Proporcionalni parametarski modeli rizika

Cvitković, Anamarija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Anamarija Cvitković

PROPORCIONALNI PARAMETARSKI MODELI RIZIKA

Diplomski rad

Voditelj rada:
Izv.prof.dr.sc.Miljenko Huzak
Neposredni voditelj:
Dr.sc.Azra Tafro

Zagreb, rujan 2018.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Zahvaljujem mentorima prof.dr.sc. Miljenku Huzaku i dr.sc. Azri Tafro na strpljenju, pomoći i vodstvu pri izradi ovog diplomskog rada.
Hvala mojim roditeljima na razumijevanju i podršci tokom studiranja.

Sadržaj

Uvod

Analiza doživljanja je skup tehnika kojima se procijenjuje i opisuje vrijeme potrebno do pojave jednog ili više određenih događaja. Naziv 'Analiza doživljenja' dolazi od prvih istraživanja gdje je događaj od interesa uglavnom bila smrt. Danas se primjenjuje i u mnogim drugim širim područjima, primjerice kod procjene vremena potrebnog za razvijanje nekih bolesti, kvara strojeva, pojave potresa, u financijama, itd. Napredak analize doživljenja u novije doba je posljedica razvoja softverskih paketa i visokih performansi računala koji su sada u mogućnosti veoma efikasno izvršiti ove zahtjevne algoritme.

U ovom radu obrađen je parametarski model proporcionalnog hazarda kojim se modeliraju podaci u analizi doživljenja. Rad je podijeljen na tri poglavlja.

U prvom poglavlju ovog rada upoznati ćemo se s konceptom doživljenja, skupom podataka koji se koriste pri analizi te s problematikom cenzuriranja i rezanja tih podataka, pri čemu ćemo objasniti razliku između ta dva pojma.

Kroz drugo poglavlje upoznati ćemo se s osnovnim pojmovima iz analize doživljenja. Definirat ćemo funkciju doživljenja, funkciju hazarda, očekivano trajanje preostalog života te ćemo dati parametarsku procjenu za te veličine.

U trećem poglavlju obrađujemo teorijsku pozadinu parametarskog proporcionalnog modela rizika te procedure koje koristimo pri zaključivanju o valjanosti modela. Također, dajemo pregled nekih od važnijih parametarskih distribucija doživljenja, eksponencijalne i Weibullove distribucije.

Na kraju trećeg poglavlja kroz primjere su prikazane pretpostavke PH modela i njihova važnost.

Za izračune i grafove u radu korišten je programski jezik R.

Poglavlje 1

Analiza doživljenja

Analiza doživljenja (eng. *survival analysis*) obuhvaća skup statističkih metoda za analizu podataka pri čemu je varijabla od interesa vrijeme do pojave određenog događaja. Pod vremenom podrazumijevamo godine, mjesece, tjedne ili dane koji prođu od početka promatranja nekog subjekta pa do trenutka pojavljivanja događaja. U praksi vrijeme do pojave određenog događaja često se naziva vrijeme neuspjeha, vrijeme doživljenja ili jednostavno vrijeme događaja.

1.1 Mjerenje i koncept doživljenja

Pretpostavimo da želimo proučavati ispitanike unutar neke skupine, odnosno proučavati pojavu nekog događaja za svakog ispitanika zasebno. Ako vrijeme do pojave događaja nije važno, događaj se može analizirati kao binarni ishod pomoću logističkog regresijskog modela. Na primjer, kada analiziramo smrtnost ispitanika povezanu s operacijom srca nije nam bitno hoće li pacijent umrijeti tijekom operacije ili će umrijeti nakon dva mjeseca provedenih u komi. Za ostale ishode, osobito kod onih bolesti koje traju dugo ili se stalno vraćaju, vrijeme do pojave događaja je od velike važnosti. Analiza koja izračuna samo ukupan broj smrtnih događaja odbacila bi vrijedne informacije i žrtvovala moć statistike.

Analiza doživljenja prvo se razvila za potrebe medicine i biologije, a potom za ekonomske, društvene i inženjerske potrebe. Osnovno značenje same riječi doživljenje proizlazi iz bilježenja podataka o vremenu smrti pojedinaca u određenoj skupini. Bilježenja se mogu provesti u ili blizu trenutka smrti, na primjer kada promatramo podatake o uzroku smrti, duljini bolesti, fizičkoj karakteristici u blizini same smrti, ili u ranijem vremenskom periodu, na primjer bilježenje podataka o spolu, dobi, povijesti bolesti, fizičkim obilježjima u ranijim razdobljima. Određene varijable, najčešće dob (vrijeme proteklo od rođenja) i/ili vrijeme proteklo od nekih drugih važnih događaja (na primjer početaka bolesti, datuma operacije i slično) smatraju se primarnim interesom. Analizom doživljenja želi se procjeniti odnos između smrtnosti i tih primarnih varijabli, odnosno procijenjuje se koliki utjecaj primarne varijable imaju na smrt ispitanika unutar grupe. Pri tome se dopušta što je moguće više utjecaja nekih drugih manje važnih varijabli. Potonje, u ovom kontekstu, nazivaju se popratnim varijablama. Popratne varijable uvedene su zbog mogućeg utjecaja

na primarne, ali nisu proučavane same za sebe. Ovisno o odnosima među varijablama koje želimo proučavati, odabiremo koje ćemo varijable uzeti za primarne, a koje za popratne varijable.

U ovom ćemo radu razmatrati općenitije situacije pa ćemo umjesto pojma "smrtnosti" uglavnom koristiti prikladniji pojam "neuspjeha". U takvom kontekstu pojedinci nisu nužno (iako mogu biti) živa bića. Oni mogu biti na primjer masovno proizvedeni uređaji, kao što su električne svjetiljke. U tom kontekstu neuspjeh može značiti nemogućnost funkcioniranja u predodređenoj ulozi.

Dakle, prvenstveno ćemo se baviti proučavanjem podataka o neuspjehu te odnosu neuspjeha i nekoliko važnih varijabli.

1.1.1 Izloženost riziku

Kalendarski vremenski period pri kojem je subjekt izložen procesu promatranja poznat je kao vrijeme promatranja. Svaki subjekt, nezavisno jedan od drugog, ulazi u proces promatranja u nekom vremenu. Vremenski period koji subjekt provede u procesu promatranja, mjeren od početka promatranja, poznat je kao subjektovo vrijeme. S druge strane, period od početka promatranja do trenutka kada je uslijedio neuspjeh subjekta naziva se vrijeme doživljenja.

U većini analiza podataka o doživljenju zainteresirani smo za proučavanje odnosa neuspjeha među skupinama subjekata pod određenim uvjetima. Jasno je, što je dulje razdoblje za koje je pojedinac pod promatranjem, to je vjerojatnije da će se prije ili kasnije uočiti neuspjeh. Usporedivost broja neuspjeha zahtijeva da se usredotočimo na jedno razdoblje promatranja. Da bismo napravili usporedbu, trebali bismo za svakog pojedinca znati koje je njegovo razdoblje izlaganja riziku, tj. kada će se za pojedinca dogoditi neuspjeh i pridodati promatranim neuspjesima. Prilikom eksperimenta razdoblje je obično dostupno za svakog pojedinca, iako kada je velika količina podataka može se upotrijebiti približna procjena podataka. Na primjer, za podatke o popisu stanovništva ovo razdoblje obično nije poznato i mora se procijeniti.

1.1.2 Upotreba teorije vjerojatnosti

Teorija vjerojatnosti je matematička disciplina koja se bavi proučavanjem slučajnih pojava, odnosno empirijskih događaja čiji ishodi nisu uvijek strogo definirani. Jedan od osnovnih alata u teoriji vjerojatnosti je eksperiment pomoću kojeg se provodi ispitivanje veze između uzroka i posljedice. Na ishod eksperimenta često utječe više uvjeta i ako se eksperiment ponavlja više puta pod jednakim uvjetima, pojavljuje se određena zakonitost unutar skupa ishoda. Teorija vjerojatnosti bavi se takvim zakonitostima uvođenjem kvantitativne mjere u obliku realnog pozitivnog broja, odnosno vjerojatnosti. Vjerojatnost procjenjuje mogućnost, odnosno nemogućnost ostvarenja ishoda. Najjednostavnije rečeno, vjerojatnost je mjera ostvarivosti slučajnog događaja.

Teorija vjerojatnosti ne oslanja se samo na empirijske i intuitivne metode već na formalnu teoriju povezanu s drugim matematičkim pojmovima. Osnovni pojmovi vjerojat-

nosti mogu se razlikovati ovisno o točki gledanja te isto tako rezultati i interpretacije rezultata mogu biti različite. Zakoni vjerojatnosti nisu uvijek jednostavni i lako razumljivi.

Budući da proučavamo odnose, prirodno ih je predstaviti preko pojma vjerojatnosti. Također, svaki događaj koji se osjeti i koji za posljedicu ima određeni učinak nosi određenu vjerojatnost. Te su vjerojatnosti podaci koji se najčešće izračunavaju naknadno i koji služe za retroaktivnu analizu. Dakle, možemo iskoristiti vrlo dobro razvijene tehnike i koncepte teorije vjerojatnosti kako bismo bolje razumjeli dobivene podatke. U svim primjenama statističkih metoda zasnovanih na vjerojatnostima, modeli i pretpostavke na kojima se temelje nisu obično u potpunosti zadovoljene. Zbog toga želimo prikazati prednosti primjene tih metoda, pogotovo kada su kompleksnije, primjerice u slučajevima kada se za analizu koristi metoda maksimalne vjerodostojnosti, o kojoj ćemo više govoriti u nastavku.

Teorija vjerojatnosti pruža osnovnu pozadinu i alate za analizu doživljenja, ali također daje velike i raznovrsne mogućnosti izlaska iz teorijskih modela u smislu odgovarajućeg deskriptivnog dogovora.

1.1.3 Vrste varijabli

Prethodno smo spomenuli varijable od interesa, najčešće varijable koje pobliže predstavljaju doživljenje ili neuspjeh,koje nazivamo primarnim. Kao što smo naveli najčešće primarne varijable su dob i vrijeme do određenog događaja. Ako odaberemo spol kao primarnu varijablu, tada je primarna varjabla binarna gdje nula označava muški, a jedinica ženski spol.

Varijable s istim ishodom, odnosno s istom vrijednosti se mogu mjeriti izravnim brojenjem u situacijama eksperimentalnog tipa ili neizravnom procjenom na temelju popisanih podataka, tzv. popisa. Najčešće se želi izračunati, odnosno procijeniti, koliko pojedinaca preživi neki promatrani period u ovisnosti o nekoliko važnih varijabli.

Kao što smo već spomenuli, uz primarne varijable, preostale izmjerene varijable od interesa su popratne varijable. Popratne varijable mogu biti zemljopisni položaj, društvena klasa (često mjerena kao indeks koji se temelji na dohotku, obrazovanju, itd.) i fizičke karakteristike kao što su krvni tlak, težina i vitalnost. Poželjno ih je uvesti ovisno o tome u kojoj mjeri utječu na pojavu neuspjeha. Mogu se uvesti analitički, uvođenjem nekog prilično jednostavnog modela ili izgradnjom zasebnih životnih tablica.

Metoda uvođenja popratnih varijabli izgradnjom zasebnih životnih tablica je sigurnija, ali se može primijeniti samo ako se popratna varijabla može definirati u nekoliko kategorija. Na primjer, uobičajno je imati odvojene životne tablice za muškarce i žene, jer praksa pokazuje različite stope smrtnosti za žene u odnosu na muškarce.

1.2 Životne tablice

Prve životne tablice ili tablice smrtnosti sastavili su John Graunt $(1662.)$, jedan od prvih demografa, i engleski astronom Sir Edmond Halley $(1678.)$. Životne tablice izgrađuju se za potrebe šireg kruga korisnika dajući cjelovitu sliku smrtnosti stanovništva, odnosno broj

osoba određene starosti unutar jedne skupine koje godišnje umru. S obzirom na njihov značaj izgrađuju se u gotovo svim zemljama svijeta.

Izrada tablica smrtnosti je jedna od najstarijih tehnika u demografskoj analizi. Tablice sadrže niz demografskih pokazatelja. Osnovni pokazatelj je vjerojatnost smrti na osnovi koje se izračunavaju sve ostale biometrijske funkcije: vjerojatnost doživljenja, broj živih, broj umrlih, očekivano trajanje života i druge. Za izradu detaljnih tablica smrtnosti potrebno je raspolagati podacima o stanovništvu, prema spolu i pojedinačnim godinama starosti, podacima o broju živorođenih prema spolu i podacima o umrlim prema spolu te pojedinačnim godinama starosti.

Osim u analizi doživljenja primjenjuju se u izradi projekcija stanovništva, definiranju neto stopa reprodukcije ženskog stanovništva, itd. Od posebne su važnosti za izračunavanje i određivanje visine premije na području životnog osiguranja, u mirovinskom i invalidskom osiguranju. Tablice omogućavaju najkompletnije i najtočnije usporedbe smrtnosti različitih populacija ili dijelova populacija. Razina i smjer promjene smrtnosti po starosti neposredno određuje dužinu očekivanog trajanja života na dan rođenja kao sintetičkog pokazatelja smrtnosti stanovništva.

1.2.1 Biometrijske funkcije u tablicama smrtnosti

U životnoj tablici x označava starost i obično se u tablicu unose samo veličine za prirodne brojeve x. Sirove vjerojatnosti smrti q'_x računaju se za ukupno, muško i žensko stanovništvo za sve dobne skupine, $x = 0, 1, \ldots, 99$.

Izglađivanje sirovih vjerojatnosti smrti je postupak transformacije istih kako bi se otklonile posljedice grešaka slučajne prirode. Te greške najčešće su posljedica nedovoljno velikih skupina živih ili umrlih u određenoj godini starosti, kao i nedovoljne točnosti osnovnih podataka, posebno o godinama starosti umrlih. Izglađene vjerojatnosti smrti označavaju se sa q_x i predstavljaju vjerojatnost da će osoba stara x godina umrijeti prije nego što dosegne starost od $x + 1$ godine. Pomoću njih izračunavaju se sve ostale vrijednosti u tablicama.

Funkcija p_x (vjerojatnost doživljenja) definirana je kao $p_x = 1 - q_x$. Predstavlja vjerojatnost da će osoba stara x godina doživjeti starost od $x + 1$ godine.

Pretpostavimo da promatramo grupu od l_0 novorođene djece, dakle djece u dobi nula. Pretpostavimo da u toj skupini nema novih rađanja, ni useljavanja, ni iseljavanja. Kako vrijeme prolazi grupa se postepeno smanjuje jer njeni članovi umiru, a ne rađaju se novi. Tablica smrtnosti je reprezentacija smrtnosti takve grupe. To je zapravo model doživljenja izražen pomoću očekivanog broja preživjelih od početnog broja l_0. Početni broj l_0 naziva se korijen tablice i predstavlja (očekivani) broj novorođenih osoba u danoj populaciji. Obično se uzima $l_0 = 100000$ ili $l_0 = 1000000$.

Sa l_x u tablici smrtnosti se označava broj osoba u danoj populaciji koje su žive u dobi x, odnosno broj članova promatrane skupine koji su stari točno x godina. Broj l_x koristimo za predviđanje smanjenja populacije jer se zbog smrtnosti smanjuje s povećanjem starosti. Prema navedenom, brojevi l_x smisao dobivaju samo u usporedbi s početnim l_0, odnosno usporedbom svakog l_x s prethodnim l_{x-1}. Prema tome, funkcija l_x (broj živih) definira se kao

$$l_x = l_{x-1} p_{x-1},$$

s početnom vrijednosti l_0.

Granična dob tablice smrtnosti označava se s ω i definira kao dob za koju vrijednost l_x postaje zanemarivo mala (u odnosu na l_0), definira se kao $l_\omega = 0$.

Funkcija d_x koja predstavalja broj umrlih definirana je kao

$$d_x = l_x - l_{x-1} = l_x q_x$$

i pokazuje koliko od broja živih osoba starosti x umire prije nego dostigne starost od $x+1$ godina.

Zbroj brojeva živih osoba N_x računamo:

$$N_x = \sum_{i=0}^{x} l_i.$$

Ovaj broj znači fiktivne zbrojeve osoba koje su od skupine l_0 istovremeno živorođenih preživjele nultu godinu, zatim onih koje su od ovih preživjele prvu, drugu itd., do stote godine. Tehnički se ovaj broj dobija kumuliranjem vrijednosti l_x.

Funkcija L_x koja predstavalja srednji broj živih definirana je kao

$$L_x = \frac{(l_x + l_{x+1})}{2}$$

i označava broj živih u starosti od x do $x+1$ godina.

Funkcija P_x predstavlja stopu doživljenja i definirana je s

$$P_x = \frac{L_{x+1}}{L_x}.$$

Dakle, stopa doživljenja je omjer osoba starih između x i $x+1$ godina koji će doživjeti starost od $x+1$ do $x+2$ godina.

Funkcija očekivano trajanje života e_x^0 definirana je kao

$$e_x^0 = \frac{1}{2}\frac{N_x}{l_x}.$$

Predstavlja očekivano trajanje života jedne osobe stare x godina pod uvjetima starosti iz perioda na koji se odnose tablice. Očekivano trajanje života jedan je od najboljih pokazatelja razvijenosti društva.

Numeričke vrijednosti biometrijskih funkcija prikazane su u odgovarajućim kolonama u tabelarnom dijelu.

1.3 Baza podataka analize doživljenja

Osnovni cilj analize doživljenja je usporedba vremena doživljenja dvije ili više promatranih skupina i otkrije razlikuju li se statistički značajno ta vremena doživljenja. Pojedinci u skupini koju proučavamo mogu biti ljudi, životinje, insekti i tako dalje. Sama skupina može se definirati na različite načine. Podaci mogu na primjer biti grupirani po zemljopisnom

položaju (stanovništvo grada ili države, pacijenti u bolnici ili u nizu bolnica) ili prema povijesti (vrsti liječenja, vrsti bolesti, zapošljavanju).

Baza podataka je skup prikupljenih podataka, osobito podataka o doživljenju i neuspjehu. Točnosti praćenja zapisa je osobito važna, budući da su relevantni neuspjesi obično rasprostranjeni tijekom vremenskih razdoblja, a svaki novi neuspjeh zahtijeva novi ulazak u taj zapis.

Podaci eksperimentalnog tipa zahtijevaju više napora u analiziranju i pohranjivanju jer se manje više detaljni zapisi moraju održavati za subjekte koje promatramo tijekom razdoblja studije. Iste varijable za pojedinačne subjekte također trebaju biti ažurirane kada je potrebno. Bitno je, da zapisi što točnije navode razdoblja tijekom kojih je pojedinac bio izložen riziku, u smisu da će se zabilježiti neuspjeh ako se to dogodilo u bilo kojem trenutku tijekom tog razdoblja, i neće se zabilježiti ako se to dogodilo u bilo kojem drugom trenutku.

Sva zapažanja trebaju biti zabilježena na prvoj točnosti koja je izvediva i smislena. Ovaj opći princip, koji je opravdan na temelju korištenja što je više moguće dostupnih informacija, od velike je važnosti u pogledu vremena opažanja smrti ili neuspjeha. Smrti se ponekad bilježe samo ako se javljaju u dužim vremenskim intervalima. Na primjer u određenom tjednu, za razliku od određenog dana ili sata.

Što su duži vremenski intervali, to je veća vjerojatnost da će jedan vremenski interval sadržavati više od jedne smrti. Ta pojava često nosi naziv "višestruke smrti" i u takvim slučajevima nije moguće odlučiti o redoslijedu smrti. Prividno jednaka vremena smrti nazivaju se vezanim očitanjima. Predložene su neke metode rješavanja veza, no sve uglavnom uključuju dodatne probleme koji se temelje na sumnjivim pretpostavkama. Ako ne postoji jasna predodžba razumnih dosljednih rezultata, među mogućim poredcima, treba zaključiti da je zapravo interval promatranja preširok i da bi ga trebalo pokušati suziti. Naravno, moguće je napraviti različite pretpostavke o rješavanju zabilježenih veza. Na primjer, ako su dvije osobe zabilježene u razdoblju od 6 mjeseci, a njihove dobi posljednjeg rođendana na početku razdoblja bile su 28 i 58 godina, razumno je da nam to daje veliku težinu mogućnosti da je druga osoba zapravo prva od dvije koja će umrijeti.

1.4 Cenzuriranje podataka

Iako je vrijeme do pojave određenog događaja neprekidna varijabla, analiza doživljenja uzima u obzir glavni analitički problem koji nazivamo cenzuriranje. Cenzuriranje se događa kada imamo djelomične informacije o vremenu doživljenja pojedinca, ali ne znamo točno vrijeme doživljenja. Drugim riječima, kad ispitujemo vrijeme do neuspjeha za određenu skupinu, na kraju vremena promatranja pojedini subjekti koji su u skupini neće doživjeti neuspjeh. Dakle, cenzuriranjem ne dobivamo potpunu informaciju o vremenu doživljenja već samo znamo da je dosegla neku vrijednost.

Tri glavna razloga za cenzuriranje podataka:

1. Kod promatranog subjekta promatrani događaj se nije dogodio unutar promatranog razdoblja.

2. Iz nekog razloga više nije moguće pratiti subjekt unutar perioda promatranja, subjekt je izgubljen tokom procesa promatranja.

3. Subjekt se povlači iz promatranja zbog smrtnog ishoda (ukoliko smrt nije promatrani događaj) ili nekog drugog razloga.

Pretpostavimo da promatramo uzorak od N subjekata. Vrijeme cenzuriranja je slučajna varijabla. Označimo s T_i stvarno vrijeme doživljenje subjekta $i = 1, 2, \ldots, N$ (vrijeme koje nas zanima u studiji), a sa C_i potencijalno cenzurirano vrijeme. Pretpostavimo da su to nezavisne slučajne varijable. Sada vrijeme proteklo od uključivanja u studiju označimo sa

$$t_i = \min(T_i, C_i).$$

Uvodimo indikator događaja za cenzuriranje, δ_i. Indikator poprima vrijednost $\delta_i = 1$ ukoliko je događaj od interesa opažen tj. ako je $T_i < C_i$, odnosno $\delta_i = 0$ u slučaju cenzuriranja, ako je $T_i \geq C_i$.

Dakle, s obzirom na prisutnost cenzuriranja često je najbolji način prikazivanja podataka u obliku uređenog para (t_i, δ_i). Jedna od osnovnih pretpostavki statističkih modela analize doživljenja jest da su podaci $\{(t_i, \delta_i), i = 1, 2, \ldots, N\}$ dobiveni u uzorku od N subjekata nezavisni. Nadalje, ukoliko su podaci cenzurirani, pretpostavlja se da su varijable T_i nezavisne, što nam u praksi daje mogućnost relativno točne procjene cenzuriranih podataka onim podacima koji su dobiveni od necenzuriranih u trenucima nakon cenzuriranja. To je zato što takvo, neinformativno cenzuriranje (vrijeme cenzuriranja nezavisno od vremena opažanih događaja) ne utječe na vjerojatnost nekog ishoda u studiji, odnosno subjekti čiji su podaci cenzurirani imaju istu vjerojatnost preživljavanja kao i ostali subjekti iz uzorka.

U praksi se prisutnost informativnog cenzuriranja pokušava izbjeći pažljivim dizajnom studije i oprezom pri njenom provođenju. Ukoliko je broj subjekata koji više nisu dio studije iz ranije navedenih razloga relativno malen, mogu se primijeniti metode analize bazirane na neinformativnom cenzuriranju.

U praksi se najčešće javljaju dva tipa cenzuriranja, tzv. tip I i tip II.

- **Cenzuriranje tipa I**

  Kod ovog tipa prikupljanja podataka pretpostavlja se da se promatranje slučajnog uzorka prekida nakon izvjesnog fiksnog vremena te t predstavlja vrijeme doživljenja poznato samo za one subjekte koji su doživljeli neuspjeh do tog vremena. Ukoliko do tog trenutka osoba nije iskusila događaj od interesa, smatrat ćemo to cenzuriranim promatranjem, jer je sasvim moguće da se događaj pojavio nakon završetka promatranja.

- **Cenzuriranje tipa II**

  Problem cenzuriranja tipa I je fiksno vrijeme cenzuriranja. Ako je fiksno vrijeme veliko, veliki su troškovi eksperimenta. Ako je fiksno vrijeme malo, može se dogoditi

da se promatra manji dio uzorka. Da bi se to izbjeglo, možemo odrediti da eksperiment završi nakon što određeni broj subjekata doživi neuspjeh, što nam i predstavlja cenzuriranje tipa II. Pretpostavimo da imamo ukupno $n \in \mathbb{N}$ podataka. Od n podataka neka $r \in \mathbb{N}$ subjekata u uzorku doživi neuspjeh, tada imamo $n - r$ cenzuriranih promatranja. Dakle, dani podaci su

$$T_1 \leq T_2 \leq \cdots \leq T_r.$$

U ovom slučaju vrijeme cenzuriranja je T_r, slučajni trenutak u kojem promatranje završava. Ova metoda je pogodna za testiranje raznih alata i oprema.

Nadalje, u praksi podaci mogu biti lijevo, desno i intervalno cenzurirani.

1.4.1 Desno cenzuriranje

Pretpostavimo da je subjekt ušao u proces promatranja u nekom vremenu t_0. Neka je subjekt doživio neuspjeh u nama nepoznatom trenutnu $t_0 + t$. Posljednja informacija s kojom raspolažemo je ta da se u trenutku $t_0 + c$ $(c, t \in \mathbb{R}_+, c < t)$ nije dogodio neuspjeh. Tada c nazivamo cenzuriranim vremenom doživljenja i pritom je riječ o desnom cenzuriranju. Ovaj način promatranja se koristi kada zbog nedostatka vremena ili velikih troškova istraživanje mora završiti prije nego što sve osobe iskuse događaj od interesa ili napuste promatranje zbog nekog razloga. Vrijeme promatranja nije isto za sve osobe, i ovisi o tome jesu li iskusile događaj od interesa ili napustile promatranje prijevremeno. Dakle, podaci su desno cenzurirani kada točno vrijeme doživljenja promatranog pojedinca postaje nepotpuno na desnoj strani promatranog perioda.

1.4.2 Lijevo cenzuriranja

Vrijeme je lijevo cenzurirano u c ukoliko je vrijeme doživljenja manje ili jednako c. Dakle, događaj od interesa za promatrani subjekt se dogodio prije njegovog ulaska u istraživanja i ne znamo točno kad se dogodio. Na primjer, promatrajmo starost djeteta u kojoj nauči neki zadatak. Ako je dijete već znalo to napraviti na početku istraživanja taj podatak će biti lijevo cenzuriran. Prema tome, lijevo cenzurirani podaci su oni kod kojih je vrijeme doživljenja promatranog subjekta nepotpuno na lijevoj strani. Pojavljuje se u medicini, na primjer kad ne znamo točno vrijeme prvog izlaganja virusu, već počinjemo pratiti osobu nakon pozitivnog testa na promatrani virus. Iako podaci mogu biti i lijevo cenzurirani, većina podataka u analizi doživljenja je desno cenzurirana.

1.4.3 Intervalno cenzuriranje

Kao kombinacija lijevog i desnog cenzuriranja javlja se pojam intervalnog cenzuriranja. Dakle, kod intervalnog cenzuriranja imamo dvostruko cenzurirane podatke.

Intervalno cenzuriranje je prikupljanje podataka unutar nekog intervala $\langle c_1, c_2 \rangle$, znači da je nepoznato vrijeme doživljenja $t_0 + t$ veće, odnosno manje od opaženog vremena

$t_0 + c_1$, odnosno $t_0 + c_2$ tj. $t_0 + c_1 < t_0 + t < t_0 + c_2$ $(c_1, c_2, t \in \mathbb{R}_+, c_1 < t < c_2)$. Točan trenutak neuspjeha znamo samo ako se dogodio unutar nekog intervala.

Ovaj tip cenzuriranja pojavljuje se u slučajevima gdje se promatranja ponavljaju nekoliko puta u različitim intervalima. Pojavljuje se u nekim medicinskim istraživanjima, gdje pacijenti dolaze na promatranja u nekoliko perioda, te se samo na promatranju može utvrditi da je nastupio događaj od interesa. Primjer takvog istraživanja je praćenje bolesnika svaka tri mjeseca da bi se otkrio povratak raka nakon operacije bolesnik.

Također, ovaj način promatranja je čest i kod ispitivanja raznih proizvoda, gdje inspekcija dolazi u periodima kako bi ispitala kvalitetu i valjanost samih proizvoda.

Osim cenzuriranja, problem pri provođenju studije može predstavljati i velika heterogenost podataka. Subjekti koji su dio studije se vrlo često razlikuju po određenim karakteristikama koje mogu imati utjecaj na konačni ishod. Iako se ovaj problem pokušava riješiti postavljanjem određenih kriterija pri ulasku u studiju, o njemu je potrebno voditi računa i pri postavljanju statističkih modela za svaku pojedinu studiju.

1.5 Odrezani podaci

Podaci u analizi doživljenja se često skupljaju u ograničenom vremenskom periodu ili unutar određenih vrijednosti. Na primjer, osiguravajuća društva često zanimaju samo štete veće od nekog određenog iznosa, tada oni nemaju nikakvih saznanja o štetama koje su manje od tog iznosa niti postoje li takve štete uopće. U takvim situacijama govorimo o rezanim podacima (eng. *truncation data*) i potrebna je primjena rezanih distribucija.

Kažemo da je uzorak podataka odrezan ukoliko se u njemu nalaze samo podaci za one osobe koje su iskusile događaj od interesa točno u unaprijed određenom intervalu $\langle Y_l, Y_d \rangle, l, d \in \mathbb{N}$. Osobe koje nisu iskusile događaj od interesa u navedenom intervalu nisu promatrane i istražitelji nemaju nikakve informacije o njima, one su 'odrezane' od istraživanja. To je i razlika između odrezanih i cenzuriranih podataka, jer smo kod cenzuriranih podataka imali barem djelomičnu informaciju i o tim osobama. Veoma je važno da znamo da su podaci s kojima radimo odrezani, jer se tada za procjene osnovnih parametara analize doživljenja koriste druge tehnike.

1.5.1 Lijevo odrezani podaci

Razanje slijeva javlja se kada subjekt (pojedinac ili predmet) ulazi u promatranje u određenoj dobi i prati se od vremena ulaska do vremena kada dolazi neuspjeh ili do cenzuriranja subjekata. Tada je Y_d beskonačan, a u obzir dolaze samo oni subjekti kojima se događaj od interesa dogodio nakon trenutka Y_l.

Ovakvu vrstu podataka možemo imati kada, primjerice, želimo procijeniti distribuciju veličina nekih sitnih čestica. Promjer tih čestica mjerimo pomoću mikroskopa, ali neke čestice su možda toliko malene da neće biti moguće niti pomoću mikroskopa izmjeriti njihov promjer. Njih ćemo zanemariti, odnosno nećemo uzeti nikakve podatke o njima. U ovom slučaju odrezali smo sve podatke vezane za čestice koje imaju promjer manji od onog kojeg može zabilježiti mikroskop.

Još jedan primjer lijevog rezanja može biti istraživanje nad stanovnicima umirovljeničkog doma. Da bi osoba mogla biti primljena u umirovljenički dom, ona mora imati dovoljan broj godina, stoga sve osobe koje su umrle ranije ne mogu biti dio istraživanja, i zato ove podatke smatramo lijevo odrezanima.

1.5.2 Desno odrezani podaci

Desno rezanje javlja se kada promatramo samo subjekte koji su doživjeli događaj od interesa.

Pretpostavimo da želimo procijeniti distribuciju udaljenosti zvijezda od zemlje. Neke zvijezde su toliko daleke da neće biti moguće izmjeriti njihovu udaljenost od zemlje, stoga njih nećemo uzeti u obzir, odnosno podaci o njima će biti odrezani.

Desno odrezani podaci pojavljuju se primjerice i u istraživanju gdje želimo procijeniti distribuciju vremena potrebnog za razvijanje neke zarazne bolesti, ukoliko se osoba zarazila transfuzijom. Tada imamo fiksiran datum kada ćemo uzimati podatke za ovo istraživanje, pa u obzir nećemo uzimati podatke za one osobe kojima je razdoblje potrebno za razvijanje bolesti veće od razdoblja koje je prošlo od dana transfuzije do datuma kada se prikupljaju podaci. Stoga, su podaci za osobe kojima se bolest nije razvila do dana uzimanja podataka desno odrezani.

Poglavlje 2

Distribucije doživljenja

2.1 Distribucija doživljenja

Podaci u analizi doživljenja, kako smo spomenuli u prethodnom poglavlju, predstavljaju vrijeme proteklo od odabrane točke početka promatranja do vremena nekog događaja koji nam je od interesa. Ova vremena su zapravo empirijska realizacija slučajne varijable T koju nazivamo vrijeme doživljenja. Određenu vrijednost koju T poprima označit ćemo s t.

Prema prethodnom poglavlju, događaj može označavati smrt, početak razvoja neke bolesti, pojavu tumora ili bilo koje iskustvo od interesa koje se može dogoditi pojedincu. U skladu s tim događaj nazivamo neuspjeh odnosno pad.

Pretpostavimo da promatramo homogenu populaciju. Neka slučajna varijabla T poprima vrijednosti u skupu $[0, +\infty\rangle$. Funkcija $F : [0, +\infty\rangle \to [0, 1]$ definirana formulom

$$F(t) = \mathbb{P}(T \leq t)$$

daje vjerojatnost da je vrijeme doživljenja manje ili jednako t i naziva se funkcijom distribucije varijable T. Funkcija distribucije je monotono neopadajuća i zadovoljava sljedeća dva svojstva:

$$F(0) = 0 \text{ i } \lim_{x\to\infty} F(t) = 1.$$

Pretpostavimo da je funkcija distribucije F derivabilna u gotovo svim točkama domene. Tada njenu derivaciju

$$\frac{d}{dt}F(t) = f(t)$$

nazivamo funkcijom gustoće slučajne varijable T.

Definicija 1 *Funkcija doživljenja $S : \mathbb{R}_+ \to [0, 1]$ slučajne varijable T definirana je sa*

$$S(t) = \mathbb{P}(T > t).$$

Funkcija doživljenja (eng. *survival function*) je vjerojatnost da slučajna varijabla T postigne vrijednost veću od t. Funkcija doživljenja se obično koristi u analizi doživljenja kao vjerojatnost da osoba živi duže od t godina ili da doživi neki drugi događaj od interesa tek nakon trenutka t.

Ako je T neprekidna slučajna varijabla, tada je i $S(t)$ neprekidna, nenegativna, padajuća funkcija. Kao što možemo vidjeti, funkcija doživljenja je komplementarna funkciji distribucije

$$S(t) = 1 - F(t),$$

te za funkciju doživljenja vrijede svojstva analogna svojstvima funkcije distribucije:

$$S(0) = 1 \text{ i } \lim_{t\to\infty} S(t) = 0.$$

Ako je $S(t)$ derivabilna, onda je funkcija gustoće

$$f(t) = -\frac{dS(t)}{dt}. \tag{2.1}$$

Stoga vrijedi

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(x)dx.$$

U analizi doživljenja spomenuta funkcija gustoće ponekad se naziva i krivulja smrtnosti, te prema definiciji zapravo predstavlja trenutačnu stopu umiranja.

Neka je sada vrijeme doživljenja T diskretna slučajna varijabla koja može poprimiti vrijednosti $\{t_j\}; t_1 < t_2 < \dots$. Tada su odgovarajuće vjerojatnosti neuspjeha u trenutku $t_j, j = 1, 2, \dots$ dane izrazom:

$$p_j = \mathbb{P}(T = t_j) > 0,\ j = 1, 2, 3, \dots . \tag{p}$$

Pomoću ovih vjerojatnosti lako je odrediti funkciju doživljenja, odnosno vjerojatnost preživljavanja nakon trenutka t.

Definicija 2 *Funkcija doživljenja diskretne slučajne varijable T je vjerojatnost da će jedinka preživjeti interval $[0, t\rangle$:*

$$S(t) = \mathbb{P}(T > t) = \sum_{t_j > t} p_j, \tag{2.2}$$

gdje su $t_j, j = 1, 2, \dots$ vrijednosti koje T može poprimiti te vrijedi $t_1 < t_2 < \dots$.

Diskretna slučajna varijabla može biti posljedica zaokruživanja rezultata istraživanja, grupiranja trenutaka u kojima je zabilježen neuspjeh u intervale i slično. U ovom slučaju $S(t)$ je također nerastuća funkcija.

Sada promatramo ove vjerojatnosti u dva uzastopna vremenska trenutka:

$$S(t_{j-1}) = p_j + p_{j+1} + \dots$$

$$S(t_j) = p_{j+1} + p_{j+2} + \dots$$

Oduzimanjem dobijemo:

$$p_j = S(t_{j-1}) - S(t_j). \tag{2.3}$$

Diskretna funkcija doživljenja $S(t)$ je padajuća step funkcija čije se vrijednosti mijenjaju samo u vremenima t_j te je, kao i funkcija doživljenja u neprekidnom slučaju, neprekidna zdesna:

$$S(t_j^-) > S(t_j) \text{ i } S(t_j^-) = S(t_{j-1}),\ \ t_{j-1} \leq t_j^- < t_j,$$

pri čemu je

$$S(t_j^-) = \lim_{t \uparrow t_j} S(t).$$

Alternativno, funkcijom doživljenja možemo smatrati vjerojatnost da će jedinka doseći interval $[t, +\infty\rangle$:

$$\widetilde{S}(t) = \mathbb{P}(T \geq t),$$

odnosno

$$\widetilde{S}(t) = S(t^-).$$

2.2 Funkcija hazarda

Uz funkciju doživljenja osnovna veličina u analizi doživljenja je funkcija hazarda (eng. *hazard function*). Interpretiramo je kao priliku da osoba dobi t doživi promatrani događaj u sljedećem trenutku. Poznata je i kao intenzitet smrtnosti u demografiji i aktuarstvu, te kao funkcija rizika ili jednostavno kao stopa hazarda.

Definicija 3 *Funkcija hazarda za neprekidnu slučajnu varijablu T, $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ definira se formulom*

$$\lambda(t) = \lim_{\triangle t \to 0^+} \frac{\mathbb{P}(t \leq T < t + \triangle t \mid T \geq t)}{\triangle t}\,. \tag{2.4}$$

Prema definiciji $\lambda(t)\Delta t$ možemo interpretirati kao "približnu" vjerojatnost da osoba dobi t doživi promatrani događaj u neposrednom sljedećem trenutku. Pomoću funkcije hazarda možemo vidjeti kako se vjerojatnost da subjekt iskusi događaj od interesa mijenja kroz vrijeme.

Izraz $\mathbb{P}(t \leq T < t + \triangle t \mid T \geq t)$ označava uvjetnu vjerojatnost, tj. vjerojatnost da slučajna varijabla T poprimi vrijednost iz intervala $\langle t, t + \Delta t]$ ako je njezina vrijednost veća od t. Iskoristimo li formulu za uvjetnu vjerojatnost događaja dobivamo sljedeće:

$$\begin{aligned}
\mathbb{P}(t \leq T < t + \triangle t \mid T \geq t) &= \frac{\mathbb{P}(t \leq T < t + \triangle t, T \geq t)}{\mathbb{P}(T \geq t)} \\
&= \frac{\mathbb{P}(t \leq T < t + \triangle t)}{\mathbb{P}(T \geq t)} \\
&= \frac{F(t + \triangle t) - F(t)}{S(t)}.
\end{aligned}$$

Uvrstimo li dobivenu jednakost u (2.4) slijedi:

$$\lambda(t) = \frac{1}{S(t)} \lim_{\triangle t \to 0} \frac{F(t + \triangle t) - F(t)}{\triangle t}.$$

Budući da je gornji limes jednak funkciji gustoće $f(t)$, funkciju hazarda možemo zapisati i u sljedećem obliku:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)}. \tag{2.5}$$

Postoji više oblika funkcije hazarda ovisno o problemu koji opisuje. Jedino zajedničko svojstvo svim funkcijama hazarda je da su nenegativne, tj. $\lambda(t) \geq 0$. Funkcija hazarda može za vrijeme pojave promatranog događaja posjedovati različite karakteristike. Primjerice, može biti konstantna, rasti, padati, može biti *bathtub* oblika (oblik 'kade') ili *humpshaped* oblika (oblik 'grbe'). Modeli sa rastućim funkcijama hazarda mogu se pojaviti kod prirodnog starenja. To je posljedica činjenice da je kod starijih ljudi veća stopa umiranja. Padajuće funkcije hazarda su rjeđe, pojavljuju se kod slučajeva kad je odmah u početku najveća vjerojatnost neuspjeha. Upotrebljavaju se kod nekim tipovima elektroničkih uređaja ili kod pacijenata koji su imali transplantacije. *Bathtub* oblik je najprikladniji za populacije koje se prate od rođenja. U ovom slučaju, u ranom periodu, intezitet (funkcija hazarda) je vrlo velik, zbog dječjih bolesti. S vremenom brzo pada, da bi se nakon toga stabilizirao, tj. postao gotovo konstantan. U posljednjoj fazi funkcija hazarda počinje rasti prateći prirodno starenje. Razni strojevi se također mogu odmah na početku pokvariti zbog neispravnih dijelova, a kasnije opet postati neispravni zbog dotrajalosti. *Humpshaped* oblik obično se koristi za modele nakon uspješnih operacija gdje je početni rizik veći (zbog infekcija, hemoragija ili drugih komplikacija), nakon čega slijedi stalan pad rizika prateći pacijentov oporavak.

Definicija 4 *Neka je dan skup $J = \{t_1, t_2, t_3, ...\}$. Diskretna funkcija hazarda, $\lambda : J \to \mathbb{R}_+$ dana je sa*

$$\lambda(t_j) = \mathbb{P}(T = t_j \mid T \geq t_j), \;\; t_j \in J.$$

Prema (p) vrijedi

$$\lambda(t_j) = \frac{p_j}{S(t_j^-)} = \frac{p_j}{p_j + p_{j+1} + \cdots},$$

gdje je $S(t_0) = 1$.

Koristeći (2.3) slijedi

$$\lambda(t_j) = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})},$$

pa imamo:

$$S(t_j) = S(t_j - 1)(1 - \lambda(t_j)), \; j = 1, 2, \ldots.$$

Stavimo li $S(t_0) = S(0) = 1$ i rekurzivno primijenimo gornju formulu za vremena $t_1 < t_2 < ...$ možemo primjetiti, da se funkcija doživljenja može napisati i kao produkt uvjetnih vjerojatnosti doživljenja

$$S(t) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})}.$$

Iz toga je lako izvesti relaciju koja povezuje funkciju hazarda i funkciju doživljenja za diskretnu slučajnu varijablu:

$$S(t) = \prod_{t_j \leq t} (1 - \lambda(t_j)).$$

Ova formula potvrđuje ono što je intuitivno jasno, da doživljenje trenutka t nužno zahtijeva i doživljenje svih trenutaka t_j između 0 i t.

Sada imamo i vjerojatnost pada u intervalu $[t_{j-1}, t_j\rangle$ koja je dana sa

$$\mathbb{P}(T = t) = \lambda(t) \prod_{t_j \leq t} (1 - \lambda(t_j)).$$

Definicija 5 *Funkcija kumulativnog hazarda $\Lambda : \mathbb{R}_+ \to \mathbb{R}_+$ pripadne funkcije distribucije F definirana je izrazom*

$$\Lambda(t) := \int_0^t \lambda(u)du, \tag{2.6}$$

gdje je λ funkcija hazarda definirana u (2.4).

U diskretnom vremenu kumulativni hazard definiramo:

$$\Lambda(t) = -\sum_{t_j \leq t} \log(1 - \lambda(t_j)).$$

Za relativno mali $\lambda(t_j)$ vrijedi da je $\log(1-\lambda(t_j))$ približno jedanak $-\lambda(t_j)$, pa kumulativni hazard postaje približno jednak izrazu:

$$\Lambda(t) = \sum_{t_j \leq t} \lambda(t_j).$$

Na osnovi funkcije hazarda možemo odrediti kojem modelu pripadaju podaci, npr. eksponencijalnom ili Weibullovom.

Primjer 6 (Weibullova funkcija) *Distribuciju čija je funkcija doživljenja dana formulom*

$$S(t) = e^{-\lambda t^\alpha},\ \lambda > 0, \alpha > 0$$

zovemo Weibullova funkcija.

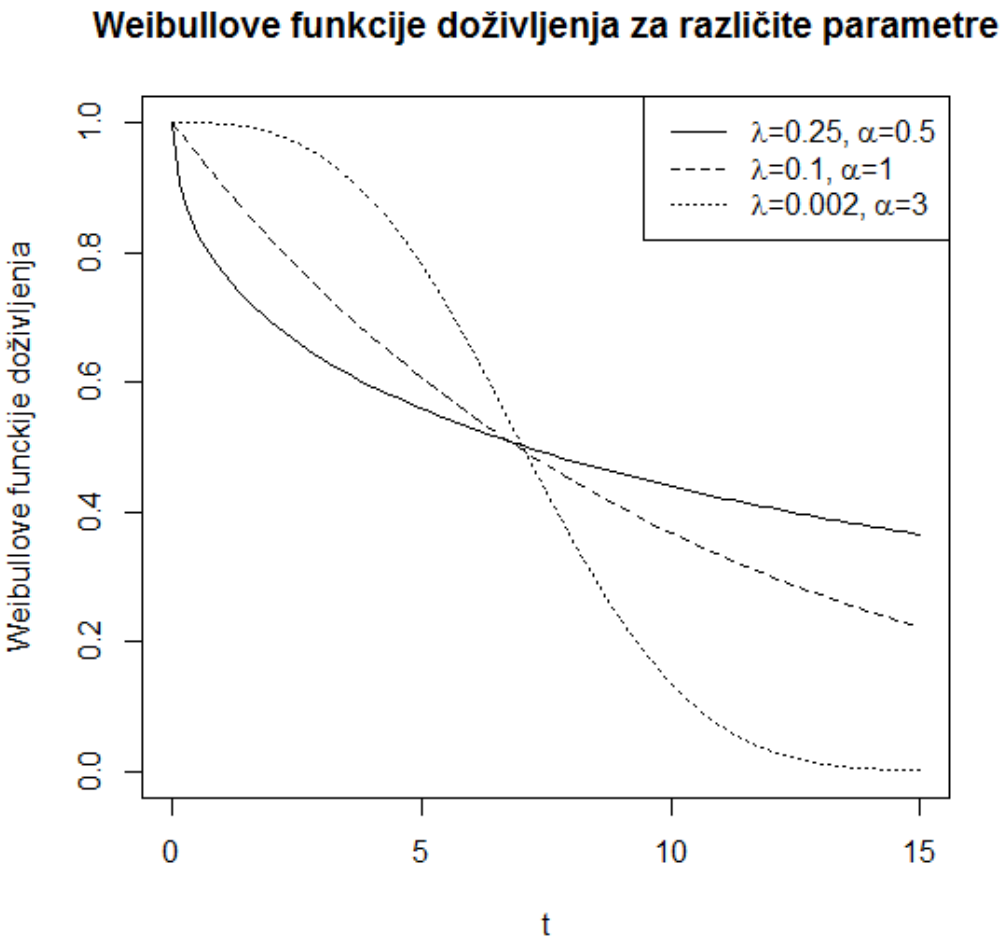
Za Weibullovu distribuciju funkcija rizika dana je izrazom

$$\lambda(t) = \lambda\alpha t^{\alpha-1},\ t \geq 0,$$

gdje je $\lambda > 0$ parametar oblika i $\alpha > 0$ parametar skaliranja.

Slika 2.1 prikazuje Weibullove funkcije doživljenja za različite parametre λ i α.

Sve krivulje doživljenja imaju jednake osnovne osobine. Monotono padajuće su funkcije čija vrijednost je 1 u 0, a 0 kada $t \to \infty$.



Slika 2.1

Funkcija rizika Weibullove distribucije je monotono rastuća za $\alpha > 1$, *padajuća za* $0 < \alpha < 1$ *i konstantna za* $\alpha = 1$.

Za Weibullovu distribuciju kumulativna funkcija rizika ima oblik

$$\Lambda(t) = \lambda t^{\alpha}$$

i ona je nelinearna funkcija od t.

Eksponencijalni model je posebni slučaj Weibullovog modela, za $\alpha = 1$.

Možemo uočiti da su vjerojatnosna funkcija gustoće $f(t)$, funkcija hazarda $\lambda(t)$ i funkcija doživljenja $S(t)$ zapravo samo različiti načini zapisivanja razdiobe vremena doživljenja T pa uspostavljamo veze između njih kako bismo iz informacija koje pruža jedna funkcija odredili informacije koje su nam potrebne.

Veze su dane sljedećim jednadžbama:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{-d\log[S(t)]}{dt}.$$

Integriranjem dobijemo

$$\int_0^t \lambda(u)du = -\log S(u)|_0^t,$$

te iz činjenice da je $S(0) = 1$, dolazimo do izraza za kumulativni hazard:

$$\Lambda(t) = -\log S(t).$$

Djelujemo li sad eksponencijalnom funkcijom, dobit ćemo relaciju koja povezuje funkciju doživljenja i funkciju hazarda (odnosno kumulativnog hazarda):

$$S(t) = \exp[-\Lambda(t)] = \exp\left[-\int_0^t \lambda(u)du\right], \tag{2.7}$$

te

$$f(t) = \lambda(t)\exp\left(-\int_0^t \lambda(u)du\right) = \lambda(t)\exp(-\Lambda(t)).$$

Iskoristimo li činjenicu da je $S(\infty) = 0$ te da $\lambda(t) \geq 0$, slijedi

$$\lim_{t\to\infty}\int_0^t \lambda(u)du = \infty.$$

Distribucija slučajne varijable T je u potpunosti određena jednom od ovih jednadžbi. Također, možemo primjetiti da su funkcija doživljenja i funkcija hazarda obrnuto proporcionalne veličine za dano t, što je veća funkcija doživljenja to je manja funkcija hazarda i obrnuto.

2.3 Očekivano trajanje života

Očekivano trajanje života, kako smo naveli u prvom poglavlju, označavamo s e_t^0 i govori nam koliko je očekivano trajanje ostatka života osobe koja ima t godina.

Definicija 7 *Očekivano trajanje ostatka života u trenutku t definiramo kao:*

$$e_t^0 := \mathbb{E}[T - t \mid T > t]$$

Ako je T neprekidna slučajna varijabla, vrijedi:

$$e_t^0 = \frac{\int_t^\infty (x-t)f(x)dx}{S(t)}.$$

Nadalje, ako iskoristimo prethodno dokazanu jednakost $f(x) = -S'(x)$, imamo:

$$\begin{aligned}\mathbb{E}[T-t|T>t]S(t) &= \int_t^\infty (x-t)f(x)dx = \\ &= -(x-t)S(t)|_t^\infty + \int_t^\infty S(x)dx = \\ &= 0 + \int_t^\infty S(x)dx.\end{aligned}$$

Sada očekivano trajanje života možemo zapisati kao

$$e_t^0 = \frac{\int_t^\infty (x-t)f(x)dx}{S(t)} = \frac{\int_t^\infty S(x)dx}{S(t)}.$$

Uvrštavanjem $t = 0$ i $S(0) = 1$ u gornju jednadžbu dobivamo formulu za očekivano ukupno trajanje života , $\mu := e_0^0$.

Očekivano ukupno trajanje života ili srednja stopa neuspjeha μ definira se kao matematičko očekivanje slučajne varijable T:

$$\mu = \mathbb{E}[T] = \int_0^\infty xf(x)dx = \int_0^\infty x\lambda(x)S(x)dx.$$

Možemo uočiti da je očekivano trajanje preostalog života jednako površini ispod krivulje funkcije doživljenja za vrijednosti veće od t, podijeljeno sa $S(t)$. S druge strane za očekivano trajanje života μ vrijedi da je ono jednako ukupnoj površini ispod krivulje funkcije doživljenja.

Sada lako možemo uspostaviti vezu između funkcije doživljenja i varijance:

$$\mathbb{E}[T^2] = \int_0^\infty x^2 f(x)dt = -x^2S(t)|_t^\infty + \int_t^\infty 2xf(x)dx = 0 + 2\int_0^\infty xf(x)dx,$$

$$\mathbb{V}ar(T) = \mathbb{E}([T-\mathbb{E}(T)]^2) = \mathbb{E}[T^2] - (\mathbb{E}[T])^2 = 2\int_0^\infty xf(x)dx - \left(\int_0^\infty S(x)dx\right)^2.$$

Definicija 8 *p-ti kvantill distribucije T je najmanji t_p takav da zadovoljava:*

$$S(t_p) \leq 1-p,$$

odnosno, $t_p = \inf\{x : S(x) \leq 1-p\}$.

Definicija 9 *Medijan životnog vijeka je 0.5-kvantil distribucije slučajne varijable T, označava se sa $T_{0.5}$.*

Za neprekidnu varijablu T vrijedi da medijan životnog vijeka zadovoljava jednadžbu

$$S(T_{0.5}) = 0.5.$$

2.4 Uvjetna vjerojatnost neuspjeha i središnja stopa

U analizi podataka o smrtnosti često nas zanima vjerojatnost da osoba koja je doživjela vrijeme t^* preživi i do vremena t za $t \geq t^*$, tj. ne moramo nužno proučavati samo vrijeme doživljenja, već možemo promatrati vrijeme do određenog događaja koji nam je od važnosti. Možemo govoriti o vjerojatnosti da se događaj dogodio u malom intervalu između t i $t + \Delta t$. Nadalje, hoćemo da ta vjerojatnost bude uvjetna s obzirom na individualno vrijeme doživljenja za neko vrijeme t. To je važno, jer ako je osoba već umrla, onda nije rizična za taj događaj. Dakle, želimo one osobe koje ulaze na početak intervala $[t, t + \Delta t\rangle$.

Uvjetna vjerojatnost neuspjeha u intervalu $[t, t + \Delta t\rangle$, uz dano doživljenje vremena (dobi) t, je približno $\lambda(t)\Delta t$, dok je za bezuvjetnu vjerojatnost neuspjeha u $[t, t + \Delta t\rangle$ približno vrijedi

$$f(t)\Delta t = S(t)\lambda(t)\Delta t.$$

Također, prema (2.5) za vjerojatnosnu funkciju gustoće vrijedi

$$f(t) = S(t)\lambda(t). \tag{2.8}$$

Uočimo da funkciju doživljenja možemo izraziti u terminima budućeg životnog vijeka

$$S(t) = \int_t^\infty f(u)du = \int_t^\infty \lambda(u)S(u)du,$$

dok je u (2.7) funkcija doživljenja izražena u teminima doživljenog.

Zbog jednostavnosti zapisa umjesto Δt pisat ćemo x. U aktuarskoj literaturi uvjetna vjerojatnost da će osoba stara t godina umrijeti unutar narednih x godina označava se s ${}_xq_t$. Sa x je prikazana duljina razdoblja tijekom kojeg se računa ova vjerojatnost. Za $x = 1$ koristit ćemo kraću oznaku q_t umjesto ${}_xq_t$.

Imamo:

$$\begin{aligned} {}_xq_t &= \mathbb{P}(t \leq T < t + x | T \geq t) = \frac{\int_t^{t+x} f(u)du}{\int_t^0 f(u)du} \\ &= \frac{\int_t^{t+x} \lambda(u)S(u)du}{\int_t^0 \lambda(u)S(u)du} = \frac{S(t) - S(t+x)}{S(t)}. \end{aligned}$$

Uočimo da je $S(t) - S(t + x)$ udio neuspjeha između dobi t i $t + x$. S druge strane, očekivana stopa smrtnosti koju smo označili sa e_t^0 dana je izrazom

$$e_t^0 = \frac{\int_t^{t+x} \lambda(u)S(u)du}{\int_t^{t+x} S(u)du} = \frac{S(t) - S(t+x)}{\int_t^{t+x} S(u)du}.$$

Slijedi da je vjerojatnost da će osoba koja je doživjela t godina preživjeti još barem x godina jednaka

$${}_xp_t = 1 -_x q_t = \frac{S(t+x)}{S(t)}.$$

Za skup točaka $t_0 = 0 < t_1 < t_2 < ... < t_k$ imamo

$$\begin{aligned} S(t_k) &= \exp\left[-\int_t^{t_k} \lambda(u)du\right] = \exp\left[-\sum_{i=0}^{k-1}\int_{t_i}^{t_{i+1}} \lambda(u)du\right] \\ &= \prod_{i=0}^{k-1} \exp\left[-\int_{t_i}^{t_{i+1}} \lambda(u)du\right] = \prod_{i=0}^{k-1} p_{x_i} = \prod_{i=0}^{k-1}(1 -_x q_t). \end{aligned} \tag{2.9}$$

Formula (2.9) je vrlo važna za izgradnju tablica smrtnosti, o kojoj smo više govorili u prvom poglavlju.

2.5 Parametarska distribucija doživljenja

S ciljem pronalaska veze između vremena doživljenja promatranih subjekata i unaprijed određenih varijabli koristi se statističko modeliranje. Ovakvim pristupom prema analizi vremena doživljenja možemo otkriti kako za pojedine grupe subjekata vrijeme doživljenja ovisi o tim varijablama.

Jedna od mogućih klasa modela koju ćemo u ovom radu koristiti u statističkoj analizi su parametarski modeli. Parametarski model (ili parametarska familija) je familija distribucija određena konačnim brojem parametara. Ukoliko se radi o neprekidnom modelu, zapisujemo ga kao:

$$\mathcal{P} = \{f(\cdot;\theta) : \theta \in \mathbf{\Theta}\},$$

gdje je f funkcija gustoće vremena doživljenja, θ vektor parametara i $\mathbf{\Theta}$ prostor dozvoljenih vrijednosti koje vektor parametara može poprimiti.

2.5.1 Postupci prilagodbe funkcije parametarske distribucije

Prilagodba parametarskih distribucija doživljenja pretpostavlja da je oblik funkcije doživljenja (tj. familija kojoj pripada) već poznat. Potrebno je procijeniti vrijednosti parametara koji se pojavljuju u matematičkoj formuli funkcije distribucije doživljenja koju promatramo. Razlikujemo dvije glavne grupe metoda koje se u primijenjenim istraživanjima standarno koriste za procjenu parametara: grafičke i analitičke. Obje grupe metoda imaju svoje prednosti i nedostatke.

Brzina i jednostavnost su prednosti grafičkih metoda. Većinom se zasnivaju na crtanju nekih funkcija funkcije hazarda ($\lambda(t)$) ili kumulativne funkcije hazarda ($\Lambda(t)$) nasuprot nekih funkcija varijable t. Funkcije su obično odabrane tako da će, ukoliko je parametarski oblik funkcije doživljenja prikladan, dobiveni graf biti približno jednak ravnoj liniji. Upotreba posebnog grafičkog papira kojeg nazivamo hazardni papir ili vjerojatnosni papir može olakšati crtanje.

Od analitičkih metoda standarno se koristi metoda maksimalne vjerodostojnosti koja se smatra vrlo pouzdanom i o njoj ćemo detaljnije govoriti u nastavku.

Iako se statističke metode često koriste kod rješavanja problema procjene parametara, rezultati dobiveni korištenjem ovih metoda nisu pouzdani u slučaju malog uzorka podataka, te se ne preporučuju za korištenje u takvim situacijama.

2.5.2 Procjena parametara

Kao što smo prethodno spomenuli, parametarski modeli su poznati do na nepoznati parametar $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}^k$. Budući da nam je parametar nepoznat cilj nam je na temelju podataka procijeniti vrijednost parametra. Za pravu vrijednost nepoznatog parametra koristimo oznaku θ, a za procjenitelja koristimo oznaku $\widehat{\theta}$. Definiciju navodimo u nastavku.

Definicija 10 *Neka je $\mathcal{P} = \{F_\theta : \theta \in \mathbf{\Theta} \subseteq \mathbb{R}^k\}$ parametarski statistički model, θ k-dimenzionalni parametar, $\mathbf{\Theta}$ prostor dozvoljenh vrijednosti parametara. Neka je $\mathbf{X} = (X_1, \ldots, X_n)$ slučajni uzorak s distribucijom iz danog modela i $t : \mathbb{R}^n \to \mathbf{\Theta}$ funkcija. Slučajan vektor $T = t(\mathbf{X})$ je procjenitelj za θ.*

2.5.2.1 Procjenitelj maksimalne vjerodostojnosti

Metoda maksimalne vjerodostojnosti (eng. *Maximum Likelihood Method*), kraće ML-metoda, jedna je od najpopularnijih općih statističkih metoda za nalaženje dobrih procjenitelja nepoznatih parametara u parametarskim modelima. Njezina prednost je u tome što se može primijeniti na većinu teorijskih distribucija, te na razne uzorke cenzuriranih podataka. Osim toga, uz određene uvjete, ova metoda ima dobra statistička svojstva kao što su konzistetnost i asimptotska normalnost. Otkriće ML-metode pripisuje se engleskom statističaru A. Fisheru $(1890. - 1962.)$.

Parametarski modeli, kao što su Weibullov ili eksponencijalni model, lako se mogu procjenjivati. Parametri se procjenjuju metodom maksimalne vjerodostojnosti za što nam je potrebno odrediti funkciju vjerodostojnosti.

Definicija 11 *Neka je $(x_1, x_2, \ldots, x_n)$ vrijednost slučajnog uzorka $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ iz parametarskog statističkog modela s pripadnom funkcijom gustoće $f(x|\theta)$, gdje je $\theta \in \Theta$ nepoznati parametar. Funkcija vjerodostojnosti $L : \Theta \to \mathbb{R}$ je zadana formulom:*

$$L(\theta) := \prod_{i=1}^{n} f(x_i|\theta).$$

Definicija 12 *Procjena metodom maksimalne vjerodostojnosti parametra θ je ona vrijednost $\widehat{\theta} \in \mathbf{\Theta}$ za koju funkcija vjerodostojnosti poprima maksimalnu vrijednost:*

$$L(\widehat{\theta}) = \max_{\theta \in \mathbf{\Theta}} L(\theta).$$

Očito je $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$. Dakle, procjenitelj maksimalne vjerodostojnosti, kraće MLE, parametra θ je statistika $\widehat{\theta}(X_1, X_2, \ldots, X_n)$.

Zbog strogog rasta logaritamske funkcije problem maksimiziranja $L(\theta)$ ekvivalentan je problemu maksimizacije logaritma funkcije vjerodostojnosti:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Ako je $l(\theta)$ derivabilna funkcija i θ k-dimenzionalni parametar s vrijednostima u otvorenom skupu $\mathbf{\Theta}$, odnosno $\theta = (\theta_1, \ldots, \theta_k)$, onda možemo odrediti ML-procjenitelj kao stacionarnu točku od l, tj. rješavajući sustav jednadžbi $\frac{\partial l(\theta)}{\partial \theta_i} = 0$ za svaki $i = 1, \ldots, k$.

Gornji račun vrijedi uz pretpostavku da podaci nisu cenzurirani. Budući da su podaci u istraživanjima najčešće cenzurirani, moramo tome prilagoditi funkciju vjerodostojnosti.

Svojstva ML-procjenitelja:

1. ML-procjenitelj ne mora biti nepristran,
2. ako je ML-procjenitelj jedinstven, onda je on funkcija bilo koje dovoljne statistike,
3. ako je ML-procjenitelj jedinstven i postoji potpuna dovoljna statistika, tada vrijedi da ako je ML-procjenitelj nepristran, onda je ujedno i najmanje varijance.

Primjer 13 (Procjena parametra metodom maksimalne vjerodostojnosti) *Neka je* $(X_1, \ldots, X_n)$ *jednostavan slučajan uzorak iz eksponencijalnog modela s parametrom* $\lambda > 0$. *Slučajna varijabla* X *iz eksponencijalne distribucije dana je sljedećom funkcijom gustoće:*

$$f(x) = \lambda e^{-\lambda x}, \text{ za } x > 0.$$

Funkcija vjerodostojnosti dana je sljedećim izrazom:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x} = \theta^n e^{-\lambda \sum_{i=1}^{n} x_i} = \lambda^n e^{-n\lambda \overline{x}}.$$

ML-procjenitelj za λ *tražimo rješavajući jednadžbu log-vjerodostojnosti:*

$$l(\lambda) = \log \lambda^n - \lambda \sum_{i=1}^{n} x_i. \tag{2.10}$$

Deriviranjem izraza (2.10) te izjednačavanjem s nula dobivamo:

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Rightarrow \lambda = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i},$$

$$l''(\lambda) = -\frac{n}{\lambda^2} < 0,$$

pa je $\widehat{\lambda} = \frac{1}{\overline{X}}$ *ML-procjenitelj za* λ *iz* $Exp(\lambda)$*-modela.*

Procjenitelji dobiveni ovom metodom imaju dobra i lako odrediva asimptotska svojstva, te su stoga posebno dobri za procjenitelje na osnovi velikih uzoraka. Nažalost ML-procjenitelj ne mora uvijek postojati, a najčešće nije ni jedinstven. Ponekad je ML-procjenitelje teško izračunati, ali mnogi računalni programi kao što su R i Matlab dostupni su za pomoć u njihovu računanju.

Poglavlje 3

Parametarski proporcionalni model hazarda

U mnogim znanstvenim istraživanjima, ali i u financijama, važno je utvrditi ovisi li neka odabrana veličina, (npr. duljina života) o drugim mjerenim veličinama npr. spolu, potrošnji nekog proizvoda, visini u odrasloj dobi, itd. Veza između takvih mjerenja je vrlo rijetko jasna i deterministička, pa je najčešće predstavljamo koristeći vjerojatnosne modele.

Veličinu od interesa modeliramo kao slučajnu varijablu koju nazivamo zavisnom varijablom ili odzivom (eng. *response*), a sva ostala mjerenja nazivamo nezavisnim varijablama ili poticajnim, te prediktorima ili kovarijatama (eng. *predictors*).

Podatke koje želimo opisati tipično reprezentiramo kao niz parova $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, gdje je y_i realizacija slučajne varijable Y_i čija razdioba ovisi o kovarijatama $\mathbf{x}_i$. Kovarijate mogu primati numeričke vrijednosti ili kategorijalne vrijednosti. U potonjem slučaju zovemo ih faktorima (npr. spol ili kategorija vozila). Također, kovarijate mogu biti proizvoljno velike dimenzije.

Problem kod promatranja populacija koje su heterogene, tj. gdje se pojedinci razlikuju po obilježjima koja mogu imati utjecaj na ishod, rješavamo uvođenjem regresijskih modela. Regresijski model uključuje utjecaj različitih faktora na zavisnu varijablu, odnosno na ishod.

Najpoznatiji takav model je jednostavna linearna regresija. Jednostavnom linearnom regresijom opisuje se odnos među pojavama za koje je svojstveno da svakom jediničnom porastu vrijednosti jedne varijable odgovara približno jednaka linearna promjena druge varijable. Pretpostavljamo da slučajna varijabla Y_i na linearan način ovisi o nekoj kovarijati X_i, tj. linearna regresija ima oblik

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

gdje je (ε_i) niz nezavisnih jednako distribuiranih slučajnih varijabli s očekivanjem 0 i varijancom σ^2. Ovaj model ima tri parametra β_0, β_1 i σ^2.

U ovom odjeljku predstavljamo jedan način generalizacije modela doživljenja na regresijski model doživljenja, tj. dozvoljavamo da uzorak bude heterogen dodavanjem vektora kovarijata $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Kao i kod drugih regresijskih modela, $\mathbf{X}$ može

predstavljati kombinaciju binarnih, polinomnih, neprekidnih, spline-proširenih, pa čak i ordinalnih kovarijata ako postoji linearna zavisnost s varijablom odziva.

3.1 Model proporcionalnog hazarda

Modeli proporcionalnog hazarda su klasa statističkih modela doživljenja. Općenito, funkcija hazarda ovisi i o vremenu i o skupu kovarijata, od kojih neke mogu biti ovisne o vremenu. Modeli proporcionalnog hazarda odvajaju te dvije komponente pa je funkcija hazarda za vrijeme neuspjeha T u trenutku t, za pojedinca čiji je vektor kovarijata $\mathbf{X}$, dana formulom:

$$\begin{aligned} \lambda(t|\mathbf{X}) &= \lambda(t)\exp(\mathbf{X}\beta) \\ &= \lambda(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n), \end{aligned}$$

gdje je $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ vektor regresijskih koeficijenata, a $\lambda(t)$ osnovna funkcija hazarda.

Ovu regresijsku formulaciju modela proporcionalnog hazarda skraćeno ćemo označavati s PH. Dakle, PH model nam daje izraz za rizik u vremenu t za subjekte sa skupom nezavisnih varijabli $\mathbf{X}$. $\mathbf{X}$ predstavlja vektor nezavisnih varijabli koje su modelirane da predvide pojedinačne rizike.

Uočavamo da je rizik u trenutku t umnožak dvije veličine. Prva veličina $\lambda(t)$ predstavlja osnovnu funkciju hazarda ili funkciju hazarda za standarni subjekt, a druga veličina je eksponencijalni izraz. Bitno svojstvo formule, a tiče se PH pretpostavke (pretpostavke o proporcionalnom hazardu) je to da je osnovna funkcija hazarda samo formulacija vremena t, a ne uključuje kovarijate. Suprotno tome, eksponencijalni izraz uključuje kovarijate, ali ne uključuje t. Za ovakve kovarijate kažem da su vremenski nezavisne varijable. U ovom radu bavit ćemo se isključivo vremenski nezavisnim kovarijatama.

Vremenski nezavisna varijabla definirana je kao varijabla čija se vrijednost za dani subjekt ne mijenja kroz vrijeme. Primjetimo da se varijable kao što su starosna dob i težina mijenjaju kroz vrijeme.

Osnovna funkcija hazarda $\lambda(t)$ ne ovisi o kovarijatama tj. dobiva se kada sve kovarijate izjednačimo sa 0 ($X_1 = X_2 = \ldots = X_n = 0$), zato je i nazivamo osnovnom funkcijom. Uvrštavanjem u PH dobivamo:

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

Dakle, model se svodi na funkciju hazarda $\lambda(t)$ za baznu vrijednost kovarijata $(X_1, X_2, \ldots, X_n) = (0, 0, \ldots, 0)$. Slijedi da $\lambda(t)$ može biti smatrana osnovnom verzijom funkcije hazarda prije uključivanja i razmatranja nezavisnih varijabli.

Za $\lambda(t)$ može se upotrijebiti bilo koja proporcionalna funkcija hazarda, a kasnije ćemo pokazati da $\lambda(t)$ može biti potpuno neodređena bez umanjivanja sposobnosti procjene β. Ovisno o tome ima li osnovna funkcija hazarda $\lambda(t)$ konstantan skalarni parametar β_0, $\mathbf{X}\beta$

može uključiti ili isključiti β_0. $\mathbf{X}\beta$ nazivamo linearni prediktor. Eksponencijalna forma koristi se zbog pozitivnosti, a izraz $\exp(\mathbf{X}\beta)$ može se nazvati relativna funkcija hazarda i u mnogim je slučajevima funkcija primarnog interesa jer opisuje relativne učinke kovarijata na doživljenje. Takav model implicira da je omjer hazarda kod dva subjekta iz populacije konstantan tijekom danog vremena i da se kovarijate ne mijenjaju tijekom vremena. Kovarijate kao konstante su poželjne budući da se najčešće radi o podacima koji se ne mijenjaju kroz vrijeme, kao što su spol i rasa.

Model PH može se također izraziti pomoću kumulativnih funkcija hazarda i funkcije doživljenja:

$$\Lambda(t|\mathbf{X}) = \Lambda(t)\exp(\mathbf{X}\beta)$$
$$S(t|\mathbf{X}) = \exp[-\Lambda(t)\exp(\mathbf{X}\beta)] = \exp[-\Lambda(t)]^{\exp(\mathbf{X}\beta)},$$

gdje je $\Lambda(t)$ osnovna kumulativna funkcija hazarda, a $S(t|\mathbf{X})$ vjerojatnost doživljenja do nekog određenog trenutka t s obzirom na vrijednosti kovarijata $\mathbf{X}$.

Propozicija 14

$$S(t|\mathbf{X}) = S(t)^{\exp(\mathbf{X}\beta)},$$

gdje je $S(t)$ osnovna funkcija doživljenja, $\exp(-\lambda(t))$.

Dokaz.

$$\begin{aligned} S(t) &= \exp\left(-\int_0^t \lambda(u|x)du\right) \\ &= \exp\left(-\exp(\mathbf{X}\beta)\int_0^t \lambda(u)du\right) \\ &= \left(\exp\left(-\int_0 t\lambda(u)du\right)\right)^{\exp(\mathbf{X}\beta)} \\ &= S(t)^{\exp(\mathbf{X}\beta)}. \end{aligned}$$

■

Propozicija 15

$$f(t|\mathbf{X}) = \exp(\mathbf{X}\beta)[S(t)]^{\exp(\mathbf{X}\beta)-1}f(t),$$

gdje je $f(t)$ funkcija gustoće (2.1).

Dokaz.

$$\begin{aligned} f(t|\mathbf{X}) &= \lambda(t|\mathbf{X})S(t|\mathbf{X}) \\ &= \lambda(t)\exp(\mathbf{X}\beta)[S(t)]^{\exp(\mathbf{X}\beta)} \\ &= \exp(\mathbf{X}\beta)[S(t)]^{\exp(\mathbf{X}\beta)-1}f(t). \end{aligned}$$

U zadanjoj jednakosti iskoristili smo (2.4). ■

Dakle, učinak povećanja udjela u kovarijatama multiplikativan je s obzirom na funkciju hazarda i kumulativnu funkciju hazarda s faktorom $\exp(\mathbf{X}\beta)$, ili ekvivalentno potenciranju funkcije doživljenja na $\exp(\mathbf{X}\beta)$.

3.2 Pregled modela i tumačenje parametara

U prethodnom poglavlju spomenuli smo prednosti log-hazarda ili log-kumulativnog hazarda koji omogućuje jednostavnije razdvajanje i provjeru dijelova distribucije i regresije. Sada navodimo dva izraza pomoću kojih možemo lineralizirati model PH s obzirom na $\mathbf{X}\beta$:

$$\begin{aligned} \log \lambda(t|\mathbf{X}) &= \log \lambda(t) + \mathbf{X}\beta \\ \log \Lambda(t|\mathbf{X}) &= \log \Lambda(t) + \mathbf{X}\beta. \end{aligned}$$

Bez obzira koji se od tri izraza modela koristi, prije nego model stavimo u upotrebu, potrebno je provjeriti njegovu adekvatnost. Prvo treba provjeriti treba li neke od uključenih varijabli transformirati prije uključivanja u model. Zatim imamo određene pretpostavke u parametarskom modelu doživljenja PH:

1. Oblik osnovnih funkcija (λ, Λ i S) treba točno odrediti.

2. Pri modeliranju funkcije hazarda odnos log-hazarda i jedne ili više varijabli, koje su praćene za svakog subjekta, je linearan.

3. Način na koji kovarijate utječu na distribuciju treba biti množenje hazarda ili kumulativnog hazarda s $\exp(\mathbf{X}\beta)$ ili ekvivalentno dodavanje $\mathbf{X}\beta$ na log-hazard ili log-kumulativni hazard za svaki t. Pretpostavlja se da je utjecaj kovarijata isti kod svih vrijednosti t jer se $\log \lambda(t)$ može odvojiti od $\mathbf{X}\beta$.

Regresijski koeficijent za X_j, β_j, predstavlja povećanje log hazarda ili log kumulativnog hazarda u bilo kojoj fiksnoj točki u vremenu, ako se X_j povećava za jednu jedinicu, a sve ostale kovarijate se održavaju konstantnim. To formalno zapisujemo na sljedeći način:

$$\beta_j = \log \lambda(t \mid X_1, X_2, \ldots, X_j + 1, X_{j+1}, \ldots, X_n) - \log \lambda(t \mid X_1, \ldots, X_j, \ldots, X_n),$$

što je ekvivalentno zapisu log omjera hazarda u vremenu t.

Regresijski koeficijent se može jednako lako napisati kao omjer hazarda u vremenu t. Omjer hazarda za $X_j + d$ prema X_j, kada su sve ostale kovarijate konstantne, je $\exp(\beta_j d)$. Stoga je učinak povećanja X_j za d zapravo povećanje hazarda događaja za faktor $\exp(\beta_j d)$ u svim vremenskim točkama, sve to uz pretpostavku da je X_j linearno povezan s $\log \lambda(t)$.

Općenito, omjer hazarda za pojedinca s prediktorskim vektorom vrijednosti $\mathbf{X}^*$ u odnosu na pojedinca s kovarijatama $\mathbf{X}$ je

$$\begin{aligned} \frac{\lambda(t|\mathbf{X}^*)}{\lambda(t|\mathbf{X})} &= \frac{\lambda(t)\exp(\mathbf{X}^*\beta)}{\lambda(t)\exp(\mathbf{X}\beta)} \\ &= \frac{\exp(\mathbf{X}^*\beta)}{\exp(\mathbf{X}\beta)} \\ &= \exp[(\mathbf{X}^* - \mathbf{X})\beta]. \end{aligned}$$

Omjer hazarda za dva pojedinca ovisi o razlici njihovih linearnih prediktora i ne ovisi o vremenu. Lakše je predstaviti omjer hazarda koji ima vrijednost veću od 1 nego omjer hazarda koji je manji od 1, a to će se dogoditi ako vrijedi

$$\lambda(t \mid \mathbf{X}^*) \geq \lambda(t \mid \mathbf{X}).$$

Pretpostavimo sada da postoji samo jedna varijabla od interesa, X_1, koja može poprimiti vrijednost 0 ili 1 (tj. binarna je). PH se može zapisati kao :

$$\begin{aligned} \lambda(t|X_1 = 0) &= \lambda(t) \\ \lambda(t|X_1 = 1) &= \lambda(t)\exp(\beta_1). \end{aligned}$$

U slučaju istinitosti pretpostavke o proporcionalnom hazardu u svakom trenutku t, imamo da je $\exp(\beta_1)$ omjer hazarda kada je $X_1 = 1$ i hazarda kada je $X_1 = 0$. Ovaj jednostavan slučaj nema pretpostavku regresije, ali pretpostavlja PH i oblik funkcije $\lambda(t)$. Ako je jedina kovarijata X_1 kontinuirana, model je oblika

$$\lambda(t \mid X_1) = \lambda(t)\exp(\beta_1 X_1).$$

Bez daljnje modifikacije (kao što je transformacija kovarijata), kada model pretpostavlja pravac za log-hazard ili za sve t, povećanje X_1 povećava funkciju hazarda za faktor $\exp(\beta_1)$ po jednoj jedinici.

3.3 Omjer hazarda, relativni rizik, razlike rizika

Drugi načini modeliranja kovarijata također se mogu navesti uz multiplikativni učinak na hazard. Na primjer, može se pretpostaviti da je učinak kovarijata pribrojen funkciji hazarda neuspjeha, umjesto da se pomnoži s faktorom. Učinak kovarijata može se također opisati preko omjera smrtnosti (relativnog rizika), razlika u riziku, omjera vjerojatnosti ili povećanja očekivanog vremena neuspjeha. Međutim, kao što je omjer vjerojatnosti prirodan način opisivanja učinka za binarnu razdiobu, tako je relativni rizik često prirodan način opisivanja utjecaja na vrijeme doživljenja. Jedan od razloga je zato što omjer hazarda može biti konstantan.

Rizik (eng. *risk*) je omjer učestalosti (incidencija) neke pojave u nekom vremenskom intervalu i broja izloženih riziku te pojave na početku intervala

$$rizik = \frac{učestalost\ pojave}{broj\ izloženih\ na\ početku}.$$

Neka je dan eksperiment u kojem želimo saznati koji je učinak nekog tretmana na ispitanike. Promatramo dvije skupine ispitanika koje se razlikuju samo po tome primaju li tretman ili ne. Grupa koja prima tretman u eksperimentu naziva se ispitna ili tretirana skupina, skupina koja prima tretman naziva se kontrolna skupina. Kontrolna skupina trebala bi se razlikovati od ispitne po istraživanoj pojavi i istraživanim kovarijatama. Ona nam pruža osnovnu liniju koja nam omogućava da vidimo ima li odabrani tretman učinak.

Faktor, koji predstavlja razliku kontrolne i ispitne skupine, je nezavisna varijabla. Nezavisna je varijabla jer ne ovisi o tome što se događa u eksperimentu. Umjesto toga, to je nešto što primjenjuje ili odabire osoba koja vrši eksperiment. Nasuprot tome, zavisna varijabla u eksperimentu je traženi odgovor, mjera rizika kojom utvrđujemo je li učinak djelovao. Zavisna varijabla ovisi o nezavisnoj, a ne obrnuto.

Omjer broja oboljelih i broja osoba pod rizikom za razvoj bolesti predstavlja mjeru koju zovemo apsolutni rizik. Za usporedbu rizika između skupina, statistika ipak odabire relativni rizik.

Relativni rizik (eng. *relative risk*), RR jednak je omjeru apsolutnog rizika osoba u izloženoj skupini i apsolutnog rizika osoba u neizloženoj skupini. Formalno, ako je π_1 vjerojatnost da se događaj dogodio u skupini 1, a π_2 vjerojatnost da se događaj dogodio u skupini 2, onda je relativni rizik

$$RR = \frac{\pi_1}{\pi_2}.$$

Razlog preferiranja relativnog rizik u odnosu na razlike rizika $\pi_1 - \pi_2$ leži u činjenici da je populacijski rizik za većinu bolesnika prilično mali, a time su i razlike manje drastične. Na primjer, ako je vjerojatnost nekog raka u jednoj 0.001, a u ostalim 0.009, razlika je 0.008 (isto kao i između 0.419 i 0.411), ali relativni rizik je 9.

Razmatranje ishoda čija se vjerojatnost u vremenu mijenja omogućava nam omjer hazarda jer koristi informacije prikupljene u različito vrijeme. Obično se koristi u kontekstu opstanaka tijekom vremena. Ako je omjer hazarda 0.5, onda je relativni rizik od umiranja u jednoj skupini polovica rizika od umiranja u drugoj skupini.

Navedene pojmove ćemo bolje predočiti sljedećim primjerom.

Primjer 16 *Pretpostavimo da su eksperimentalni podaci zabilježeni tijekom eksperimenta i nalaze se u tablici 4.3.1. Tablica prikazuje podatke o razini doživljenja, razlici i relativnom riziku za tri hipotetske vrste ispitanika. Za svaku vrstu ispitanika A, B i C odabrani su uzorci koji čine ispitnu skupinu (T) i oni koji čine kontrolnu skupinu (C), te se proučava pojava smrti u obje skupine. To je eksperiment u kojoj su dvije skupine, sa i bez faktora, praćene dovoljno dugo da se smrt pojavljuje u brojevima dovoljno velikim da se statistički testovi obave sa što većom točnošću.*

Subjekt	*Doživljenje 5-godina*		*Razlike*	*Relativni rizik*
	C	*T*		
A	*0.98*	*0.99*	*0.01*	*0.01/0.02=0.5*
B	*0.80*	*0.89*	*0.09*	*0.11/0.2=0.55*
C	*0.25*	*0.50*	*0.25*	*0.5/0.75=0.67*

(4.3.1)

Pretpostavljamo da subjekti A, B i C imaju sve veći rizični čimbenik. Na primjer, dob ispitanika na početku može biti 30 *godina za subjekt A,* 50 *za subjekt B i* 70 *godina za subjekt C.*

Podatke u tablici ćemo pobliže objasniti na primjeru subjekta A. Iz tablice vidimo da kod subjekta A, u ispitnoj skupini od 100 *ispitanika imamo* 98 *koji su preživljeli* 5 *godina, a u kontrolnoj od* 100 *ispitanika* 99*. Iz danih vjerojatnosti izračunata je razlika rizika i*

relativni rizik. Razlika rizika za subjekt A jednaka je $(1 - 0.99) - (1 - 0.98) = 0.1$*, a relativni rizik* $(1 - 0.99)/(1 - 0.98) = 0.5$*.*

Pretpostavimo da tretman utječe na hazard konstantom 0.5 *(tj. koristi se PH i konstantan omjer hazarda* 0.5*). Prema tablici možemo uočiti da razlika i relativni rizik ovise o opstanku kontrolnog subjekta. Kontrolni subjekt koji ima "dobre" vrijednosti kovarijata ostavit će malo mjesta za bolju prognozu od tretmana.*

Omjer hazarda je osnova za opisivanje mehanizma učinka (nekog tretmana). U gore navedenom primjeru, razumno je da tretman utječe na svaki subjekt smanjenjem njegove opasnosti od smrti za faktor 2, iako manje bolesne osobe imaju niske razlike u smrtnosti.

3.4 Eksponencijalni i Weibullov model

Neka $\mathbf{X}\beta$ u PH modelu označava linearnu kombinaciju predikatora, bez slobodnog člana β_0.

3.4.1 Eksponencijalni regresijski model doživljenja

Eksponencijalni model je osnovni parametarski model u analizi doživljenja. Pretpostavimo da je svako vrijeme eksponencijalno distribuirano s parametrom λ. Tada je očekivanje od T jednako $\mathbb{E}[T] = 1/\lambda$.

Korištenjem PH formulacije, eksponencijalni regresijski model doživljenja može se prikazati na sljedeći način :

$$\begin{aligned} \lambda(t|\mathbf{X}) &= \lambda \exp(\mathbf{X}\beta), \\ S(t|\mathbf{X}) &= \exp[-\lambda t \exp(\mathbf{X}\beta)] = \exp(-\lambda t)^{\exp(\mathbf{X}\beta)}. \end{aligned}$$

Parametar λ može se smatrati kao antilogoritam izraza budući da se taj model može napisati kao $\lambda(t|\mathbf{X}) = \exp[(\log \lambda) + \mathbf{X}\beta]$.

Učinak $\mathbf{X}$ na očekivanje ili medijan vremena neuspjeha:

$$\begin{aligned} \mathbb{E}[T|\mathbf{X}] &= \frac{1}{\lambda \exp(\mathbf{X}\beta)}, \\ T_{0.5}|\mathbf{X} &= \frac{\log 2}{\lambda \exp(\mathbf{X}\beta)}. \end{aligned}$$

Navedeni eksponencijalni regresijski model može se napisati u nekom drugom obliku koji je više numerički stabilan tako da prikažemo parametar λ pomoću odsječka modela $\mathbf{X}\beta$, odnosno $\lambda = \exp(\beta_0)$. Izraz $\exp(\beta_0)$ možemo promatrati kao stopu doživljenja u kontrolnoj skupini koju identificiramo kovarijatom $\mathbf{X} = 0$. Ta veličina, pomnožena s faktorom koji ovisi o vrijednosti neke kovarijate, daje stopu hazarda u skupini koja je identificirana tom kovarijatom. Model pretpostavlja da je hazard različitih skupina proporcionalan.

Nakon redefiniranja $\mathbf{X}\beta$ tako da uključuje β_0, λ se može ispustiti u svim navedenim formulama u ovom odjeljku.

3.4.2 Weibullov model regresije

Weibullov model (distribucija) jedan je od najčešće korištenih statističkih modela u teoriji pouzdanosti i teoriji životnog vijeka. Model je nazvan po švedskom fizičaru Waloddi Weibullu $(1887. - 1979.)$.

Već smo spomenuli da je osnovna funkcija hazarda za Weibullov model dana izrazom $\lambda(t) = \alpha\gamma t^{\gamma-1}$, $t \geq 0$, $\gamma, \alpha > 0$, a osnovna kumulativna funkcija hazarda $\Lambda(t) = \alpha t^{\gamma}$. Omjer $\lambda(t)/\Lambda(t) = \gamma/t$ ovisi o nepoznatom parametru γ.

Model Weibullove regresije određen je jednom od sljedećih funkcija, uz pretpostavku da $\mathbf{X}\beta$ ne sadrži slobodni član β_0 :

$$\begin{aligned} \lambda(t|\mathbf{X}) &= \alpha\gamma t^{\gamma-1}\exp(\mathbf{X}\beta), \\ \Lambda(t|\mathbf{X}) &= \alpha t^{\gamma}\exp(X\beta), \\ S(t|X) &= \exp[-\alpha t^{\gamma}\exp(\mathbf{X}\beta)] = [\exp(-\alpha t^{\gamma})]^{\exp(\mathbf{X}\beta)}, \end{aligned}$$

gdje su α, γ i β nepoznati parametri.

Parametar α u homogenom Weibullovom modelu zamijenjen je s $\alpha\exp(\mathbf{X}\beta)$. Medijan vremena doživljenja dan je s

$$T_{0.5}|\mathbf{X} = \{\log 2/[\alpha\exp(\mathbf{X}\beta)]\}^{1/\gamma}.$$

Kao i kod eksponencijalnog modela, parametar α može biti ispušten i zamijenjen s $\exp(\beta_0)$ ako je preslikavanje β_0 dodano u $\mathbf{X}\beta$. Iz numeričkih razloga ponekad je korisno napisati Weibullov PH model kao:

$$S(t|\mathbf{X}) = \exp(-\Lambda(t|\mathbf{X})),$$

gdje je

$$\Lambda(t|\mathbf{X}) = \exp(\gamma\log t + \mathbf{X}\beta).$$

3.4.3 Procjena

Parametarski modeli kao štu su Weibullov ili eksponencijalni model lako se mogu procjenjivati u okviru generaliziranih linearnih modela. Parametri λ i β procjenjuju se metodom maksimalne vjerodostojnosti za što nam je potrebno odrediti funkciju vjerodostojnosti. Točnije procjenjuju se maksimiziranjem log funkcije vjerodostojnosti koju smo konstruirali u prethodnom poglavlju. Jedina razlika je umetanje $\exp(X_i\beta)$ u funkciju vjerodostojnosti:

$$l = \log L = \sum_{Y_i\ necenzuriran} \log[\lambda(Y_i)\exp(X_i\beta)] - \sum_{i=1}^{n}\Lambda(Y_i)\exp(X_i\beta).$$

Funkcija vjerodostojnosti za PH model se često naziva i parcijalna funkcija vjerodostojnosti jer ona razmatra samo vjerojatnosti za one subjekte kod kojih se dogodio događaj i ne razmatra vjerojatnosti za subjekte koji su cenzurirani.

Jednom kada se $\widehat{\beta}$, ML-procjenitelj od β, izračuna uz procjene standardnih pogrešaka velikih uzoraka, tada se može izračunati i procjena omjera hazarda te njihov pouzdani interval. Ako sa s označimo procijenjenu standardnu pogrešku od $\widehat{\beta}_j$, $(1-\alpha)100\%$ pouzdani

interval za omjer hazarda $\frac{\lambda(t|X_j+1)}{\lambda(t|X_j)}$ dan je sa $\exp(\widehat{\beta}_j \pm zs)$, pri čemu z predstavlja $(1-\frac{\alpha}{2})$ -kvantil za standardnu normalnu distribuciju.

Kad se procjenjuju parametri osnovne funkcije hazarda, može se izvesti ML-procjenitelj za $\lambda(t)$, $\widehat{\lambda}(t)$. ML-procjenitelj od $\lambda(t|\mathbf{X})$, funkcije hazarda kao funkcije od varijable t i kovarijate X, dan je sa

$$\widehat{\lambda}(t|\mathbf{X}) = \widehat{\lambda}(t)\exp(\mathbf{X}\widehat{\beta}).$$

ML-procjenitelj od $\Lambda(t)$, $\widehat{\Lambda}(t)$, može se izvesti iz integrala od $\widehat{\lambda}(t)$ u odnosu na t. Tada se ML-procjenitelj od $S(t|\mathbf{X})$ može izvesti na sljedeći način:

$$\widehat{S}(t|\mathbf{X}) = \exp[-\widehat{\Lambda}(t)\exp(\mathbf{X}\widehat{\beta})].$$

Za Weibullov model, ML-procjenitelje parametara funkcija hazarda α i γ označavamo sa $\widehat{\alpha}$, odnosno $\widehat{\gamma}$. ML-procjenitelji od $\lambda(t|\mathbf{X})$, $\Lambda(t|\mathbf{X})$ i $S(t|\mathbf{X})$ za ovaj model su:

$$\begin{aligned}
\widehat{\lambda}(t|\mathbf{X}) &= \widehat{\alpha}\widehat{\gamma}t^{\widehat{\gamma}-1}\exp(\mathbf{X}\widehat{\beta}) \\
\widehat{\Lambda}(t|\mathbf{X}) &= \widehat{\alpha}\widehat{\gamma}t^{\widehat{\gamma}}\exp(\mathbf{X}\widehat{\beta}) \\
\widehat{S}(t|\mathbf{X}) &= \exp[-\widehat{\Lambda}(t|\mathbf{X})].
\end{aligned}$$

Pouzdani intervali za $S(t|\mathbf{X})$ se dobivaju primjenom genalizirane matrice kako bi dobili procjenu standardne pogreške od $\log[\widehat{\lambda}(t|\mathbf{X})]$ iz procijenjene informacijske matrice svih parametara hazarda i regresije. Prema tome, pouzdani interval za $\widehat{S}$ biti će u obliku

$$\widehat{S}(t|\mathbf{X})^{\exp(\pm zs)}.$$

ML-procjenitelj od β i parametri hazarda vode izravno do ML-procjenitelja očekivanja i medijana duljine života. Za Weibullov model ML-procjenitelj medijana duljine života za dani $\mathbf{X}$ je

$$\widehat{T}_{0.5}|\mathbf{X} = \{\log 2/[\widehat{\alpha}\exp(\mathbf{X}\widehat{\beta})]\}^{1/\widehat{\gamma}}.$$

Za eksponencijalni model, ML-procjenitelj očekivane životne duljine za subjekta s vrijednostima kovarijate $\mathbf{X}$ dan je s

$$\widehat{\mathbb{E}}[T|X] = [\widehat{\lambda}\exp(X\widehat{\beta})]^{-1},$$

gdje je $\widehat{\lambda}$ ML-procjenitelj od λ.

3.5 Provjera pretpostavki PH modela

U potpoglavlju 3.2 naveli smo tri pretpostavke parametarskog proporcionalnog modela. Kao što ime govori, temeljna pretpostavka na kojoj model počiva je upravo ona o proporcionalnom riziku. Takva pretpostavka podrazumijeva da su funkcije hazarda jednake do na množenje konstantom, učinak dane kovarijate ne mijenja tijekom vremena. Ako se prekrši ova pretpostavka, PH model je nevažeći, te su potrebne složenije analize. U ovom potpoglavlju ćemo detaljnije opisati koji odnosi bi trebali biti zadovoljeni u modelu, kako bi se otkrilo odstupanje od pretpostavki.

3.5.1 PH model s jednom binarnom varijablom

Prvo pretpostavimo model PH s jednim binarnim prediktorom X_1. Tada je PH model oblika

$$\lambda(t|X_1) = \lambda(t)\exp(\beta_1 X_1).$$

Kako je X_1 binarna varijabla imamo

$$\begin{aligned}\lambda(t|X_1 = 0) &= \lambda(t),\\ \lambda(t|X_1 = 1) &= \lambda(t)\exp(\beta_1).\end{aligned}$$

Slijedi da je omjer hazarda jednak

$$\frac{\lambda(t|X_1 = 1)}{\lambda(t|X_1 = 0)} = \exp(\beta_1),$$

odnosno

$$\log\frac{\lambda(t|X_1 = 1)}{\lambda(t|X_1 = 0)} = \beta_1.$$

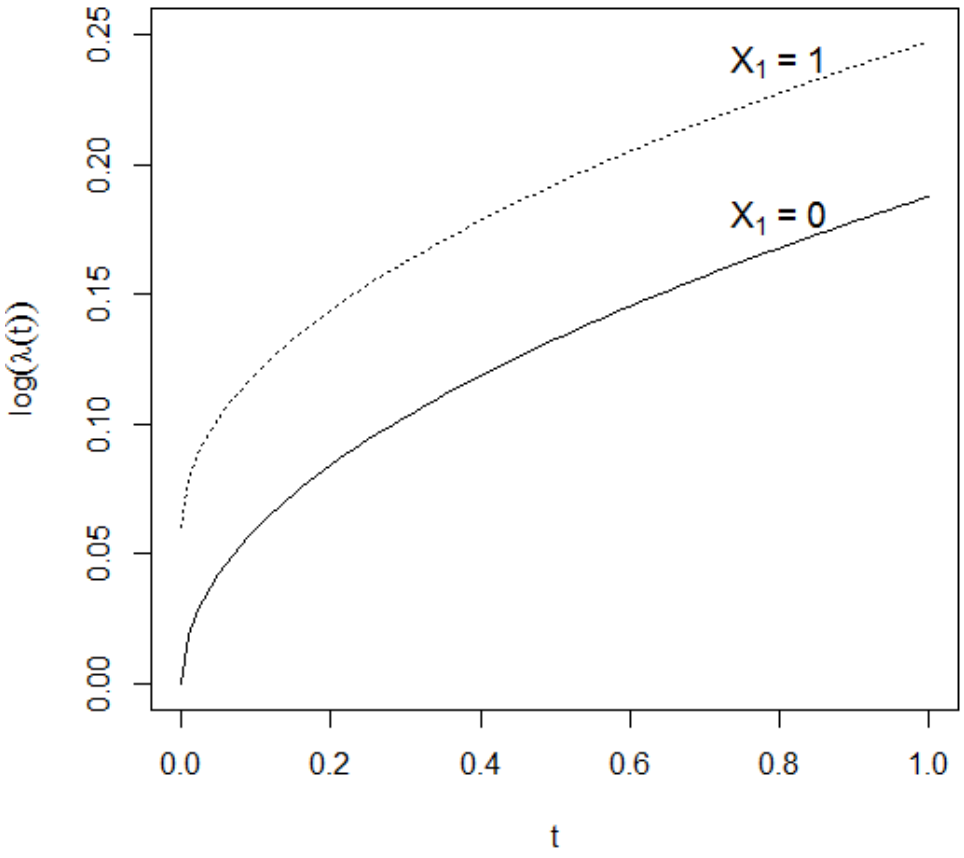
U ovom slučaju omjer hazarda je konstantan obzirom na vrijeme, a time su i funkcije hazarda proporcionalne. Dakle, uz pretpostavku oblika $\lambda(t)$ imamo ispunjenu i PH pretpostavku.

Na slici 3.1 grafički je ilustrirano kada su pretpostavke ispunjene. Na grafičkom prikazu možemo uočiti da je udaljenost između funkcija hazarda konstantna i iznosi β_1.

U posebnom slučaju, kada je $\lambda(t)$ Weibullova imamo

$$\begin{aligned}\log\lambda(t \mid X_1 = 0) &= \log(\alpha\gamma) + (\gamma - 1)\log t,\\ \log\lambda(t \mid X_1 = 1) &= \log(\alpha\gamma) + (\gamma - 1)\log t + \beta_1.\end{aligned}$$

U ovom slučaju, ako umjesto t na x-osi nacrtamo $\log t$, onda će obje krivulje biti linearne. Primjetimo također, da ako nema ovisnosi između X i doživljenja ($\beta_1 = 0$), krivulje biti će približno jednake. Odstupanja od jednakosti su zanemariva i ne predstavljaju problem PH modelu.



Slika 3.1: PH MODEL S JEDNIM BINARNI PREDIKTOROM

3.5.2 PH model s jednom neprekidnom varijablom

Promatramo model PH kada je jedini prediktor neprekidan:

$$\lambda(t \mid X_1) = \lambda(t) \exp(\beta_1 X_1).$$

Odnos između logaritamske funkcije hazarda i neprekidnog prediktora je linearan

$$\log \lambda(t \mid X_1) = \log \lambda(t) + \beta_1 X_1.$$

Osim pretpostavki PH modela i pretpostavke za oblik $\lambda(t)$ imamo i pretpostavku linearnosti, osim ako je drukčije navedeno. Odstupanja od linearnosti za određenu neprekidnu kovarijatu ukazuju na to da varijabla nije ispravno uklopljena u model. Nelinearnost ne stvara problem za kategoričke varijable.

Promatramo li funkcije hazarda za vremena t_1 i t_2 takva da je $t_1 \neq t_2$ imamo

$$\begin{aligned} \lambda(t_1|X_1) &= \lambda(t_1)\exp(\beta_1 X_1), \\ \lambda(t_2|X_1) &= \lambda(t_2)\exp(\beta_1 X_1). \end{aligned}$$

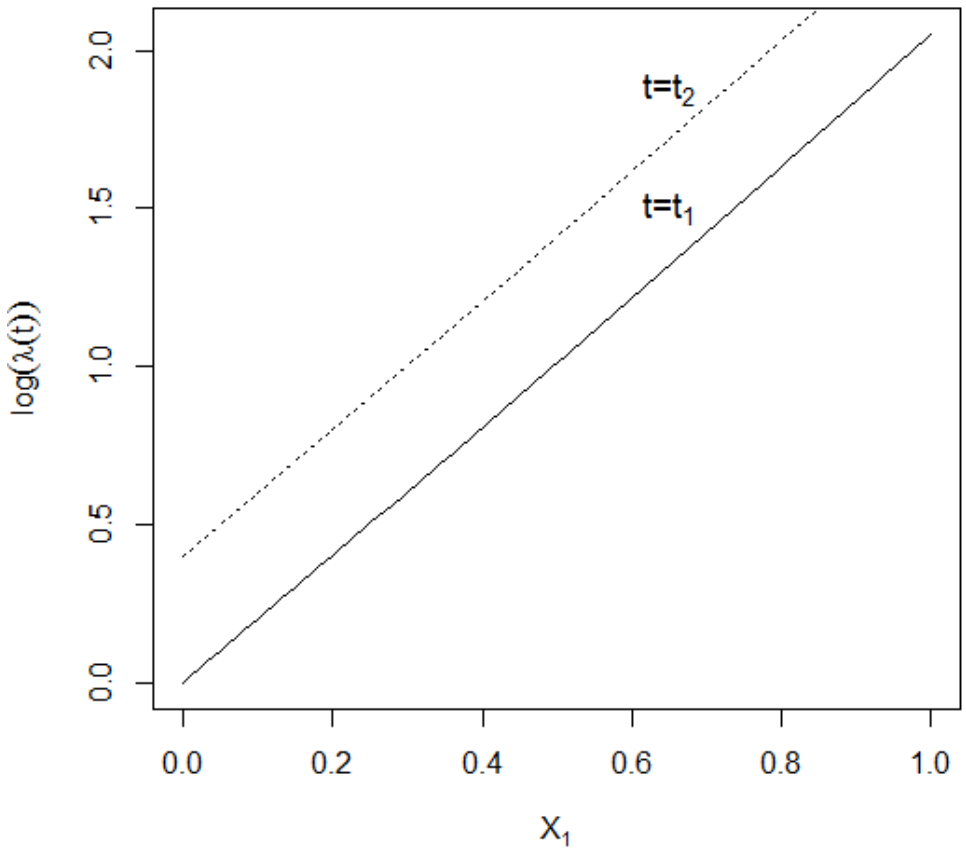
Slijedi, omjer hazarda jednak

$$\frac{\lambda(t_1|X_1)}{\lambda(t_2|X_1)} = \frac{\lambda(t_1)}{\lambda(t_2)},$$

odnosno

$$\log\frac{\lambda(t_1|X_1)}{\lambda(t_2|X_1)} = \log\frac{\lambda(t_1)}{\lambda(t_2)}.$$

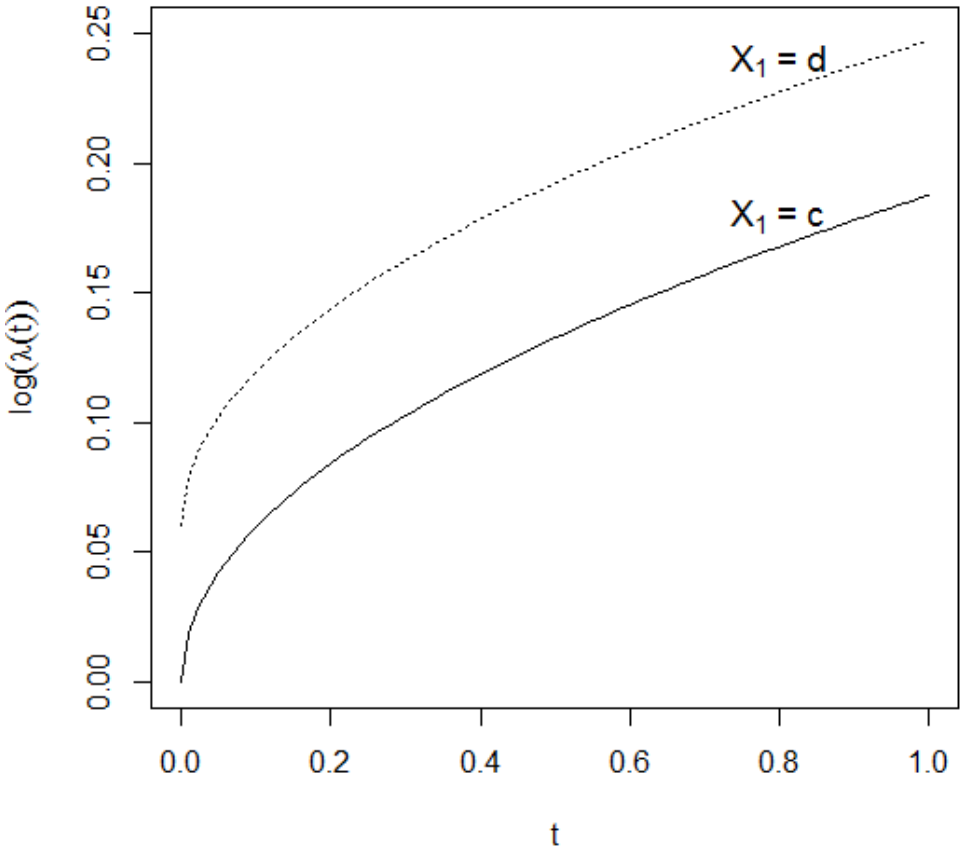
Odnosi koji se u ovom modelu moraju održati prikazani su grafički na slikama 3.2 i 3.3.



Slika 3.2: PH MODEL S JEDNOM NEPREKIDNOM VARIJALOM

Prema grafu na slici 3.2, za $X_1 \in [0,1]$ izraz $\log(\lambda(t))$ mijenja se linearno za svaki fiksni t, a razlika izraza za dvije različite vrijednosti t_1 i t_2 je konstantna i iznosi $\log\frac{\lambda(t_1)}{\lambda(t_2)}$.

Analogno vrijedi ako umjesto funkcije hazarda promatramo kumulativnu funkciju hazarda, tj. tada je razlika jednaka $\log \frac{\Lambda(t_1)}{\Lambda(t_2)}$.



Slika 3.3: PH MODEL S JEDNIM NEPREKIDNIM PREDIKTOROM

Sada ćemo promatrati odnos između sljedeće dvije funkcije hazarda:

$$\begin{aligned} \lambda(t|X_1 = c) &= \lambda(t)\exp(\beta_1 c), \\ \lambda(t|X_1 = d) &= \lambda(t)\exp(\beta_1 d). \end{aligned}$$

Omjer hazarda za dane dvije krivulje je

$$\frac{\lambda(t|X_1 = d)}{\lambda(t|X_1 = c)} = \exp((d-c)\beta_1),$$

odnosno

$$\log \frac{\lambda(t|X_1 = d)}{\lambda(t|X_1 = c)} = (d - c)\beta_1.$$

U ovom PH modelu, oblik funkcije za zadanu vrijednost X_1 mora biti u skladu s pretpostavljenom funkcijom $\lambda(t)$, te odnos između dvije krivulje mora se održati za bilo koje dvije vrijednosti c i d od X_1, kao što je prikazano na slici 3.3. Također, na grafu uočavamo da je udaljenost između funkcija hazarda jednaka $(d - c)\beta_1$.

Kao i u 3.5.1, za Weibullov model sve funkcije moraju biti linearne u $\log t$.

3.5.3 PH model s dvije kovarijate

Kada postoji više prediktora, PH pretpostavka može se prikazati na način sličan slikama 3.1 i 3.3. Sada ćemo promatrati PH model s dvje kovarijate, X_1 i X_2:

$$\lambda(t|X) = \lambda(t) \exp(\beta_1 X_1 + \beta_2 X_2)$$

Neka je X_1 binarna, a X_2 neprekidna kovarijata, te pretpostavimo da nema zavisnosti između te dvije varijable.

Dakle, promatramo:

$$\begin{aligned} \lambda(t|X_1 = 0) &= \lambda(t) \exp(\beta_2 X_2) \\ \lambda(t|X_1 = 1) &= \lambda(t) \exp(\beta_1 + \beta_2 X_2). \end{aligned}$$

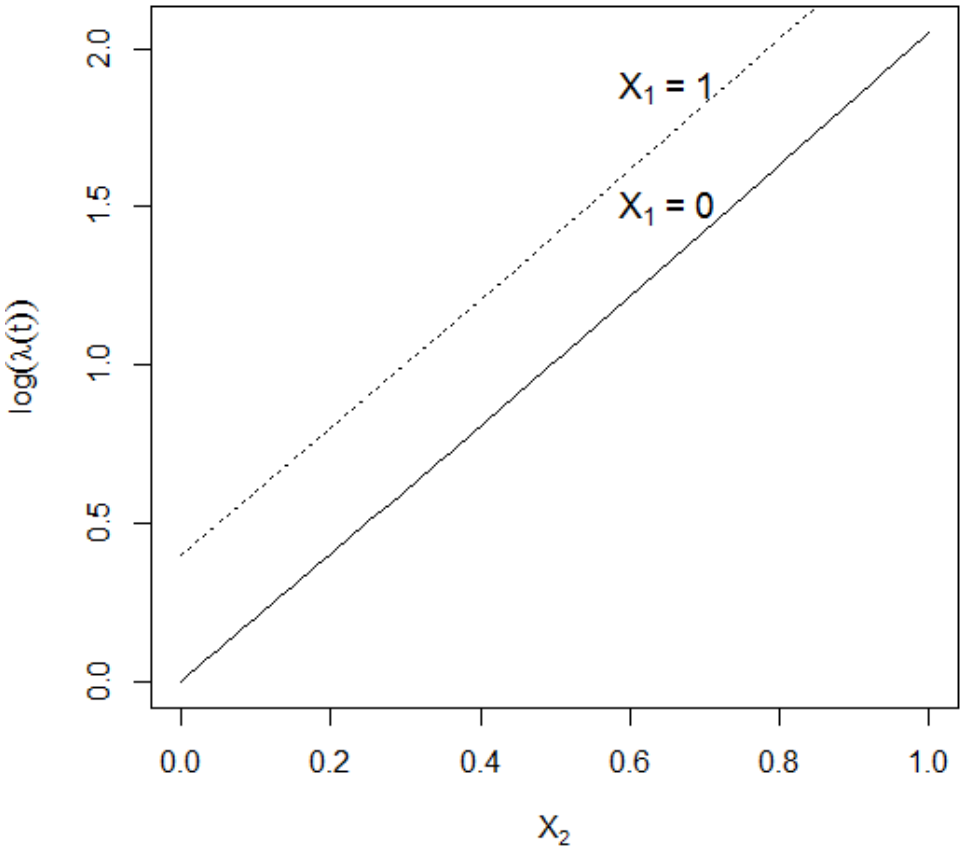
Omjer hazarda jednak je

$$\frac{\lambda(t|X_1 = 0)}{\lambda(t|X_1 = 1)} = \exp(\beta_1),$$

odnosno

$$\log \frac{\lambda(t|X_1 = 0)}{\lambda(t|X_1 = 1)} = \beta_1.$$

Grafički prikaz promatranog modela na slici 3.4 pokazuje da se odnos između logaritamske funkcije hazarda i kovarijate X_2 mora održati pri svakom trenutku t, tj. da bi PH pretpostavka bila ispunjena, ako pretpostavimo linearnost za X_2 i da nema zavisnosti između X_1 i X_2. Udaljenost između pravaca na grafu je β_1, a nagib pravaca je β_2.



Slika 3.4: PH MODEL S DVIJE KOVARIJATE

U svim navedenim primjerima PH modela funkciju hazarda možemo zamijeniti s kumulativnom funkcijom hazarda.

3.6 Prednosti PH modela

PH model baziran je na pretpostavci proporcionalnog hazarda i ne uzima se određena distribucija vjerojatnosti za vrijeme doživljenja. Ključni razlog popularnosti PH modela leži u činjenici da se, unatoč tome da je funkcija osnovnog hazarda neodređena, dobre ocjene koeficijenata regresije i prilagođene funkcije doživljenja mogu izvesti za širok spektar podataka. Drugim riječima, PH je robustan model.

Općenito robustnost PH modela i njegov specifičan oblik je atraktivan iz nekoliko razloga. Kao što smo spomenuli, PH model je umnožak osnovne funkcije hazarda koja sadrži t i eksponencijalnog izraza koji sadrži kovarijate, a ne sadrži t. Eksponencijalni dio formule privlačan je jer osigurava nenegativne ocjene prilagođenog modela. Ako bismo umjesto

eksponencijalnog dijela imali linearnu funkciju od $\mathbf{X}$, onda bi mogli dobiti nenegativne ocjene hazarda što nije dozvoljeno. Budući da se po definiciji vrijednosti bilo koje funkcije hazarda mogu kretati između 0 i $+\infty$, želimo da i drugi dio formule bude nenegativan. Hazard za svaki promatrani subjekt je fiksna proporcija hazarda za bilo koji drugi subjekt i konstantna je u vremenu. Bitna osobina PH modela je što, iako funkcija osnovnog hazarda nije određena, možemo procijeniti parametre β u eksponencijalnom dijelu modela, a oni su nam potrebni da bi procijenili utjecaj varijabli od interesa. Omjer hazarda se također računa bez ocjene osnovne funkcije hazarda.

Također, možemo primjetiti da se i funkcija hazarda $\lambda(t|\mathbf{X})$ te odgovarajuća funkcija doživljenja $S(t|\mathbf{X})$ mogu ocjenjeniti za PH model čak iako osnovna funkcija hazarda nije određena. Iz toga slijedi da uz minimum pretpostavki možemo dobiti primarne informacije iz analize doživljenja, a to su omjer hazarda i funkcija doživljenja.

Još jedan razlog zbog koje je PH model popularan je upravo to što ima prioritet nad logističkim modelom kad imamo informacija o vremenu doživljenja i kada postoji cenzuriranje podataka. PH model koristi više informacija tj. i vrijeme doživljenja za razliku od logističkog modela koji razmatra samo opcije je li se događaj dogodio ili ne, te zanemaruje vrijeme doživljenja i cenzuriranje.

Bibliografija

[1] Frank E.Harrell, Jr., *Regression Modeling Strategies*, Springer, New York, 2001.

[2] Regina C. Elandt-Johnson, Norman L. Johnson, *Survival Models and Data Analysis*, Wiley, New York, 1999.

[3] J. P. Klein, M. L. Moeschberger, *Survival Analysis: Tehniques for Censored and Truncate Data*, Springer-Verlag, New York, 1997.

Sažetak

Analiza doživljenja je skup tehnika kojima se procjenjuje i opisuje vrijeme potrebno do pojave jednog ili više određenih događaja. U radu uvodimo osnovne pojmove analize doživljenja kao što su cenzuriranje i rezanje podataka, funkcija doživljenja, funkcija hazarda, kumulativna funkcija hazarda, očekivano trajanje života, itd. One su detaljno definirane u prva dva poglavlja. Prije njih navedene su osnovne definicije i pojmovi vjerojatnosti i statistike. U prvom poglavlju objašnjeni su tipovi podataka koji se obrađuju u analizi preživljavanja. Podaci mogu biti cenzurirani i odrezani. Postoji više vrsta svakog tipa, tako imamo lijevo i desno cenzurirane podatke, lijevo i desno odrezane podatke, intervalno cenzurirane podatke, te kombinacije. Veoma je bitno da prepoznamo s kakvim oblikom podataka radimo, jer se ovisno o njima mijenja i tehnika kojom procjenjujemo.

Upoznajemo se s procjeniteljem maksimalne vjerodostojnosti za funkciju doživljenja, kojom procjenjujemo promatrane veličine.

Detaljno obrađujemo teorijsku pozadinu parametarskog proporcionalnog modela rizika, stavljajući naglasak na Weibullov i eksponencijalni model doživljenja. Parametarski proporcionalni model rizika pretpostavlja da funkcija ovisi o prediktorskim varijablama na specifičan način.

U konačnici, koristeći se principima definiranim u teorijskom dijelu, pobliže opisujemo pretpostavke PH modela te ističemo prednosti modela.

Summary

Survival analysis is a set of methods for analysing time duration until one or more events happen. This paper introduces basic concepts of survival analysis such as censoring and truncation data, survival function, hazard function, the cumulative hazard function, the mean residual life function etc. They are considered in the second chapter. In the first chapter, we deal with the types of survival data. There are left and right censored data, left and right truncated data, interval censored data, and combination of them. It is important to know which type of data we have, because depending on that, we choose the techniques for estimation.

We define maximum likelihood estimation of the survival function.

The theoretical background for parametric proportional hazard model is developed. Weibull and exponential survival model are given as examples of the most commonly used specific models.

Ultimately, using the principles defined in the theoretical part, we described the assumptions and prominent advantages of PH model.

Životopis

Rođena sam 16.siječnja.1994. godine u Ljubuškom, Bosna i Hercegovina. Osnovnu školu Silvija Strahimira Kranjčevića pohađala sam u Mostaru, Bosna i Hercegovina. Nakon završene osnovne škole 2008. godine upisujem opći smjer gimanzije fra. Grge Martića također u Mostaru. Nakon položene državne mature 2012. godina upisujem preddiplomski studij Matematke na Prirodoslovno-matematičkom fakultetu u Splitu. Uspješno završavam preddiplomski studij izradom završnog rada pod nazivom Dobro utemeljeni skupovi iz područja Teorije skupova. Nakon stečene titule prvostupnika, 2016. godina upisujem diplomski studij Financijske i poslovne matematike na Prirodoslovno-matematičkom fakultetu u Zagrebu. Krajem diplomskog studija odradila sam studensku praksu u IT odjelu Erste banke u Zagrebu, te svoje radno iskustvo ću nastaviti kao pripravnik u spomenutoj banci.