ relativno mala u odnosu na $L(\beta, \delta|D_0)^{a_0}$. Ako za c_0 uzmemo male vrijednosti, funkcija $\pi_0(\beta|c_0)$ će dominirati u formuli (3.18). Vrijednost skalara c_0 će ovisiti o strukturi podataka i veličini broja a_0. Kao i za izbor parametara $(\mu_{a_0}, \sigma^2_{a_0})$, i za izbor vrijednosti c_0 se preporuča uzeti više različitih vrijednosti, te promatrati kako koja od tih vrijednosti utječe na aposteriornu distribuciju.
Za $\pi_0(\delta|\theta_0)$, vrijednost θ_0 biramo tako da njena vrijednost bude relativno mala u odnosu na $L(\beta, \delta|D_0)^{a_0}$. Uzimamo f_{0i} da bude proporcionalna širini intervala, tj. $f_{0i} \propto s_i - s_{i-1}$ i $g_{0i} \to 0$ za $i = 1, 2, ..., M$. Uzimanjem malih vrijednosti za g_{0i} stvaramo neinformativnu gama apriornu distribuciju, a uzimanjem f_{0i} na iznad opisani način dopuštamo da varijanca od δ_i ovisi o širini i-tog intervala.

3.3.3 Aposteriorna distribucija

Zajednička aposteriorna gustoća za $(\beta, \delta, a_0|D)$ za trenutno istraživanje je dana sa

$$\begin{aligned}\pi(\beta, \delta, a_0|D) \propto &\left[\prod_{j=1}^{M}\left\{\exp\{-\delta_j(a_j + b_j)\}\prod_{k\in\mathcal{D}_j}\{1 - \exp(-\eta_k T_j)\}\right\}\right] \\ &\times L(\beta, \delta|D_0)^{a_0}\pi_0(\beta|c_0)\pi_0(\delta|\theta_0)\pi(a_0|\alpha_0, \lambda_0),\end{aligned} \tag{3.23}$$

gdje su η_k, a_i, b_i i T_j dani formulama (3.14), (3.15) i (3.16), i $L(\beta, \delta|D_0)$ je funkcija vjerodostojnosti zadana kao u (3.13) samo uz to da je D na svim mjestima zamijenjeno sa D_0. U formuli (3.23) i za $L(\beta, \delta|D)$ i za $L(\beta, \delta|D_0)$ koristimo istu podjelu varijable vremena na intervale $(s_{j-1}, s_j]$, zato da bi δ imala isto značenje u podacima starog i novog istraživanja. Točke s_i određujemo na sljedeći način:

- spajamo $\{y_i, i = 1, 2, ..., n\}$ i $\{y_{0i}, i = 1, 2, ..., n_0\}$ zajedno, koje onda označimo sa $\{y^*_i, i = 1, 2, ..., n + n_0\}$,
- intervale $(s_{j-1}, s_j]$ biramo tako da u njima ima jednak broj smrti i broj cenzuriranih opservacija za kombinirana vremena y^*_i.

Izabrani s_i je $\frac{i}{M}$-ti kvantil vremena y^*_i, gdje je M ukupni broj intervala. Interval biramo tako da u svakom od njih bude po barem jedna cenzurirana opservacija i jedna smrt iz y_{0i} i iz y_i.

Poglavlje 4

Generalizacija Coxovog modela

4.1 Uzimanje uzoraka iz aposteriorne distribucije Gibbsovom metodom uzorkovanja

Gibbsova metoda uzorkovanja pripada MCMC metodama, te je ujedno i jedna od najboljih metoda za uzimanje uzoraka. Dobila je ime po američkom znanstveniku Josiah Willard Gibbsu, no formalno je predstavljena od strane američkih matematičara Stuarta i Donalda Gemana 1984. godine. U osnovnoj verziji, Gibbsova metoda je poseban slučaj Metropolis-Hastingsovog algoritma, no u proširenom slučaju ona se može smatrati općim okvirom za uzorkovanje iz velikog skupa varijabli, uzorkovajući svaku varijablu ili u nekim slučajevima grupe varijabli. Gibbsova metoda uzorkovanja je primjenjiva kada zajednička distribucija nije poznata eksplicitno ili ako je teško uzorkovati iz nje direktno, no poznata je uvjetna distribucija svake varijable, kao što je slučaj u Bayesovoj statistici.
Neka je $\theta = (\theta_1, \theta_2, ..., \theta_p)'$ p-dimenzionalni vektor parametara i neka je $\pi(\theta|D)$ njegova aposteriorna funkcija gustoće uz dane podatke D.

Koraci osnovne sheme Gibbsovog algoritma:

1. Odaberi proizvoljan vektor $\theta_0 = (\theta_{1,0}, \theta_{2,0}, ..., \theta_{p,0})'$ i stavi $i = 0$.
2. Generiraj $\theta_{i+1} = (\theta_{1,i+1}, \theta_{2,i+1}, ..., \theta_{p,i+1})'$ na sljedeći način:
   - $\theta_{1,i+1} \sim \pi(\theta_1|\theta_{2,i}, ..., \theta_{p,i}, D)$,
   - $\theta_{2,i+1} \sim \pi(\theta_2|\theta_{1,i+1}, \theta_{3,i}..., \theta_{p,i}, D)$,
   - $\vdots$
   - $\theta_{p,i+1} \sim \pi(\theta_p|\theta_{1,i+1}, \theta_{2,i+1}..., \theta_{p-1,i+1}, D)$.

3. Stavi $i = i + 1$ i vrati se na korak 2.

Ovim načinom smo dobili uzorke koji dolaze iz aproksimacije zajedničke distribucije svih varijabli. Za "dovoljno" veliki i ta generirana distribucija je približno jednaka aposteriornoj distribuciji $\pi(\theta|D)$. Također, gledajući samo generirani skup $\theta_{k,i}$, za $k = 1, ..., p$, on aproksimira marginalnu distribuciju $\pi(\theta_k|\theta_1, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_p, D)$.

4.2 Generalizacija Coxovog modela

Ovaj model je generalizacija modela po dijelovima konstantnog hazarda. Radi jednostavnosti, pretpostavimo da postoji samo jedna kovarijata. Pretpostavimo također da je funkcija po dijelovima konstantnog hazarda dana sa $h(y|x) = \lambda_k \exp(x\beta_k)$ za $y \in I_k$, gdje je $I_k = (s_{k-1}, s_k]$ za $k = 1, 2, ..., J$, te da je x jednodimenzionalan. Neka su i

- $\lambda_k \sim \mathcal{G}(\eta_k, \gamma_k)$ za $k = 1, ..., J$ nezavisni,
- $\beta_{k+1}|\beta_0, \beta_1, ..., \beta_k \sim N(\beta_k, \omega_k^2)$ za $k = 0, ..., J - 1$ i
- β_k-ovi su nezavisni od $\lambda = (\lambda_1, \lambda_2, ..., \lambda_J)'$.

Pretpostavljamo da su hiperparametri η_k, γ_k, ω_k i β_0 poznati unaprijed. Ovaj model je diskretna verzija neparametarskog modela hazarda sa diskretnom verzijom gama procesa za funkciju osnovnog hazarda $h_0(\cdot)$, gdje je $\frac{\eta_k}{\gamma_k}$ matematičko očekivanje, a $\frac{\eta_k}{\gamma_k^2}$ varijanca od apriorne distribucije λ_k. Za "dovoljno" male intervale $I_k = (s_{k-1}, s_k]$ diskretna verzija će se neznatno razlikovati od stvarnog neprekidnog gama procesa. Diskretni autokorelirani apriorni proces za β_k dozvoljava da se kovarijate mijenjaju tijekom vremena. Za vrijednosti regresijskih koeficijenata β u susjednim intervalima I_{k-1} i I_k se očekuje da budu "blizu", te da se njihova međusobna nezavisnost smanjuje što su intervali više međusobno udaljeni. Hiperparametri ω_k se mogu koristiti kao pokazatelji za moguću promjenu regresijskih koeficijenata. Na primjer, za β_{k+1} možemo očekivati, sa 95% pouzdanošću, da bude u intervalu $[\beta_k - 1.96\omega_k\ , \beta_k + 1.96\omega_k]$.

Označimo sa $D_0 = \{(x_i, (a_{l_i}, a_{r_i}]), i = 1, 2, ..., n\}$ skup opservacija za n pojedinaca. Za vrijeme doživljenja y_i i-tog pojedinca znamo da je u intervalu $(a_{l_i}, a_{r_i}]$, te $a_{l_i} < a_{r_i}$ su dvije točke uzete iz mreže točaka $(a_1, ..., a_J)$, ne nužno uzastopne. Sa x_i označimo kovarijatu za i-tog pojedinca. Sa $D = \{(x_i, y_i), i = 1, 2, ..., n\}$ označimo potpune podatke, te neka je $\theta = (\beta', \lambda')'$, gdje je $\beta = (\beta_1, ..., \beta_n)$. Distribucija od $(y_i|x_i, \theta)$ je po dijelovima eksponencijalna pa je funkcija vjerodostojnosti za potpune podatke dana sa

$$L(\theta|D) \propto \prod_{k=1}^{J} \left[\lambda_k^{d_k} \exp\left\{ -\lambda_k \sum_{j \in \mathcal{R}_j} \exp(x_j\beta_k) \triangle_{jk} \right\} \exp\left(\sum_{j \in \mathcal{D}_k} x_j\beta_k \right) \right], \tag{4.1}$$

gdje je $\mathcal{R}_k$ skup pojedinaca koji su pod rizikom od smrti u trenutku a_{k-1}, $\triangle_{jk} = \min(y_j, a_k) - a_{k-1}$, $\mathcal{D}_k$ skup pojedinaca koji su umrli u intervalu $I_k = (a_{k-1}, a_k]$, a d_k je njihov broj. Za uzorkovanje iz aposteriorne distribucije od θ koristimo Gibbsovu metodu uzorkovanja:

$$(\lambda_k|\lambda^{(-k)}, \beta, D) \sim \mathcal{G}\left(\eta_k + d_k, \gamma_k + \sum_{j\in\mathcal{R}_k} \exp(x_j\beta_k)\triangle_{jk}\right), \tag{4.2}$$

$$\pi(\beta_k|\beta^{(-k)}, \lambda, D) \propto \phi(\beta_k|\mu_k, \sigma_k^2) \exp\left(-\lambda_k \sum_{j\in\mathcal{R}_k} \exp(x_j\beta_k)\triangle_{jk}\right), \tag{4.3}$$

gdje je $\lambda^{(-k)} = (\lambda_1, ..., \lambda_{k-1}, \lambda_{k+1}, ..., \lambda_J)$ i analogno $\beta^{(-k)} = (\beta_1, ..., \beta_{k-1}, \beta_{k+1}, ..., \beta_J)$. $\phi(\cdot|\mu_k, \sigma_k^2)$ je funkcija gustoće od $N(\mu_k, \sigma_k^2)$, kojoj su parametri

$$\mu_k = \frac{\left(\sum_{j\in\mathcal{D}_k} x_j\right)\omega_k^2\omega_{k-1}^2 + \beta_{k-1}\omega_k^2 + \beta_{k+1}\omega_{k-1}^2}{\omega_k^2 + \omega_{k-1}^2},$$

$$\sigma_k^2 = \frac{\omega_k^2\omega_{k-1}^2}{\omega_k^2 + \omega_{k-1}^2},$$

za $k = 1, 2, ..., J$ i $\beta_{J+1} = 0$.

U Coxovom modelu su svi β_k jednaki zajedničkom β.
Uvjetna distribucija od $[y_i|\theta, D_o]$ je po dijelovima eksponencijalna sa parametrima $\exp(x_i\beta_k)\lambda_k$, te je zadana sa

$$\begin{aligned} f(y_i|\theta, D_o) = &\left[1 - \exp\left\{-\sum_{l=l_i+1}^{r_i} \lambda_l \exp(x_i\beta_l)\tilde{\triangle}_l\right\}\right]^{-1} \lambda_k \exp(x_i\beta_k) \\ &\times \exp\left\{-\sum_{l=l_i+1}^{r_i} [\lambda_l \exp(x_i\beta_l)\tilde{\triangle}_l - \lambda_k \exp(x_i\beta_k)(y_i - a_{k-1})]\right\}, \end{aligned} \tag{4.4}$$

za $y_i \in I_k$, $l_i + 1 \leq k \leq r_i$ i $\tilde{\triangle}_l = a_l - a_{l-1}$. Gornja formula je produkt funkcija gustoće multinomijalne i eksponencijalne distribucije.

4.3 Primjer

Studija o raku dojke

Ova studija o raku dojke provedena je od strane Finkelsteina i Wolfea 1985. godine. Objekti promatranja su bile pacijentice koje su liječene od raka dojke u Joint Center for Radiation Therapy u Bostonu od 1976. do 1980. godine. Podaci se sastoje od dva skupa podataka. Prvi skup podataka se odnosi na 46 pacijentica kojima je rak dojke dijagnosticiran u ranoj fazi, te su liječene radioterapijom. Dok se drugi skup podataka odnosi na 48 pacijentica koje su liječene i radioterapijom i kemoterapijom. I jednim i drugim skupinama pacijentica su nadzirane promjene kroz preglede u bolnici svakih 4 do 6 mjeseci. Za pacijentice koje su geografski bile udaljenije od bolnice ti pregledi su vršeni i rjeđe. Liječnici su bilježili, na skali od 0 do 3 (0-nema promjene, 1-minimalna promjena, 2-umjerena promjena, 3-ozbiljna promjena), razliku u fizičkom izgledu pacijentica s obraćanjem pozornosti na oteklinu na dojci, proširenje krvnih žila, povlačenje dojke i općenito izgled pacijentica. Pošto su pacijentice propustile neke od svojih posjeta, podaci o promjenama su zapisani u obliku $(a, b]$, gdje je a broj mjeseca u kojem je pacijentica došla na pregled i do tada nije bilo nikakvih promjena, a b je broj mjeseca u kojem je došlo do promjene u tkivu. Pošto su se posjeti različitih pacijentica klinici dogodili u različitim vremenima, cenzurirani intervali u podacima se često preklapaju. Podaci o pacijenticama su dani u idućim tablicama.

Radioterapija			Kemoterapija i radioterapija		
(45,-]	(25,37]	(37,-]	(8,12]	(0,5]	(30,34]
(6,10]	(46,-]	(0,5]	(0,22]	(5,8]	(13,-]
(0,7]	(26,40]	(18,-]	(24,31]	(12,20]	(10,17]
(46,-]	(46,-]	(24,-]	(17,27]	(11,-]	(18,21]
(46,-]	(27,34]	(36,-]	(17,23]	(33,40]	(4,9]
(7,16]	(36,44]	(5,11]	(24,30]	(31,-]	(11,-]
(17,-]	(46,-]	(19,35]	(16,24]	(13,39]	(14,19]
(7,14]	(36,48]	(17,25]	(13,-]	(19,32]	(4,8]
(37,44]	(37,-]	(24,-]	(11,13]	(34,-]	(34,-]
(0,8]	(40,-]	(32,-]	(16,20]	(13,-]	(30,36]
(4,11]	(17,25]	(33,-]	(18,25]	(16,24]	(18,24]
(15,-]	(46,-]	(19,26]	(17,26]	(35,-]	(16,60]
(11,15]	(11,18]	(37,-]	(32,-]	(15,22]	(35,39]
(22,-]	(38,-]	(34,-]	(23,-]	(11,17]	(21,-]
(46,-]	(5,12]	(36,-]	(44,48]	(23,32]	(11,20]
(46,-]			(14,17]	(10,35]	(48,-]

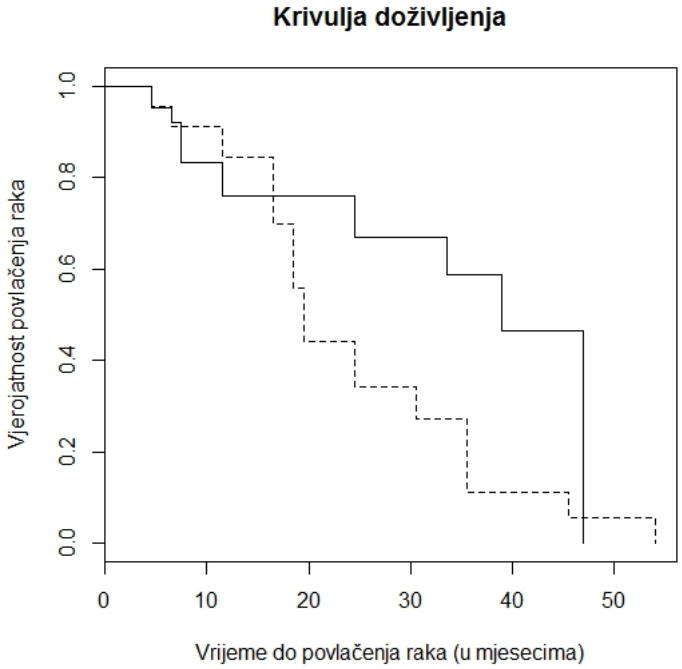
Kod na u ovom primjeru je napravljen u programskom jeziku R, pomoću paketa dynsurv [3]. Sami podaci ove studije su nalaze u R-u pod nazivom bcos, u sklopu tog paketa, te se sastoje od 3 varijable: left, right i trt. Left označava lijevu granicu intervala podataka $(a, b]$, right desnu, uz napomenu da ako je podatak desno cenzuriran, desna granica intervala je jednaka $+\infty$ (Inf). Treća varijabla je trt koja ima vrijednosti Rad ili RadChem, ovisno o tome da li je pacijentica liječena radioterapijom ili kemoterapijom i radioterapijom. Varijabla granice nam daje granice intervala $I_k = (s_{k-1}, s_k], k = 1, ..., 40$, uz napomenu da je $s_0 = 0$.

```
> podaci = bcos
> podaci
   left right     trt
1    45   Inf     Rad
2     6    10     Rad
3     0     7     Rad
4    46   Inf     Rad
5    46   Inf     Rad
6     7    16     Rad
7    17   Inf     Rad
8     7    14     Rad
9    37    44     Rad
10    0     8     Rad
11    4    11     Rad
12   15   Inf     Rad
13   11    15     Rad
14   22   Inf     Rad
15   46   Inf     Rad
16   46   Inf     Rad
17   25    37     Rad
18   46   Inf     Rad
19   26    40     Rad
20   46   Inf     Rad
21   27    34     Rad
22   36    44     Rad
23   46   Inf     Rad
24   36    48     Rad
25   37   Inf     Rad
26   40   Inf     Rad
27   17    25     Rad
28   46   Inf     Rad
29   11    18     Rad
30   38   Inf     Rad
31    5    12     Rad
32   37   Inf     Rad
33    0     5     Rad
34   18   Inf     Rad
35   24   Inf     Rad
36   36   Inf     Rad
37    5    11     Rad
38   19    35     Rad
39   17    25     Rad
40   24   Inf     Rad
41   32   Inf     Rad
42   33   Inf     Rad
43   19    26     Rad
44   37   Inf     Rad
45   34   Inf     Rad
46   36   Inf     Rad
47    8    12 RadChem
48    0    22 RadChem
49   24    31 RadChem
50   17    27 RadChem
51   17    23 RadChem
52   24    30 RadChem
53   16    24 RadChem
54   13   Inf RadChem
55   11    13 RadChem
56   16    20 RadChem
57   18    25 RadChem
58   17    26 RadChem
59   32   Inf RadChem
```

```
   60   23   Inf RadChem              78   10    35 RadChem
   61   44    48 RadChem              79   30    34 RadChem
   62   14    17 RadChem              80   13   Inf RadChem
   63    0     5 RadChem              81   10    17 RadChem
   64    5     8 RadChem              82    8    21 RadChem
   65   12    20 RadChem              83    4     9 RadChem
   66   11   Inf RadChem              84   11   Inf RadChem
   67   33    40 RadChem              85   14    19 RadChem
   68   31   Inf RadChem              86    4     8 RadChem
   69   13    39 RadChem              87   34   Inf RadChem
   70   19    32 RadChem              88   30    36 RadChem
   71   34   Inf RadChem              89   18    24 RadChem
   72   13   Inf RadChem              90   16    60 RadChem
   73   16    24 RadChem              91   35    39 RadChem
   74   35   Inf RadChem              92   21   Inf RadChem
   75   15    22 RadChem              93   11    20 RadChem
   76   11    17 RadChem              94   48   Inf RadChem
   77   22    32 RadChem
```

```
> granice = bcos.grid
> granice
 [1]  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
[23] 26 27 30 31 32 33 34 35 36 37 38 39 40 44 45 46 48 60

> form=survfit(Surv(left, right,type="interval2") ~ trt, data = podaci)
> summary(form)
Call: survfit(formula = Surv(left, right, type = "interval2") ~ trt,
    data = podaci)

                trt=Rad
 time n.risk  n.event survival std.err lower 95% CI upper 95% CI
  4.5   46.0 2.13e+00    0.954  0.0310        0.895        1.000
  6.5   43.9 1.53e+00    0.920  0.0399        0.845        1.000
  7.5   42.3 4.08e+00    0.832  0.0552        0.730        0.947
 11.5   38.3 3.25e+00    0.761  0.0629        0.647        0.895
 15.5   35.0 2.41e-15    0.761  0.0629        0.647        0.895
 17.5   35.0 6.53e-09    0.761  0.0629        0.647        0.895
 24.5   35.0 4.26e+00    0.668  0.0694        0.545        0.819
 25.5   30.7 3.32e-08    0.668  0.0694        0.545        0.819
 33.5   30.7 3.75e+00    0.587  0.0726        0.460        0.748
```

```
 34.5   27.0 9.09e-07    0.587  0.0726        0.460        0.748
 36.5   27.0 1.68e-02    0.586  0.0726        0.460        0.747
 39.0   27.0 5.55e+00    0.466  0.0735        0.342        0.635
 42.0   21.4 2.80e-03    0.466  0.0735        0.342        0.635
 47.0   21.4 2.14e+01    0.000     NaN           NA           NA

                trt=RadChem
 time n.risk  n.event survival std.err lower 95% CI upper 95% CI
  4.5  48.00 2.08e+00   0.9567  0.0294       0.9008        1.000
  6.5  45.92 2.08e+00   0.9134  0.0406       0.8373        0.997
 11.5  43.84 3.32e+00   0.8442  0.0523       0.7476        0.953
 12.5  40.52 1.30e-08   0.8442  0.0523       0.7476        0.953
 16.5  40.52 6.99e+00   0.6987  0.0662       0.5803        0.841
 18.5  33.54 6.75e+00   0.5580  0.0717       0.4338        0.718
 19.5  26.78 5.57e+00   0.4420  0.0717       0.3217        0.607
 21.5  21.22 8.87e-07   0.4420  0.0717       0.3217        0.607
 22.5  21.22 6.10e-09   0.4420  0.0717       0.3217        0.607
 23.5  21.22 9.79e-14   0.4420  0.0717       0.3217        0.607
 24.5  21.22 4.80e+00   0.3421  0.0685       0.2310        0.506
 30.5  16.42 3.39e+00   0.2714  0.0642       0.1707        0.431
 31.5  13.03 2.51e-03   0.2714  0.0642       0.1707        0.431
 33.5  13.03 1.39e-02   0.2711  0.0642       0.1705        0.431
 34.5  13.01 2.18e-11   0.2711  0.0642       0.1705        0.431
 35.5  13.01 7.71e+00   0.1105  0.0452       0.0495        0.247
 45.5   5.30 2.65e+00   0.0552  0.0330       0.0171        0.178
 54.0   2.65 2.65e+00   0.0000     NaN           NA           NA

> plot(form,main="",xlab="Vrijeme do povlačenja raka (u mjesecima)",
+ylab="Vjerojatnost povlačenja raka",lty=1:2)
> title(main="Krivulja doživljenja")
```



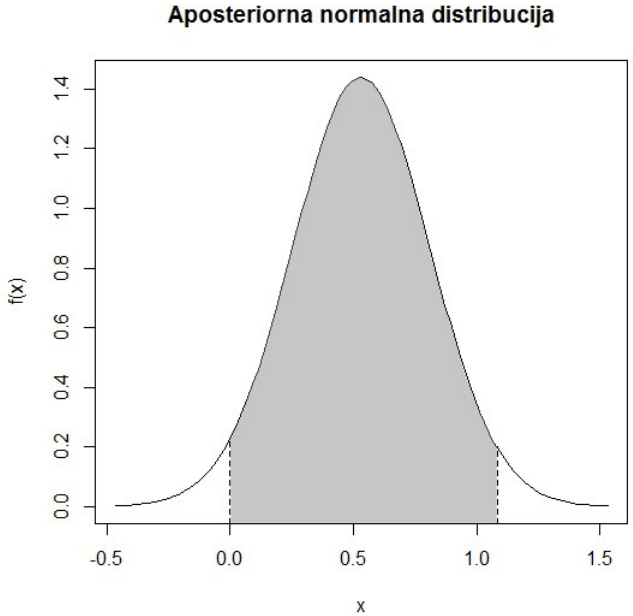
Na grafu su prikazane krivulje doživljenja za grupu pacijentica koje su liječene samo radioterapijom (puna linija) i za grupu pacijentica koje su liječene i radioterapijom i kemoterapijom (isprekidana linija).
Sada ćemo, koristeći 10.000 iteracija Gibbsove metode uzorkovanja, naći aposteriornu distribuciju parametra za dane podatke uz sljedeću apriornu distribuciju. Dakle, pretpostavimo da λ_k ima gama distribuciju sa očekivanjem $\frac{\eta_k}{\gamma_k}$=0.5 i varijancom $\frac{\eta_k}{\gamma_k^2}$=1.25. Iz toga slijedi da su hiperparametri gama distribucije $(0.2, 0.4)$, tj. $\lambda_k \sim \mathcal{G}(\eta_k, \gamma_k) = \mathcal{G}(0.2, 0.4)$, $k = 1, ..., J$. Za apriornu distribuciju za β_k uzmemo standardnu normalnu distribuciju, $\beta_k \sim N(0, 1)$, $k = 1, ..., J$. Koristit ćemo "TimeIndependent" model, što znači da ćemo pretpostavit da su $\mu_1 = \mu_2 = ... = \mu_J$, gdje je μ parametar aposetriorne distribucije $N(\mu, \sigma^2)$.

```
> formula = Surv(left, right, type="interval2") ~ trt
> set.seed(462)
> fit0 = bayesCox(formula, podaci, granice, out="Cox.txt",
+model="TimeIndep",
+base.prior=list(type="Gamma", shape=0.2, rate=0.4),
+coef.prior=list(type="Normal", mean=0, sd=1),
+gibbs=list(iter=10000, burn=20, thin=1, verbose=TRUE, nReport=5))
```

```
> coef(fit0)
             Low       Mid     High Time        Cov     Model
1  -0.004374406 0.5312973 1.083398    0 trtRadChem TimeIndep
2  -0.004374406 0.5312973 1.083398    4 trtRadChem TimeIndep
3  -0.004374406 0.5312973 1.083398    5 trtRadChem TimeIndep
4  -0.004374406 0.5312973 1.083398    6 trtRadChem TimeIndep
5  -0.004374406 0.5312973 1.083398    7 trtRadChem TimeIndep
6  -0.004374406 0.5312973 1.083398    8 trtRadChem TimeIndep
7  -0.004374406 0.5312973 1.083398    9 trtRadChem TimeIndep
8  -0.004374406 0.5312973 1.083398   10 trtRadChem TimeIndep
9  -0.004374406 0.5312973 1.083398   11 trtRadChem TimeIndep
10 -0.004374406 0.5312973 1.083398   12 trtRadChem TimeIndep
11 -0.004374406 0.5312973 1.083398   13 trtRadChem TimeIndep
12 -0.004374406 0.5312973 1.083398   14 trtRadChem TimeIndep
13 -0.004374406 0.5312973 1.083398   15 trtRadChem TimeIndep
14 -0.004374406 0.5312973 1.083398   16 trtRadChem TimeIndep
15 -0.004374406 0.5312973 1.083398   17 trtRadChem TimeIndep
16 -0.004374406 0.5312973 1.083398   18 trtRadChem TimeIndep
17 -0.004374406 0.5312973 1.083398   19 trtRadChem TimeIndep
18 -0.004374406 0.5312973 1.083398   20 trtRadChem TimeIndep
19 -0.004374406 0.5312973 1.083398   21 trtRadChem TimeIndep
20 -0.004374406 0.5312973 1.083398   22 trtRadChem TimeIndep
21 -0.004374406 0.5312973 1.083398   23 trtRadChem TimeIndep
22 -0.004374406 0.5312973 1.083398   24 trtRadChem TimeIndep
23 -0.004374406 0.5312973 1.083398   25 trtRadChem TimeIndep
24 -0.004374406 0.5312973 1.083398   26 trtRadChem TimeIndep
25 -0.004374406 0.5312973 1.083398   27 trtRadChem TimeIndep
26 -0.004374406 0.5312973 1.083398   30 trtRadChem TimeIndep
27 -0.004374406 0.5312973 1.083398   31 trtRadChem TimeIndep
28 -0.004374406 0.5312973 1.083398   32 trtRadChem TimeIndep
29 -0.004374406 0.5312973 1.083398   33 trtRadChem TimeIndep
30 -0.004374406 0.5312973 1.083398   34 trtRadChem TimeIndep
31 -0.004374406 0.5312973 1.083398   35 trtRadChem TimeIndep
32 -0.004374406 0.5312973 1.083398   36 trtRadChem TimeIndep
33 -0.004374406 0.5312973 1.083398   37 trtRadChem TimeIndep
34 -0.004374406 0.5312973 1.083398   38 trtRadChem TimeIndep
35 -0.004374406 0.5312973 1.083398   39 trtRadChem TimeIndep
36 -0.004374406 0.5312973 1.083398   40 trtRadChem TimeIndep
37 -0.004374406 0.5312973 1.083398   44 trtRadChem TimeIndep
```

```
38 -0.004374406 0.5312973 1.083398   45 trtRadChem TimeIndep
39 -0.004374406 0.5312973 1.083398   46 trtRadChem TimeIndep
40 -0.004374406 0.5312973 1.083398   48 trtRadChem TimeIndep
41 -0.004374406 0.5312973 1.083398   60 trtRadChem TimeIndep

> low=coef(fit0)[1,1]
> low
[1] -0.004374406
> mean=coef(fit0)[1,2]
> mean
[1] 0.5312973
> high=coef(fit0)[1,3]
> high
[1] 1.083398
```

Dobivene vrijednosti Low i High su $2,5\%$ i $97,5\%$ kvantili očekivanja aposteriorne distribucije, tj. to je 95% pouzdani interval za parametar matematičkog očekivanja μ (Mid). Dakle, za aposteriornu distribuciju smo dobili parametar $\mu = \mu_1 = ... = \mu_J$, dok varijancu možemo dobiti iz 95% pouzdanog intervala.
Aposteriorna distribucija : $\pi(\beta|D) \sim N(\ 0.5312973\ ,\ 0.277493^2\ )$.
95% pouzdani interval za matematičko očekivanje μ : $[\ -0.004374406\ ,\ 1.083398\ ]$.



Napravimo sada i "TimeVarying" model, u kojem se pretpostavlja da parametri očekivanja aposteriorne distribucije $\mu_1, \mu_2 ..., \mu_J$ ne moraju biti jednaki u svakom intervalu I_k. Apriorna distribucija za λ_k, $k = 1, ..., J$, je kao i u "TimeIndependent" modelu $\mathcal{G}(0.2, 0.4)$, dok je apriorna distribucija za β_k zadana autoregresivnim procesom reda 1 ($AR(1)$). Taj proces je dan formulom

$$\beta_k = \phi\beta_{k-1} + \sigma Z_k, k = 1, ..., J,$$

gdje su Z_k nezavisne jednako distribuirane slučajne varijable sa standardnom normalnom distribucijom ($N(0, 1)$),te $\sigma \in \mathbb{R}$, $\sigma > 0$ i $\phi \in \mathbb{R}$. Označimo sa μ_k i ω_k parametre normalne aposteriorne distribucije za β_k. Tada je

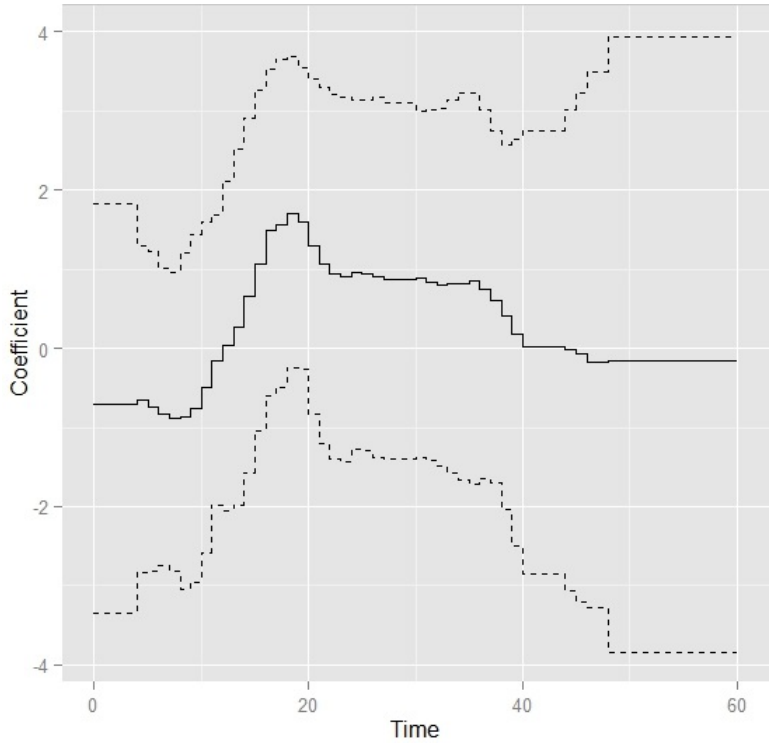
$$\beta_k \sim N(\mu_k, \omega_k^2) = N(\phi\mu_{k-1}, \phi^2\omega_{k-1}^2 + \sigma^2),$$

uz pretpostavku $\beta_0 \equiv \mu_0 \sim N(\mu_0, 0)$.

```
> fit1 = bayesCox(formula, podaci, granice, out="tvCox.txt",
+model="TimeVarying",
+base.prior=list(type="Gamma", shape=0.2, rate=0.4),
+coef.prior=list(type="AR1", sd=1),
+gibbs=list(iter=50000, burn=20, thin=1, verbose=TRUE, nReport=5))
> coef(fit1)
          Low           Mid      High Time        Cov       Model
1  -3.3474890 -0.714534718 1.8213930    0 trtRadChem TimeVarying
2  -3.3474890 -0.714534718 1.8213930    4 trtRadChem TimeVarying
3  -2.8383005 -0.660349807 1.3000095    5 trtRadChem TimeVarying
4  -2.8243435 -0.750350281 1.2251928    6 trtRadChem TimeVarying
5  -2.7441552 -0.831900159 1.0128875    7 trtRadChem TimeVarying
6  -2.8241955 -0.883885012 0.9517257    8 trtRadChem TimeVarying
7  -3.0551315 -0.874128429 1.2074008    9 trtRadChem TimeVarying
8  -2.9601985 -0.755941768 1.4433980   10 trtRadChem TimeVarying
9  -2.5851860 -0.496467102 1.6061093   11 trtRadChem TimeVarying
10 -1.9769650 -0.151980223 1.6825488   12 trtRadChem TimeVarying
11 -2.0453807  0.030929882 2.1053125   13 trtRadChem TimeVarying
12 -1.9756352  0.269036860 2.5110855   14 trtRadChem TimeVarying
13 -1.5831430  0.651030669 2.9098058   15 trtRadChem TimeVarying
14 -1.0415815  1.065208259 3.2641430   16 trtRadChem TimeVarying
15 -0.5939706  1.487793702 3.5377513   17 trtRadChem TimeVarying
16 -0.4969564  1.565937273 3.6589830   18 trtRadChem TimeVarying
17 -0.2473885  1.712782190 3.6884038   19 trtRadChem TimeVarying
18 -0.2657171  1.595222454 3.5459558   20 trtRadChem TimeVarying
19 -0.8253821  1.288589213 3.4133370   21 trtRadChem TimeVarying
```

```
20 -1.1962067  1.066153167 3.3074272   22 trtRadChem TimeVarying
21 -1.3976310  0.942944712 3.2133065   23 trtRadChem TimeVarying
22 -1.4274407  0.908491829 3.1731670   24 trtRadChem TimeVarying
23 -1.2728505  0.964718976 3.1310180   25 trtRadChem TimeVarying
24 -1.2983805  0.950274173 3.1367303   26 trtRadChem TimeVarying
25 -1.3785193  0.912596617 3.1684568   27 trtRadChem TimeVarying
26 -1.3996992  0.876793064 3.0975610   30 trtRadChem TimeVarying
27 -1.3747902  0.890193359 2.9902152   31 trtRadChem TimeVarying
28 -1.4209127  0.839509112 3.0088705   32 trtRadChem TimeVarying
29 -1.4895005  0.795526494 3.0367165   33 trtRadChem TimeVarying
30 -1.5834535  0.812151129 3.1349465   34 trtRadChem TimeVarying
31 -1.6712665  0.820798478 3.2288205   35 trtRadChem TimeVarying
32 -1.7180015  0.857695997 3.2200678   36 trtRadChem TimeVarying
33 -1.6517192  0.745372256 3.0201040   37 trtRadChem TimeVarying
34 -1.7045440  0.599953592 2.7510050   38 trtRadChem TimeVarying
35 -2.0441870  0.410935491 2.5802110   39 trtRadChem TimeVarying
36 -2.5005085  0.172779233 2.6388813   40 trtRadChem TimeVarying
37 -2.8464227  0.025858764 2.7534573   44 trtRadChem TimeVarying
38 -3.0697905 -0.009856522 3.0104223   45 trtRadChem TimeVarying
39 -3.2005005 -0.075181724 3.2293618   46 trtRadChem TimeVarying
40 -3.2708080 -0.168129000 3.4860170   48 trtRadChem TimeVarying
41 -3.8359527 -0.156522465 3.9370753   60 trtRadChem TimeVarying
```

Iz tablice možemo očitati vrijednosti Mid, Low i High koji predstavljaju matematičko očekivanje, te $2,5\%$ i $97,5\%$ kvantile parametra očekivanja aposteriorne distribucije, koji se mijenjaju kroz vrijeme ($k = 1, 2, ..., J$) isto kao i očekivanje Mid. Iduća slika prikazuje 95% pouzdane intervale za μ_k, $k = 1, ..., J$. Puna crta predstavlja parametar očekivanja kroz vrijeme, a isprekidane linije su pouzdani interval.


4
2
0
-2
-4
Coefficient
0
20
40
60
Time

Bibliografija

[1] Casella George, George Edward I. : *Explaining the Gibbs Sampler*, The American Statistician, **46** ,167–174 (1992)

[2] Chen Ming-Hui, Ibrahim Joseph G., Sinha Debajyoti : *Bayesian Survival Analysis*, Springer (2001)

[3] Chen Ming-Hui, Yan Jun, Wang Xiaojing : *Package 'dynsurv'*, *Dynamic models for survival data* (2014)
http://cran.r-project.org/web/packages/dynsurv/index.html

[4] Christian P. Robert : *The Bayesian Choice: A Decision-Theoretic Motivation*, Springer New York (1994)

[5] Finkelstein Dianne M., Wolfe Robert A.: *A Semiparametric Model for Regression Analysis of Interval-Censored FailureTime Data*, Biometrics, **41** , 942–944 (1985)

[6] Sarapa Nikola : *Teorija vjerojatnosti*, Školska knjiga (2002)

[7] Zhang Daowen : predavanja sa NCSU (poglavlje 6)
http://www4.stat.ncsu.edu/~dzhang2/st745/index.html

[8] Bayesov pristup: Prednosti i nedostaci
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_introbayes_sect006.htm

Sažetak

Modele doživljenja je općenito dosta teško za modelirati. Bayesovski pristup analizi doživljenja uvelike olakšava modeliranje problema, pogotovo u biomedicini u kojoj i nalazimo najveću primjenu poluparametarskih modela kojima smo se bavili u ovom radu. Za svaki od modela u radu iznijeli smo načine dolaska do apriorne distribucije, izveli smo funkciju vjerodostojnosti i aposteriornu distribuciju.
U bayesovskom pristupu analizi doživljenja veliku ulogu ima i Gibbsova metoda uzorkovanja, koja je jedna od MCMC metoda. Ona nam omogućava dobivanje aposteriorne distribucije za model.
Velika prednost bayesovskog pristupa je da nam on omogućava da iz starijih istraživanja izvučemo neke pretpostavke za model novog istraživanja, ukoliko su ona slična. Klinička istraživanja su najbolji primjer toga, primjerice studije o različitim bolestima kao što su rak, AIDS i slične. Naravno, postoje i nedostaci ovog pristupa. Jedan od glavnih je to što se bazira na subjektivnosti statističara. Odabir krive apriorne distribucije može rezultirati krivim rezultatima.
Bayesovski pristup u analizi doživljenja je jako korisno oružje, pogotovo od kada su razvijeni jaki statistički softveri koji olakšavaju implementaciju tih modela, no ono se mora znati i pravilno koristiti.

Summary

It is well known that survival models are generally quite hard to fit. Bayesian paradigm in survival analysis greatly eases semiparametric models, especially in biomedicine where the semiparametic models, that are evaluated in this paper, are mostly used. For each of the models in this paper, we presented ways to reach a priori, we performed the likelihood function and the posterior distribution.
Gibbs sampler, which is one of the MCMC sampling algorithms, plays an important role in the Bayesian paradigm in survival analysis. It allows us to obtain the posterior distributions for the model.
The great advantage of Bayesian paradigm is that it allows us to draw some assumptions from earlier research and, if they are similar, use them for the model of the new research. This can be seen in clinical studies where, for example, researchers study variety of diseases such as cancer, AIDS and similar. However, there are some disadvantages to this approach. The greates flaw is that it is based on the statisticans subjectivity. If the wrong prior distribution is chosen, wrong results can occur.
Bayesian paradigm in survival analysis can be a powerfull tool, especially when there is a good softwear available that eases the implementation of this models. However, it is required to know how to use it properly.

Životopis

Rođena sam 1. veljače 1989. godine u Karlovcu. Osnovnu školu sam završila 2003. godine. Uz tu školu sam paralelno završila i osnovnu glazbenu školu 2002. godine. Nakon toga upisujem se u Gimnaziju Karlovac koju završavam 2007. godine. Odmah nakon završetka srednje škole upisujem preddiplomski sveučilišni studij matematike na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Godine 2012. upisujem diplomski sveučilišni studij Matematička statistika.