UNIVERZITET U BEOGRADU
MATEMATIČKI FAKULTET

Lidija Tomanić

KOKSOV MODEL PREŽIVLJAVANJA

master rad

Beograd, 2023.

Mentor:

dr Bojana Milošević, vanredni profesor
Univerzitet u Beogradu, Matematički fakultet

Članovi komisije:

dr Marko Obradović, docent
Univerzitet u Beogradu, Matematički fakultet

dr Marija Cuparić, docent
Univerzitet u Beogradu, Matematički fakultet

Datum odbrane: _____

Naslov master rada: Koksov model preživljavanja

Rezime: Ovaj master rad istražuje Koksov model preživljavanja, koji je jedan od ključnih alata u analizi podataka za proučavanje vremena do nekog određenog događaja.

U prvom delu rada, opisani su osnovni koncepti analize preživljavanja, uključujući definicije ključnih pojmova kao što su funkcija preživljavanja, rizik od događaja i cenzurisanje podataka. Nakon toga, detaljno je objašnjen koncept Kaplan-Majerove krive, koja je ocena funkcije raspodele slučajne veličine od interesa u prisustvu cenzurisanih podataka i omogućava poređenje različitih grupa.

Glavni deo rada posvećen je Koksovom modelu preživljavanja, koji je statistički model koji pruža mogućnost istraživanja uticaja različitih faktora na vreme do događaja. Opisan je matematički okvir modela, uključujući Koksovu proporcionalnu stopu rizika, ocenu parametara i interpretaciju rezultata. Prikazane su i metode za validaciju modela i testiranje pretpostavki.

Nakon teorijskog opisa, rad se fokusira na primenu Koksovog modela na realnim podacima. Korišćenjem softverskog programa R, prikazana je implementacija modela na stvarnim podacima, sa ciljem da se identifikuju značajni faktori koji utiču na vreme do događaja.

Kroz teorijski opis i praktičnu implementaciju u softverskom programu R, rad pruža uvid u statističku analizu vremena do događaja i nudi korisne smernice za primenu ovih metoda u budućim istraživanjima.

Za kraj bih želela da se zahvalim svom mentoru, dr Bojani Milošević, zato što me je zainteresovala za ovu temu, kao i za mnogobrojne sugestije koje su značajno doprinele kvalitetu rada.

Ključne reči: analiza preživljavanja, vreme do događaja, rizik od događaja, Kaplan-Majerova kriva

Sadržaj

Glava 1

Analiza preživljavanja

Ovo poglavlje predstavlja uvod u analizu preživljavanja i obrađuje sledeće teme: problem koji se rešava analizom preživljavanja, ciljeve te analize, ključne pojmove i terminologiju koja se koristi u njoj.

Analiza preživljavanja predstavlja skup statističkih procedura za koje je promenljiva od interesa upravo vreme dok se događaj ne pojavi. Pod vremenom podrazumevamo godine, mesece, nedelje, dane, sate itd. koji prođu od početka posmatranja nekog subjekta, pa do momenta pojavljivanja događaja. Kada kažemo događaj, mislimo na neki specifičan događaj koji može zadesiti pojedinca, kao što su smrt, pojava bolesti, povratak bolesti nakon razdoblja remisije, oporavak (kao što je povratak na posao) ili bilo koja druga značajna pojava. Pretpostavljamo da posmatramo jedan događaj od interesa, ali može biti i više od jednog, kao što je smrt od različitih uzroka ili pojavljivanje različitih bolesti. U slučaju posmatranja više događaja od interesa problem nazivamo problem višestrukog rizika. U analizi preživljavanja, obično se vremenska veličina naziva vremenom preživljavanja jer ukazuje na vreme tokom kog je pojedinac „preživeo" u određenom periodu praćenja. Događaj se naziva neuspehom, jer se interesuje za događaj koji obuhvata smrt, pojavu bolesti ili neko drugo negativno iskustvo pojedinca. Međutim, vreme preživljavanja može biti i vreme povratka na posao nakon hirurške intervencije, gde neuspeh predstavlja pozitivan događaj.

Označimo sa T slučajnu veličinu koja predstavlja vreme preživljavanja osobe. Budući da T označava vreme, njene moguće vrednosti su nenegativni brojevi. Sa t ćemo označiti realizovanu vrednost slučajne veličine T. Važi

$$P(T \leq t) = \int_0^t f(s)\, ds, \qquad 0 \leq t < \infty.$$

Dalje, neka je D slučajna veličina koja predstavlja indikator da li se događaj od interesa zaista desio. Tačnije, $D = 1$ ako se događaj desio tokom perioda studije, ili $D = 0$ ako je vreme preživljavanja cenzurisano pre kraja perioda studije. Važno je napomenuti da ako osoba ne doživi događaj (neuspeh) tokom trajanja studije, cenzura postaje jedina preostala mogućnost za praćenje preživljavanja te osobe. To znači da je $D = 0$ samo ako se jedna od sledećih situacija dogodi: osoba preživi do kraja studije, osoba bude izgubljena tokom praćenja ili se povuče tokom trajanja studije.

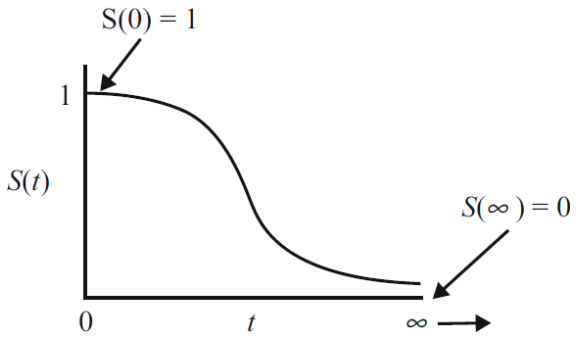
Funkcija preživljavanja $S(t)$ predstavlja verovatnoću da slučajna veličina prekorači specifično vreme t, tj.

$$S(t) = P(T > t).$$

Funkcija preživljavanja je grafički prikazana na Slici 1.1, a preuzeta je iz [8].

Osobine funkcije preživljavanja:

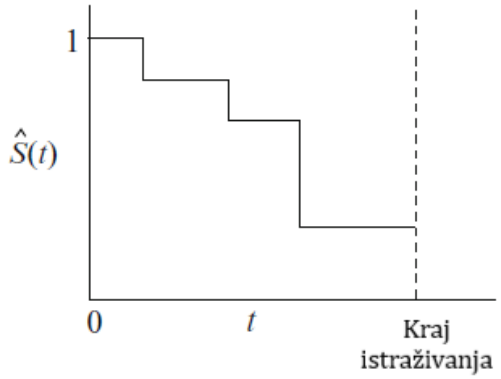
- nerastuća funkcija;
- na početku studije, tj. za $t = 0$, važi $S(0) = 1$, što znači da, pošto se još nije desio događaj, verovatnoća preživljavanja nakon vremena 0 iznosi 1;
- za $t = \infty jeS(\infty) = 0$, što znači da, ako bi period studije bio neograničen, na kraju niko ne bi preživeo, tako da kriva preživljavanja mora da opadne do 0.



Slika 1.1: Funkcija preživljavanja

U praksi, kada radimo s konkretnim podacima, ocena funkcije preživljavanja je Kaplan-Majerova kriva, o čemu će više reči biti u narednom poglavlju. Štaviše, budući da period studije nikada nije beskonačan pošto mogu postojati konkurentni rizici za neuspeh, postoji mogućnost da se određeni događaj neće desiti. Ocenjena

funkcija preživljavanja, označena kao $\hat{S}$ na grafikonu, možda neće doseći nulu na kraju studije, što je prikazano na Slici 1.2.



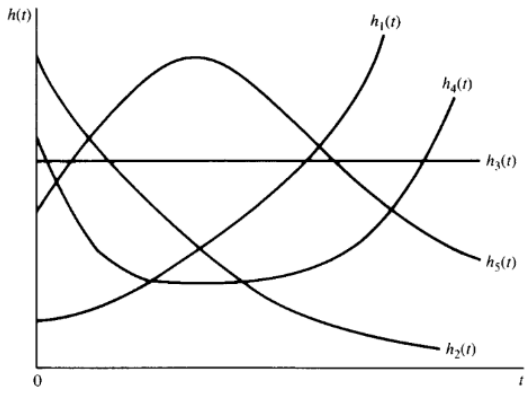
Slika 1.2: Ocenjena kriva preživljavanja

Često se u analizi preživljavanja koristi i stopa rizika (poznata i kao stopa hazarda), označena sa $h(t)$, koja se definiše na sledeći način:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t}.$$

Stopa rizika $h(t)$ pruža trenutni potencijal za pojavu događaja po jedinici vremena, pod uslovom da se događaj nije desio pre vremena t. Ponekad se stopa rizika naziva i uslovna stopa preživljavanja. To je zato što je brojilac u formuli za stopu rizika uslovna verovatnoća da vreme preživljavanja bude između t i Δt, pod uslovom da je vreme preživljavanja T veće ili jednako od t. Stopa rizika se deli sa Δt, koja predstavlja malu vremensku jedinicu. Vrednosti stope rizika su u opsegu između 0 i ∞. Bitno je naglasiti da se, za razliku od funkcije preživljavanja koja se fokusira na nepojavljivanje događaja, stopa rizika fokusira na pojavu događaja. Na neki način, stopa rizika može se smatrati suprotnom stranom informacija koje pruža funkcija preživljavanja. Na Slici 1.3 su prikazani različiti grafici stope rizika, a slika je preuzeta iz [13]. Karakteristike stope rizika:

- nenegativna;
- nema gornju granicu.



Slika 1.3: Različiti grafici stope rizika

Funkcija rizika (poznata i kao funkcija hazarda ili hazardna funkcija) se definiše kao odgovarajući integral:

$$H(t) = \int_0^t h(s)\, ds, \qquad t > 0.$$

Postoji jasna veza između funkcije preživljavanja i stope rizika, tj. ako je poznata $S(t)$ lako možemo da odredimo $h(t)$ i obrnuto. Prema definiciji, funkcija preživljavanja $S(t)$ daje verovatnoću da osoba preživi duže od vremena t. To znači da je verovatnoća preživljavanja u intervalu $[t, t+dt]$ jednaka verovatnoći da osoba preživi duže od vremena t, a manje od vremena $t+dt$, što možemo uz korišćenje uslovne verovatnoće zapisati kao:

$$P(t \leq T < t + \Delta t \,|\, T \geq t) = \frac{S(t) - S(t+dt)}{S(t)}.$$

Sada, prema definiciji, stopa rizika $h(t)$ daje trenutnu stopu događaja smrti na vremenskoj osi, pod uslovom da je osoba preživela do vremena t. U matematičkim terminima, možemo je definisati kao:

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \,|\, T \geq t)}{dt}.$$

Dalje je

$$h(t) = \lim_{dt \to 0} \frac{S(t) - S(t+dt)}{dt} \cdot \frac{1}{S(t)}.$$

Koristeći osobine limesa i funkcije logaritma, dobijamo

$$h\,(t) = -\frac{d}{dt} \log S\,(t)$$

tj.

$$\int_0^t h\left(x\right) dx \;=\; -\ln S\left(t\right).$$

Konačno je

$$S\left(t\right) = e^{-\int_0^t h(x)dx} = e^{-H(t)}.$$

Nakon što smo razmotrili stopu rizika i njen odnos sa funkcijom preživljavanja, definišemo osnovne ciljeve analize preživljavanja:

- određivanje i tumačenje funkcije preživljavanja i rizika;
- poređenje funkcije preživljavanja i rizika;
- procena odnosa između objašnjavajućih promenljivih i vremena preživljavanja.

Kod većine analiza preživljavanja često se susrećemo sa cenzuriranim podacima, što otežava primenu standardnih statističkih metoda. Cenzurisanje se pojavljuje kada imamo delimičnu informaciju o pojavi događaja ali ne znamo tačno vreme pojavljivanja. Uopšteno, postoje tri razloga zbog kojih se pojavljuje cenzurisanje:

- događaj od interesa se nije pojavio pre kraja perioda posmatranja;
- osoba koja se prati može biti izgubljena tokom perioda posmatranja;
- osoba može napustiti studiju zbog smrti (ako smrt nije povezana sa događajem od interesa) ili nekog drugog konkurentnog rizika kao što je nepovoljna reakcija na lek.

Sve ove situacije dovode do toga da ne možemo tačno utvrditi vreme preživljavanja osobe u studiji. Ukoliko ne znamo tačno vreme preživljavanja osobe, što se događa kada se studija završi ili kada osoba ne može da bude praćena ili se povuče iz studije, govorimo o desno cenzurisanim podacima. Naravno, postoje i levo cenzurisani podaci, koji se javljaju kada je vreme preživljavanja pacijenata nekompletno sa leve strane perioda posmatranja. Na primer, posmatramo grupu pacijenata zaraženih HIV virusom. Proces posmatranja stanja pacijenta može da počne od momenta kada je osoba postala pozitivna na HIV virus, ali mi uglavnom ne znamo tačno vreme koje je proteklo od momenta prvog izlaganja riziku zaraze pa do momenta kada je ustanovljeno postojanje HIV virusa u organizmu. Prema tome, vreme preživljavanja je cenzurisano sa leve strane. Ovakva vrsta cenzurisanja podataka se ređe sreće u praksi. Podaci analize preživljavanja takođe mogu biti intervalno cenzurirani, što se može dogoditi ako je stvarno vreme preživljavanja subjekta unutar određenog vremenskog intervala. Primetimo da intervalno cenzurisanje zapravo obuhvata i desno i

levo cenzurisanje kao specijalne slučajeve. Označimo sa T_1 početno vreme preživljavanja, odnosno vreme kada je subjekt uveden u istraživanje ili rizik počeo. Sa druge strane, T_2 predstavlja vreme događaja preživljavanja, tj. vreme kada je događaj poput smrti ili nekog drugog ishoda nastupio. Levo cenzurisani podaci se javljaju kada je vrednost T_1 jednaka 0, a T_2 poznata gornja granica vremena preživljavanja. Nasuprot tome, desno cenzurisani podaci nastaju kada je vrednost T_2 beskonačno, a T_1 poznata donja granica vremena preživljavanja. Kao što je pomenuto, cenzurisanje je jedna od specifičnih karakteristika podataka o vremenu neuspeha. Primetimo da cenzurisani podaci predstavljaju specifičnu vrstu nedostajućih podataka, budući da cenzurisana posmatranja i dalje pružaju neke parcijalne informacije, dok nedostajuća posmatranja ne pružaju nikakve informacije o promenljivoj od interesa. Pod desno cenzuriranim podacima podrazumevamo da je vreme neuspeha od interesa posmatrano tačno ili je veće od vremena cenzurisanja. Očigledno je da je važno razumeti način na koji se dešava desno cenzuriranje kako bismo pravilno analizirali podatke o vremenu neuspeha koji su desno cenzurirani. Veoma važno pitanje prilikom analize preživljavanja uz prisustvo cenzuriranja jeste da li je mehanizam cenzuriranja nezavisan ili ne. Pod tim se podrazumeva da stopa neuspeha ili rizika (opasnost, hazard) ostaje ista za subjekte koji još uvek učestvuju u studiji i za subjekte koji su cenzurisani. Ovo se može formalno izraziti formulom:

$$\lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t, Y(t) = 1)}{\Delta t},$$

gde je T slučajna veličina koja predstavlja vreme preživljavanja osobe, a $Y(t) = 1$ znači da subjekt nije doživeo neuspeh niti je bio cenzurisan pre vremena t. Prethodni izraz precizno definiše uslov za nezavisnost mehanizma cenzuriranja u kontekstu analize preživljavanja. U okviru modela slučajne cenzure, gore navedeni uslov je ekvivalentan

$$\lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t, C(t) \geq t)}{\Delta t},$$

gde C predstavlja cenzurisanu promenljivu.

U narednom poglavlju opisano je kako oceniti i prikazati krive preživljavanja koristeći Kaplan-Majerovu metodu, koju ćemo radi lakše notacije nadalje obeležavati sa KM. Kaplan-Majerova metoda je posebno korisna za analizu cenzuriranih podataka o vremenu neuspeha jer uzima u obzir cenzurisana posmatranja i omogućava procenu verovatnoće preživljavanja u različitim vremenskim tačkama. Takođe, opisano je kako testirati da li dve ili više krivih preživljavanja procenjuju istu krivu. Najpopularniji test za takvu svrhu naziva se test logaritmovanih rangova.

Glava 2

Kaplan-Majerova kriva preživljavanja i test logaritmovanih rangova

2.1 Kaplan-Majerova kriva preživljavanja

Neparametarska ocena funkcije preživljavanja za vreme neuspeha $t_{(f)}$ je data sa

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \cdot P(T > t_{(f)} | T \geq t_{(f)}). \tag{2.1}$$

Izraz se oslanja na koncept uslovne verovatnoće. Konkretno, $P(T > t_{(f)}|T \geq t_{(f)})$ označava uslovnu verovatnoću da će događaj neuspeha nastupiti nakon trenutka $t_{(f)}$, pri čemu se uzima u obzir da je preživljavanje veće ili jednako $t_{(f)}$. Ova formula ima za svrhu iterativno izračunavanje ocene funkcije preživljavanja za svaki vremenski interval $t_{(f)}$ koristeći prethodno izračunate vrednosti. Drugim rečima, izraz predstavlja verovatnoću preživljavanja do prethodnog trenutka neuspeha $t_{(f-1)}$, pomnoženu sa uslovnom verovatnoćom preživljavanja do vremena $t_{(f)}$, pod uslovom da je preživljavanje do tada najmanje $t_{(f)}$.

Alternativno, gore navedeni izraz može biti prikazan i kao granična vrednost proizvoda ukoliko umesto $\hat{S}(t_{(f-1)})$ stavimo proizvod razlomaka koji ocenjuju uslovne verovatnoće u momentu $t_{(f-1)}$ i ranije

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} P(T > t_{(i)} | T \geq t_{(i)}).$$

KM formula se dokazuje korišćenjem formule za uslovnu verovatnoću. Neka je događaj A definisan da osoba ostaje živa najmanje do trenutka $t_{(f)}$, a događaj B da

osoba ostaje živa nakon trenutka $t_{(f)}$, tj. $A = \{T \geq t_{(f)}\}$, a $B = \{T > t_{(f)}\}$, važi $A \cap B = B$, tj. $P(A \cap B) = P(B) = S\left(t_{(f)}\right)$. Kako je $t_{(f)}$ sledeće vreme neuspeha posle $t_{(f-1)}$ znači da nema neuspeha između $t_{(f)}$ i $t_{(f-1)}$ pa je $P(A) = P\left(T > t_{(f-1)}\right) = S\left(t_{(f-1)}\right)$. Dalje imamo $P(B|A) = P\left(T > t_{(f)} \middle| T \geq t_{(f)}\right)$. Koristeći formulu uslovne verovatnoće $P(A \cap B) = P(A) \cdot P(B|A)$ dobijamo $S\left(t_{(f)}\right) = S\left(t_{(f-1)}\right) \cdot P\left(T > t_{(f)} \middle| T \geq t_{(f)}\right)$.
Neka u trenutku $t_{(j)}$ postoji n_j subjekata u skupu rizika i njih d_j je ostvarilo događaj, tada je ocena verovatnoće neuspeha d_j/n_j, a ocena verovatnoće preživljavanja je $1 - d_j/n_j$, odnosno

$$P(T > t_{(j)}|T \geq t_{(j)}) = \frac{n_j - d_j}{n_j}.$$

KM ocena funkcije preživljavanja u trenutku t je data sa

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \frac{n_i - d_i}{n_i}$$

i $\hat{S}(t) = 1$ za $t < t_{(1)}$.

2.2 Disperzija KM krive

Preciznost KM ocene (2.1) se ocenjuje određivanjem standardne greške te ocene, koja je kvadratni koren njene disperzije. Dakle, problem se svodi na određivanje disperzije KM krive. Disperzija KM krive je određena sa

$$D(\hat{S}(t)) = D(\prod_{i:t_{(i)} \leq t} \frac{n_i - d_i}{n_i}) = D(\prod_{i:t_{(i)} \leq t} \hat{p}_i),$$

gde je $\hat{p}_i = \frac{n_i - d_i}{n_i}$.
Da bi se olakšao račun [1], posmatra se logaritam ocene funkcije preživljavanja:

$$D(\ln \hat{S}(t)) = D(\sum_{i:t_{(i)} \leq t} \ln \hat{p}_i) = \sum_{i:t_{(i)} \leq t} D(\ln \hat{p}_i),$$

pri čemu se pretpostavlja da su događaji unutar populacije nezavisni.
Dakle, problem se svodi na određivanje $D(\ln \hat{p}_i)$, a za rešavanje tog problema koristi se delta metoda, koju u narednom odeljku detaljno objašnjavamo.

[1] Disperzija sume nezavisnih slučajnih veličina je jednaka sumi njihovih pojedinačnih disperzija.

Delta metoda

Delta metoda se oslanja na Tejlorov razvoj prvog reda [2] funkcije f slučajne promenljive X u okolini $\mu = E(X)$, pri čemu treba da važi $f'(\mu) \neq 0$. Ovom metodom se aproksimativno određuje disperzija $f(X)$ izrazom

$$f(X) \approx f(\mu) + f'(\mu)(X - \mu),$$

gde je

$$f'(\mu) = \frac{\partial f(X)}{\partial X}|_{X=\mu}.$$

Dalje je

$$D(f(X)) \approx D(f(\mu) + f'(\mu)(X - \mu)) = (f'(\mu))^2 D(X - \mu) = (f'(\mu))^2 D(X).$$

Ocena disperzije delta metodom je

$$\hat{D}(f(X)) = f'(\hat{\mu}))^2 \hat{\sigma}^2,$$

gde je $\hat{\sigma}^2$ ocena disperzije $D(X)$ i $\hat{\mu}$ ocena očekivanja $E(X)$.

Vratimo se na traženje disperzije KM ocene funkcije preživljavanja, tj. $D(\ln \hat{p}_i)$. Da bi se dobila disperzija ocene, pretpostavlja se da su opservacije u skupu rizika u trenutku $t_{(i)}$ nezavisne opservacije iz Bernulijeve raspodele sa konstantnom verovatnoćom p_i. Pod ovom pretpostavkom, ocena ove verovatnoće je $\hat{p}_i$ sa disperzijom $\hat{p}_i(1 - \hat{p}_i)/n_i$. Primenom delta metode, ocena disperzije je

$$\hat{D}(\ln \hat{p}_i) = \frac{1}{\hat{p}_i^2} \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i} = \frac{d_i}{n_i(n_i - d_i)}.$$

Dakle, ocena disperzije logaritma KM ocene delta metodom je

$$\hat{D}(\ln \hat{S}(t)) = \sum_{i:t_{(i)} \leq t} \hat{D}(\ln \hat{p}_i) = \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

Da bi se odredila ocena disperzije KM ocene funkcije preživljavanja, ponovo se primenenjuje delta metoda. Ako je $f(X) = \exp(X)$, odnosno $\hat{S}(t) = \exp(\ln(\hat{S}(t))$. Tada je Tejlorov razvoj

$$\exp(X) \approx \exp(\mu) + \exp(\mu)(X - \mu),$$

a aproksimativna ocena disperzije

$$\hat{D}(\exp(X)) = (\exp(\hat{\mu}))^2 \hat{\sigma}^2.$$

[2] Da bismo primenili Tejlorov razvoj funkcija mora biti glatka u okolini tačke razvoja.

Sledi da je ocena disperzije KM ocene

$$\hat{D}(\hat{S}(t) = (\hat{S}(t))^2 \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

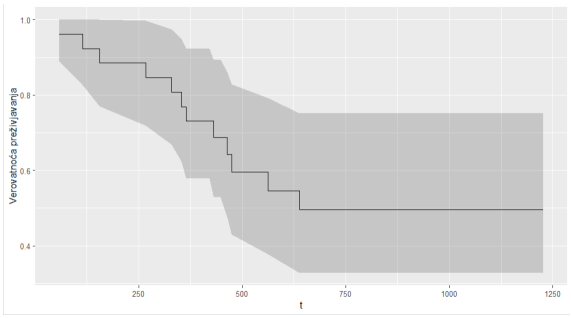
Ova ocena je poznata kao Grinvudova formula za ocenu disperzije KM ocene funkcije preživljavanja. Zaključno, Grinvudova formula za ocenu disperzije KM ocene funkcije preživljavanja pruža važan alat za ocenu preciznosti u realnim situacijama. U narednom odeljku razmatraćemo primene ove formule i njen značaj u analizi preživljavanja.

2.3 Intervali poverenja za KM krive

Za velike uzorke KM ocena ima aproksimativno normalnu raspodelu [11], pa ćemo izračunati krajeve $(1 - \alpha)100\%$ intervala poverenja za funkciju preživljavanja u trenutku t:

$$\hat{S}(t) \pm Z_{1-\alpha/2}\sqrt{\hat{D}(\hat{S}(t)},$$

gde je $Z_{1-\alpha/2}$ $1 - \alpha/2$-i kvantil standardne normalne raspodele. Ova ideja se može vizualizovati putem Slike 2.1. Međutim, problem ovog pristupa jeste da su krajevi intervala često izvan 0 i 1 [10], odnosno $\hat{S} \in [0, 1]$, a $\hat{S}(t) \pm Z_{1-\alpha/2}\sqrt{\hat{D}(\hat{S}(t)} \in [-\infty, +\infty]$, kao i normalnost KM ocene ako obim uzorka nije veliki. Sa ciljem rešavanja ovih problema, u radu [11] predlaže se da se ocenjivanje intervala poverenja zasniva na funkciji $\ln[-\ln \hat{S}(t)]$. Prednost ove funkcije je u tome što ona uzima vrednosti u intervalu $(-\infty, +\infty)$. Ovu transformaciju ćemo u nastavku označavati sa $\ln - \ln$. Ocena disperzije $\ln - \ln$ KM ocene funkcije preživljavanja, koju dobijamo



Slika 2.1: Grafički prikaz intervala poverenja

primenom delta metode na $X = \ln \hat{S}(t)$, je

$$\hat{D}(\ln(-\ln \hat{S}(t))) = \frac{1}{(\ln \hat{S}(t))^2} \sum_{i:t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}.$$

Krajevi $(1-\alpha)100\%$ interval poverenja za $\ln - \ln$ funkcije preživljavanja u trenutku t su dati sa

$$\ln(-\ln \hat{S}(t)) \pm Z_{1-\alpha/2}\sqrt{\hat{D}(\ln(-\ln \hat{S}(t)))},$$

gde je $Z_{1-\alpha/2}$ $(1-\alpha/2)$ kvantil standardne normalne raspodele. Ako sa c_1 i c_2 označimo donji i gornji kraj ovog intervala poverenja, onda je $(1-\alpha)100\%$ interval poverenja za $\hat{S}(t)$

$$(\exp(-e^{c_2}), \exp(-e^{c_1})).$$

Interval poverenja nije definisan za $\hat{S}(t) = 0$ ili $\hat{S}(t) = 1$. U tom slučaju, preporučuje se korišćenje $(0, 0)$ ili $(1, 1)$ kao intervala poverenja ako je potrebno grafički, u suprotnom izostaviti interval poverenja za te tačke. Dakle, interval poverenja važi samo za one vrednosti vremena za koje je KM ocena definisana, što je u osnovi posmatrani domen vremena preživljavanja.

Borgan i Listol [2] su pokazali da je $\ln - \ln$ interval poverenja za funkciju preživljavanja bolji od običnog, linearnog intervala. $\ln - \ln$ interval poverenja daje pouzdanu ocenu za uzorke malog obima, čak i u 50% cenzurisanja. Drugim rečima, ovaj pristup pruža ocene koje su veoma bliske stvarnoj vrednosti parametra. To znači da ovaj interval ima visok stepen tačnosti u proceni stvarne vrednosti parametra koristeći raspoložive podatke. Međutim, za veoma velike uzorke, ove dve metode su ekvivalentne, što znači da ne postoji značajna razlika u njihovoj efikasnosti. Takođe se primećuje da $\ln - \ln$ intervali poverenja nisu simetrični oko tačkaste ocene funkcije preživljavanja, što ukazuje na potrebu za prilagođavanjem prilikom interpretacije rezultata.

2.4 Test logaritmovanih rangova

Interesuje nas kako oceniti da li su KM krive za dve ili više grupa statistički ekvivalentne. Najpopularniji metod za to je test logaritmovanih rangova, poznat i kao Mantel-Koksov test.

Da su dve KM krive statistički ekvivalentne znači da na osnovu postupka testiranja koji upoređuje dve krive nemamo dokaze koji bi ukazivali da su stvarne (populacijske) krive preživljavanja različite. Test logaritmovanih rangova je specifičan

oblik χ^2 testa koji se koristi u analizi preživljavanja za upoređivanje preživljavanja između grupa. Test logaritmovanih rangova koristi Kaplan-Majerovu ocenu (2.1) funkcije preživljavanja za svaku grupu i zatim upoređuje realizovane brojeve događaja između grupa koristeći statistički test baziran na χ^2 raspodeli. U narednom odeljku ćemo ga detaljnije prikazati.

Test logaritmovanih rangova za dve grupe

Za svako vreme neuspeha, $t_{(f)}$, u celom skupu podataka, tabelarno se prikazuje broj ispitanika (m_{if}) koji su doživeli neuspeh u tom trenutku, razdvojeno po grupama $i = 1, 2$. Zatim se prikazuju brojevi ispitanika (n_{if}) koji su još uvek pod rizikom za neuspeh u tom vremenu, takođe razdvojeni po grupama $i = 1, 2$. Prethodnu tabelu proširujemo tako što dodajemo ocene očekivane vrednosti i razlike između opaženih i ocene očekivanih vrednosti za svaku grupu u svakom vremenu neuspeha. Za svaku grupu, izraz za ocene očekivane vrednosti se računa kao udeo ukupnog broja ispitanika koji su u riziku u tom trenutku, pomnožen ukupnim brojem neuspeha za obe grupe:

$$e_{if} = \left(\frac{n_{if}}{n_{1f} + n_{2f}} \right) \cdot (m_{1f} + m_{2f}), \qquad i = 1, 2.$$

U slučaju dve grupe, statistika logaritmovanih rangova se računa na sledeći način:

$$L = \frac{(\sum_{i=1}^{2} (O_i - E_i))^2}{\sum_{i=1}^{2} D(O_i - E_i)},$$

gde O_i predstavlja opažene vrednosti, a E_i očekivane vrednosti u grupi i. Nulta hipoteza koja se testira je da nema razlike između dve krive preživljavanja. Pri ovoj nultoj hipotezi, L je približno raspodeljena kao χ^2 sa jednim stepenom slobode. Stoga se p-vrednost za test logaritmovanih rangova određuje iz tablica χ^2 raspodele.

Test logaritmovanih rangova za više grupa

Test logaritmovanih rangova se može koristiti i za upoređivanje tri ili više krivih preživljavanja. Nulta hipoteza u ovom uopštenom slučaju je da su sve krive preživljavanja iste. Iako se ista tablična struktura koristi za izračunavanje kada postoji više od dve grupe, test statistika je matematički složenija i uključuje i disperziju i kovarijacije zbrojenih opaženih minus očekivanih rezultata za svaku grupu. Pogodna

matematička formula se izražava u matričnom obliku.
Neka je $i = 1, 2, \ldots, G$ i $f = 1, 2, \ldots, k$, gde je G broj grupa, a k broj različitih vremena neuspeha. n_{if} je broj ispitanika u riziku i-te grupe u f-tom trenutku neuspeha, a m_{if} broj ispitanika koji su doživeli neuspeh u i-toj grupi u f-tom trenutku neuspeha. Tada je očekivani broj neuspeha i-te grupe u f-tom trenutku neuspeha

$$e_{if} = \frac{n_{if}}{n_f} \cdot m_f,$$

gde su $n_f = \sum_{i=1}^{G} n_{if}$ i $m_f = \sum_{i=1}^{G} m_{if}$. Označimo razliku između opaženih i očekivanih brojeva događaja za i-tu grupu kao

$$O_i - E_i = \sum_{f=1}^{k} (m_{if} \ - \ n_{if}).$$

Formule za disperziju i kovarijaciju izražene su kao:

$$D(O_i - E_i) = \sum_{f=1}^{k} \frac{n_{if} \ (n_f \ - \ n_{if}) \ m_{if} \ (n_f \ - \ m_f)}{n_f^2 (n_f \ - \ 1)},$$

$$Cov(O_i - E_i, \ O_l - E_l) = \sum_{f=1}^{k} \frac{- \ n_{if} \ n_{lf} \ m_f \ (n_f \ - \ m_f)}{n_f^2 (n_f \ - \ 1)}.$$

Označimo vektor razlika između opaženih i očekivanih vrednosti kao

$$\mathbf{d} \ = \ (O_1 - E_1, \ O_2 - E_2, \ ..., O_{G-1} - E_{G-1})',$$

a matricu V

$$\mathbf{V} = ((v_{il})),$$

gde je $v_{ii} = D(O_i - E_i)$ i $v_{il} = Cov(O_i - E_i, \ O_l - E_l)$ za $i = 1, 2, \ldots, G-1$ i $l = 1, 2, \ldots, G-1$. Dobijamo statistiku

$$L = \ \mathbf{d}/\mathbf{V}^{-1}\mathbf{d},$$

koja pod pretpostavkom nulte hipoteze da sve G grupe dele zajedničku krivu preživljavanja, ima χ^2 raspodelu sa $G-1$ stepenom slobode.

Kaplan-Majerova kriva i test logaritmovanih rangova predstavljaju statističke metode za ocenu i upoređivanje preživljavanja među različitim grupama. U sledećem poglavlju ćemo istražiti Koksov model, koji proširuje ovu analizu omogućavajući istovremeno razmatranje uticaja više faktora na preživljavanje.

Glava 3

Koksov PH model

Pri proučavanju modela preživljavanja, analiziraju se dve osnovne komponente: osnovna stopa rizika koja opisuje kako se rizik menja tokom vremena i uticaj parametara koji opisuju varijaciju rizika u odnosu na nezavisne promenljive. Dejvid Koks je primetio da, ukoliko se pretpostavi da je stopa rizika proporcionalna nekoj drugoj funkciji koja opisuje kako se rizik menja tokom vremena, može se proceniti uticaj parametara bez određivanja same funkcionalne forme stope rizika. Ovaj pristup analizi podataka preživljavanja naziva se primena Kokovog modela proporcionalnih rizika.

Koksov model proporcionalnih rizika je najjednostavniji matematički model za ocenu krivih preživljavanja kada se istovremeno razmatra nekoliko objašnjavajućih promenljivih. Ključna pretpostavka Koksovog modela jeste pretpostavka proporcionalnih rizika (PH): kada se prediktori ne menjaju tokom vremena, odnos rizika između bilo koja dva posmatranja je konstantan u odnosu na vreme.
U modelu proporcionalnih rizika, stopa rizika je data sa:

$$h_{x(\cdot)} = h_0(t)\psi(x(t)), \qquad t \in [0, +\infty)$$

gde je $\psi(x(t))$ pozitivna funkcija zavisna od promenljive $x(t)$, koja se menja tokom vremena, a $h_0(t)$ osnovna (bazna) stopa rizika. Ako je funkcija ψ nepoznata, tada se radi o neparametarskom modelu. Međutim, u većini slučajeva, funkcija ψ se parametrizuje kao $\psi(x(t)) = \exp(\beta^T x)$, gde je $\beta = (\beta_1, \ldots, \beta_p)$. Na ovaj način definisani model je semiparametarski ili poluparametarski model. Ako je h_0 iz neke poznate familije raspodele, tada govorimo o parametarskom modelu. Često korišćene poznate familije raspodele u analizi preživljavanja obuhvataju eksponencijalnu raspodelu, Vejbulovu raspodelu i log-normalnu raspodelu.

Dakle, Koksov PH model se obično zapisuje u obliku sledeće formule:

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} X_i\beta_i}, \qquad \mathbf{X} = (X_1, X_2, ..., X_p). \tag{3.1}$$

Ovaj model daje izraz za stopu rizika u vremenu t za pojedinca s određenom specifikacijom skupa objašnjavajućih veličina označenih kao $\mathbf{X}$. Drugim rečima, $\mathbf{X}$ predstavlja vektor (skup) prediktora koje se modeliraju kako bi se predvideo rizik pojedinca. Koksov model podrazumeva da je stopa rizika proizvod dve funkcije. Prva od tih veličina, $h_0(t)$, naziva se osnovna (bazna) stopa rizika. Druga veličina je eksponencijalni izraz $e^{\sum_{i=1}^{p} X_i\beta_i}$ koja je nenegativna. Vrednosti bilo koje stope rizika su između nule i beskonačnosti, odnosno stopa rizika je uvek nenegativna. Ako bi, na primer, umesto eksponencijalnog izraza, deo modela sa $\mathbf{X}$-ovima bio linearan, mogu se dobiti negativne ocene hazarda, što nije dopušteno. Važna osobina formule (3.1), koja se odnosi na pretpostavku proporcionalnih rizika, jeste da je bazna stopa rizika funkcija vremena t, ali ne uključuje prediktore. Nasuprot tome, eksponencijalni izraz prikazan ovde uključuje prediktore, ali ne uključuje t. Prediktori ovde se nazivaju prediktori vremenski nezavisne promenljive. Ipak, moguće je posmatrati prediktore koji uključuju t. Takvi prediktori nazivaju se vremenski zavisne promenljive. Ako se uzimaju u obzir veličine zavisne od vremena, Koksov model se i dalje može koristiti, ali takav model više ne zadovoljava pretpostavku proporcionalnih rizika i naziva se proširenim Koksovim modelom. Primeri vremenski nezavisnih promenljivih su pol i pušački status. Pušački status se može menjati kroz vreme, naime pacijent koji je registrovan kao pušač mogao je tokom lečenja prestati sa tom aktivnošću za stalno i postati bivši pušač i obrnuto. Ono što se lako zapaža je da se promenljive kao što su starosno doba i težina menjaju kroz vreme, ali može biti veoma zgodno tretirati takve promenljive kao vremenski nezavisne, ukoliko se njihova vrednost ne menja drastično kroz vreme.

Važne osobine Koksovog modela

- Primetimo da, ukoliko su svi X-evi jednaki nuli, Koksova formula se svodi na stopu osnovnog rizika, jer je

  $$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} X_i\beta_i} = h_0(t)e^0 = h_0(t).$$

  To se odnosi i na cenzurisane podatke jer kada su svi prediktori jednaki nuli, to implicira da subjekat nema nijedan od prediktora koji bi uticao na preži-

vljavanje. Ovo nam omogućuje da uključimo cenzurisane podatke u analizu i procenimo osnovnu stopu rizika bez obzira na status cenzurisanja.

- Koksov model je delimično parametarski (semiparametarski) model. Ovo je bitno jer ne zahteva potpune informacije o osnovnoj stopi rizika kako bi se procenili koeficijente za prediktore. Umesto toga, model procenjuje odnos relativnih rizika na osnovu dostupnih podataka, što ga čini fleksibilnim za analizu preživljavanja sa cenzurisanim podacima.
- Iako osnovna stopa rizika nije precizno određena, mogu se dobiti prilično dobre ocene regresionih koeficijenata, stope rizika i krivih preživljavanja za različite situacije s podacima. Drugim rečima, Koksov proporcionalni model rizika je robusni model, što znači da će rezultati dobijeni korišćenjem Koksovog modela biti približni rezultatima ispravnog parametarskog modela.
- Važno je napomenuti da se stopa rizika $h(t, X)$ i odgovarajuće krive preživljavanja $S(t, X)$ mogu odrediti za Koksov model, čak i ako osnovna stopa rizika nije određena. Dakle, uz pomoć Koksovog modela i minimalnog broja pretpostavki, mogu se dobiti osnovne informacije koje su potrebne, a to su funkcija rizika i funkcija preživljavanja.
- Cenzurisani podaci se tretiraju kao delimični podaci o preživljavanju, gde stopa rizika uzima u obzir subjekte koji su doživeli događaj (nezavisno od toga da li su cenzurisani ili ne) i one koji su cenzurisani. Model koristi delimične informacije o preživljavanju kako bi procenio koeficijente i relativne rizike za svaki prediktor.
- Koksov model koristi više informacija, vremena preživljavanja, u odnosu na logistički model koji uzima u obzir binarni $(0, 1)$ ishod i zanemaruje vremena preživljavanja i cenzuriranje. Ovo je važno jer Koksov model može bolje proceniti uticaj prediktora na preživljavanje, uzimajući u obzir vreme do događaja i cenzuriranje.

3.1 Količnik hazarda

Količnik hazarda se definiše kao količnik rizika dva subjekta. Individue koje se porede se razlikuju po vrednostima nezavisnih promenljivih koje ih karakterišu.

Ocenu količnika rizika možemo zapisati kao:

$$\widehat{HR} = \frac{\hat{h}(t,\ \mathbf{X}^*)}{\hat{h}(t,\ \mathbf{X})},$$

gde vektori $\mathbf{X}^* = (X_1^*, X_2^*, ..., X_p^*)$ i $\mathbf{X} = (X_1, X_2, ..., X_p)$ označavaju skupove prediktora koji karakterišu jedinku. Najčešće se uzima da je brojilac veći od imenioca, tj. da je količnik rizika veći od jedan, odnosno:

$$\hat{h}(t,\ \mathbf{X}^*) > \hat{h}(t,\ \mathbf{X}).$$

Stoga, X-evi se kodiraju tako da grupi sa većim rizikom, obično neizloženoj grupi odgovara $\mathbf{X}^*$, a grupi sa manjim rizikom odgovara $\mathbf{X}$. Drugi način da zapišemo gore navedenu formulu jeste korišćenjem formule Koksovog modela, tj.

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\widehat{h_0}(t)e^{\sum_{i=1}^p X_i^*\beta_i}}{\widehat{h_0}(t)e^{\sum_{i=1}^p X_i\beta_i}} = \frac{e^{\sum_{i=1}^p X_i^*\beta_i}}{e^{\sum_{i=1}^p X_i\beta_i}} = e^{\sum_{i=1}^p (X_i^* - X_i)\beta_i}.$$

3.2 Pretpostavke PH modela

Zahteva se da je količnik hazarda konstantan u vremenu, tj. da je rizik jednog subjekta proporcionalan riziku drugog subjekta, gde je konstanta srazmernosti nezavisna od vremena. Konačni izraz za količnik hazarda ne zavisi od vremena:

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\widehat{h_0}(t)e^{\sum_{i=1}^p X_i^*\beta_i}}{\widehat{h_0}(t)e^{\sum_{i=1}^p X_i\beta_i}} = e^{\sum_{i=1}^p (X_i^* - X_i)\beta_i}.$$

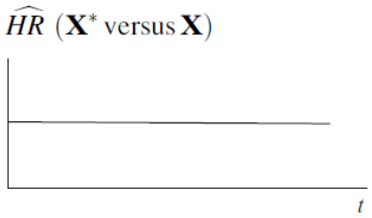
Označimo ovu konstantu sa $\hat{\theta}$, tj. imamo $\hat{\theta} = e^{\sum_{i=1}^p (X_i^* - X_i)\beta_i}$. Odnos rizika možemo izraziti na sledeći način:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}.$$

Ovaj izraz nam omogućava da povežemo vrednosti stopa rizika. Ovime dobijamo izražen oblik za stopu rizika sa prediktorima $\mathbf{X}^*$ kao:

$$\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X}).$$

Izraz kaže da je stopa rizika za jednog pojedinca proporcionalna stopi rizika za drugog pojedinca, pri čemu konstanta proporcionalnosti $\hat{\theta}$ ne zavisi od vremena. Grafički, ovaj izraz kaže da se ocenjeni količnik rizika koji upoređuje bilo koja dva pojedinca prikazuje kao konstanta tokom vremena, što je ilustrovano na Slici 3.1.



Slika 3.1: Grafički prikaz količnika hazarda u slučaju Koksovog modela

3.3 Ocenjivanje parametera Koksovog modela

U ovom poglavlju prikazaćemo metod maksimalne verodostojnosti za ocenjivanje parametara modela.

Pretpostavimo da imamo podatke o preživljavanju nezavisnih pojedinaca, koji mogu biti cenzurisani, pri čemu je stopa rizika data sa (3.1). Označimo vremena preživljavanja kao $t_1 < t_2 < \ldots < t_k$, gde k označava ukupan broj neuspeha. Pretpostavljamo da nikoja dva pojedinca ne doživljavaju neuspeh istovremeno, što znači da svaki trenutak t_j ima tačno jednog pojedinca koji doživljava neuspeh. S obzirom na to da pojedinci mogu imati različite stope rizika, važno je pratiti koji pojedinac doživljava neuspeh u trenutku t_j i koji su pojedinci bili u riziku pre tog trenutka t_j. Samo poznavanje ukupnog broja neuspeha i ukupnog broja pojedinaca u riziku pre trenutka t_j (kako je bilo kod Kaplan-Majerove krive) nije dovoljno. Zato koristimo oznake I_j za indeks pojedinca koji doživljava neuspeh u trenutku t_j, i R_j za skup indeksa pojedinaca koji su bili pod rizikom pre tog trenutka t_j. Obično I_j nazivamo „označeni" pojedinac, a R_j „skup pod rizikom".

Ocene parametara metodom maksimalne verodostojnosti za Koksov model dobijaju se maksimizacijom funkcije verodostojnosti koja se obično označava sa L ili $L(\beta)$, gde β označava skup nepoznatih parametara. Funkcija verodostojnosti za Koksov model se često naziva i parcijalna funkcija verodostojnosti jer ona razmatra samo verovatnoće za one subjekte kod kojih se desio događaj i ne razmatra verovatnoće za subjekte koji su cenzurisani. Takva parcijalna funkcija verodostojnosti se može zapisati kao proizvod nekoliko funkcija verodostojnosti, na primer ako je bilo k neuspeha

$$L = L_1 \times L_2 \times ... \times L_k = \prod_{j=1}^{k} L_j = \prod_{j=1}^{k} \frac{\exp(X_{I_j}\beta_j)}{\sum_{m \in R_j} \exp(X_{R_m}\beta_j)}.$$

L_j je funkcija verodostojnosti za j-to vreme neuspeha, a subjekti kod kojih postoji

rizik da se desi događaj u vremenu t_j čine grupu rizika i označavaju sa R_j. Skup grupa rizika se smanjuje kako se vreme povećava. Iako se parcijalna funkcija verodostojnosti fokusira na subjekte koji su doživeli neuspeh, informacija o vremenu preživljavanja pre cenzuriranja koristi se za subjekte koji su cenzurirani. Drugim rečima, osoba koja je cenzurirana nakon f-tog vremena neuspeha je deo skupa rizika koji se koristi za izračunavanje L_f, iako je ta osoba kasnije cenzurisana. Sledeći korak je maksimizacija te funkcije. To se obično postiže maksimizacijom prirodnog logaritma L, zbog jednostavnijeg računa. Postupak maksimizacije se izvodi računanjem parcijalnih izvoda logaritma L u odnosu na svaki parametar u modelu, a zatim rešavanjem sistema jednačina.

$$\frac{\partial \ln L}{\partial \beta_i} = 0, \qquad i = 1, ..., p \text{ (p je broj parametara).}$$

Za računanje ocena maksimalne verodostojnosti često se primenjuje Njutn-Rapsonov algoritam. Ovaj postupak se sastoji od niza iteracija, pri čemu se pretpostavlja da će se ocena poboljšavati sa svakom narednom iteracijom. Često, ali ne uvek, konvergira ka ciljanim ocenama maksimalne verodostojnosti.

Na primer, pretpostavimo da su podaci preživljavanja za pojedince $1, 2, 3$ i 4, redom, $5, 3, 12$ i $10+$, gde $+$ označava cenzurisanu vrednost. Tada su uređeni trenuci neuspeha $(t_1, t_2, t_3) = (3, 5, 12)$, odgovarajuće oznake pojedinaca su $(I_1, I_2, I_3) = (2, 1, 3)$, a posmatrani skupovi rizika su $R_1 = \{1, 2, 3, 4\}$, $R_2 = \{1, 3, 4\}$ i $R_3 = \{3\}$. Tada je parcijalna funkcija verodostojnosti data sa

$$L = \frac{\psi_2}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \times \frac{\psi_1}{\psi_1 + \psi_3 + \psi_4} \times \frac{\psi_3}{\psi_3} = \frac{\psi_2 \psi_1}{(\psi_1 + \psi_2 + \psi_3 + \psi_4)(\psi_1 + \psi_3 + \psi_4)},$$

gde je ψ_i odgovarajuća eksponencijalna funkcija koja se pojavljuje u (3.1).

3.4 Ocenjene krive preživljavanja

Dve osnovne veličine koje nas zanimaju iz analize preživljavanja su ocena količnika hazarda i ocene krivih preživljavanja. Pošto je opisan postupak za računanje ocene količnika rizika, sada ćemo se osvrnuti na ocenjivanje krivih preživljavanja koristeći Koksov model. Kada se Koksov model koristi za ocenjivanje krive preživljavanja, krive preživljavanja se dobijaju tako da budu prilagođene objašnjavajućim promenljivama u modelu. Otuda i potiče naziv prilagođene krive preživljavanja. Kao i KM krive, i one su stepenastog oblika.

Iz stope rizika za Koksov PH model

$$h(t, \mathbf{X}) \;=\; h_0(t) e^{\sum_{i=1}^{p} X_i \beta_i}$$

može se dobiti kriva preživljavanja

$$S(t, \mathbf{X}) = (S_0(t))^{e^{\sum_{i=1}^{p} X_i \beta_i}}.$$

Ova formula je osnova za određivanje krivih preživljavanja. Izraz za ocenjenu krivu preživljavanja je

$$\hat{S}(t, \mathbf{X}) = (\hat{S}_0(t))^{e^{\sum_{i=1}^{p} X_i \hat{\beta}_i}}.$$

Ocene $\hat{S}_0(t)$ i $\hat{\beta}_i$ se mogu dobiti korišćenjem raznih softverskih alata koji imaju ugrađene funkcije, ali vrednosti za X-eve moraju biti određene od strane istraživača, kako bi program mogao da izračuna ocene za funkcije preživljavanja. Na primer, paket `survival` je osnovni za analizu preživljavanja u R. Funkcija `coxph()` se koristi za ocenu parametara Koksovog modela koristeći metodu parcijalne funkcije verodostojnosti. Ova metoda koristi Njutn-Rapsonov algoritam za pronalaženje ocena. Nakon toga, možete koristiti funkciju `survfit()` koja koristi Kaplan-Majerovu metodu za ocenu funkcije preživljavanja i generisanje krivih preživljavanja. Takođe, za pronalaženje standardnih grešaka, funkcija `survfit()` koristi Grinvudovu formulu.

Neophodno je posvetiti posebnu pažnju evaluaciji pretpostavki kako bi se osigurala tačna interpretacija rezultata. Tome će više pažnje biti posvećeno u narednom poglavlju.

Glava 4

Evaluacija pretpostavki PH modela

U ovom poglavlju opisane su tri metode za proveru pretpostavki PH modela - grafički metod, primena testova saglasnosti sa modelom (GOF) i metode vremenski zavisnih promenljivih. Metode vremenski zavisnih promenljivih će biti detaljnije opisane u jednom od narednih poglavlja.

4.1 Grafički metod

Postoje dva tipa grafičkih metoda koji se često primenjuju u analizi preživljavanja.

Najznačajniji među njima uključuje upoređivanje ocenjenih $-\ln-\ln$ krivih preživljavanja za različite grupe promenljivih koje se ispituju. Kada se ove krive odrede, njihova paralelnost ukazuje na ispunjenje pretpostavke proporcionalnih rizika. $-\ln-\ln$ kriva preživljavanja je transformacija ocenjene krive preživljavanja koja se dobija dvostrukim uzimanjem prirodnog logaritma ocenjene verovatnoće preživljavanja. Matematički, $-\ln-\ln$ krivu zapisujemo kao $-\ln(-\ln\hat{S})$. Primetimo da je logaritam $\hat{S}$ (jer je $\hat{S}$ verovatnoća) uvek negativan broj. Budući da možemo logaritmovati samo pozitivne brojeve, ubacujemo minus prvog logaritam pre nego što uzmemo drugi logaritam. Vrednost za $-\ln(-\ln\hat{S})$ može biti pozitivna ili negativna. Ekvivalentan zapis zapisu -$\ln(-\ln\hat{S})$ je $-\ln(\int_0^t h(u)\,du)$. Ovaj rezultat dolazi iz formule $S(t) = \exp(-\int_0^t h(u)\,du)$ koja povezuje funkciju preživljavanja i funkciju rizika. Dalje ćemo ilustrovati kako je moguće proveriti pretpostavku proporcionalnih rizika analizom paralelnosti $-\ln-\ln$ krivih. Da bismo to postigli, potrebno je detaljno objasniti izraz koji se dobija nakon primene logaritma na $\hat{S}$. Počinjemo od formule za krivu preživljavanja koja odgovara stopi rizika za Koksov PH model.

Podsetimo se da je

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i},$$
$$S(t, \mathbf{X}) = [S_0(t)]^{e^{\sum_{i=1}^{p} \beta_i X_i}},$$

gde $S_0(t)$ označava baznu funkciju preživljavanja koja odgovara baznoj stopi rizika $h_0(t)$. Na prethodni izraz sada ćemo primeniti logaritam:

$$\ln S(t, \mathbf{X}) = e^{\sum_{i=1}^{p} \beta_i X_i} \cdot \ln S_0(t).$$

Budući da je $S_0(t)$ verovatnoća koja se kreće između 0 i 1, a logaritam brojeva između 0 i 1 je negativan, kako bismo primenili drugi logaritam, moramo koristiti negativni znak. Primenom drugog logaritma dobijamo:

$$\begin{aligned} \ln(-\ln S(t, \mathbf{X})) &= \ln\left(-e^{\sum_{i=1}^{p} \beta_i X_i} \cdot \ln S_0(t)\right), \\ &= \ln(e^{\sum_{i=1}^{p} X_i \beta_i}) + \ln(-\ln S_0(t)), \\ &= \textstyle\sum_{i=1}^{p} X_i \beta_i + \ln(-\ln S_0(t)). \end{aligned}$$

Radi doslednosti, uobičajena praksa je da se stavi minus ispred drugog logaritma kako bismo dobili izraz $-\ln -\ln$ koji je prikazan ovde. Ipak, neki softverski paketi ne koriste drugi minus znak. Konačno je

$$-\ln(-\ln S(t, \mathbf{X})) = -\sum_{i=1}^{p} X_i \beta_i - \ln(-\ln S_0(t)).$$

Neka $\mathbf{X}_1 = (X_{11}, X_{12}, \ldots, X_{1p})$ predstavlja vrednosti prediktora za jednog pojedinca, dok $\mathbf{X}_2 = (X_{21}, X_{22}, \ldots, X_{2p})$ označava vrednosti prediktora za drugog pojedinca. Tada se odgovarajuće $-\ln -\ln$ krive za ove pojedince dobijaju na sledeći način

$$-\ln(-\ln S(t, \mathbf{X}_1)) = -\sum_{i=1}^{p} X_{1i} \beta_i - \ln(-\ln S_0(t)),$$

$$-\ln(-\ln S(t, \mathbf{X}_2)) = -\sum_{i=1}^{p} X_{2i} \beta_i - \ln(-\ln S_0(t)).$$

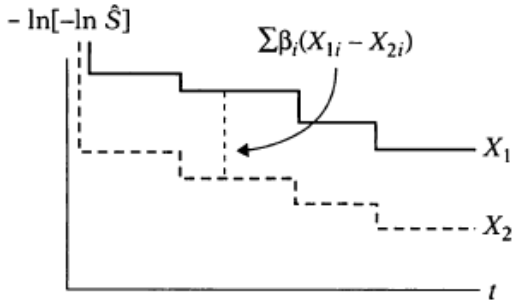
Oduzimanjem prve $-\ln -\ln$ krive od druge dobijamo izraz koji je linearna razlika odgovarajućih vrednosti prediktora oba pojedinca. Važno je primetiti da ova razlika ne uključuje vremensku promenljivu t. Imamo

$$\ln(-\ln S(t, \mathbf{X}_1)) - \ln(-\ln S(t, \mathbf{X}_2)) = \sum_{i=1}^{p} (X_{1i} - X_{2i})\beta_i,$$

odnosno

$$\ln(-\ln S(t, \mathbf{X}_1)) = \ln(-\ln S(t, \mathbf{X}_2)) + \sum_{i=1}^{p} (X_{1i} - X_{2i})\beta_i.$$

Gornji izraz kaže da ako koristimo Koksov PH model i grafički prikažemo ocenjene $-\ln-\ln$ krive preživljavanja za pojedince na istom grafikonu, dve krive će biti približno paralelne. Razdaljina između ove dve krive je linearni izraz koji uključuje razlike u vrednostima prediktora, ali ne uključuje vreme. Primetimo, ako je vertikalna udaljenost između dve krive konstantna, tada su krive paralelne, što je ilustrovano na Slici 4.1.



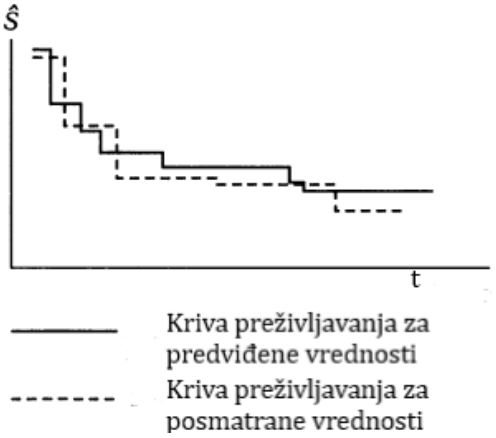
Slika 4.1: Grafički metod primene $-\ln-\ln$ krive

Postoje dva pristupa kada je reč o empirijskim graficima. Prvi pristup se odnosi na prikazivanje $-\ln-\ln$ krivih preživljavanja na osnovu Kaplan-Majerove ocene, pri čemu se ne uzima u obzir postojanje Koksovog modela. Drugi pristup, s druge strane, omogućava prikazivanje $-\ln-\ln$ krivih preživljavanja koje su prilagođene za prediktore za koje već pretpostavljamo da zadovoljavaju pretpostavku proporcionalnih rizika. Važno je napomenuti da ovi prediktori nisu uključeni u Koksov model koji se trenutno ispituje.

Glavni problem je kako odlučiti „koliko je paralelno dovoljno paralelno?". Ova odluka može biti vrlo subjektivna za određeni skup podataka, posebno ako je veličina skupa relativno mala. U radu [8] je preporučeno korišćenje konzervativne strategije pri donošenju ove odluke tako što ćemo pretpostaviti da je PH pretpostavka ispunjena osim ako postoji snažan dokaz o neparalelnosti $-\ln-\ln$ krivih. Još jedan problem se odnosi na to kako kategorizovati neprekidnu promenljivu. Ako se izabere previše kategorija, podaci se „razređuju" u svakoj kategoriji, što otežava poređenje različitih krivih. Takođe, jedna kategorizacija na primer u tri grupe može

dati drugačiju grafičku sliku od drugačije kategorizacije u tri grupe. Zbog toga, preporučuje se da broj kategorija bude razumno mali (npr. dve ili tri, ako je moguće) i da izbor kategorija bude smislen, pružajući prihvatljivu ravnotežu u brojevima. Još jedan problem pri korišćenju $-\ln-\ln$ kriva preživljavanja odnosi se na to kako proceniti pretpostavku proporcionalnih rizika istovremeno za više promenljivih. Interakcije među prediktorima takođe mogu biti izazov pri korišćenju $-\ln-\ln$ krivih preživljavanja. Kada se analizira više prediktora, posebno kada postoji sumnja da interakcije između njih mogu uticati na preživljavanje, interpretacija rezultata može postati složenija. Uvođenje interakcija može dovesti do promena u obliku i nagibu krivih preživljavanja, što dodatno komplikuje donošenje zaključaka o ispunjenju PH pretpostavke. Ova situacija dodatno naglašava važnost pažljive analize i razmatranja potencijalnih interakcija među prediktorima tokom istraživanja $-\ln-\ln$ krivih preživljavanja. Jedna strategija za istovremene poređenja je kategorizacija svih promenljivih posebno, formiranje kombinacija kategorija, a zatim upoređivanje $-\ln-\ln$ krivih za sve kombinacije na istom grafikonu. Mana je što će se podaci ponovno „razrediti" kako broj kombinacija postaje čak umereno velik. Takođe, čak i ako postoji dovoljan broj za svaku kombinovanu kategoriju, često je teško utvrditi koje promenljive su odgovorne za bilo kakvu neparalelnost koja može biti pronađena.

Alternativni grafički pristup je poređenje posmatranih i ocenjenih krivih preživljavanja. Posmatrane krive se dobijaju za grupe promenljive koja se ocenjuje, bez uključivanja ove promenljive u PH model. Ocenjene krive se dobijaju kada se ova promenljiva uključi u PH model. Ako su posmatrane i predviđene krive bliske, tada je pretpostavka proporcionalnih rizika razumna. Ovde ćemo opisati samo metodu koja podrazumeva korišćenje Kaplan-Majerovih krivih za dobijanje posmatranih grafikona. Prvo moramo stratifikovati naše podatke prema grupama prediktora koji se ocenjuje. Zatim dobijamo posmatrane grafikone tako što posebno izračunavamo Kaplan-Majerove krive za svaku kategoriju. Da bismo dobili „očekivane" grafikone, prilagođavamo Koksov PH model koji sadrži prediktor koji se ocenjuje. Očekivane grafike dobijamo tako što posebno zamenjujemo vrednost za svaku kategoriju prediktora u formulu za ocenjenu krivu preživljavanja, čime dobijamo posebnu ocenjenu krivu preživljavanja za svaku kategoriju, kao što je prikazano na Slici 4.2.



Slika 4.2: Grafički metod poređenja posmatranih i ocenjenih krivih preživljavanja

Za poređenje posmatranih i očekivanih grafikona, prikazujemo obe grupe na istom grafikonu. Ako su posmatrani i očekivani grafikoni „bliski" jedan drugom za svaku kategoriju prediktora koji se procenjuje, možemo zaključiti da je PH pretpostavka zadovoljena. Međutim, ako jedna ili više kategorija pokazuje značajne razlike između posmatranih i očekivanih grafikona, zaključujemo da je PH pretpostavka narušena.

Očigledna mana ovog pristupa je odlučivanje „koliko blizu je blizu" kada upoređujemo posmatrane i očekivane krive za određenu kategoriju. To je analogno odlučivanju „koliko paralelne su paralelne" kada upoređujemo $-\ln-\ln$ krive preživljavanja. Ovde preporučujemo da se PH pretpostavka smatra neprihvaćenom samo kada su posmatrani i očekivani grafikoni značajno različiti. Prilikom korišćenja grafika posmatranih naspram očekivanih krivih za procenu PH pretpostavke kod neprekidne promenljive, posmatrani grafici se dobijaju na isti način kao kod kategoričkih promenljivih, formiranjem stratuma na osnovu kategorija neprekidne promenljive i zatim dobijanjem KM krivih za svaku kategoriju. Međutim, kod neprekidnih prediktora postoje dve opcije za izračunavanje očekivanih grafika. Jedna opcija je korišćenje Koksovog PH modela koji sadrži $k-1$ pomoćnih (engl. dummy) promenljivih koje označavaju kategorije. Očekivani grafik za određenu kategoriju se dobija kao prilagođena kriva preživljavanja zamenom vrednosti pomoćnih promenljivih koje definišu tu kategoriju u formulu za procenjenu krivu preživljavanja

$$h(t, \mathbf{X}_c) = h_0(t) e^{\sum_{i=1}^{k-1} X_{ci}\beta_i},$$

$$\hat{S}(t, \mathbf{X}_c) = \left[\hat{S}_0(t)\right]^{e^{\sum_{i=1}^{p} X_{ci}\hat{\beta}_i}},$$

gde je $\mathbf{X}_c = (X_{c1}, X_{c2}, \ldots, X_{ck-1})$. Druga opcija je korišćenje Koksovog PH modela koji sadrži neprekidni prediktor koji se procenjuje. Očekivani grafici se zatim dobijaju kao prilagođene krive preživljavanja tako što se specificiraju vrednosti prediktora koje razlikuju kategorije, na primer, kada se koriste prosečne vrednosti prediktora za svaku kategoriju.

Pored grafičkih metoda za proveru proporcionalnosti, takođe se često koriste statističke tehnike, poput testova saglasnosti sa modelom, kako bi se dublje istražila adekvatnost Koksovog modela proporcionalnih rizika.

4.2 Testovi saglasnosti sa modelom

Drugi pristup za procenu pretpostavke proporcionalnih rizika uključuje testove saglasnosti sa modelom (engl. goodness-of-fit; GOF). Prvo ćemo izračunati Šenfeldove reziduale za svaki prediktor u modelu tako što od posmatrane vrednosti oduzmemo težinski prosek koji su još uvek u riziku u trenutku t.
Neka je $\mathbf{X}_i$ p-dimenzionalni vektor prediktora za i-tog pojedinca i β p-dimenzionalni vektor koeficijenata. $Y_j(t) = I(T_j > t)$ ukazuje da li je j-ti subjekt još uvek pod rizikom (živ) u vremenu t, a $\gamma_i(\beta, t)$ je mera rizika subjekta i, tj. $\gamma_i(\beta, t) = e^{\beta' X_i(t)} \equiv \gamma_i(t)$. Neka je $x(\beta, s)$ težinski prosek $\mathbf{X}$ nad onim opservacijama koje su još uvek pod rizikom u vremenu s, sa težinama $Y_i(s)\gamma_i(s)$:

$$\overline{x}(\beta, t_k) = \frac{\sum_{i=1}^{n} Y_i(s)\gamma_i(s)X_i(s)}{\sum_{i=1}^{n} Y_i(s)\gamma_i(s)}.$$

Neka d označava ukupan broj događaja. Neka $\mathbf{X}_{(k)}$ predstavlja vektor prediktora za odgovarajućeg pojedinca sa događajem u k-tom događaju vremena t_k, gde je $k = 1, 2, \ldots, d$. Nadalje, neka R_k označava skup rizika u vremenu t_k, što je skup svih pojedinaca koji su još uvek u riziku u trenutku t_k. Tada se Šenfeldov rezidual definiše kao:

$$r_k(\hat{\beta}) = X_{(k)} - E(X_{(k)}|R_k)$$

što, kada nema ponavljanja, je $r_k(\beta) = X_{(k)} - \overline{x}(\beta, t_k)$. U praksi, mi zamenjujemo β sa $\hat{\beta}$ i to označavamo $\hat{r_k}$.
Ako je ispunjena pretpostavka proporcionalnih rizika, tada je $E(\hat{r_k}) \approx 0$. Stoga će grafikon Šenfeldovih reziduala biti raspršen oko 0 u odnosu na vremena događaja.
Težinski Šenfeldovi reziduali [9] se mogu dobiti množenjem standardnih Šenfeldovih reziduala sa inverzom ocenjene kovarijantne matrice r_i, označenim kao $\hat{D}(r_i)$:

$$r_i^* = [\hat{D}(r_i)]^{-1} r_i.$$

Težinski Šenfeldovi reziduali imaju bolju dijagnostičku moć nego obični Šenfeldovi reziduali.

Ako se stopa rizika menja tokom vremena, tada je

$$\beta_j \equiv \beta_j + \theta_j g_j(t),$$

gde θ objašnjava kako se odnosi među osobinama menjaju tokom vremena i g_j prediktivni proces. Test statistika koja se koristi je

$$T = (\sum G_k \hat{r}_k)^T D^{-1} (\sum G_k \hat{r}_k),$$

sa

$$D = \sum G_k \hat{V}_k G_k^T - (\sum G_k \hat{V}_k)(\sum \hat{V}_k)^{-1} (\sum G_k \hat{V}_k)^T,$$

gde je $\hat{V}_k$ ocenjena disperzija β u trenutku t_k i G $p \times p$ dijagonalna matrica gde je $G_{jj}(t) = g_j(t)$. Test statistika ima približno χ^2 raspodelu sa p stepeni slobode. Ideja iza statističkog testa je da ako PH pretpostavka važi za određeni prediktor, tada Šenfeldovi reziduali za taj prediktor neće biti povezani sa vremenom preživljavanja. Ova p-vrednost se koristi za procenu PH pretpostavke za taj prediktor. Nesignifikantna (tj. velika) p-vrednost, na primer veća od $0,05$, nam ukazuje da ne postoji dovoljno jak dokaz da se odbaci nulta hipoteza.

Istraživač može doneti objektivniju odluku koristeći statistički test, nego što je obično moguće kada se koriste bilo koji od dva prethodno opisana grafička pristupa. Međutim, grafički pristup omogućava istraživaču da otkrije specifična odstupanja od PH pretpostavke; istraživač može videti šta se dešava na grafiku. Stoga preporučujemo da prilikom procene PH pretpostavke istraživač koristi i grafičke procedure i statističko testiranje pre donošenja finalne odluke.

Glava 5

Postupak stratifikacije Koksovog modela

Stratifikovani Koksov model je varijacija standardnog Koksovog modela koja se koristi kada podaci imaju više različitih grupa (stratuma). U ovom modelu, svaka grupa (stratum) tretira se zasebno, a parametri Koksovog modela se ocenjuju za svaki stratum posebno.

Ako se ustanovi da stopa rizika za neki od prediktora ne zadovoljava pretpostavku proporcionalnih rizika u Koksovom modelu, što se može identifikovati kroz primenu bilo koje od opisanih metoda u prethodnom odeljku, moguće je razmotriti primenu stratifikovanog Koksovog modela. Stratifikovani Koksov model omogućava kontrolu prediktora koji krše pretpostavku proporcionalnosti rizika putem stratifikacije. U ovom modelu, prediktori koji zadovoljavaju pretpostavku PH modela se uključuju u model kao i obično, dok se prediktor koji zahteva stratifikaciju ne uključuje među ostale prediktore, već se tretira kao nezavisan faktor. Podsetimo se, Koksov PH model ima osobinu da je količnik stopa rizika dva subjekta sa prognostičkim faktorima ili kovarijantama konstantan (ne menja se sa vremenom). Ovo znači da je odnos rizika od neuspeha za dva subjekta isti bez obzira na to koliko su dugo živeli.

Pretpostavljamo da k promenljivih ne zadovoljava pretpostavke Koksovog PH modela proporcionalnog rizika, dok p promenljivih zadovoljava. Označimo promenljive koje ne zadovoljavaju pretpostavke kao $Z_1, Z_2, \ldots, Z_k$, a promenljive koje zadovoljavaju pretpostavke kao $X_1, X_2, \ldots, X_p$. Za izvođenje postupka stratifikacije Koksovog modela, definišemo jednu novu promenljivu, koju nazivamo Z^*, koristeći Z promenljive za stratifikaciju. To postižemo formiranjem kategorija za svaki Z_i,

uključujući i one koji su neprekidne promenljive. Zatim formiramo kombinacije kategorija, i te kombinacije su naši stratumi. Ovi stratumi predstavljaju kategorije nove promenljive Z^*. Uopšteno, stratifikovana promenljiva Z^* ima k^* kategorija, gde je k^* ukupan broj kombinacija formiran nakon kategorizacije svake od Z_i promenljive.

Opšti oblik stope rizika za stratifikovani Koksov model je

$$h_g\left(t, \mathbf{X}\right) = h_{0_g}\left(t\right) \exp(X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p).$$

Ova formula sadrži indeks g koji označava g-ti stratum. Promenljiva Z^* nije eksplicitno uključena u model, jer se koristi samo za kreiranje stratuma, ali X-evi za koje se pretpostavlja da zadovoljavaju pretpostavke modela jesu.

Na primer, neka je T vreme do recidiva za pacijenta, C indikator cenzuriranja ($C = 1$ ako je događaj zabeležen, $C = 0$ ako je cenzuriran), X_1 godine pacijenata, X_2 nivo markera tumora, a X_3 vrsta primenjenog tretmana. Tada je stopa rizika u Koksovom modelu:

$$h(t) = h_0(t) \cdot \exp(X_1\beta_1 + X_2\beta_2 + X_3\beta_3),$$

gde je $h_0(t)$ bazna stopa rizika, a β_1, β_2 i β_3 su koeficijenti regresije za odgovarajuće promenljive. Ukoliko pretpostavimo da nivo markera tumora (X_2) ne ispunjava pretpostavke proporcionalnosti rizika, primenićemo stratifikaciju. Prvo, kategorizujemo X_2 u K grupa. Kombinujući promenljive, formiramo odgovarajuće stratume Z_g^*. Možemo definisati stratifikovani Koksov model:

$$h_g(t) = h_{0_g}(t) \cdot \exp(X_1\beta_1 + X_3\beta_3),$$

gde je $h_{0_g}(t)$ bazna stopa rizika za g-ti stratum. Ovim pristupom, nivo markera tumora (X_2) se tretira kao nezavisan faktor unutar svakog stratuma, što omogućava modeliranje nelinearnih efekata i kompenzaciju za nesklad s pretpostavkom proporcionalnosti rizika.

Bazna stopa rizika $h_{0_g}(t)$ može biti različita za svaki od stratuma, ali koeficijenti $\beta_1, \beta_2, \ldots, \beta_p$ su isti za sve stratume. To znači da svaki stratum ima svoju baznu stopu rizika koja nije nužno ista za sve stratume. To omogućava prilagođavanje različitim osnovnim rizicima događaja u različitim grupama. Međutim, treba napomenuti da, iako su koeficijenti regresije $\beta_1, \ldots, \beta_p$ isti za sve stratume, to se ne odnosi na sve promenljive. Koeficijenti se odnose samo na one promenljive koje su deo zajedničkog modela (bez interakcija) i nisu specifične za stratume. Ova karakteristika je poznata kao pretpostavka odsustva interakcija. Pretpostavka odsustva

interakcija podrazumeva da promenljive koje se koriste za stratifikaciju ne interaguju sa X-ovima u modelu. Budući da se bazne stope rizika razlikuju za svaki stratum, funkcije preživljavanja će biti različite. Međutim, budući da su koeficijenti uz X-ove isti za svaki stratum, ocene količnika hazarda će biti iste za sve stratume.

Maksimizujemo (parcijalnu) funkciju verodostojnosti L koja se dobija množenjem funkcija verodostojnosti za svaki stratum, gde L-ovi označavaju funkcije verodostojnosti za različite stratume, pri čemu se svaki od ovih L-ova dobija iz odgovarajuće stope rizika:

$$L = L_1 \times L_2 \times \ldots \times L_{k^*}.$$

Ukoliko bismo izostavili pretpostavku bez interakcija, model možemo zapisati kao

$$h_g(t, \mathbf{X}) = h_{0_g}(t) \exp(X_1\beta_{1_g} + X_2\beta_{2_g} + \ldots + X_p\beta_{p_g}).$$

Primetimo da u modelu sa interakcijama, svaki koeficijent ima indeks g koji označava g-ti stratum, što ukazuje da su koeficijenti različiti za svaki od stratuma. Alternativni način zapisa modela sa interakcijama uključuje proizvode koji uključuju promenljivu Z^* sa svakim od prediktora. Međutim, da bismo ispravno napisali ovaj model, moramo koristiti $k^* - 1$ pomoćne promenljive koje razlikuju kategorije promenljive Z^*; takođe, svaka od ovih pomoćnih promenljivih, koje označavamo kao $Z_1^*, Z_2^*, \ldots, Z_{k^*-1}^*$ treba biti uključena u proizvod sa svakim od X-ova. Tada je

$$\begin{aligned} h_g(t, \mathbf{X}) = \; & h_{0_g}(t) \exp(X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p \\ & + (Z_1^* \cdot X_1)\beta_{11} + \ldots + (Z_1^* \cdot X_p)\beta_{p1} \\ & + (Z_2^* \cdot X_1)\beta_{12} + \ldots + (Z_2^* \cdot X_p)\beta_{p2} \\ & + \ldots + (Z_{k^*-1}^* \cdot X_1)\beta_{1k^*-1} + \ldots + (Z_{k^*-1}^* \cdot X_p)\beta_{pk^*-1}). \end{aligned}$$

Da bismo proverili pretpostavku bez interakcije, možemo izvršiti test količnika verodostojnosti koji upoređuje (redukovani) model bez interakcije sa (potpunim) interakcijskim modelom. Nulta hipoteza je da pretpostavka bez interakcije zadovoljena. Test statistika se dobija određivanjem razlike između logaritma funkcije verodostojnosti za modele bez interakcije i interakcije. Ova test statistika, ako je nulta hipoteza tačna, ima približnu χ^2 raspodelu. Broj stepeni slobode je $p(k^* - 1)$, gde p označava broj X-ova, a k^* označava broj kategorija koje čine Z^*.

Glava 6

Proširenje Koksovog modela proporcionalnih rizika za vremenski zavisne prediktore

Važna karakteristika Koksovog modela, koja se odnosi na pretpostavku proporcionalih rizika, jeste da bazna stopa rizika zavisi od vremena t, ali ne uključuje prediktore X-eve, dok eksponencijalni izraz uključuje predikore, ali ne uključuje vreme. Prediktori u ovom slučaju nazivaju se vremenski nezavisnim prediktorima. Međutim, moguće je razmotriti prediktore koji uključuju vreme. Takvi prediktori nazivaju se vremenski zavisnim promenljivama. Ako se razmatraju vremenski zavisni prediktori, Koksov model se i dalje može koristiti, ali takav model više ne zadovoljava pretpostavku proporcionalnih rizika i naziva se prošireni Koksov model. U ovom poglavlju ćemo razmotriti vremenski zavisne promenljive i odgovarajući prošireni Koksov model.

Vremenski zavisna promenljiva se definiše kao bilo koja veličina čija vrednost za određeni subjekt može varirati tokom vremena. Većina definisanih promenljivih je oblika vremenski nezavisne promenljive pomnožene vremenom ili nekom funkcijom vremena. Ako znamo vrednost nezavisnog prediktora za određeni subjekt, sve vrednosti vremenske promenljive su potpuno definisane tokom određenog vremenskog intervala istraživanja. Npr, nivo obrazovanja može biti predstavljena kao vremenski nezavisna promenljiva pomnožena vremenom (godine). Još jedan tip vremenski zavisne promenljive je „interna" promenljiva. Dakle, razmatramo promenljive čije se vrednosti mogu menjati tokom vremena za svakog ispitanika; a u slučaju internih promenljivih, razlog za promenu vrednosti zavisi od (internih) karakteristika ili po-

našanja određenog pojedinca. Primer takve promenljive je status pušenja u vremenu t. S druge strane, promenljiva se naziva eksternom promenljivom ako se njena vrednost menja pretežno zbog spoljnih karakteristika okruženja koje mogu istovremeno uticati na više osoba. Primeri sporedne promenljive su indeks zagađenosti vazduha u vremenu t za određeno geografsko područje, status zaposlenja u vremenu t, ako glavni razlog za to da li neko ima zaposlenje ili ne zavisi više od ekonomskih uslova nego od individualnih karakteristika. Postoje i delimično interne i sporedne promenljive.

U situaciji analize preživljavanja koja uključuje i vremenski nezavisne i vremenski zavisne prediktore, možemo formulisati prošireni Koksov model koji uključuje oba tipa prediktora:

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp(\sum_{i=1}^{p_1} X_i \beta_i + \sum_{j=1}^{p_2} \delta_j X_j(t)).$$

Kao i kod Koksovog PH modela, prošireni model sadrži baznu stopu rizika $h_0\left(t\right)$ koja se množi eksponencijalnom funkcijom. Međutim, u proširenom modelu, eksponencijalni deo sadrži vremenski nezavisne prediktore, označene sa X_i, i vremenski zavisne prediktore, označene sa $X_j(t)$. Celi skup prediktora u vremenu t je označen kao $\mathbf{X}(t)$.

Metode za izvođenje statističkih zaključaka su iste kao i kod PH modela, tj. mogu se koristiti testovi zasnovani na količniku verodostojnosti i metode intervala poverenja za veliki uzorak.

Jedna važna pretpostavka proširenog Koksovog modela je da efekat vremenski zavisne promenljive $X_j(t)$ na verovatnoću preživljavanja u vremenu t zavisi od vrednosti te promenljive upravo u istom vremenu t, a ne od vrednosti u ranijem ili kasnijem vremenu. Važno je napomenuti da iako vrednosti promenljive $X_j(t)$ mogu se menjati tokom vremena, model pruža samo jedan koeficijent za svaku vremenski zavisnu promenljivu u modelu. Dakle, u vremenu t, postoji samo jedna vrednost promenljive $X_j(t)$ koja ima uticaj na rizik, ta vrednost se može menjati u vremenu t. Ipak, moguće je modifikovati definiciju vremenski zavisne promenljive kako bi se omogućio „efekat kašnjenja" (engl. lag-time effect). Kako bi se ilustrovala ideja efekta kašnjenja, pretpostavimo, na primer, da se razmatra status zaposlenja, meren nedeljno i označen kao $EMP(t)$, kao vremenski zavisna promenljiva. Tada, prošireni Koksov model koji ne uzima u obzir efekat kašnjenja pretpostavlja da efekat statusa zaposlenja na verovatnoću preživljavanja u nedelji t zavisi od posmatrane vrednosti ove promenljive u istoj nedelji t, a ne, na primer, u prethodnoj nedelji. Međutim,

ako bismo omogućili, recimo, kašnjenje od jedne nedelje, status zaposlenja može biti modifikovan tako da hazardni model u vremenu t predviđa status zaposlenja u nedelji $t-1$. Dakle, promenljiva $EMP(t)$ se zamenjuje u modelu promenljivom $EMP(t-1)$.

Uopšteno govoreći, prošireni Koksov model se može alternativno napisati kako bi se omogućila modifikacija vremenski zavisne promenljive od interesa u skladu s efektom kašnjenja. Ako označimo sa L_j vrednost kašnjenja određenu za vremenski zavisnu promenjivu j, tada se „prošireni model sa kašnjenjem" može zapisati kao

$$h\left(t,\mathbf{X}\left(t\right)\right)=h_0\left(t\right)\exp(\sum_{i=1}^{p_1}X_i\beta_i+\sum_{j=1}^{p_2}\delta_jX_j(t-L_j)).$$

Napomena: Promenljiva $X_j(t)$ u prethodnoj verziji proširenog modela sada je zamenjena promenljivom $X_j(t-L_j)$.

Sada opisujemo formulu za količnik hazarda koji proizlazi iz proširenog Koksovog modela. Najvažnija karakteristika ove formule je da pretpostavka proporcionalnih stopa rizika više nije ispunjena. Opšta formula za količnik hazarda u proširenom Koksovom modelu je prikazana ovde

$$\widehat{HR}\left(t\right)=\frac{\hat{h}\left(t,\mathbf{X}^*\left(t\right)\right)}{\hat{h}\left(t,\mathbf{X}\left(t\right)\right)}=\exp(\sum_{i=1}^{p_1}(X_i^*-X_{\_i})\hat{\beta}_i+\sum_{j=1}^{p_2}\hat{\delta}_j(X_j^*\left(t\right)-X_j(t))).$$

Ova formula opisuje odnos rizika u određenom trenutku t i zahteva definisanje prediktora dva subjekta u tom trenutku t. Ova dva skupa se označavaju kao $\mathbf{X}^*(t)$ i $\mathbf{X}(t)$. Dva skupa prediktora, $\mathbf{X}^*(t)$ i $\mathbf{X}(t)$, identifikuju dve specifikacije u trenutku t za kombinovani skup prediktora koji sadrže kako vremenski nezavisne tako i vremenski zavisne promenljive. Individualne komponente za svaki skup prediktora su $\mathbf{X}^*=(X_1^*,X_2^*,...,X_{p_1}^*,X_1^*\left(t\right),X_2^*\left(t\right),...,X_{p_2}^*\left(t\right))$ i $\mathbf{X}=(X_1,X_2,...,X_{p_1},X_1(t),X_2(t),...,X_{p_2}(t))$. Budući da opšta formula za količnik hazarda uključuje razlike u vrednostima vremenski zavisnih promenljivih u trenutku t, ovaj količnik hazarda je funkcija vremena. Stoga, uopšteno gledano, prošireni Koksov model ne zadovoljava pretpostavku proporcionalnih rizika ako je bilo koji δ_j različit od nule. Važno je napomenuti da u formuli za količnik hazarda, koeficijent $\hat{\delta}_j$ koji se odnosi na razliku u vrednostima j-te vremenski zavisne promenljive nije sam po sebi vremenski zavisan uzimajući u obzir sva vremena u kojim je ova promenljiva izmerena u istraživanju.

Sada ćemo opisati kako koristiti prošireni Koksov model kako bismo proverili pretpostavku proporcionalnih rizika za vremenski nezavisne promenljive i procenili

efekat promenljive, koja ne zadovoljava pretpostavku proporcionalnih rizika.
Ako skup podataka za naše istraživanje sadrži nekoliko, recimo p, vremenski nezavisnih promenljivih, možda bismo želeli da prilagodimo Koksov PH model koji sadrži svaku od ovih promenljivih

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp(\sum_{i=1}^{p} X_i \beta_i).$$

Međutim, kako bismo procenili da li je takav PH model prikladan, možemo proširiti ovaj model definisanjem nekoliko izvedenih veličina koji uključuju svaku vremenski nezavisnu promenljivu sa nekom funkcijom vremena. Drugim rečima, ako se i-ta vremenski nezavisna promenljiva označava kao X_i, tada možemo definisati i-tu proizvodnu veličinu kao $X_i \cdot g_i(t)$, gde je $g_i\left(t\right)$ neka funkcija vremena za i-tu promenljivu. Prošireni Koksov model koji istovremeno uzima u obzir sve vremenski nezavisne promenljive od interesa prikazan je

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp(\sum_{i=1}^{p} X_i \beta_i + \sum_{i=1}^{p} \delta_i X_i g_i(t)).$$

Prilikom korišćenja proširenog modela, ključna odluka je oblik koji funkcije $g_i\left(t\right)$ treba da imaju. Najjednostavniji oblik za $g_i\left(t\right)$ je da su sve $g_i\left(t\right)$ identično jednake nuli u bilo kom trenutku; to je još jedan način izražavanja originalnog PH modela koji ne sadrži vremenski zavisne promenljive. Drugi izbor za $g_i\left(t\right)$ je da je $g_i\left(t\right) = t$. To implicira da za svaki X_i u modelu postoji odgovarajuća vremenski zavisna promenljiva u obliku $X_i \cdot t$. S druge strane, pretpostavimo da želimo da se fokusiramo na određenu vremenski nezavisnu promenljivu, recimo promenljivu X_L. Tada bi $g_i\left(t\right)$ bilo jednako t za $i = L$, ali bi bilo jednako nuli za sve ostale i. Odgovarajući prošireni Koksov model bi tada sadržao samo jednu proizvodnu veličinu $X_L \cdot t$. Još jedan izbor za $g_i(t)$ je logaritam od t, tako da će odgovarajuće vremenski zavisne promenljive biti oblika $X_i \cdot \ln t$. Još jedan izbor bi bio da $g_i(t)$ bude Hevisajdova funkcija oblika $g\left(t\right) = \begin{cases} 1 & , t \geq t_0 \\ 0 & , t < t_0 \end{cases}$.

Nulta hipoteza je da su svi δ_i jednaki nuli. Pod nultom hipotezom model se svodi na PH model. Koristitimo test količnika verodostojnosti

$$X_L = 2(\ln L_{proširenCM} - \ln L_{PHCM}), \tag{6.1}$$

gde je $L_{proširenCM}$ parcijalna funkcija verodostojnosti za prošireni Koksov model, a L_{PHCM} parcijalna funkcija verodostojnosti za PH model. Test statistika koja se tako

dobija ima približno χ^2 raspodelu sa p stepeni slobode pod nultom hipotezom, gde p označava broj parametara koji se postavljaju na nulu pod H_0 [8]. Ako se utvrdi da je test odstupa od nulte hipoteze, tj. p-vrednost je manja od unapred određenog nivoa značajnost, možemo zaključiti da PH pretpostavka nije zadovoljena za barem jedan od prediktora u modelu. Da bismo utvrdili koji prediktor(i) ne zadovoljava(ju) PH pretpostavku, možemo postupiti s eliminacijom proizvoda unazad sve dok se ne postigne konačni model.

Sada ćemo opisati upotrebu Hevisajdove funkcije ili odskočne funkcije (engl. heaviside function). Kao jednostavan primer, pretpostavimo da model sadrži samo jedan nezavisan prediktor koji ne zavisi od vremena, tačnije, status izloženosti E, primenljiva koja uzima vrednosti 0 i 1. Kada se koristi ova funkcija, formula za količnik hazarda daje konstantne količnike hazarda za različite vremenske intervale. Neka je

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp(\beta E + \delta E g(t)).$$

Za $t \geq t_0$ važi $g\left(t\right) = 1$, pa je $E \cdot g\left(t\right) = E$, tj. stopa rizika je

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp((\beta + \delta) E).$$

Odgovarajući količnik hazarda postaje:

$$\widehat{HR} = \exp(\hat{\beta} + \hat{\delta}).$$

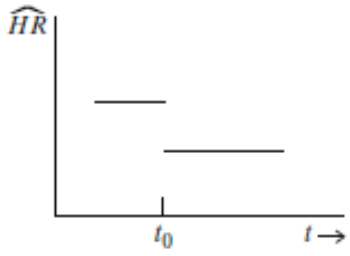
Za $t < t_0$ važi $g\left(t\right) = 0$, pa je $E \cdot g\left(t\right) = 0$, tj. stopa rizika je:

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp(\beta E).$$

Odgovarajući količnik hazarda postaje:

$$\widehat{HR} = \exp(\hat{\beta}).$$

Dakle, pokazali smo da upotreba jedne odskočne funkcije rezultira proširenim Koksovim modelom koji daje dve vrednosti količnika hazarda, pri čemu svaka vrednost ostaje konstantna tokom fiksnog vremenskog intervala (vidi Sliku 6.1).



Slika 6.1: Količnik hazarda kad je $g(t)$ odskočna funkcija [8]

Postoji način da se ovaj model zapiše koristeći dve odskočne funkcije u istom modelu. Ovaj alternativni model je

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp(\delta_1\left(E \cdot g_1\left(t\right)\right) + \delta_2\left(E \cdot g_2\left(t\right)\right))$$

Dve odskočne funkcije označene su $g_1\left(t\right)$ i $g_2\left(t\right)$. Svaka od ovih funkcija se nalazi u modelu kao deo proizvodne veličine sa promenljivom izloženosti E.
Funkcija $g_1\left(t\right)$ ima vrednost $g_1\left(t\right) = \begin{cases} 1 & , t \geq t_0 \\ 0 & , t < t_0 \end{cases}$, dok funkcija $g_2\left(t\right)$ ima oblik $g_2\left(t\right) = \begin{cases} 1 & , t < t_0 \\ 0 & , t \leq t_0 \end{cases}$.
Alternativni model i odgovarajući količnik hazarda mogu se izraziti na sledeći način. Za vreme $t \geq t_0$, funkcije $g_1\left(t\right)$ i $g_2\left(t\right)$ imaju vrednosti $g_1\left(t\right) = 1$ i $g_2\left(t\right) = 0$. U ovom slučaju, stopa rizika postaje

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp\left(\delta_1\left(E \cdot 1\right) + \delta_2\left(E \cdot 0\right)\right) = h_0\left(t\right) \exp\left(\delta_1 E\right),$$

a odgovarajući količnik hazarda je

$$\widehat{HR} = \exp(\hat{\delta}_1).$$

Za vreme $t < t_0$, funkcije $g_1\left(t\right)$ i $g_2\left(t\right)$ imaju vrednosti $g_1\left(t\right) = 0$ i $g_2\left(t\right) = 1$. Sada je stopa rizika

$$h\left(t, \mathbf{X}\left(t\right)\right) = h_0\left(t\right) \exp\left(\delta_1\left(E \cdot 0\right) + \delta_2\left(E \cdot 1\right)\right) = h_0\left(t\right) \exp\left(\delta_2 E\right),$$

a količnik hazarda

$$\widehat{HR} = \exp(\hat{\delta}_2).$$

Matematički gledano, ove vrednosti su iste kao one dobijene iz originalnog modela koji sadrži samo jednu odskočnu funkciju. Drugim rečima, $\hat{\delta}_1$ u alternativnom modelu je jednako $\hat{\beta} + \hat{\delta}$ u originalnom modelu (koji sadrži jednu odskočnu funkciju), dok je $\hat{\delta}_2$ u alternativnom modelu jednako $\hat{\beta}$ u originalnom modelu. Dakle, videli smo da se odskočne funkcije mogu koristiti kako bi se dobili ocenjeni količnici hazarda koji ostaju konstantni unutar svakog od dva odvojena vremenska intervala praćenja. Takođe možemo proširiti upotrebu odskočnih funkcija kako bismo dobili nekoliko različitih procena količnika hazarda koji ostaju konstantni unutar nekoliko vremenskih intervala.

Kao i kod jednostavnijeg Koksovog PH modela, regresioni koeficijenti u proširenom Koksovom modelu se ocenjuju korišćenjem metode maksimalne verodostojnosti. Ocene se dobijaju maksimizacijom (delimične) funkcije verodostojnosti L.

Međutim, računanje za prošireni Koksov model je složenije u odnosu na Koksov PH model, jer su skupovi rizika koji se koriste za formiranje funkcije verodostojnosti složeniji sa vremenski zavisnim promenljivama.

Glava 7

Analiza preživljavanja pacijenata sa dijabetičkom retinopatijom

7.1 Uvod u istraživanje

Dijabetička retinopatija je komplikacija kod pacijenata koji imaju šećernu bolest (diabetes mellitus) koja često može dovesti do gubitka vida.

Posmatramo studiju o 197 pacijenata koji su nasumično odabrani kao pacijenti sa visokim rizikom od dijabetičke retinopatije. Svaki pacijent je podvrgnut laserskom tretmanu na jednom oku, dok drugo oko nije tretirano. Dakle, skup podataka ima dva unosa za svakog pacijenta. Vreme koje je prošlo od početka tretmana do trenutka kada je vidna oštrina opala ispod 5/200 u dve uzastopne posete predstavlja događaj od interesa za svako oko. Važno je napomenuti da postoji ugrađeno kašnjenje od otprilike 6 meseci između početka tretmana i početka merenja vidne oštrine (posete su bile svaka 3 meseca). Vrednosti preživljavanja u ovom skupu podataka predstavljaju stvarno vreme do gubitka vida u mesecima, sa oduzimanjem minimalnog mogućeg vremena do događaja (6 i po meseci). Cenzurisanje se dogodilo usled smrti, prestanka praćenja pacijenata ili završetka studije. Događaj od interesa je gubitak vida.

Analiza preživljavanja i statistička obrada podataka izvršene su primenom softverskog okruženja R (verzija 4.3.1) i korišćenjem paketa `survival` [16].

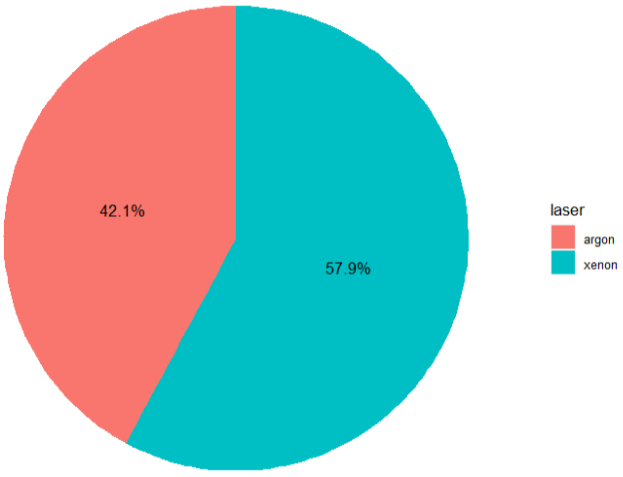
7.2 Opis skupa podataka

Koristimo ugrađenu bazu podataka `diabetic` paketa `survival`. Baza podataka sadrži 394 opservacije za 197 pacijenata.

- `id` jedinstveni id subjekta;
- `laser` tip korišćenog lasera, xenon i argon;
- `age` sa koliko godina je dijagnostikovan dijabetes;
- `eye` oko koje je lečeno, right i left;
- `trt` 0 = kontrolno oko, 1 = lečeno oko;
- `risk` skor rizika za oko vrednosti od 6 do 12;
- `time` vreme do gubitka vida ili poslednjeg praćenja;
- `status` 0 = cenzurisan; 1 = gubitak vida.

7.3 Analiza skupa podataka

Kao što je već pomenuto, posmatramo 197 pacijenata i za svakog pacijenta pratimo oba oka, pri čemu se jedno oko tretira laserom, a drugo ne. Dakle, imamo 394 jedinstvene vrste u bazi, od kojih je 239 cenzurisano. Takođe, na Slici 7.1 je prikazano da se xenon laser češće koristi u odnosu na argon laser.



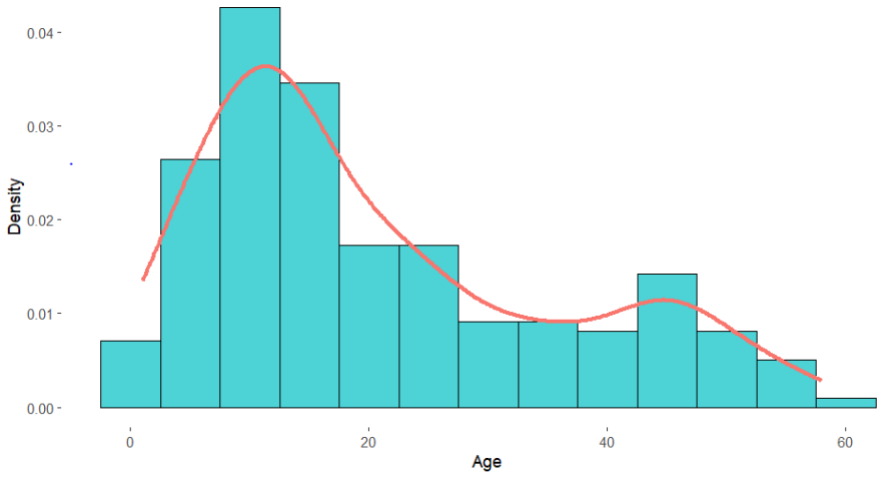
Slika 7.1: Pitica primenjenog lasera

Tabela 7.1 prikazuje primenjene laserske tretmane na svako oko (levo i desno) u skupu podataka. Ova tabela omogućava vizualnu procenu učestalosti primenjenih laserskih tretmana, što omogućava upoređivanje između oba oka. Primećujemo da se desno oko tretira češće, kao i da se laser `xenon` nešto učestalije primenjuje.

	left	right
xenon	47	67
argon	42	41

Tabela 7.1: Tabela kontingencije primenjenih laserskih tretmana po oku

Na Slici 7.2 prikazan je histogram koji prikazuje raspodelu godina u kojima se javlja dijabetes. Medijana iznosi 16, dok je srednja vrednost 20.78.

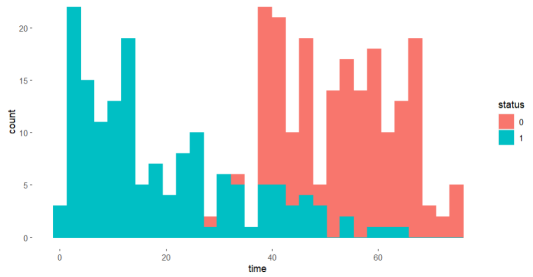


Slika 7.2: Histogram godina u kojima se javlja dijabetes

Dijabetes se deli na dva tipa, tip 1 i tip 2, u zavisnosti od godina kada se javlja. Tip 1 se javlja pre 20. godine i naziva se još juvenilni dijabetes, dok se tip 2 naziva adultni dijabetes. Kako bismo bolje razumeli naše podatke, kreiraćemo još jedan prediktor `type` koji će predstavljati kategorizaciju neprekidne promenljive `age`. Novi prediktor će uzimati dve vrednosti: `juvenile` i `adult`, a prelomna tačka će biti postavljena na 20. U našem skupu podataka, primećujemo da je više slučajeva juvenilnog dijabetesa, tačnije oko 60%.

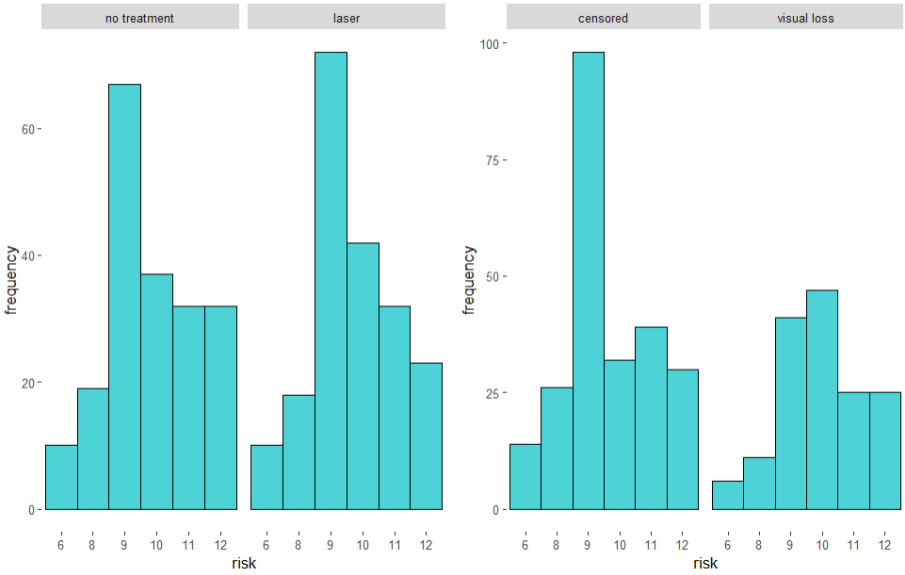
Slika 7.3, nam pruža vizuelnu raspodelu vrednosti prediktora `time` za svaku od kategorija prediktora `status`. Ono što možemo primetiti jeste da se nakon 40-te

godine povećava verovatnoća gubitka vida, što ćemo kasnije proveriti i primenom Koksovog modela.



Slika 7.3: Histogram promenljive `time` prema kategorijama promenljive `status`

Histogrami apsolutnih frekvencija na Slici 7.4 prikazuju raspodelu rizika grupisanu po kategorijama `trt` i `status`, redom. Hipoteza koju možemo testirati u narednom poglavlju je da li tretman laserom utiče na ocenu rizika, kao i da li ocena rizika može predvideti gubitak vida.

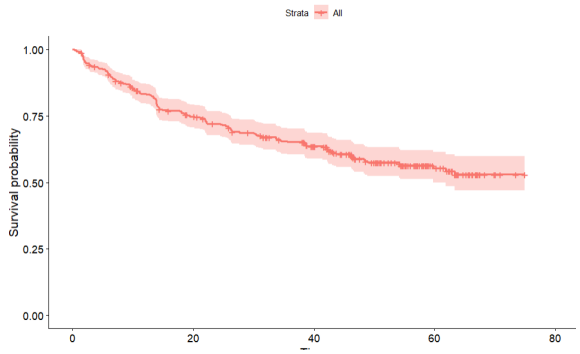


Slika 7.4: Histogram promeljive `risk` prema kategorijama promenljivih `trt` i `status` redom

7.4 Analiza preživljavanja

Prvo ćemo pokušati uočiti razlike u vremenima preživljavanja koristeći odgovarajuće Kaplan-Majerove krive.

Na osnovu ocenjene Kaplan-Majerove krive preživljavanja koristeći podatke iz kolona `time` i `status`, može se zaključiti da je početna verovatnoća preživljavanja veoma visoka. Međutim, kako vreme prolazi, verovatnoća preživljavanja pacijenata postepeno opada (vidi Sliku 7.5 i Tabelu 7.2). Ovo ukazuje na potrebu za daljim istraživanjem i eventualno preduzimanjem mera kako bi se poboljšalo preživljavanje očiju pacijenata tokom vremena.



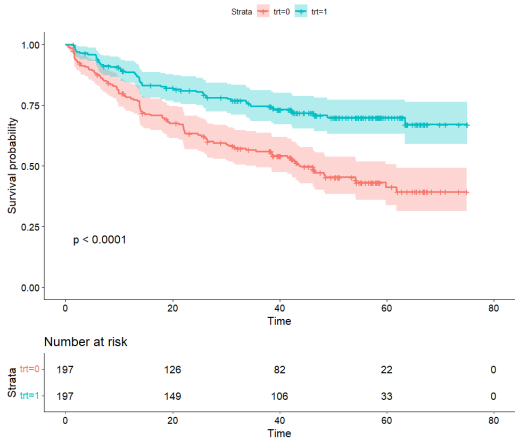
Slika 7.5: Kaplan-Majerova kriva preživljavanja očiju za pacijente sa dijabetesom

Tabela 7.2: `Call: survfit(formula = Surv(time, status) 1, data = diabetic_as_tibble)`

Time	n.risk	n.event	Survival	Std. Err	Lower 95% CI	Upper 95% CI
0.3	394	1	0.997	0.00253	0.993	1.000
8.3	335	49	0.871	0.01700	0.839	0.905
16.3	286	38	0.770	0.02154	0.729	0.814
24.3	259	18	0.721	0.02307	0.677	0.768
32.3	227	18	0.670	0.02436	0.624	0.720
40.3	185	11	0.636	0.02518	0.589	0.688
48.3	128	13	0.585	0.02695	0.534	0.640
56.3	84	4	0.564	0.02791	0.512	0.622
64.3	36	3	0.531	0.03235	0.471	0.598
72.3	5	0	0.531	0.03235	0.471	0.598

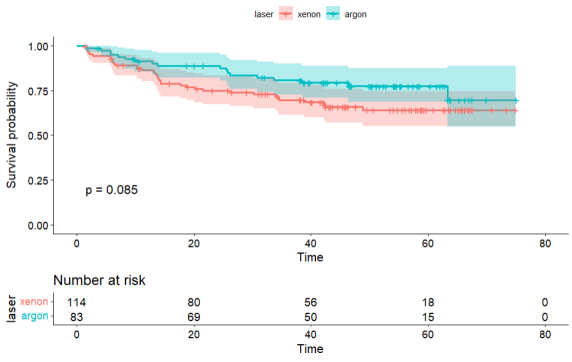
Na Slici 7.6 Kaplan-Majerovih krivih primećuje se veća verovatnoća preživljavanja očiju pacijenata tretiranih laserom u poređenju sa očima koje nisu tretirane. Takođe, kako vreme raste, dve Kaplan-Majerove krive se sve više razdvajaju, što

ukazuje na to da su pozitivni efekti tretmana bolji što duže pacijent ostaje u remisiji. Navedeno je potvrđeno rezultatima testa logaritmovanih rangova, gde se dobija vrednost statistike $\chi^2 = 22.2$ i vrlo niska p-vrednost ($p = 2e - 06$), što ukazuje na statistički značajnu razliku u preživljavanju između ove dve grupe, u korist grupe tretirane laserom. Ostaje da proverimo koji laser daje bolje rezultate u lečenju.



Slika 7.6: Kaplan-Majerova kriva preživljavanja za lečene i nelečene oči pacijenata

Sada ćemo izvršiti analizu preživljavanja koristeći podatke o vremenu i statusu, ali fokusirano samo na oko pacijenta koje je tretirano laserom. U testu će učestvovati 197 pacijenata (114 u grupi `laser=xenon` i 83 u grupi `laser=argon`). Primetili smo da pacijenti tretirani laserom `argon` imaju veću verovatnoću preživljavanja u poređenju sa pacijentima tretiranim laserom `xenon` tokom posmatranog vremenskog perioda (vidi Sliku 7.7).



Slika 7.7: Kaplan-Majerova kriva preživljavanja očiju pacijenata tretiranih laserom

Da bismo proverili da li postoji statistički značajna razlika u preživljavanju izme-

đu grupa, primenili smo test logaritmovanih rangova. Vrednost statistike χ^2 iznosi 3, a test se vrši sa 1 stepenom slobode. Što je vrednost statistike veća, to ukazuje na veću razliku između grupa. p-vrednost testa logaritmovanih rangova iznosi 0.09 što je niže od nivoa značajnosti od 0.1, pa ćemo odbaciti nultu hipotezu o jednakosti preživljavanja između grupa. S obzirom na ove rezultate, detaljnije ćemo proučiti podatke kako bismo bolje razumeli eventualne razlike u preživljavanju između pacijenata tretiranih različitim laserima.

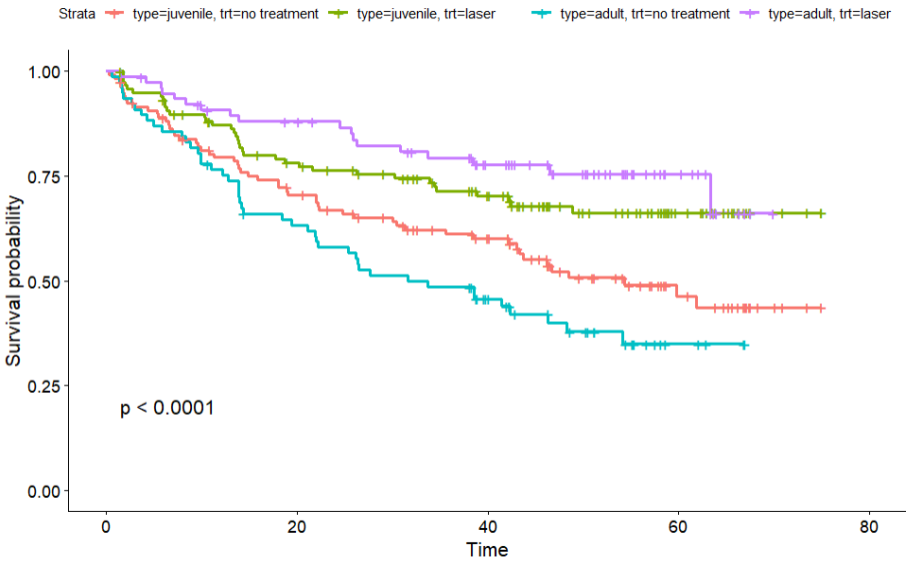
Analiza preživljavanja je izvršena na podacima o vremenu i statusu pacijenata tretiranih različitim laserima i na različitim očima:

- Pacijenti tretirani laserom `xenon` na levom oku: tokom posmatranog vremenskog perioda, verovatnoća preživljavanja se smanjuje, ali je i dalje relativno visoka i iznosi 0.766;
- Pacijenti tretirani laserom `xenon` na desnom oku: tokom vremenskog perioda, verovatnoća preživljavanja opada i iznosi 0.555;
- Pacijenti tretirani laserom `argon` na levom oku: tokom vremenskog perioda, verovatnoća preživljavanja je relativno visoka i iznosi 0.890;
- Pacijenti tretirani laserom `argon` na desnom oku: tokom vremenskog perioda, verovatnoća preživljavanja opada i iznosi 0.670.

Ukupno gledano, rezultati ukazuju na razlike u preživljavanju među ovim grupama pacijenata tretiranih različitim tipovima lasera i sa različitim okom. Da bismo proverili statističku značajnost ovih razlika, primenili smo test logaritmovanih rangova. Vrednost statistike χ^2 iznosi 11.6, a test se vrši sa 3 stepena slobode. p-vrednost testa logaritmovanih rangova iznosi 0.009, što je statistički značajno. Ova niska p-vrednost sugeriše da postoje statistički značajne razlike u preživljavanju između ovih grupa pacijenata. Dakle, ovi rezultati ukazuju na to da je verovatnoća preživljavanja pacijenata značajno različita u zavisnosti od tipa lasera i oka koji su korišćeni u njihovom tretmanu.

Preostaje nam da istražimo verovatnoću preživljavanja pacijenata, uzimajući u obzir njihovu ocenu rizika i da li su bili podvrgnuti terapiji. Analizom grafika i testa logaritmovanih rangova, potvrđeno je postojanje značajne statističke razlike u stopama preživljavanja pacijenata, što je u vezi sa njihovom ocenom rizika i primenjenom terapijom. Pacijenti sa različitim ocenama rizika i oni koji su primili terapiju pokazali su različite stope preživljavanja, što ukazuje na važnost ocene rizika i terapije u prognozi njihovog preživljavanja.

Potvrđeno je postojanje razlike u preživljavanju u zavisnosti od toga da li je pacijent lečen ili ne. Dalje, istražujemo da li postoji razlika u preživljavanju u odnosu na primenjenu terapiju kod pacijenata koji su dobili dijabetes pre ili posle 20-te godine. Na osnovu Kaplan-Majerovih krivih, zaključujemo da pacijenti sa adultnim dijabetesom koji nisu lečeni imaju nižu verovatnoću preživljavanja u odnosu na pacijente sa juvenilnim dijabetesom koji nisu lečeni. Primećujemo da lečeni pacijenti sa adultnim dijabetesom imaju bolju prognozu u odnosu na lečene pacijente sa juvenilnim dijabetesom (vidi Sliku 7.8). Međutim, nakon 60-tih godina, razlika u preživljavanju između ova dva tipa dijabetesa postaje neprimetna.



Slika 7.8: Kaplan-Majerova kriva preživljavanja u zavisnosti od terapije i tipa dijabetesa

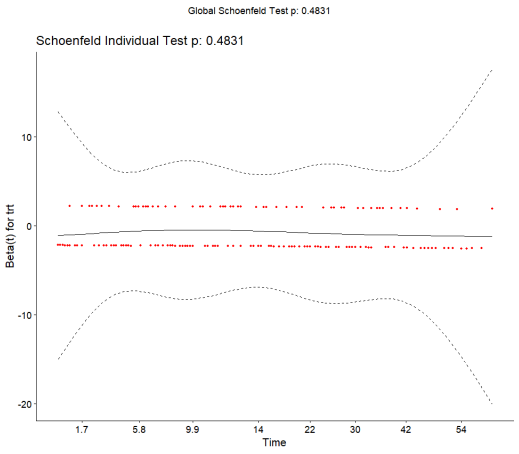
Kaplan-Majerove krive su dobre za vizualizaciju razlika u preživljavanju između dve kategorije, a test logaritmovanih rangova je koristan za utvrđivanje da li postoje razlike u preživljavanju između različitih grupa. Međutim, ne odgovaraju nam na pitanje koliko je jedna grupa subjekata u većem riziku u odnosu na drugu. Da bismo dobili takve odgovore, koristimo Koksov model s proporcionalnim rizicima za analizu podataka. Koksov model omogućava nam bolje razumevanje veze između prediktora i preživljavanja te kvantifikaciju uticaja tih prediktora na rizik preživljavanja u različitim grupama subjekata.

Definišemo Koksov model koji zavisi od prediktora `trt`. Na osnovu p-vrednosti od $1.39e-07$, zaključujemo da postoji statistički značajna razlika u riziku između

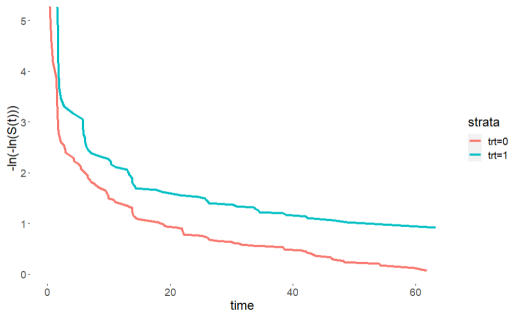
očiju pacijenata koji su koristili laser. Preciznije, oko pacijenata na koje je primenjen laser za 46% je manje verovatno da će doživeti događaj u poređenju sa okom pacijenata koje nije lečeno. Podsetimo se da smo testom logaritmovanih rangova dobili p-vrednost $p = 2e - 06$.

Da bismo testirali da li navedeni Koksov model zadovoljava pretpostavku proporcionalnosti, koristimo test saglasnosti koji se zasniva na Šenfeldovim rezidualima. Postupak je sledeći:

- Definišemo Koksov PH model i dobijemo Šenfeldove reziduale za svaki prediktor;
- Kreiramo promenljivu koja rangira redosled grešaka. Subjekt koji ima prvi (najraniji) događaj dobija vrednost 1, sledeći dobija vrednost 2, itd.
- Koristeći test zasnovan na ponderisanim Šenfeldovim rezidualima, ispitujemo korelaciju između promenljivih kreiranih u prvom i drugom koraku. Nulta hipoteza je da je korelacija između Šenfeldovih reziduala i rangiranog vremena otkaza nula.



Slika 7.9: Šenfeldovi reziduali - provera pretpostavke proporcionalnosti za promenljivu `trt`

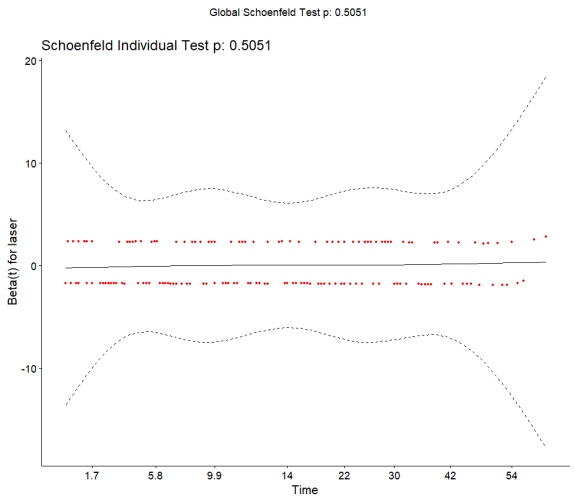


Slika 7.10: $-\ln-\ln$ krive - provera pretpostavke proporcionalnosti za promenljivu `trt`

Rezultati za `trt` pokazuju χ^2 statistiku od 0.492 sa 1 stepenom slobode i p-vrednošću od 0.48, što znači da test nije statistički značajan, tj. `trt` zadovoljava pretpostavku proporcionalnih rizika. Na grafiku Šenfeldovih reziduala puna linija predstavlja glatki splajn koji se uklapa u dijagram. Isprekidane linije označavaju

+/- 2-standardne greške oko uklapanja. Iz analize grafika možemo zaključiti da nema uočljivih obrazaca sa vremenom. Da li je ispunjena pretpostavka proporcionalnih rizika možemo da proverimo i poređenjem $-\ln-\ln$ krivih. Dodatno, poređenjem $-\ln-\ln$ krivih, primećujemo da se dve krive ne seku, stoga, vodeći se konzervativnim pristupom, ponovo zaključujemo da `trt` zadovoljava pretpostavku proporcionalnih rizika. Više detalja možete pronaći na Slici 7.10.

Definišemo Koksov model koji zavisi od prediktora `laser`. Na osnovu p-vrednosti od 0.89 dobijamo da ne postoji razlika da li se koristi argon ili xenon laser, tj. da je rizik za xenon laser veći za 2% u odnosu na argon laser. Da bismo testirali da li navedeni Koksov model zadovoljava pretpostavke proporcionalnosti, koristimo test saglasnosti koji se zasniva na Šenfeldovim rezidualima. Rezultati za `laser` pokazuju χ^2 statistiku od 0.444 sa 1 stepenom slobode i p-vrednošću od 0.51, što znači da test nije statistički značajan, tj. možemo da zaključimo da su zadovoljene pretpostavke proporcionalnosti rizika za `laser` (koji ima dva vrednosti, argon i xenon, predstavljena sa dve trake na grafikonu 7.11).



Slika 7.11: Šenfeldovi reziduali - provera pretpostavke proporcionalnosti za promenljivu `laser`

Ostali prediktori (`type`, `age`, `eye` i `risk`) zadovoljavaju pretpostavku proporcionalnih rizika jer su p-vrednosti za test proporcionalnih rizika (izvedenih pomoću funkcije cox.zph) veće od 0.05. Međutim, postoji statistički značajna razlika u riziku između grupe sa niskim rizikom i grupe sa visokim rizikom (`risk` promenljiva). Naime, p-vrednost za `risk` je 0.0137, što je manje od uobičajenog nivoa značajnosti

od 0.05. Ovo ukazuje da postoji verovatno značajna razlika u riziku između ovih grupa i da `risk` ima statistički značajan uticaj na preživljavanje. S druge strane, `eye` promenljiva ima p-vrednost blizu 0.05(0.0531). U ovakvim situacijama, može biti korisno obratiti pažnju na ovu promenljivu i dalje istražiti njen uticaj na preživljavanje.

Konačan Koksov model, koji je prikazan u Tabeli 7.3 predstavlja regresiju sa sledećim prediktorima: `eye`, `risk` i `trt`. Promenljive su statistički značajne u modelu, jer imaju p-vrednosti manje od 0.05. Iz Tabele 7.3 zaključujemo da povećanje prediktora `risk` za 1 jedinicu dovodi do povećanja rizika za 15.3%, što je u skladu sa rezultatima prikazanim na Slici 7.4. Takođe, pacijenti koji su lečeni laserom imaju 55.9% manji rizik od gubitka vida. Desno oko ima veći rizik u poređenju sa levim oko što ukazuje na razliku u preživljavanju u zavisnosti od oka.

Tabela 7.3: `Call: coxph(formula = Surv(time, status) eye + risk + trt, data = diabetic_as_tibble, cluster = id)`

	coef	exp(coef)	se(coef)	robust se	z	p
eyeright	0.34619	1.41366	0.16272	0.14420	2.401	0.0164
risk	0.14215	1.15275	0.05563	0.05919	2.402	0.0163
trtlaser	-0.81812	0.44126	0.16983	0.15331	-5.337	9.47e-08

Likelihood ratio test=33.89 on 3 df, p=2.093e-07
n= 394, number of events= 155

AIC (engl. Akaike Information Criterion) je informacioni kriterijum koji se koristi za procenu kvaliteta modela i njegova svrha je da balansira preciznost modela i njegovu složenost. Model sa nižim AIC vrednostima smatra se boljim, jer ukazuje na bolje prilagođavanje podataka sa manje parametara u modelu. AIC se računa kao $AIC = -2\ln(L) + 2k$ gde je $\ln(L)$ logaritam verovatnoće najverovatnijeg procenjenog modela, a k je broj parametara u modelu. BIC (engl. Bayesian Information Criterion) je sličan AIC-u, ali dodatno uzima u obzir veličinu uzorka i kažnjava modele sa većim brojem parametara. BIC favorizuje jednostavnije modele od AIC-a. BIC se izračunava kao $BIC = -2\ln(L) + k \cdot \ln(n)$ gde je n broj promatranja u uzorku, a ostale oznake imaju isto značenje kao kod AIC-a. Obično se preferira model sa nižim AIC ili BIC vrednostima, jer takav model pruža balans između preciznosti i jednostavnosti.

Među svim modelima koje smo konstruisali u Dodatku, onaj koji uključuje pre-

diktore `eye`, `risk` i `trt` ima najmanje AIC i BIC vrednosti, što ukazuje na to da je taj model najbolje prilagođen podacima i da je verovatno najbolji među svim modelima koje smo razmatrali.

C-indeks (engl. concordance) je statistička mera koja se često koristi u analizi preživljavanja, posebno u Koksovom proporcionalnom hazard modelu. Ona se koristi za procenu koliko dobro model prepoznaje i rangira pacijente prema njihovoj verovatnoći preživljavanja. C-indeks upoređuje parove pacijenata koji imaju slične vrednosti prediktora, tj. koji su u sličnim rizicima, i gleda koliko dobro model može da predvidi koji će od ta dva pacijenta preživeti duže. Ako model tačno predviđa da će pacijent sa većim rizikom preživeti kraće od pacijenta sa manjim rizikom, to se smatra skladnim (saglasnim) parom. Ako model pogrešno predviđa ishod, to se naziva neskladnim (nesaglasnim) parom. Vrednosti konkordancije kreću se od 0 do 1, pri čemu vrednost 1 znači da model savršeno predviđa redosled preživljavanja. Naš konačan model ima c-indeks od 0.5882, što ukazuje na umereno dobru prediktivnu sposobnost našeg modela.

7.5 Zaključak

Dijabetička retinopatija je poremećaj mrežnjače kod pacijenata sa dijabetesom melitusom. Kod osoba sa dijabetesom, visok nivo šećera u krvi postepeno začepljuje krvne sudove, što dovodi do smanjenog dotoka krvi u retinu i može izazvati dijabetičku retinopatiju. Stoga je važno sprovoditi analizu dijabetičke retinopatije kod dijabetičkih pacijenata kako bi se sprečila slepilo.

Terapija se čini efikasnom, pri čemu je ovaj efekat mnogo izraženiji kod odraslih osoba sa dijabetesom u poređenju sa osobama sa juvenilnim dijabetesom. Rezultati Koksovog modela pokazali su da nisu sve nezavisne promenljive značajne u modelu. Budući da neke nezavisne promenljive nisu značajne u modelu u ovom istraživanju, može se dodati više nezavisnih promenljivih u model kako bi se dobio bolji model.

Glava 8

Zaključak

U radu smo obuhvatili teorijsku i praktičnu analizu Koksovog proporcionalnog modela u kontekstu analize preživljavanja. Kroz teorijski deo, stekli smo dublje razumevanje osnovnih principa Koksovog modela i njegove primene u analizi preživljavanja. U praktičnom delu, primenili smo Koksov model na stvarne podatke i istražili njegovu efikasnost u modeliranju rizika preživljavanja u odnosu na različite faktore.

Koksov proporcionalni model predstavlja moćan alat koji omogućava modeliranje preživljavanja uzimajući u obzir više nezavisnih promenljivih. Njegove prednosti proističu iz sposobnosti da se nosi sa cenzurisanim podacima, koji su često prisutni u analizi preživljavanja. Osim toga, ovaj model omogućava procenu uticaja više faktora na preživljavanje istovremeno, što daje celovitu sliku o faktorima koji utiču na ishod. Koksov proporcionalni model ne zahteva stroge pretpostavke o distribuciji vremena do događaja, što ga čini robustnim i pogodnim za analizu različitih tipova podataka. Međutim, važno je napomenuti da pretpostavka proporcionalnih rizika može biti nerealna u nekim situacijama, te je potrebno pažljivo proceniti njen adekvatan izbor.

Još jedna prednost Koksovog modela je sposobnost procene osnovne stope rizika i funkcije preživljavanja čak i kada sama osnovna stopa rizika nije potpuno određena. Ovo nam pruža bitne informacije o preživljavanju i uticaju faktora od interesa na ishod, bez potrebe za dodatnim pretpostavkama.

Ipak, važno je uzeti u obzir i ograničenja modela kako bismo pravilno interpretirali rezultate i donosili odluke u praksi. Interpretacija Koksovog modela može biti izazovna, jer količnici hazarda predstavljaju relativne rizike i zahtevaju pažljivo tumačenje kako bi se doneli relevantni zaključci.

Zaključno, rezultati ovog istraživanja potvrđuju da Koksov proporcionalni model

predstavlja snažan i fleksibilan alat u analizi preživljavanja, sa širokom primenom u medicini, epidemiologiji, sociologiji i drugim oblastima istraživanja. Razumevanje prednosti i ograničenja ovog modela omogućava nam bolje donošenje odluka u analizi preživljavanja i pravilno tumačenje rezultata istraživanja.

Glava 9

Dodatak

```
library(survival)
library(knitr)
library(tibble)
library(ggplot2)
library(dplyr)
library(scales)
library(survminer)
library(gridExtra)
library(Hmisc)
library(MASS)

data(diabetic, package="survival")
?diabetic
str(diabetic)
class(diabetic)
dim(diabetic)
head(diabetic)

class(diabetic)
diabetic_as_tibble = as_tibble(diabetic)
head(diabetic_as_tibble)

length(unique(diabetic$id))
```

```
surv_obj <- Surv(diabetic$time, diabetic$status)
sum(surv_obj[, "status"] == 0)

df <- table(diabetic$laser)
percentages <- prop.table(df) * 100
pie <- ggplot(data.frame(laser = as.character(names(df)), freq
    = as.numeric(df), percent = percentages), aes(x = "", y =
   freq, fill = laser)) +
  geom_col(width = 1) +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust = 0.5)) +
  labs(fill = "laser",
       x = NULL,
       y = NULL) +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank(),
        axis.title = element_blank(),
        panel.grid = element_blank())
pie + coord_polar(theta = "y", start = 0) +
  geom_text(aes(label = paste0(round(percentages, 1), "%")),
     position = position_stack(vjust = 0.5), size = 4) +
  theme(panel.background = element_rect(fill = "white"))

df1 = diabetic[diabetic$trt == 1,]
count_data <- count(df1, eye, laser, name = "count")
ggplot(count_data, aes(x = eye, y = count, fill = laser, label
    = count)) +
  geom_bar(stat = "identity") +
  geom_text(position = position_stack(vjust = 0.5), color = "
     black") +
  labs(x = "eye", y = "count", fill = "laser") +
  theme(panel.background = element_rect(fill = "white"))

ggplot(diabetic, aes(x = age, y = ..density..)) +
  geom_histogram(binwidth = 5, fill = "#00BFC4", color = "
     black", alpha = 0.7) +
  geom_density(color = "#F8766D", size = 1.2) +
```

```
  labs(x = "Age", y = "Density") +
  theme(panel.background = element_rect(fill = "white"))

mean(diabetic$age)
hist(diabetic$age)
ggplot(diabetic, aes(age)) + geom_histogram(bins = 20)


diabetic["status"] = as.factor(diabetic$status)
g_hist <- ggplot(diabetic, aes(x = time, color = status, fill
   = status))
g_hist + geom_histogram(position = "identity") +
  theme(panel.background = element_rect(fill = "white"))

plot1 <- ggplot(diabetic, aes(x = factor(risk))) +
  geom_bar(width = 1, fill = "#00BFC4", color = "black", alpha
      = 0.7, stat = "count") +
  labs(x = "risk", y = "frequency") +
  theme(panel.background = element_rect(fill = "white")) +
  facet_wrap(~ factor(trt, labels = c("no treatment", "laser")
     ), nrow = 1)
plot2 <- ggplot(diabetic, aes(x = factor(risk))) +
  geom_bar(width = 1, fill = "#00BFC4", color = "black", alpha
      = 0.7, stat = "count") +
  labs(x = "risk", y = "frequency") +
  theme(panel.background = element_rect(fill = "white")) +
  facet_wrap(~ factor(status, labels = c("censored", "visual
     loss")), nrow = 1)
grid.arrange(plot1, plot2, ncol = 2)

diabetic_as_tibble$type =  cut(diabetic$age, breaks = c(0, 20,
    Inf), labels = c("juvenile", "adult"))

s = Surv(diabetic$time, diabetic$status)
class(s)
s
```

```
s_fit = survfit(Surv(time, status) ~ 1, data = diabetic_as_
    tibble)
s_fit
range(diabetic$time)
summary(s_fit,times = seq(0.30, 74.97, 8))
ggsurvplot(s_fit)

s1_fit = survfit(Surv(time, status) ~ trt, data = diabetic_as_
    tibble)
s1_fit
summary(s1_fit, times = seq(0.30, 74.97, 8))
ggsurvplot(s1_fit, conf.int = TRUE, pval = TRUE, risk.table =
    TRUE,
           risk.table.height = .25)
surv_obj = with(diabetic_as_tibble, Surv(time, status))
log_rank_test = survdiff(surv_obj ~ trt, data = diabetic_as_
    tibble)
print(log_rank_test)

filtered_data = subset(diabetic_as_tibble, trt == 1)
s2_fit = survfit(Surv(time, status) ~ laser, data = filtered_
    data)
s2_fit
summary(s2_fit, times = seq(0.30, 74.97, 8))
ggsurvplot(s2_fit, conf.int = TRUE, pval = TRUE, risk.table =
    TRUE,
           legend.labs = c("xenon", "argon"), legend.title = "
              laser",
           risk.table.height = .25)
surv_obj = with(filtered_data, Surv(time, status))
log_rank_test = survdiff(surv_obj ~ laser, data = filtered_
    data)
print(log_rank_test)

s3_fit = survfit(Surv(time, status) ~ laser + eye, data =
    filtered_data)
s3_fit
```

```
summary(s3_fit, times = seq(0.30, 74.97, 8))
ggsurvplot(s3_fit, conf.int = FALSE, pval = TRUE, risk.table =
    TRUE,
           risk.table.height = .25)
surv_obj = with(filtered_data, Surv(time, status))
log_rank_test = survdiff(surv_obj ~ laser + eye, data =
   filtered_data)
print(log_rank_test)

s4_fit = survfit(Surv(time, status) ~ risk, data = diabetic_as
   _tibble)
s4_fit
range(diabetic$time)
summary(s4_fit, times = seq(0.30, 74.97, 8))
ggsurvplot(s4_fit, conf.int = FALSE, pval = TRUE, risk.table =
    TRUE,
           risk.table.height = .25)
surv_obj = with(diabetic_as_tibble, Surv(time, status))
log_rank_test = survdiff(surv_obj ~ risk, data = diabetic_as_
   tibble)
print(log_rank_test)

s5_fit = survfit(Surv(time, status) ~ risk + trt, data =
   diabetic_as_tibble)
s5_fit
summary(s5_fit, times = seq(0.30, 74.97, 8))
ggsurvplot(s5_fit, conf.int = FALSE, pval = TRUE, risk.table =
    TRUE,
           risk.table.height = .25)
surv_obj = with(diabetic_as_tibble, Surv(time, status))
log_rank_test = survdiff(surv_obj ~ risk + trt, data =
   diabetic_as_tibble)
print(log_rank_test)

s6_fit = survfit(Surv(time, status) ~ type + trt, data =
   diabetic_as_tibble)
s6_fit
```

```
summary(s6_fit, times = seq(0.30, 74.97, 8))
ggsurvplot(s6_fit, conf.int = FALSE, pval = TRUE, risk.table =
    FALSE)
surv_obj = with(diabetic_as_tibble, Surv(time, status))
log_rank_test = survdiff(surv_obj ~ type + trt, data =
   diabetic_as_tibble)
print(log_rank_test)

diabetic_as_tibble$trt <- factor(diabetic_as_tibble$trt,
   levels = c(0, 1), labels = c("no treatment", "laser"))

cox_trt = coxph(Surv(time, status) ~ trt, cluster = id, data=
   diabetic_as_tibble)
print(cox_trt)
survdiff(Surv(time, status) ~ trt, data=diabetic_as_tibble)
test.ph = cox.zph(cox_trt)
print(test.ph)
ggcoxzph(test.ph)
m = survfit(Surv(time, status) ~ trt, cluster = id, data=
   diabetic_as_tibble)
s = summary(m)
s_table = data.frame(s$strata, s$time, s$n.risk, s$n.event, s$
   n.censor, s$surv, s$lower, s$upper)
s_table = s_table %>%
  rename(strata = s.strata, time = s.time, surv = s.surv,
     lower = s.lower, upper = s.upper) %>%
  mutate(negloglogsurv = -log(-log(surv)))
ggplot(s_table, aes(x = time, y = negloglogsurv, color =
   strata)) +
  geom_line(size = 1.25) +
  theme(text = element_text(size = 16),
        plot.title = element_text(hjust = 0.5), panel.
           background = element_rect(fill = "white")) +
  ylab("-ln(-ln(S(t)))")
AIC(cox_trt)
BIC(cox_trt)
```

```
cox_laser = coxph(Surv(time, status) ~ laser, cluster = id,
   data = diabetic_as_tibble)
print(cox_laser)
survdiff(Surv(time, status) ~ laser, data = diabetic_as_tibble
   )
test.ph = cox.zph(cox_laser)
print(test.ph)
ggcoxzph(test.ph)
AIC(cox_laser)
BIC(cox_laser)

cox_age = coxph(Surv(time, status) ~ age, cluster = id, data =
    diabetic_as_tibble)
print(cox_age)
test.ph = cox.zph(cox_age)
print(test.ph)
ggcoxzph(test.ph)
AIC(cox_age)
BIC(cox_age)

cut(diabetic$age, breaks = c(0, 20, Inf))

cox_type = coxph(Surv(time, status) ~ type, cluster = id, data
    = diabetic_as_tibble)
print(cox_type)
test.ph = cox.zph(cox_type)
print(test.ph)
ggcoxzph(test.ph)
AIC(cox_type)
BIC(cox_type)


cox_eye = coxph(Surv(time, status) ~ eye, cluster = id, data =
    diabetic_as_tibble)
print(cox_eye)
test.ph = cox.zph(cox_eye)
print(test.ph)
```

```
ggcoxzph(test.ph)
m = survfit(Surv(time, status) ~ eye, diabetic_as_tibble)
s = summary(m)
s_table = data.frame(s$strata, s$time, s$n.risk, s$n.event, s$
   n.censor, s$surv, s$lower, s$upper)
s_table = s_table %>%
  rename(strata=s.strata, time=s.time, surv=s.surv, lower=s.
     lower, upper=s.upper) %>%
  mutate(negloglogsurv=-log(-log(surv)))
ggplot(s_table, aes(x=time, y=negloglogsurv, color=strata)) +
  geom_line(size=1.25) +
  theme(text=element_text(size=16),
        plot.title=element_text(hjust=0.5)) +  ylab("-ln(-ln(S
           (t)))")
AIC(cox_eye)
BIC(cox_eye)

cox_risk = coxph(Surv(time, status) ~ risk, cluster = id, data
    = diabetic_as_tibble)
print(cox_risk)
test.ph = cox.zph(cox_risk)
print(test.ph)
ggcoxzph(test.ph)
AIC(cox_risk)
BIC(cox_risk)


cox = coxph(Surv(time, status) ~ eye + risk + trt, cluster =
   id, data=diabetic_as_tibble)
print(cox)
test.ph = cox.zph(cox)
print(test.ph)
ggcoxzph(test.ph)
print(logLik(cox))
print(AIC(cox))
print(BIC(cox))
s_fit = survfit(cox, data = diabetic_as_tibble)
```

```
ggsurvplot(s_fit, data = diabetic_as_tibble, risk.table = TRUE
    , conf.int = TRUE)
c_index <- rcorr.cens(diabetic_as_tibble$status, predict(cox))
print(c_index)
```

Dodatak 9.1: Analiza preživljavanja u statističkom softveru R

Bibliografija

[1] AL Blair, DR Hadden, JA Weaver, DB Archer, PB Johnston, and CJ Maguire. The 5-year prognosis for vision in diabetes. *The Ulster medical journal*, 49(2):139, 1980.

[2] Ørnulf Borgan and Knut Liestøl. A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics*, pages 35–41, 1990.

[3] Elizabeth R Brown, Joseph G Ibrahim, and Victor DeGruttola. A flexible b-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, 61(1):64–73, 2005.

[4] David Collett. *Modelling survival data in medical research*. CRC press, 2023.

[5] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[6] David M Diez and Maintainer David M Diez. Package 'oisurv', 2013.

[7] Donald S Fong, Lloyd Aiello, Thomas W Gardner, George L King, George Blankenship, Jerry D Cavallerano, Fredrick L Ferris III, Ronald Klein, and American Diabetes Association. Diabetic retinopathy. *Diabetes care*, 26(suppl_1):s99–s102, 2003.

[8] M Gail, K Krickeberg, JM Samet, A Tsiatis, and W Wong. Statistics for biology and health series editors. *Atlanta: Springer*, 2012.

[9] Patricia M Grambsch and Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[10] David W Hosmer Jr, Stanley Lemeshow, and Susanne May. *Applied survival analysis: regression modeling of time-to-event data*, volume 618. John Wiley & Sons, 2011.

[11] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[12] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[13] Elisa T Lee and John Wang. *Statistical methods for survival data analysis*, volume 476. John Wiley & Sons, 2003.

[14] Stanley Lemeshow, Susanne May, and David W Hosmer Jr. *Applied survival analysis: regression modeling of time-to-event data*. John Wiley & Sons, 2011.

[15] Mark Stevenson and IVABS EpiCentre. An introduction to survival analysis. *EpiCentre, IVABS, Massey University*, 2009.

[16] Terry M Therneau and Thomas Lumley. Package 'survival'. *R Top Doc*, 128(10):28–33, 2015.

[17] Yishu Xue and Elizabeth D Schifano. Diagnostics for the cox model. *Communications for statistical Applications and Methods*, 24(6):583–604, 2017.

Biografija autora

Lidija Tomanić rođena je 16. decembra 1996. godine u Kraljevu. Nakon završetka srednje Medicinske škole u Kraljevu 2015. godine, Lidija je odlučila nastaviti svoje obrazovanje na polju matematike. Tako je iste godine upisala Matematički fakultet Univerziteta u Beogradu na modulu Statistika, aktuarska i finansijska matematika. Diplomirala je na Matematičkom fakultetu Univerziteta u Beogradu 2020. godine. Trenutno je student master programa na Matematičkom fakultetu Univerziteta u Beogradu.