Coxov model proporcionalnih hazarda

Kolarek, Eva

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Eva Kolarek

COXOV MODEL PROPORCIONALNIH HAZARDA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, veljača 2023.

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 mentoru prof. dr. sc. Miljenku Huzaku na stručnim i kvalitetnim savjetima prilikom izrade ovoga rada.
Veliko hvala mojim roditeljima na iznimnoj podršci, razumijevanju i odricanju čime su omogućili moje obrazovanje.
Ovaj rad posvećujem Dariu kao zahvalu na beskrajnu podršku, vjeru i prisutnost u svakom trenutku.

Sadržaj

Uvod

Veliki broj kliničkih istraživanja uključuje praćenje vremena proteklog do pojave nekog događaja od interesa kao što je ozdravljenje, smrt, manifestacija bolesti, reakcija na lijek i slično. Bilježeno vrijeme predstavlja vrlo bitnu komponentu daljnje statističke analize koja se provodi u svrhu dobivanja vrijednih informacija iz prikupljenih podataka, kako bi se u konačnici povećala efikasnost liječenja pacijenata. U ovom radu opisat će se neki od alata i statističkih modela koji se koriste prilikom analize vremena proteklog do nekog trenutka, a koje generalno nazivamo **analizom doživljenja**.

Polazeći od činjenice da se analiza doživljenja inicijalno primjenjivala za analizu podataka iz područja medicine, događaj od interesa najčešće je predstavljao smrt osobe. Pritom se na temelju prikupljenih podataka nastojalo procijeniti vrijeme do kojeg će pacijent živjeti. To je motiviralo uvođenje pojma **funkcije doživljenja** koja daje vjerojatnost da osoba određenih karakteristika ne doživi smrt do nekog trenutka.

Osim na područje medicine, analiza doživljenja je zbog praktičnosti i velike informativne vrijednosti proširila svoj raspon na područje bankarstva, aktuarstva, strojarstva te velikog broja inženjerskih područja. No, pritom se pojam krivulje doživljenja zadržao neovisno o događaju koji se promatra. Stoga, nju općenito smatramo krivuljom koja nam daje informaciju o vjerojatnosti da se događaj od interesa ne dogodi do nekog trenutka, a kao takva predstavlja jedan od osnovnih alata analize doživljenja.

Podaci koji sadrže informaciju o vremenu do pojave nekog događaja pogodni su za analizu i putem **funkcije hazarda** koja nam daje informaciju o riziku realizacije događaja. Stoga je cilj statističke obrade takvih podataka upravo procjena navedenih funkcija na temelju opaženog uzorka. Općenito postoje razne metode njihove procjene, a neke od njih navest ćemo u ovom radu. Najprije ćemo predstaviti njihovu procjenu koja uključuje informaciju o vremenu do pojave događaja od interesa, a zatim ćemo opisati procjenu koja dodatno uključuje i informaciju o riziku pojave tog događaja. Osim njihove procjene, jedan od ciljeva je dati mjeru utjecaja pojedinih varijabli na funkciju doživljenja. No, navedeni ciljevi zahtjevaju postavljanje nekog matematičkog modela.

Za matematičko modeliranje problema u kontekstu analize doživljenja često su prikladni modeli proporcionalnog hazarda. Proporcionalnost hazarda općenito predstavlja situaciju u kojoj se utjecaj određenih varijabli na funkciju rizika ne mijenja kroz vrijeme.

U toj klasi modela ističe se **Coxov model proporcionalnih hazarda** koji je glavna tema ovog rada. Motivacija za njegovu konstrukciju bilo je utvrđivanje razlika u preživljavanju po različitim grupama pacijenata uz korištenje prognostičkih faktora koje u matematičkim terminima nazivamo kovarijatama ili prediktivnim varijablama. Prema tome, u slučaju provođenja kliničke studije, od velike je koristi prikupiti što više informacija o pacijentima koje se potencijalno mogu modelirati putem kovarijata, kako bi se dala mjera njihovog utjecaja na vjerojatnost doživljenja. Iako se pomoću Coxovog modela može procijeniti vjerojatnost doživljenja, njegova vrijednost leži u činjenici što daje procjenu omjera rizika bez poznavanja cijelog oblika modela. To ga očigledno čini vrlo primjenjivim u praksi, a budući da se najčešće koristi samo za procjenu omjera rizika, naziva se i modelom relativnog rizika.

U ovom radu najprije će biti opisana teorijska pozadina analize doživljenja kako bi se definirali osnovni pojmovi potrebni za razumijevanje Coxovog modela. Zatim će se dati detaljan prikaz konstrukcije i svojstava Coxovog modela, a pratit će ga metode provjere njegovih pretpostavki. Na samom kraju razmotrit će se slučaj u kojem je glavna pretpostavka Coxovog modela narušena, što će dati poticaj za opis proširenog Coxovog modela. Navedena teorijska pozadina bit će ilustrirana primjenom na stvarnim podacima o liječenju, prilikom čega će rezultati biti dobiveni uz korištenje programskog paketa R. Pripadni programski kod naveden je kao dodatak na kraju rada.

Poglavlje 1

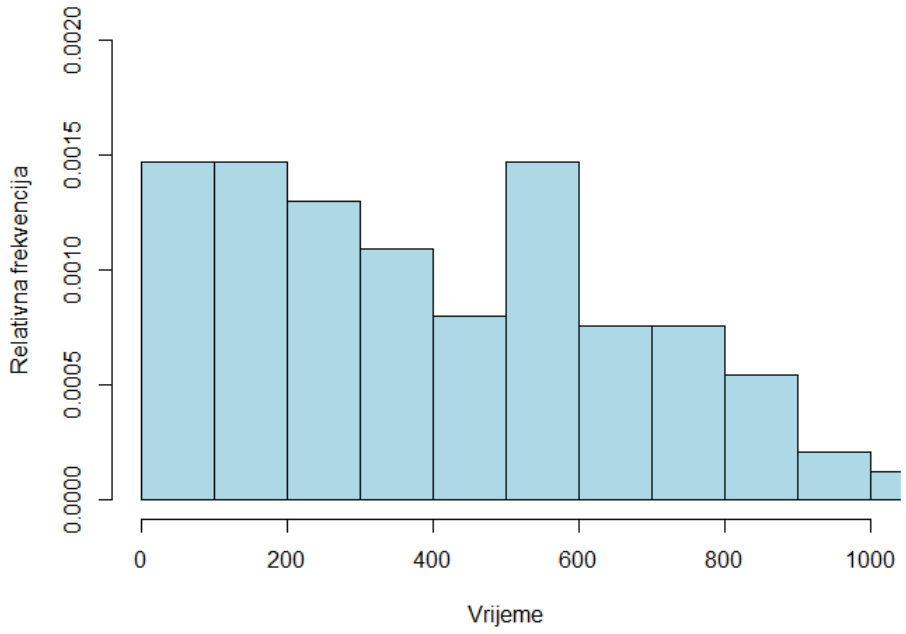
Analiza doživljenja

Analiza doživljenja predstavlja skup statističkih procedura za analizu podataka kod kojih je u središtu interesa vrijeme do pojave nekog događaja. U tu svrhu definira se vrijeme proteklo do pojave događaja od interesa kao slučajna varijabla T koju nazivamo **vrijeme doživljenja**. Pritom pod pojmom "događaj" podrazumijevamo smrt, pojavu bolesti, ozdravljenje ili bilo koji događaj od interesa koji se može dogoditi pojedincu. Realizaciju varijable T označavamo s t, a kako se radi o vremenu, T može poprimiti vrijednosti iz skupa $\mathbb{R}_+$. Osim bilježenja vrijednosti varijable T, prilikom provođenja kliničke studije uobičajeno je odrediti interval u kojem ih bilježimo. Budući da je odabrani interval često ograničen, nerijetko se događaj od interesa ne dogodi u promatranom intervalu. Posljedično, točna informacija o vremenu doživljenja nam je nepoznata. No, usprkos tome, ipak nam je dostupan podatak da se događaj **nije** dogodio u promatranom intervalu. Za takav podatak o vremenu doživljenja kažemo da je **cenzuriran**. Cenzurirane podatke također želimo uključiti prilikom procjene funkcije doživljenja, zbog čega je, osim vremena doživljenja, poželjno bilježiti informaciju o tome je li se događaj od interesa dogodio ili je cenzuriran. To nas potiče na definiranje dihotomne slučajne varijable d na sljedeći način:

$$d = \begin{cases} 1, & \text{ako se zbio događaj} \\ 0, & \text{inače (događaj je cenzuriran).} \end{cases}$$

Motivacija za koncipiranje analize doživljenja leži u činjenici što podaci o vremenu doživljenja nisu pogodni za analizu putem klasičnih statističkih procedura. Naime, vrijeme doživljenja T tipično ne prati simetričnu distribuciju, zbog čega nije razumno pretpostavljati da ono prati normalnu distribuciju. Ta činjenica onemogućava nam korištenje mnogih statističih alata koji su namijenjeni upravo za normalno distribuirane podatke. Kada bismo prikazali podatke o vremenu doživljenja histogramom u svrhu naslućivanja njegove

distribucije, u praksi bismo dobili asimetričnu raspodjelu podataka. Kao primjer[1], na slici 1.1 prikazan je histogram vremena doživljenja koji nam sugerira asimetričnost distribucije varijable T.



Slika 1.1: Primjer distribucije vremena doživljenja T

Osim toga, klasična statistička analiza podataka uključuje deskriptivnu analizu u svrhu određivanja mjere centralne tendencije. Pritom se u kontekstu analize doživljenja prirodno nameće prosječno vrijeme doživljenja. Međutim, prilikom njegovog računanja, upravo će nam cenzurirani podaci procjenu činiti lošom. Naime, ukoliko je podatak o vremenu doživljenja cenzuriran, to znači da se događaj zbio u nekom trenutku nakon zabilježenog vremena t. Stoga, kada bismo računali prosjek zabilježenih vremena doživljenja uključivši i cenzurirane podatke, dobili bismo vrijednost koja je manja od stvarnog prosjeka vremena doživljenja. Prema tome, prosječno vrijeme zabilježenih vremena doživljenja neće nam dati pouzdane zaključke. Zato definiramo **funkciju doživljenja** $S : \mathbb{R}_+ \to [0, 1]$ koja predstavlja jedan od elementarnih pojmova analize doživljenja. Ako je F funkcija distribucije

[1] Histogram prikazuje podatke o vremenu doživljenja iz tablice 2.2

varijable T, funkciju doživljenja definiramo sa

$$S(t) := \mathbb{P}(T > t) = 1 - F(t). \tag{1.1}$$

Dakle, ona predstavlja vjerojatnost da se događaj od interesa nije dogodio do vremena t. Teoretski, funkcija S poprima beskonačno mnogo vrijednosti u intervalu $[0, 1]$, no u praksi poprima konačan broj vrijednosti zbog čega je stepenasta funkcija. Funkcija doživljenja koristan je alat ukoliko nam je u središtu interesa vrijeme u kojem se događaj od interesa **nije** dogodio. S druge strane, ukoliko želimo dobiti informaciju o vremenu kada se događaj dogodio, promatramo funkciju hazarda koja nam daje informaciju o riziku ili hazardu realizacije događaja od interesa. **Funkcija hazarda** definira se sa

$$h(t) := \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}. \tag{1.2}$$

Prirodno se postavlja pitanje odnosa između funkcije doživljenja i funkcije hazarda. Pokazat ćemo da postoji jasna veza između njih. Prema formuli uvjetne vjerojatnosti [2], zbog neprekidnosti varijable T i (1.1) vrijedi

$$\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t) = \frac{\mathbb{P}(t \leq T \leq t + \Delta t)}{\mathbb{P}(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{S(t)}.$$

Tada formulu za funkciju hazarda (1.2) možemo zapisati u obliku

$$h(t) = \lim_{\Delta t \to 0} \left(\frac{F(t + \Delta t) - F(t)}{\Delta t} \right) \cdot \frac{1}{S(t)}. \tag{1.3}$$

Nadalje, primijetimo da vrijedi

$$\frac{d}{dt} F(t) = \lim_{\Delta t \to 0} \left(\frac{F(t + \Delta t) - F(t)}{\Delta t} \right)$$

pa odavde uz (1.3) slijedi

$$h(t) = \left(\frac{d}{dt} F(t) \right) \cdot \frac{1}{S(t)} = \frac{\frac{d}{dt}(1 - S(t))}{S(t)} = -\frac{d}{dt} \left(\ln S(t) \right). \tag{1.4}$$

Iz (1.4) lako slijedi da za funkciju S vrijedi

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

[2] $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Dakle, jednadžbama (1.4) i (1.5) dana je veza između funkcije hazarda i funkcije doživljenja.

Jedan od ciljeva analize doživljenja je procjena funkcije S u svrhu usporedbe vjerojatnosti doživljenja po grupama subjekata sa zajedničkim svojstvima. Vrlo poznata metoda njene procjene je Kaplan-Meierova metoda koju ćemo opisati u nastavku.

Pretpostavimo da smo zabilježili ukupno n realizacija vremena doživljenja T, od kojih imamo l različitih. Označimo ih sa $t_1, \ldots, t_l$ i bez smanjenja općenitosti pretpostavimo da vrijedi $t_1 < t_2 < \ldots < t_l$. Za $k \in \{1, \ldots, l\}$ označimo sa n_k broj osoba za koje se potencijalno može zbiti događaj od interesa u trenutku t_k. Takve osobe u kontekstu analize doživljenja nazivamo **rizičnim** osobama. Nadalje, označimo sa m_k broj osoba za koje se zbio događaj u trenutku t_k, te sa

$$d_k := \begin{cases} 1, & \text{ako se događaj zbio u } t_k \\ 0, & \text{ako je događaj cenzuriran u } t_k. \end{cases}$$

Prema definiciji funkcije doživljenja imamo:

$$S(t_k) = \mathbb{P}(T > t_k) = \mathbb{P}\left(\{T > t_k\} \cap \{T \geq t_k\}\right)$$

pa korištenjem formule uvjetne vjerojatnosti, imamo

$$S(t_k) = \mathbb{P}\left(T > t_k \mid T \geq t_k\right) \mathbb{P}\left(T \geq t_k\right). \tag{1.6}$$

Budući da se u t_{k-1} i t_k zbio događaj od interesa, vrijedi

$$\mathbb{P}(t_{k-1} < T < t_k) = 0$$

pa je zato

$$\mathbb{P}(T \geq t_k) = \mathbb{P}(t_{k-1} < T < t_k) + \mathbb{P}(T \geq t_k) = \mathbb{P}(T > t_{k-1}) = S(t_{k-1}). \tag{1.7}$$

Sada iz (1.6) i (1.7) slijedi

$$S(t_k) = S(t_{k-1}) \mathbb{P}\left(T > t_k \mid T \geq t_k\right). \tag{1.8}$$

Nadalje, vjerojatnost $\mathbb{P}\left(T > t_k \mid T \geq t_k\right)$ procjenjuje se sa izrazom

$$\frac{n_k - m_k \cdot d_k}{n_k}$$

pa je zbog (1.8) procjena od S dana sa

$$\widehat{S}(t_k) = \widehat{S}(t_{k-1}) \left(\frac{n_k - m_k \cdot d_k}{n_k} \right), \tag{1.9}$$

pri čemu je $\widehat{S}(t_0) = 1$. Procjenu $\widehat{S}$ danu u (1.9) nazivamo **Kaplan-Meierovom procjenom** funkcije doživljenja S. Uočimo da u slučaju cenzuriranog podatka u trenutku t_k vrijedi $m_k \cdot d_k = 0$ pa je

$$\widehat{S}(t_k) = \widehat{S}(t_{k-1}),$$

što znači da će u trenutku t_k procjena funkcije doživljenja ostati ista kao i u prethodnom trenutku t_{k-1}. Koristeći formulu (1.9) možemo pobliže predočiti ideju metode na kraćem primjeru. Pretpostavimo da smo zabilježili vremena doživljenja $t_1, \ldots, t_5$ za koje vrijedi $0 \leq t_1 < t_2 < \ldots < t_5$. Neka su dobiveni podaci prikazani u tablici 1.1, na temelju kojih želimo procijeniti funkciju doživljenja $S(t)$ te ju zatim grafički prikazati.

Vrijeme doživljenja T	d_k	m_k	n_k
t_1	1	1	$n_1 = 7$
t_2	1	3	$n_2 = n_1 - m_1 = 7 - 1 = 6$
t_3	1	1	$n_3 = n_2 - m_2 = 6 - 3 = 3$
t_4	0	1	$n_4 = n_3 - m_3 = 3 - 1 = 2$
t_5	1	1	$n_5 = n_4 - m_4 = 2 - 1 = 1$

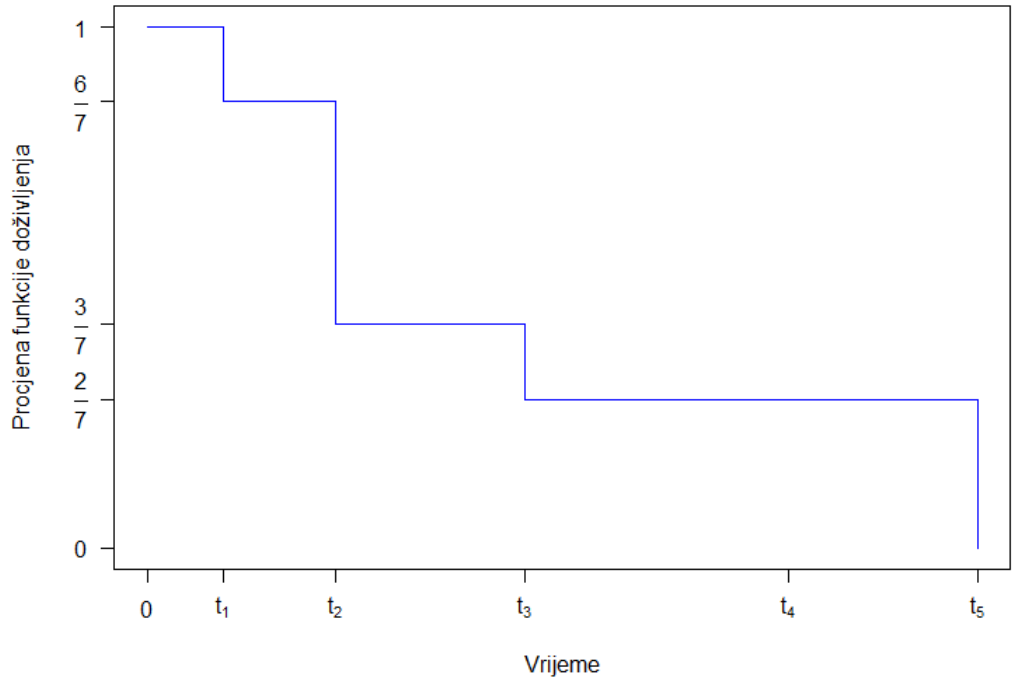
Tablica 1.1: Tablica podataka

Prema formuli (1.9) računamo:

$$\begin{aligned}
\widehat{S}(t_1) &= 1 \cdot \frac{7 - 1 \cdot 1}{7} = \frac{6}{7} \\
\widehat{S}(t_2) &= \widehat{S}(t_1) \cdot \frac{6 - 3 \cdot 1}{6} = \frac{6}{7} \cdot \frac{1}{2} = \frac{3}{7} \\
\widehat{S}(t_3) &= \widehat{S}(t_2) \cdot \frac{3 - 1 \cdot 1}{3} = \frac{3}{7} \cdot \frac{2}{3} = \frac{2}{7} \\
\widehat{S}(t_4) &= \widehat{S}(t_3) \cdot \frac{2 - 1 \cdot 0}{2} = \widehat{S}(t_3) = \frac{2}{7} \\
\widehat{S}(t_5) &= \widehat{S}(t_4) \cdot \frac{1 - 1 \cdot 1}{1} = 0.
\end{aligned}$$

Sada ćemo pomoću programskog paketa R dati grafički prikaz procijenjene funkcije doživljenja, pri čemu ćemo kao ulaznu tablicu podataka koristiti tablicu 1.1. Dobiveni rezultat prikazan je na slici 1.2, na kojoj možemo vidjeti da se procijenjene vrijednosti funkcije doživljenja podudaraju s prethodno izračunatim vrijednostima $\widehat{S}(t_1), \ldots, \widehat{S}(t_5)$.

Funkcije doživljenja koristan su alat za usporedbu vjerojatnosti doživljenja po grupama subjekata. Naime, ukoliko za svaku grupu odredimo pripadnu funkciju doživljenja, u mogućnosti smo dati grubi zaključak o postojanju razlike u vremenu doživljenja između njih. Za testiranje statistički značajne razlike između dvije funkcije doživljenja prikladan



Slika 1.2: Kaplan-Meierova procjena funkcije doživljenja S(t)

je **test log-rangova** sa nultom hipotezom

$$H_0 : \text{funkcije doživljenja } S_1 \text{ i } S_2 \text{ su jednake.}$$

Pripadna testna statistika konstruirana je na sljedeći način. Pretpostavimo da smo zabilježili realizacije $t_1, \ldots, t_n$ slučajne varijable T, od kojih je l različitih. Bez smanjenja općenitosti možemo pretpostaviti da su to vremena $t_1, \ldots, t_l$. Želimo usporediti dvije funkcije doživljenja: jednu za grupu 1, a drugu za grupu 2. Stoga za $t \in \{t_1, \ldots, t_l\}$ označimo

sa

$$\begin{aligned} n_{1,t} &:= \text{broj rizičnih osoba u grupi 1 neposredno prije trenutka } t \\ &\quad \text{za koje se može zbiti događaj u trenutku } t, \\ n_{2,t} &:= \text{broj rizičnih osoba u grupi 2 neposredno prije trenutka } t \\ &\quad \text{za koje se može zbiti događaj u trenutku } t, \\ m_{1,t} &:= \text{broj osoba u grupi 1 za koje se zbio događaj u trenutku } t, \\ m_{2,t} &:= \text{broj osoba u grupi 2 za koje se zbio događaj u trenutku } t. \end{aligned}$$

Ako vrijedi nulta hipoteza, tada je očekivani broj osoba iz grupe 1 za koje se zbio događaj u trenutku t jednak

$$e_{1,t} = \frac{m_{1,t} + m_{2,t}}{n_{1,t} + n_{2,t}} \cdot n_{1,t}.$$

Pritom

$$\frac{m_{1,t} + m_{2,t}}{n_{1,t} + n_{2,t}}$$

predstavlja vjerojatnost da se zbio događaj u trenutku t. Slično, očekivani broj osoba iz grupe 2 za koje se zbio događaj u trenutku t dan je sa

$$e_{2,t} = \frac{m_{1,t} + m_{2,t}}{n_{1,t} + n_{2,t}} \cdot n_{2,t}.$$

Sada stavimo

$$O_1 - E_1 := \sum_{i=1}^{l} (m_{1,t_i} - e_{1,t_i})$$

$$O_2 - E_2 := \sum_{i=1}^{l} (m_{2,t_i} - e_{2,t_i})$$

i definiramo [3] testnu statistiku

$$\frac{(O_2 - E_2)^2}{\mathbb{V}\text{ar}\,(O_2 - E_2)}, \tag{1.10}$$

gdje je

$$\mathbb{V}\text{ar}\,(O_2 - E_2) = \sum_{i=1}^{l} \frac{n_{1,t_i} n_{2,t_i} (m_{1,t_i} + m_{2,t_i}) (n_{1,t_i} + n_{2,t_i} - m_{1,t_i} - m_{2,t_i})}{(n_{1,t_i} + n_{2,t_i})^2 (n_{1,t_i} + n_{2,t_i} - 1)}.$$

[3] primijetimo da vrijedi $O_1 - E_1 = -(O_2 - E_2)$

Statistiku (1.10) nazivamo statistikom **log-rangova**, a pokazuje se da za nju asimptotski vrijedi

$$\frac{(O_2 - E_2)^2}{\mathbb{V}\text{ar}\,(O_2 - E_2)} \overset{H_0}{\sim} \chi^2(1).$$

S obzirom na ishod ove testne statistike određuje se pripadna p-vrijednost, a na temelju koje dajemo zaključak o odbacivanju hipoteze testa log-rangova.

Primijetimo da su nam za Kaplan-Meierovu procjenu funkcije doživljenja bili potrebni samo podaci o trenutku događaja te informacija o tome je li se događaj zbio ili je cenzuriran. Međutim, u kliničkim studijama česte su situacije u kojima imamo mogućnost prikupljanja dodatnih informacija koje potencijalno utječu na vjerojatnost doživljenja subjekata. Kako bismo uključili i takve informacije, potreban nam je model koji će ih obuhvaćati te nam dati mogućnost procjene funkcije doživljenja i ostalih veličina koje pritom prirodno slijede iz konstrukcije modela. Navedene zahtjeve obuhvaća Coxov model, pomoću kojeg ćemo moći kvantificirati utjecaj pojedinih varijabli na vjerojatnost doživljenja.

Poglavlje 2

Coxov model

2.1 Osnovna formula

Kako bismo analizirali podatke o doživljenju, jedan od mogućih pristupa je prilagodba nekog od parametarskih statističkih modela. Općenito, parametarski model je familija distribucija određena konačnim brojem parametara, a neke od uobičajenih distribucija u takvom modelu su Weibullova distribucija, eksponencijalna distribucija kao poseban slučaj Weibullove, log-normalna te generalizirana gama distribucija. Jedna od karakteristika parametarskih modela u analizi doživljenja je ta da vrijeme doživljenja prati distribuciju koja nam je poznata. Međutim, u praksi je često bilo vrlo teško odrediti tu distribuciju, a posljedično i funkciju hazarda, zbog čega je prirodno bilo ići u smjeru konstrukcije modela koji neće imati nikakve pretpostavke na distribuciju vremena doživljenja. U tu svrhu, D. Cox je 1972. godine predstavio model koji ne podrazumijeva poznavanje distribucije vremena doživljenja, poznat pod nazivom **Coxov regresijski model**. Sama konstrukcija modela temelji se na pretpostavci da je omjer funkcija hazarda između dva subjekta (ili dvije grupe) u fiksnom trenutku t konstantan, ili ekvivalentno, da je funkcija hazarda jedne osobe proporcionalna funkciji hazarda druge osobe za fiksni trenutak t. Kako bi se naglasila važnost navedene pretpostavke u samoj konstrukciji modela, Coxov model često se naziva i **Coxov model proporcionalnih hazarda**. Pomoću ove pretpostavke, u ovom ćemo poglavlju dati opis strukture Coxovog modela, počevši od definicije klasičnog modela proporcionalnog hazarda.

Stoga pretpostavimo da imamo podatke o n osoba koje su raspoređene u dvije grupe. Nadalje, neka grupa 1 predstavlja ljude koji nisu liječeni, a grupa 2 ljude koji su liječeni. Označimo sa h_0 funkciju hazarda osoba iz grupe 1, a sa h_i funkciju hazarda i-te osobe iz grupe 2. Zanima nas je li se funkcija hazarda osobe iz grupe 2 promijenila relativno na funkciju hazarda osobe iz grupe 1. Drugim riječima, zanima nas je li primjena lijeka utjecala na vjerojatnost doživljenja. Polazeći od pretpostavke da je omjer funkcija hazarda

između grupa proporcionalan, model proporcionalnog hazarda možemo pisati u obliku

$$\frac{h_i(t)}{h_0(t)} = \lambda, \forall t, \tag{2.1}$$

pri čemu je $\lambda \in \mathbb{R}$. Omjer u formuli (2.1) nazivamo **relativnim hazardom** ili, češće, **omjerom rizika** (engl. *Hazard Ratio*). Prema tome, u središtu interesa nam je odrediti upravo oblik skalara λ. Zato, neka je X indikatorska varijabla definirana sa

$$X = \begin{cases} 1, & \text{ako je osoba liječena (pripada grupi 2)} \\ 0, & \text{ako osoba nije liječena (pripada grupi 1).} \end{cases}$$

Sa x_i označimo vrijednost varijable X za osobu $i \in \{1, \ldots, n\}$. Budući da je prirodan zahtjev da je relativni hazard λ nenegativan, razumno je staviti $\lambda = \exp(\beta \cdot x_i)$, za neki $\beta \in \mathbb{R}$. Tada se omjer rizika (2.1) može zapisati u obliku

$$\frac{h_i(t)}{h_0(t)} = \exp(\beta \cdot x_i),$$

odnosno, funkcija hazarda za osobu i može se napisati u obliku

$$h_i(t) = h_0(t) \cdot \exp(\beta \cdot x_i). \tag{2.2}$$

Prema tome, ako osoba i nije tretirana lijekom, tada je pripadna funkcija hazarda jednaka h_0. S druge strane, za osobu tretiranu lijekom, funkcija hazarda je oblika (2.2). Ovim pristupom, $\exp(\beta)$ dat će nam informaciju o tome koliko je puta hazard osobe tretirane lijekom veći ili manji u odnosu na osobu koja nije tretirana lijekom. Taj omjer zapravo kvantificira efikasnost lijeka.

Navedena ideja proširuje se na slučaj kada funkcija hazarda h_i ovisi o vrijednostima $x_1, \ldots, x_p$ nezavisnih varijabli $X_1, \ldots, X_p$, pri čemu su X_j, $j = 1, \ldots, p$ varijable bilo kakvog tipa. Označimo sa $\mathbf{x_i} = (x_{1i}, \ldots, x_{pi})$ vrijednosti kovarijata i-te osobe te sa h_0 funkciju hazarda osobe čije su sve kovarijate jednake nuli. Tada, zbog pretpostavke proporcionalnog hazarda dvije osobe, funkciju hazarda i-te osobe možemo pisati u obliku

$$h_i(t) = h_0(t) \cdot \lambda(\mathbf{x_i}).$$

Funkcija $\lambda(\mathbf{x_i})$ može se interpretirati kao hazard u vremenu t za osobu čije su kovarijate jednake $\mathbf{x_i}$, relativno na hazard osobe čije su sve kovarijate jednake nuli. Sada ostaje pitanje kojeg je oblika funkcija $\lambda(\mathbf{x_i})$. Kao i ranije, prirodno ju je pisati u obliku

$$\lambda(\mathbf{x_i}) = \exp(S_i),$$

pri čemu je

$$S_i = \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \ldots + \beta_p \cdot x_{pi}$$

linearna kombinacija kovarijata koju nazivamo **linearnom komponentom** modela ili **skor rizika** za i-tu osobu. Na ovaj način dobivamo generalnu formulu Coxovog modela proporcionalnih hazarda koja je oblika

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i\right), \tag{2.3}$$

gdje je $\mathbf{X} = (X_1, \ldots, X_p)$ vektor kovarijata. Nadalje, kako bismo ga prikazali u terminima linearnog modela, često ga reprezentiramo sa

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \sum_{i=1}^{p} \beta_i \cdot X_i.$$

Dakle, formula (2.3) daje nam izraz za hazard u trenutku t koji pripada osobi s danim kovarijatama $\mathbf{X}$. Ona nam govori da je funkcija hazarda jednaka produktu dvije veličine.

Prva veličina je funkcija $h_0(t)$ koju nazivamo **funkcija osnovnog hazarda** (engl. *Baseline Hazard Function)*. Dodijeljen joj je takav naziv jer uvrštavanjem kovarijata $(X_1, \ldots, X_p) = (0, \ldots, 0)$ u formulu (2.3) dobivamo

$$h(t, \mathbf{X}) = h_0(t).$$

U tom slučaju, u modelu nije prisutna niti jedna kovarijata pa je Coxov model reduciran na funkciju osnovnog hazarda. Prema tome, h_0 se može smatrati početnim hazardom neke osobe, prije nego što smo u obzir uzeli poznate dodatne informacije sadržane u pripadnim kovarijatama $\mathbf{X}$.

Druga veličina je $\exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i\right)$ koja ovisi o kovarijatama $X_1, \ldots, X_p$. Njih još nazivamo i prediktivnim varijablama jer pomoću njih želimo procijeniti i predvidjeti ponašanje funkcije hazarda za osobu sa upravo tim kovarijatama. Ovaj dio formule osigurava da funkcija hazarda bude nenegativna, odnosno da vrijedi $0 \leq h(t) < \infty$. Budući da varijable $X_1, \ldots, X_p$ ne mijenjaju svoje vrijednosti tijekom vremena, nazivamo ih vremenski nezavisnim kovarijatama. Posljedično, $\exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i\right)$ ne ovisi o vremenu t.

Dakle, specifičnost formule (2.3) je upravo u tome što funkcija h_0 ovisi samo o vremenu t, a ne i o kovarijatama $X_1, \ldots, X_p$, dok s druge strane drugi dio formule ovisi samo o kovarijatama, ali ne i o vremenu t.

Velika vrijednost Coxovog modela sastoji se u tome što poznavanje funkcije osnovnog hazarda ne mora biti nužno kako bismo dobili bitne informacije o samim podacima. Preciznije, bez određivanja funkcionalne forme funkcije h_0 možemo procijeniti omjere rizika. Dakle, oblik osnovne funkcije hazarda $h_0(t)$ općenito ne mora biti poznat, što ga čini vrlo primjenjivim u praksi. Prema tome, od interesa nam je procijeniti samo koeficijente $\beta_1, \ldots, \beta_p$ koji nam daju mjeru utjecaja kovarijata $X_1, \ldots, X_p$ na logaritam omjera rizika. S

obzirom da funkcionalna forma Coxovog modela općenito nije u cjelini poznata, za Coxov model kažemo da je **semiparametarski model**.

Nadalje, pokazalo se da, iako nam osnovni hazard h_0 nije poznat, Coxov model daje vrlo dobre procjene regresijskih koeficijenata, omjera hazarda od interesa i prilagođene krivulje doživljenja za različite varijacije setova podataka koji su nam dostupni.

Prirodno pitanje koje se postavlja je ima li Coxov model prednosti u odnosu na druge modele. Odgovor je potvrdan ukoliko uspoređujemo Coxov model sa modelom logističke regresije, u slučaju kada nam je uz same podatke poznata i informacija o vremenu doživljenja. Naime, u tom slučaju Coxov model ima prednost jer uzima u obzir vrijeme doživljenja i cenzuriranje podataka, dok logistička regresija uključuje samo vremena doživljenja, ali ne i cenzuriranje. Osim toga, pokazalo se da ukoliko je pravi model neki od poznatijih parametarskih modela (primjerice, Weibullov model), da će Coxov model dati vrlo slične procjene kao i pravi model. Iako općenito postoje statistički testovi kojima se može testirati prate li podaci neki konkretan parametarski model, ukoliko nismo sigurni o kojem modelu se radi, preporučljivo je koristiti Coxov model kako bismo izbjegli potencijalno pogrešnu pretpostavku o obliku funkcije osnovnog hazarda.

2.2 Procjena parametara i funkcija vjerodostojnosti

Polazeći od činjenice da je Coxov model semiparametarski model, dovoljno je procijeniti parametre $\beta_1, \ldots, \beta_p$ kako bismo dobili procjenu funkcije doživljenja i omjera hazarda. Jedna od najpoznatijih metoda u statistici za procjenu regresijskih parametara je metoda najveće vjerodostojnosti (engl. *Maximum Likelihood Method*) jer ona rezultira procjeniteljima koji imaju dobra asimptotska svojstva, a što je prigodno za studije u kojima su dostupni veći uzorci. Općenito, u jednodimenzionalnom slučaju, za opaženi uzorak $\mathbf{t} = (t_1, \ldots, t_n)$ slučajne varijable T čija je gustoća $f(\mathbf{x}|\theta)$, **vjerodostojnost** parametra θ je funkcija $L : \Theta \to \mathbb{R}$ definirana sa

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

Prema tome, u našem višedimenzionalnom slučaju imamo $\theta = (\beta_1, \ldots, \beta_p)$, a sama metoda sastoji se u traženju parametara $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ koji maksimiziraju funkciju vjerodostojnosti. Drugim riječima, tražimo parametre $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ za koje vrijedi

$$L((\widehat{\beta}_1, \ldots, \widehat{\beta}_p)|\mathbf{t}) = \max_{(\beta_1,\ldots,\beta_p)\in\Theta} L((\beta_1, \ldots, \beta_p)|\mathbf{t}).$$

Dakle, funkcija vjerodostojnosti daje vjerojatnost da se dogodi realizacija opaženog uzorka. Međutim, bitno je uočiti da određivanje funkcije vjerodostojnosti zahtijeva poznavanje

distribucije zavisne varijable (u našem slučaju to je slučajna varijabla T), što se kosi sa činjenicom da Coxov model ne pretpostavlja poznavanje distribucije te varijable. Iz tog razloga, ovakav pristup određivanja funkcije vjerodostojnosti nije u potpunosti dobar pa je funkcija vjerodostojnosti u Coxovom modelu konstruirana na drugačiji način. Naime, umjesto poznavanja distribucije zavisne varijable, Coxova funkcija vjerodostojnosti bazira se na opaženom poretku događaja od interesa. Dakle, ključna ideja je iskoristiti informaciju o tome da znamo kojim se točno redoslijedom događao događaj od interesa, pri čemu se pretpostavlja da u intervalima između dva vremena događaja nemamo nikakvu dodatnu informaciju o utjecaju kovarijata na vrijeme doživljenja jer je u pripadnom intervalu funkcija h_0 jednaka nuli. To znači da taj interval ne daje nikakvu dodatnu informaciju o parametrima $\beta_1, \ldots, \beta_p$. Samu ideju prezentirat ćemo na jednostavnijem primjeru. Neka je dan Coxov model

$$h(t) = h_0(t) \cdot \exp(\boldsymbol{\beta}' \cdot \mathbf{X}), \tag{2.4}$$

pri čemu je $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ i $\mathbf{X} = (X_1, \ldots, X_p)'$. Pretpostavimo da promatramo n osoba i da imamo informacije o njima prikazane u tablici 2.1.

Osoba	Vrijeme	d	$\mathbf{X}$
$osoba_1$	t_1	1	$\mathbf{x}^1$
$osoba_2$	t_2	1	$\mathbf{x}^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$osoba_{k-1}$	t_{k-1}	1	$\mathbf{x}^{k-1}$
$osoba_k$	t_k	0	$\mathbf{x}^k$
$osoba_{k+1}$	t_{k+1}	1	$\mathbf{x}^{k+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$osoba_n$	t_n	1	$\mathbf{x}^n$

Tablica 2.1: Tablica podataka

Pritom varijabla Vrijeme predstavlja vrijeme kada se događaj od interesa dogodio (dakle, realizaciju slučajne varijable T), varijabla d bilježi je li se događaj dogodio ili je cenzuriran u pripadnom trenutku, a varijabla $\mathbf{X}$ predstavlja poznate informacije o osobi (dakle, vektor kovarijata). Uz to, radi jednostavnijeg opisa ideje, možemo pretpostaviti da su vremena događaja $t_1, \ldots, t_n$ poredana uzlazno i da je za $osobu_k$ vrijeme doživljenja cenzurirano, odnosno da nam nije poznat podatak u kojem trenutku nakon t_k se dogodio događaj od interesa, dok ostali podaci nisu cenzurirani. Coxova funkcija vjerodostojnosti dat će nam vjerojatnost da se događaj od interesa dogodio upravo ovim redoslijedom osoba.

Odredimo sada vjerojatnost da se događaj dogodio najprije za $osobu_1$. Definirajmo

$$L_1 := \mathbb{P}\left(\text{događaj se dogodio za } \mathbf{osobu_1} \text{ u trenutku } t_1 | \text{ događaj se dogodio u trenutku } t_1\right),$$

pa korištenjem formule uvjetne vjerojatnosti dobivamo

$$L_1 = \frac{\mathbb{P}\left(\text{događaj se dogodio za } osobu_1 \text{ u trenutku } t_1\right)}{\mathbb{P}(\text{događaj se dogodio u trenutku } t_1)}. \tag{2.5}$$

Budući da su događaji

$$\{\text{događaj se dogodio u trenutku } t_1 \text{ za osobu } i\}, \text{ za } i = 1, \ldots, n$$

međusobno disjunktni, nazivnik u formuli (2.5) jednak je

$$\sum_{r \in R} \mathbb{P}\left(\text{događaj se dogodio za } \mathbf{osobu_r} \text{ u trenutku } t_1\right), \tag{2.6}$$

pri čemu je R skup indeksa rizičnih osoba u trenutku procjene t_1, odnosno osoba za koje se mogao dogoditi događaj u trenutku t_1. Primjerice, kod računanja vjerojatnosti L_1, skup R sadrži indekse svih n osoba jer su sve izložene riziku od događaja, odnosno vrijedi $|R| = n$.

Dakle, iz (2.5) i (2.6) imamo

$$L_1 = \frac{\mathbb{P}\left(\text{događaj se dogodio za } osobu_1 \text{ u trenutku } t_1\right)}{\sum_{r \in R} \mathbb{P}\left(\text{događaj se dogodio za } osobu_r \text{ u trenutku } t_1\right)}. \tag{2.7}$$

Zamijenimo li trenutak t_1 u formuli (2.7) sa intervalom $[t_1, t_1 + \Delta t_1]$ te podijelimo dobiveno sa $\frac{\Delta t_1}{\Delta t_1}$, dobivamo

$$L_1 = \frac{\mathbb{P}\left(\text{događaj se dogodio za } osobu_1 \text{ u intervalu } [t_1, t_1 + \Delta t_1]\right)/\Delta t_1}{\sum_{r \in R} \mathbb{P}\left(\text{događaj se dogodio za } osobu_r \text{ u intervalu} [t_1, t_1 + \Delta t_1]\right)/\Delta t_1}, \tag{2.8}$$

pa puštanjem limesa $\Delta t_1 \to 0$ u (2.8), iz definicije funkcije hazarda dobivamo

$$\lim_{\Delta t_1 \to 0} L_1 = \frac{h_1(t_1)}{\sum_{r \in R} h_r(t_1)}.$$

Prema tome, vjerojatnost $L_1 = L_1(\boldsymbol{\beta})$ je dana sa

$$L_1(\boldsymbol{\beta}) = \frac{\cancel{h_0(t_1)} \cdot \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^1)}{\sum_{r \in R} \cancel{h_0(t_1)} \cdot \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r)} = \frac{\exp(\boldsymbol{\beta}' \cdot \mathbf{x}^1)}{\sum_{r \in R} \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r)}. \tag{2.9}$$

Uočimo da se u (2.9) više ne javlja član $h_0(t_1)$, što je posljedica pretpostavke Coxovog modela da je osnovni hazard $h_0(t_1)$ za sve osobe jednak. Ovdje, dakle, vidimo važnost navedene pretpostavke proporcionalnog hazarda jer nam ona omogućava procjenu funkcije

vjerodostojnosti bez poznavanja osnovne funkcije hazarda. Na analogan način bismo dobili da je vjerojatnost da se za $osobu_2$ događaj dogodio drugi po redu jednaka

$$L_2(\boldsymbol{\beta}) := \frac{\exp(\boldsymbol{\beta}' \cdot \mathbf{x}^2)}{\sum_{r \in R} \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r)},$$

pri čemu je sada broj rizičnih osoba jednak $|R| = n-1$ jer se za $osobu_1$ događaj već dogodio pa više nije u skupu rizičnih osoba. Postupak nastavljamo dalje dok ne dođemo do $osobe_k$. Naime, budući da imamo cenzurirani podatak, nije ispravno procjenjivati vjerojatnost da je $osoba_k$ doživjela događaj k-ta po redu, jer ne znamo točno vrijeme kada se događaj dogodio za tu osobu. Jedina informacija koju imamo jest da se događaj nije dogodio za $osobu_k$ do trenutka t_k. Nakon toga očito je da podaci za $osobu_k$ više nisu bilježeni, što znači da će kod procjene vjerojatnosti za $osobu_{k+1}$ broj rizičnih osoba biti za jedan manji, odnosno vrijedit će $|R| = n - k$. Postupak nastavljamo sve do $osobe_n$, pri čemu vjerojatnosti procjenjujemo samo za one osobe sa necenzuriranim podacima.

Prema tome, vjerojatnost da se događaj dogodio redoslijedom iz tablice 2.1 dana je formulom

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{J} L_i(\boldsymbol{\beta}) = \prod_{i=1}^{J} \frac{\exp(\boldsymbol{\beta}' \cdot \mathbf{x}^i)}{\sum_{r \in R(t_i)} \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r)}, \tag{2.10}$$

pri čemu je J ukupan broj osoba za koje se dogodio događaj, dok smo sa $R(t_i)$ označili skup indeksa preostalih rizičnih osoba neposredno do trenutka t_i. Formula (2.10) predstavlja funkciju vjerodostojnosti Coxovog modela (2.3). Dakle, Coxova funkcija vjerodostojnosti podrazumijeva vjerojatnosti samo za one subjekte za koje se događaj dogodio te ne uključuje vjerojatnosti za cenzurirane subjekte. Zbog toga formulu (2.10) nazivamo još i **parcijalnom funkcijom vjerodostojnosti**.

Preostaje maksimizirati funkciju vjerodostojnosti (2.10), što je ekvivalentno maksimizaciji funkcije log-vjerodostojnosti

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{J} \left[\ln\left(\exp(\boldsymbol{\beta}' \cdot \mathbf{x}^i)\right) - \ln\left(\sum_{r \in R(t_i)} \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r) \right) \right].$$

Procjenitelje $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ maksimalne vjerodostojnosti parametara $(\beta_1, \ldots, \beta_p)$ dobivamo kao rješenje sustava jednadžbi

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = 0, \text{ za } j = 1, \ldots, p. \tag{2.11}$$

Jednadžbe u sustavu (2.11) još nazivamo i **jednadžbama skora**. Vrijednosti $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ dobivamo Newton-Raphsonovom metodom koja je implementirana u gotovo svakom programskom paketu namijenjenom statističkoj analizi podataka.

Prema tome, procijenjenu funkciju hazarda pišemo u obliku

$$\widehat{h}(t, \mathbf{X}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \widehat{\beta_i} \cdot X_i\right). \tag{2.12}$$

Za dobivene procjene $\widehat{\beta_i}, i = 1, \ldots, p$, granice aproksimativnog intervala pouzdanosti $100 \cdot (1 - \alpha)\%$ jednake su

$$\widehat{\beta_i} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta_i}\right)}, \tag{2.13}$$

pri čemu je $\sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta_i}\right)} =: s_{\widehat{\beta_i}}$ standardna greška procjene $\widehat{\beta_i}$, a $z_{\frac{\alpha}{2}}$ predstavlja $(1 - \frac{\alpha}{2})$-kvantil standardne normalne distribucije. Najčešće nas zanima 95% pouzdani interval, pa prema tome vrijedi $z_{\frac{\alpha}{2}} = z_{0.975} = 1.96$. Kod ispisa procedura u programskom jeziku, osim procjene parametra β_i, ispisana je i standardna greška $s_{\widehat{\beta_i}}$ pripadne procjene. Pomoću nje možemo lako formulom (2.13) odrediti i aproksimativni interval pouzdanosti $100 \cdot (1 - \alpha)\%$. Ukoliko određeni interval ne sadrži vrijednost 0, tada zaključujemo da pripadni procijenjeni parametar $\widehat{\beta_i}$ nije značajno jednak nuli, odnosno, kažemo da je parametar $\widehat{\beta_i}$ statistički značajan. Preciznije, nulta hipoteza $H_0 : \beta_i = 0$, za neki $i \in \{1, \ldots, n\}$ testira se Waldovim testom, s pripadnom testnom statistikom

$$\frac{\widehat{\beta_i}}{s_{\widehat{\beta_i}}} \overset{H_0}{\sim} N(0, 1).$$

Pripadna p-vrijednost testa također je prikazana u izlaznoj tablici programskog paketa R. Pritom je važno napomenuti da testiranje hipoteze $H_0 : \beta_i = 0$, za neki $i \in \{1, \ldots, n\}$ podrazumijeva testiranje statističke značajnosti parametra β_i u modelu koji sadrži sve kovarijate $X_1, \ldots, X_p$. Prema tome, vrijednost $\frac{\widehat{\beta_i}}{s_{\widehat{\beta_i}}}$ ovisit će o tome koje kovarijate smo uključili u model. Dakle, zaključak o statističkoj značajnosti parametra $\widehat{\beta_i}$ ovisit će također o kovarijatama koje su uključene u model. Posljedično, može se dogoditi situacija u kojoj je neki parametar $\widehat{\beta_i}$ statistički značajan kada je u modelu jedan set kovarijata, dok sa drugim setom kovarijata neće biti statistički značajan.

2.3 Omjer rizika

Jednom kada odredimo parametre $\widehat{\beta_1}, \ldots, \widehat{\beta_p}$, možemo dati usporedbu funkcija hazarda dvije osobe s različitim kovarijatama kako bismo dobili informaciju koliko je puta veći hazard jedne u odnosu na drugu osobu. Tu informaciju nazivamo omjer rizika, kojeg definiramo sa

$$OR := \frac{h_1\left(t|\mathbf{X}^1\right)}{h_2\left(t|\mathbf{X}^2\right)},$$

pri čemu su h_1 i h_2 funkcije hazarda prve, odnosno druge osobe, a $\mathbf{X}^1 = (X_1^1, \ldots, X_p^1)$ i $\mathbf{X}^2 = (X_1^2, \ldots, X_p^2)$ pripadne kovarijate prve i druge osobe, redom. U praksi, procjena omjera rizika $\widehat{OR}$ jednaka je omjeru procijenjenih funkcija hazarda, odnosno vrijedi

$$\widehat{OR} = \frac{\widehat{h_1}\left(t|\mathbf{X}^1\right)}{\widehat{h_2}\left(t|\mathbf{X}^2\right)}. \tag{2.14}$$

Koristeći formulu Coxovog modela (2.3) i formulu (2.14), omjer rizika možemo napisati u obliku

$$\widehat{OR} = \frac{\cancel{\widehat{h_0}(t)} \cdot \exp\left(\sum_{i=1}^p \widehat{\beta}_i \cdot X_i^1\right)}{\cancel{\widehat{h_0}(t)} \cdot \exp\left(\sum_{i=1}^p \widehat{\beta}_i \cdot X_i^2\right)} = \exp\left(\sum_{i=1}^p \widehat{\beta}_i \cdot \left(X_i^1 - X_i^2\right)\right). \tag{2.15}$$

Primijetimo da omjer rizika $\widehat{OR}$ ne ovisi o trenutku t, što znači da je omjer rizika konstantan kroz vrijeme. Ovo svojstvo daje nam ideju kako provjeriti je li zadovoljena pretpostavka proporcionalnog hazarda, o čemu će detaljnije biti govora u Poglavlju 3.

2.4 Procjena osnovnog hazarda i funkcije doživljenja

Coxov model ponajprije je definiran kako bismo mogli procijeniti omjere rizika bez poznavanja funkcije hazarda čime se izbjegavaju greške prilikom procjene te funkcije. Međutim, ponekad nam informacije o omjeru hazarda nisu dovoljne kako bismo dobili tražene zaključke. Stoga ćemo u ovom poglavlju predočiti metodu procjene funkcije hazarda Coxovog modela, a samim time i pripadne funkcije doživljenja, koja u tom slučaju uključuje i vrijednosti kovarijata.

Ponajprije, procijenjena funkcija hazarda dana je sa

$$\widehat{h}(t) = \widehat{h}_0(t) \cdot \exp\left(\widehat{\boldsymbol{\beta}}' \cdot \mathbf{X}\right), \tag{2.16}$$

pri čemu je $\mathbf{X} = \left(X_1, \ldots, X_p\right)'$ vektor kovarijata. Integriranjem (2.16) imamo

$$\int_0^t \widehat{h}(u)du = \exp\left(\widehat{\boldsymbol{\beta}}' \cdot \mathbf{X}\right) \int_0^t \widehat{h}_0(u)du$$

pa odavde i iz jednakosti (1.5) lako slijedi da je procjena funkcije doživljenja dana sa

$$\widehat{S}(t, \mathbf{X}) = \left[\widehat{S}_0(t)\right]^{\exp\left(\widehat{\boldsymbol{\beta}}' \cdot \mathbf{X}\right)}. \tag{2.17}$$

Jedna od poznatijih procjena osnovnog kumulativnog hazarda je **Breslowljeva procjena** (vidi [1]) dana sa

$$\int_0^t \widehat{h}_0(u)du = \sum_{t_{(i)} \leq t} \frac{d_i}{\sum_{r \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r)},$$

gdje su $t_{(1)} < t_{(2)} < \ldots < t_{(J)}$ vremena necenzuriranih događaja. Stoga je procjena osnovne funkcije doživljenja dana sa

$$\widehat{S}_0(t) = \exp\left(- \sum_{t_{(i)} \leq t} \frac{d_i}{\sum_{r \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \cdot \mathbf{x}^r)} \right). \tag{2.18}$$

2.5 Primjena na podacima

Kliničkom studijom provedenoj u Australiji 1991. godine dobiveni su podaci o pacijentima u dvije klinike koje se razlikuju po načinu liječenja i kvaliteti tretmana pacijenata. U studiji je sudjelovalo ukupno 238 ispitanika, od čega 163 u prvoj klinici, a 75 u drugoj klinici, tijekom koje su pacijenti primali određenu dozu metadona. To je snažan analgetik koji se uobičajeno koristi za otklanjanje znakova apstinencijske krize ovisnika te koji održava normalno stanje pacijenta. Glavni cilj studije je usporediti vrijeme boravka pacijenata u klinikama i ustanoviti postoji li razlika u duljini boravka između klinika. Osim toga, želi se usporediti vrijeme odlaska pacijenata iz klinike ovisno o njihovim karakteristikama. Pritom pacijent može otići iz klinike iz dva osnovna razloga:

1. Pacijent prema procjeni liječnika mora uzimati vrlo malu količinu metadona, zbog čega više nije potrebno njegovo boravljenje u klinici;

2. Pacijent zbog neprihvatljivog ponašanja ugrožava ostale pacijente.

Prema tome, događaj koji promatramo je odlazak iz klinike, dok je slučajna varijabla od interesa vrijeme provedeno u klinici. Stoga, kako bismo dobili tražene zaključke, prikladno je koristiti teoriju analize doživljenja. Osim informacija o vremenu odlaska iz klinike, dane su dodatne informacije o svakoj osobi, što nam prirodno nameće prilagodbu Coxovog modela na podacima za modeliranje navedenog problema. Informacije o sudionicima dane su sljedećim varijablama, čije su vrijednosti prikazane u tablici 2.2.

- Varijabla **Klinika** je dihotomna varijabla koja poprima vrijednost 1 ukoliko je osoba liječena u klinici 1, dok poprima vrijednost 2 ukoliko je osoba liječena u klinici 2.

- Varijabla **Status** je dihotomna varijabla koja poprima vrijednost 1 ukoliko je zabilježeno da je osoba otišla iz klinike u promatranom intervalu. S druge strane,

poprima vrijednost 0 ukoliko nemamo informaciju o tome je li osoba u promatranom intervalu otišla iz klinike, odnosno, u slučaju kada imamo cenzurirani podatak. Prema tome, vrijednost 0 uključuje i one ispitanike koji su iz nekog razloga morali napustiti kliniku s visokom dozom metadona (primjerice, preseljenje u neku drugu kliniku, smrt i slično).

- Varijabla **Vrijeme** je neprekidna varijabla koja predstavlja vrijeme koje je proteklo od početka studije do odlaska osobe iz klinike. Vrijeme je mjereno u danima pa poprima vrijednosti iz skupa $\mathbb{N}_0$.

- Varijabla **Ustanova** je dihotomna varijabla koja poprima vrijednosti iz skupa $\{0, 1\}$, pri čemu 1 označava da je osoba bila prethodno u zatvoru, dok 0 označava da nema zatvorski dosje. Informacija o zatvorskom dosjeu bilježi se iz razloga što prethodno boravljenje u zatvoru potencijalno može uzrokovati neprihvatljivo ponašanje pacijenata, što posljedično može utjecati na odlazak iz klinike.

- Varijabla **Doza** je neprekidna varijabla koja označava maksimalnu količinu metadona (u miligramima po danu) koja je dana ispitaniku tijekom studije. Ukoliko je pacijentu dana relativno velika doza metadona, to ukazuje da osoba nije spremna otići iz klinike. Ukoliko je količina doze relativno mala, postoji mogućnost da osoba izađe iz klinike. Varijabla Doza poprima poprima vrijednosti iz skupa $\langle 0, +\infty \rangle$ jer je svakom pacijentu dana doza metadona.

Klinika	**Status**	**Vrijeme**	**Ustanova**	**Doza**	**Klinika**	**Status**	**Vrijeme**	**Ustanova**	**Doza**
1	1	428	0	50	1	1	275	1	55
1	1	262	0	55	1	1	183	0	30
1	1	259	1	65	1	1	714	0	55
1	1	438	1	65	1	0	796	1	60
1	1	892	0	50	1	1	393	1	65
1	0	161	1	80	1	1	836	1	60
1	1	523	0	55	1	1	612	0	70
1	1	212	1	60	1	1	399	1	60
1	1	771	1	75	1	1	514	1	80
1	1	512	0	80	1	1	624	1	80
1	1	209	1	60	1	1	341	1	60
1	1	299	0	55	1	0	826	0	80
1	1	262	1	65	1	0	566	1	45
1	1	368	1	55	1	1	302	1	50
1	0	602	0	60	1	1	652	0	80
1	1	293	0	65	1	0	564	0	60
1	1	394	1	55	1	1	755	1	65
1	1	591	0	55	1	0	787	0	80
1	1	739	0	60	1	1	550	1	60
1	1	837	0	60	1	1	612	0	65
1	0	581	0	70	1	1	523	0	60
1	1	504	1	60	1	1	785	1	80
1	1	774	1	65	1	1	560	0	65

Nastavak na idućoj stranici

Klinika	**Status**	**Vrijeme**	**Ustanova**	**Doza**	**Klinika**	**Status**	**Vrijeme**	**Ustanova**	**Doza**
1	1	160	0	35	1	1	482	0	30
1	1	518	0	65	1	1	683	0	50
1	1	147	0	65	1	1	563	1	70
1	1	646	1	60	1	1	899	0	60
1	1	857	0	60	1	1	180	1	70
1	1	452	0	60	1	1	760	0	60
1	1	496	0	65	1	1	258	1	40
1	1	181	1	60	1	1	386	0	60
1	0	439	0	80	1	0	563	0	75
1	1	337	0	65	1	0	613	1	60
1	1	192	1	80	1	0	405	0	80
1	1	667	0	50	1	0	905	0	80
1	1	247	0	70	1	1	821	0	80
1	1	821	1	75	1	0	517	0	45
1	0	346	1	60	1	1	294	0	65
1	1	244	1	60	1	1	95	1	60
1	1	376	1	55	1	1	212	0	40
1	1	96	0	70	1	1	532	0	80
1	1	522	1	70	1	1	679	0	35
1	0	408	0	50	1	0	840	0	80
1	0	148	1	65	1	1	168	0	65
1	1	489	0	80	1	0	541	0	80
1	1	205	0	50	1	0	475	1	75
1	1	237	0	45	1	1	517	0	70
1	1	749	0	70	1	1	150	1	80
1	1	465	0	65	2	1	708	1	60
2	0	713	0	50	2	0	146	0	50
2	1	450	0	55	2	0	555	0	80
2	1	460	0	50	2	0	53	1	60
2	1	122	1	60	2	1	35	1	40
2	0	532	0	70	2	0	684	0	65
2	0	769	1	70	2	0	591	0	70
2	0	769	1	40	2	0	609	1	100
2	0	932	1	80	2	0	932	1	80
2	0	587	0	110	2	1	26	0	40
2	0	72	1	40	2	0	641	0	70
2	0	367	0	70	2	0	633	0	70
2	1	661	0	40	2	1	232	1	70
2	1	13	1	60	2	0	563	0	70
2	0	969	0	80	2	0	1052	0	80
2	0	944	1	80	2	0	881	0	80
2	1	190	1	50	2	1	79	0	40
2	0	884	1	50	2	1	170	0	40
2	1	286	0	45	2	0	358	0	60
2	0	326	1	60	2	0	769	1	40
2	1	161	0	40	2	0	564	1	80
2	1	268	1	70	2	0	611	1	40
2	1	322	0	55	2	0	1076	1	80
2	0	2	1	40	2	0	788	0	70
2	0	575	0	80	2	1	109	1	70
2	0	730	1	80	2	0	790	0	90
2	0	456	1	70	2	1	231	1	60
2	1	143	1	70	2	0	86	1	40
2	0	1021	0	80	2	0	684	1	80
2	1	878	1	60	2	1	216	0	100
2	0	808	0	60	2	1	268	1	40
2	0	222	0	40	2	0	683	0	100
2	0	496	0	40	2	1	389	0	55

Nastavak na idućoj stranici

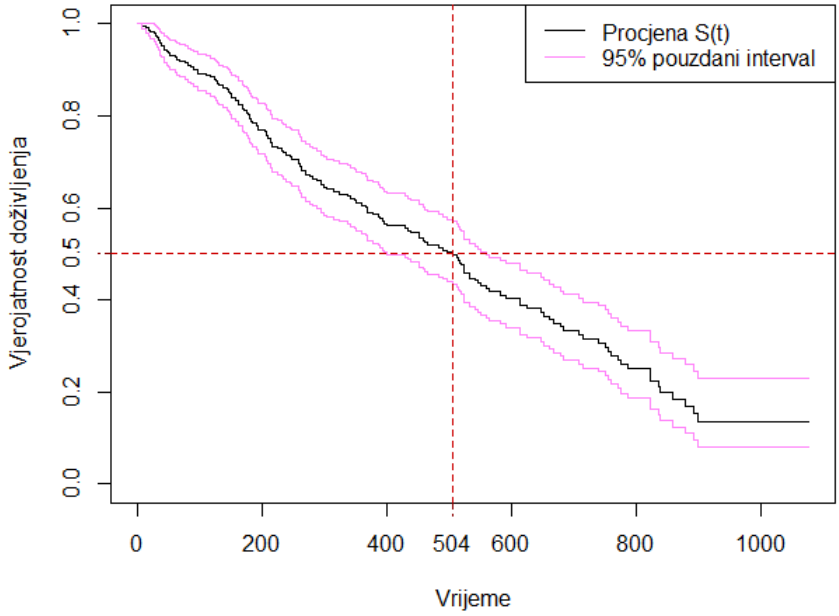
Klinika	**Status**	**Vrijeme**	**Ustanova**	**Doza**	**Klinika**	**Status**	**Vrijeme**	**Ustanova**	**Doza**
1	1	126	1	75	1	1	17	1	40
1	1	350	0	60	2	0	531	1	65
1	0	317	1	50	1	0	461	1	75
1	1	37	0	60	1	1	167	1	55
1	1	358	0	45	1	1	49	0	60
1	1	457	1	40	1	1	127	0	20
1	1	7	1	40	1	1	29	1	60
1	1	62	0	40	1	0	150	1	60
1	1	223	1	40	1	0	129	1	40
1	0	204	1	65	1	1	129	1	50
1	1	581	0	65	1	1	176	0	55
1	1	30	0	60	1	1	41	0	60
1	0	543	0	40	1	0	210	1	50
1	1	193	1	70	1	1	434	0	55
1	1	367	0	45	1	1	348	1	60
1	0	28	0	50	1	0	337	0	40
1	0	175	1	60	2	1	149	1	80
1	1	546	1	50	1	1	84	0	45
1	0	283	1	80	1	1	533	0	55
1	1	207	1	50	1	1	216	0	50
1	0	28	0	50	1	1	67	1	50
1	0	62	1	60	1	0	111	0	55
1	1	257	1	60	1	1	136	1	55
1	0	342	0	60	2	1	41	0	40
2	0	531	1	45	1	0	98	0	40
1	1	145	1	55	1	1	50	0	50
1	0	53	0	50	1	0	103	1	50
1	0	2	1	60	1	1	157	1	60
1	1	75	1	55	1	1	19	1	40
1	1	35	0	60	2	0	394	1	80
1	1	117	0	40	1	1	175	1	60
1	1	180	1	60	1	1	314	0	70
1	0	480	0	50	1	0	325	1	60
2	1	280	0	90	1	1	204	0	50
2	1	366	0	55	2	0	531	1	50
1	1	59	1	45	1	1	33	1	60
2	1	540	0	80	2	0	551	0	65
1	1	90	0	40	1	1	47	0	45

Tablica 2.2: Tablica podataka

Sada ćemo provesti analizu doživljenja koristeći teorijsku pozadinu opisanu u prethodnim poglavljima. Najprije ćemo prikazati Kaplan-Meierove procjene krivulja doživljenja, zatim ćemo prilagoditi Coxov model na podatke kako bismo odredili omjere rizika, i na kraju usporediti početnu Kaplan-Meierovu procjenu funkcije doživljenja sa dobivenom Coxovom krivuljom doživljenja.

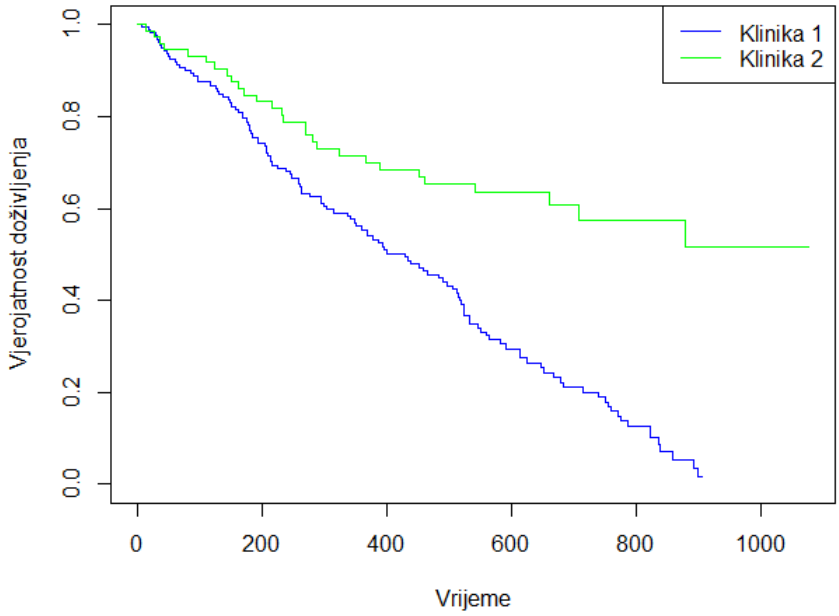
Kao što je već spomenuto, jedna od bitnijih informacija u analizi doživljenja koja nam daje vrlo sažetu informaciju o podacima je funkcija doživljenja. Budući da je događaj od interesa koji promatramo odlazak iz klinike, vrijednost funkcije doživljenja u vremenu t predstavljat će vjerojatnost da osoba nije izašla iz klinike do trenutka t, odnosno vjerojatnost da je osoba boravila u klinici barem t dana. Na slici 2.1 dana je pripadna Kaplan-

Meierova procjena funkcije doživljenja. Dakle, ona ne obuhvaća nikakve informacije o pacijentima osim vremena njihovog odlaska i informaciju o tome je li pripadni podatak cenzuriran ili nije. Drugim riječima, uzima u obzir samo varijable Vrijeme i Status. Osim toga, na istoj slici vidimo da je funkcija doživljenja stepenasta funkcija koja nam daje informaciju o medijalnom vremenu doživljenja jednakom 504 dana, s intervalom pouzdanosti 95% jednakim $[399, 560]$. Dakle, možemo zaključiti da je vjerojatnost da osoba ostane u klinici barem 504 dana jednaka 50%.



Slika 2.1: Kaplan-Meierova procjena funkcije doživljenja s medijalnim vremenom doživljenja

Budući da nas zanima u kojoj je klinici vjerojatnost duljeg boravka pacijenata veća, usporedit ćemo krivulje doživljenja po klinikama koje su prikazane na slici 2.2. Na njoj uočavamo da se krivulje doživljenja poprilično razlikuju. Preciznije, krivulja doživljenja za Kliniku 1 poprima manje vrijednosti od krivulje doživljenja za Kliniku 2, u svakom danu t. To znači da je vjerojatnost da osoba boravi t dana u Klinici 2 veća nego u Klinici 1, za svaki fiksni dan t. Dakle, grafički prikaz sugerira da postoji razlika u krivuljama doživljenja između klinika, no tu tvrdnju valja provjeriti statističkim testom. Prikladan test



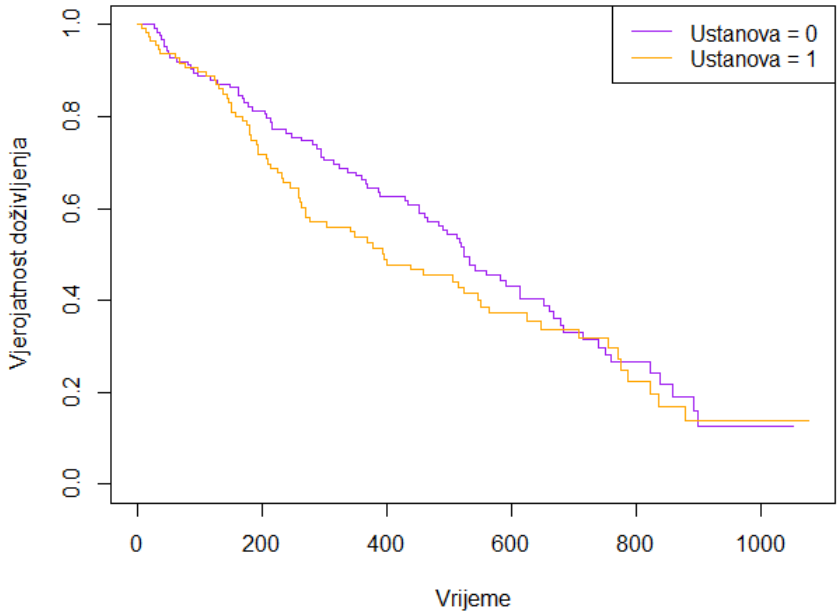
Slika 2.2: Usporedba funkcija doživljenja po klinikama

za testiranje hipoteze

$$H_0 : \text{ne postoji razlika u krivuljama doživljenja među klinikama}$$

je test log-rangova, a pripadna testna statistika je jednaka 27.9, s pripadnom p-vrijednosti 10^{-7}. Prema tome, na svim standardnim razinama značajnosti (1%,5% i 10%) odbacujemo hipotezu H_0 u korist alternativne. Stoga zaključujemo da postoji statistički značajna razlika među krivuljama doživljenja klinika. Zato ima smisla u nastavku promatrati omjer rizika osoba iz navedene dvije klinike, kako bismo kvantificirali tu razliku.

Osim toga, mogli bismo se pitati ima li razlike u krivuljama doživljenja ukoliko pacijente podijelimo u dvije skupine, prema tome imaju li zatvorski dosje ili ne. Prikaz usporedbe pripadnih funkcija doživljenja dan je na slici 2.3. Ona nam sugerira da je prisutna razlika u funkcijama doživljenja, no pripadna testna statistika testa log-rangova jednaka je 1.3 , što rezultira pripadnom p-vrijednosti jednakom 0.3. Prema tome, ne možemo sa sigurnošću tvrditi da postoji statistički značajna razlika u krivuljama doživljenja po varijabli Ustanova. Zato u nastavku nema smisla promatrati pripadne omjere rizika ovisno o toj varijabli.



Slika 2.3: Usporedba funkcija doživljenja po varijabli Ustanova

Sada ćemo primijeniti Coxov model na podatke, uzimajući pritom u obzir varijable Status, Vrijeme, Ustanova i Doza. Dakle, nećemo uzimati u obzir informaciju o tome što su pacijenti bili liječeni u dvije različite klinike. Coxov model koji će uključivati i varijablu Klinika opisat ćemo u drugom dijelu rada.

Promatramo Coxov model

$$h(t, (\text{Ustanova}, \text{Doza})) = h_0(t) \cdot \exp(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza}).$$

Pitamo se je li utjecaj varijabli Ustanova i Doza na vjerojatnost boravka u klinici statistički značajan te kolike su pripadne procjene koeficijenata β_1 i β_2. Navedene procjene, njihove standardne greške i rezultati testova značajnosti parametara dani su u tablici 2.3.

Prema dobivenim rezultatima, procijenjeni Coxov model je dan sa

$$h(t, (\text{Ustanova}, \text{Doza})) = h_0(t) \cdot \exp(0.189745 \cdot \text{Ustanova} - 0.036087 \cdot \text{Doza}).$$

Dakle, procjena parametra β_1 jednaka je 0.189745, no pripadna p-vrijednost je veća od svake razumne razine značajnosti, na temelju čega zaključujemo da varijabla Ustanova nije

Varijabla	Koeficijent	$e^{\text{Koeficijent}}$	Standardna greška koeficijenta	Z statistika	P-vrijednost
Ustanova	0.189745	1.208941	0.164274	1.155	0.248
Doza	-0.036087	0.964557	0.006001	-6.013	$1.82 \cdot 10^{-9}$

Tablica 2.3: Rezultati prilagodbe Coxovog modela

statistički značajna. Nadalje, parametar $\beta_2 = -0.036087$ pokazuje se statistički značajnim pa možemo zaključiti da promjena vrijednosti varijable Doza utječe na vjerojatnost doživljenja osobe. Preciznije, zbog negativnog predznaka navedenog koeficijenta zaključujemo da se vjerojatnost odlaska iz klinike povećava ukoliko smanjujemo dozu metadona.

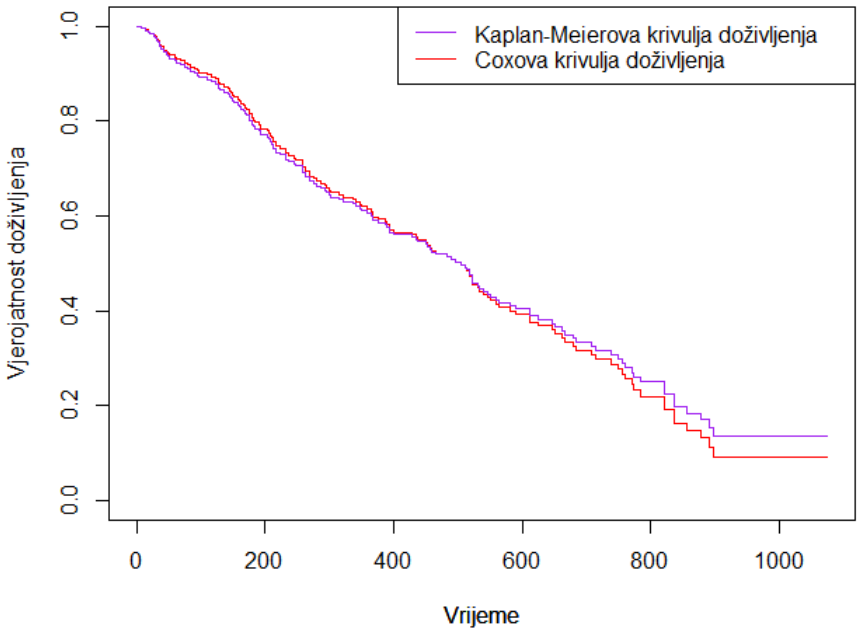
Na temelju dobivenog koeficijenta, pomoću formule (2.15) možemo usporediti vjerojatnosti odlazaka iz klinike za dvije osobe s konkretno zadanim vrijednostima kovarijata Doza i Ustanova. Konkretno, ako želimo procijeniti koliko se puta povećala ili smanjila vjerojatnost odlaska iz klinike ukoliko se osobe razlikuju za 1 miligram u dozi metadona, te nemaju zatvorski dosje, računamo

$$\frac{h(t, (\text{Ustanova} = 0, \text{Doza=y+1}))}{h(t, (\text{Ustanova} = 0, \text{Doza=y}))} = \frac{\exp\left(0.189745 \cdot 0 - 0.036087 \cdot (y+1)\right)}{\exp\left(0.189745 \cdot 0 - 0.036087 \cdot y\right)} =$$

$$= \exp\left(-0.036087 \cdot (y + 1 - y)\right) = \exp\left(-0.036087\right) = 0.96456,$$

pri čemu je $y \in \langle 0, +\infty \rangle$ proizvoljna vrijednost varijable Doza. Uočimo da je dobivena vrijednost upravo vrijednost dobivena u izlaznoj tablici u stupcu $e^{\text{Koeficijent}}$. Prema tome, zaključujemo da osoba koja ima veću dozu metadona za 1 miligram, ima otprilike 3.5% manju vjerojatnost odlaska iz klinike nego osoba s manjom dozom. Ovaj rezultat je u skladu sa činjenicom da su osobe s malom dozom metadona puštene iz klinike.

Na kraju, na slici 2.4 prikazana je funkcija doživljenja dobivena primjenom Coxovog modela, uz korištenje programskog paketa R. Pritom je korištena Breslowljeva procjena osnovne funkcije doživljenja (2.18). Kako bismo ju usporedili sa Kaplan-Meierovom procjenom na početku, prikazujemo obje dobivene krivulje na istom grafu. Naime, primjećujemo da se krivulje doživljenja gotovo ne razlikuju, što sugerira da je prilagođeni Coxov model dobar model za opis podataka sadržanih u varijablama Ustanova i Doza.



Slika 2.4: Usporedba funkcija doživljenja

Poglavlje 3

Pretpostavka proporcionalnog hazarda

Valjana analiza podataka uključuje provjeru zadovoljenosti pretpostavki modela kojeg odabiremo. Kako bi prilagodba Coxovog modela bila prikladna, potrebno je provjeriti jesu li ispunjene sljedeće pretpostavke:

1. Linearna veza logaritma hazarda i kovarijata u modelu;
2. Proporcionalnost funkcija hazarda između dvije osobe (ili grupe ljudi) za fiksno vrijeme. Drugim riječima, omjer funkcija hazarda mora biti konstantan kroz vrijeme.

U ovom poglavlju opisat ćemo drugu pretpostavku jer ona predstavlja temeljnu pretpostavku Coxovog modela. Osvrnemo li se na prethodna poglavlja, prirodno slijede dva načina za provjeru pretpostavke proporcionalnog hazarda. Prvi od njih uključuje analizu funkcije omjera hazarda svake dvije osobe kao funkciju vremena. Ukoliko bi se pokazalo da ovako konstruirana funkcija nije konstantna, mogli bismo zaključiti da je narušena pretpostavka proporcionalnog hazarda. Drugi način uključuje grafičku usporedbu funkcija hazarda svake dvije osobe. U slučaju presjeka funkcija, zaključili bismo da vrijedi neproporcionalnost hazarda. Međutim, navedena dva pristupa zahtijevaju procjenu osnovne funkcije hazarda za svaku osobu, što izbjegavamo jer Coxov model odabiremo upravo kako bismo izbjegli procjenu funkcije h_0. Zato se generalno koriste tri pristupa za testiranje navedene pretpostavke. Najčešće korišten je grafički pristup koji uključuje analizu log-log krivulja doživljenja. Drugi se temelji na statističkom testu koji je poznat pod nazivom *zph* test, dok treći uključuje dodavanje vremenski ovisnih kovarijata u model.

U prethodnom poglavlju, kod prilagodbe Coxovog modela na podatke, nismo provjeravali navedenu pretpostavku. Sada ćemo pokazati da je prilagodba bila opravdana, a osim toga, ilustrirat ćemo da varijabla Klinika, koju smo izostavili iz modela, narušava upravo tu pretpostavku. Jedan od mogućih pristupa rješavanju ovakvog problema, u kojem neka varijabla narušava pretpostavku proporcionalnog hazarda, opisat ćemo u Poglavlju 4.

3.1 Log-log krivulje

Najpoznatiji način provjere pretpostavke proporcionalnog hazarda je analiza log-log krivulja. Metoda uključuje usporedbu krivulja čija je formula dana sa

$$\ln(-\ln\left(\widehat{S}(t)\right),$$

i koje nazivamo **log-log krivuljama**. Dakle, log-log krivulje su u principu transformacije procijenjenih funkcija doživljenja $\widehat{S}(t)$.

Ilustrirajmo zašto su upravo log-log krivulje praktičan alat za detekciju neispunjavanja pretpostavke proporcionalnog hazarda. U (2.17) dana je formula procijenjene funkcije doživljenja

$$\widehat{S}(t, \mathbf{X}) = \left(\widehat{S}_0(t)\right)^{\exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i\right)}. \tag{3.1}$$

Kako je $\widehat{S}_0(t) \in [0, 1]$, logaritmiranjem jednakosti (3.1) dobivamo

$$\ln\left(\widehat{S}(t, \mathbf{X})\right) = \ln\left(\widehat{S}_0(t)\right) \cdot \exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i\right) \in \langle -\infty, 0]\,. \tag{3.2}$$

Budući da je funkcija ln definirana na intervalu $\langle 0, +\infty\rangle$, kako bismo mogli logaritmirati jednadžbu (3.2), najprije ju množimo sa -1, pa zatim logaritmiranjem dobivamo

$$\begin{aligned} \ln(-\ln\left(\widehat{S}(t, \mathbf{X})\right)) &= \ln\left(-\ln\left(\widehat{S}_0(t)\right)\right) + \ln\left(\exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i\right)\right) = \\ &= \ln\left(-\ln\left(\widehat{S}_0(t)\right)\right) + \sum_{i=1}^{p} \beta_i \cdot X_i. \end{aligned} \tag{3.3}$$

Pretpostavimo sada da želimo usporediti dvije log-log krivulje

$$l_1 := \ln\left(-\ln\left(\widehat{S}\left(t, \mathbf{X}^1\right)\right)\right) \tag{3.4}$$

i

$$l_2 := \ln\left(-\ln\left(\widehat{S}\left(t, \mathbf{X}^2\right)\right)\right), \tag{3.5}$$

pri čemu je $\mathbf{X}^1 = (X_1^1, \ldots, X_p^1)$ vektor kovarijata osobe iz prve kategorije, a $\mathbf{X}^2 = (X_1^2, \ldots, X_p^2)$ vektor kovarijata osobe iz druge kategorije. Oduzimanjem (3.4) i (3.5) te korištenjem (3.3) dobivamo

$$l_1 - l_2 = \cancel{\ln\left(-\ln\left(\widehat{S}_0(t)\right)\right)} - \cancel{\ln\left(-\ln\left(\widehat{S}_0(t)\right)\right)} + \sum_{i=1}^{p} \beta_i \cdot \left(X_i^1 - X_i^2\right). \tag{3.6}$$

Uočimo da u (3.6) pretpostavljamo da su osnovne funkcije doživljenja $\widehat{S}_0(t)$ za obje osobe jednake, što znači da su im funkcije osnovnog hazarda također jednake. Jednadžbu (3.6) možemo ekvivalentno zapisati u obliku

$$l_1 = l_2 + \sum_{i=1}^{p} \beta_i \cdot \left(X_i^1 - X_i^2\right). \tag{3.7}$$

Rezultat (3.7) nam kaže da se log-log krivulje razlikuju za vrijednost

$$\sum_{i=1}^{p} \beta_i \cdot \left(X_i^1 - X_i^2\right) \in \mathbb{R},$$

koja ne ovisi o vremenu t. Stoga možemo zaključiti da pretpostavka proporcionalnog hazarda povlači činjenicu da se log-log krivulje razlikuju za konstantu. Prema tome, paralelnost log-log krivulja ukazuje na zadovoljenost pretpostavke proporcionalnog hazarda, dok neparalelnost ukazuje na njeno narušavanje.

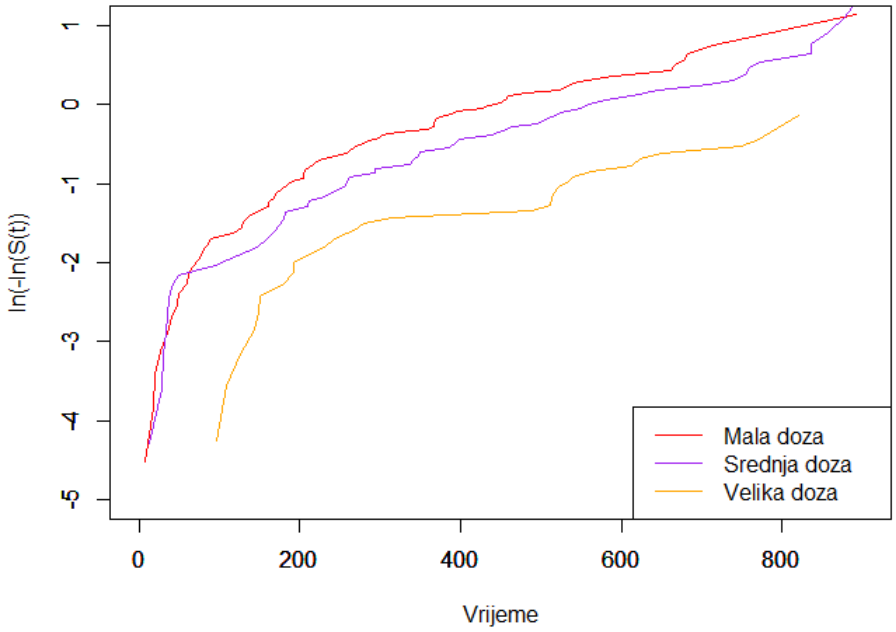
Sada ćemo ovom metodom provjeriti jesmo li opravdano u prethodnom poglavlju prilagodili Coxov model. Pritom će dobivene log-log krivulje uključivati one kovarijate koje zadovoljavaju pretpostavku proporcionalnog hazarda, a neće uključivati kovarijatu čije ispunjenje te pretpostavke ispitujemo. Kako bismo provjerili da neka kovarijata X_i zadovoljava pretpostavku proporcionalnog hazarda, potrebno je usporediti pripadne log-log krivulje po kategorijama iste. Najprije ispitujemo zadovoljava li varijabla Doza navedenu pretpostavku. Budući da je varijabla Doza neprekidna varijabla, grafička usporedba krivulja za svaku vrijednost varijable Doza bila bi vrlo nepregledna. Zato je u slučaju neprekidnih varijabli čest slučaj kategorizacija vrijednosti na više kategorija. U našem slučaju vrijednosti varijable Doza dijelimo na tri kategorije tako da u svakoj kategoriji imamo približno jednak broj podataka. Promatrat ćemo kategorije prikazane u tablici 3.1.

Velika doza	Srednja doza	Mala doza
$[20, 55]$	$\langle 55, 65]$	$\langle 65, 110]$

Tablica 3.1: Kategorije varijable Doza

Pripadne log-log krivulje po navedenim kategorijama vidimo na slici 3.1. Dobivene krivulje su približno paralelne, što nam ukazuje na činjenicu da varijabla Doza zadovoljava pretpostavku proporcionalnog hazarda.

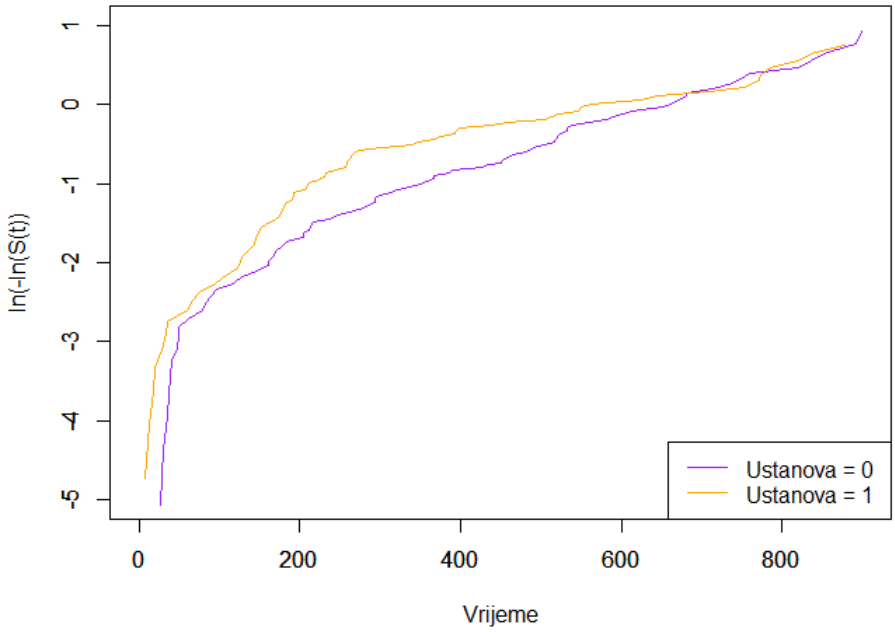
Nadalje, na slici 3.2 vidimo usporedbu log-log krivulja po kategorijama varijable Ustanova, pri čemu je u procjenama krivulja uključena varijabla Doza jer ona zadovoljava pretpostavku proporcionalnog hazarda. Na grafičkom prikazu vidimo da su dobivene log-log krivulje ugrubo paralelne, sa prisutnim manjim područjima presjeka. Budući da nemamo



Slika 3.1: Usporedba log-log krivulja po kategorijama varijable Doza

očitu neparalelnost, možemo reći da varijabla Ustanova zadovoljava pretpostavku proporcionalnog hazarda. Međutim, bilo bi poželjno ispitati tu pretpostavku statističkim testom kako bismo potvrdili zaključak.

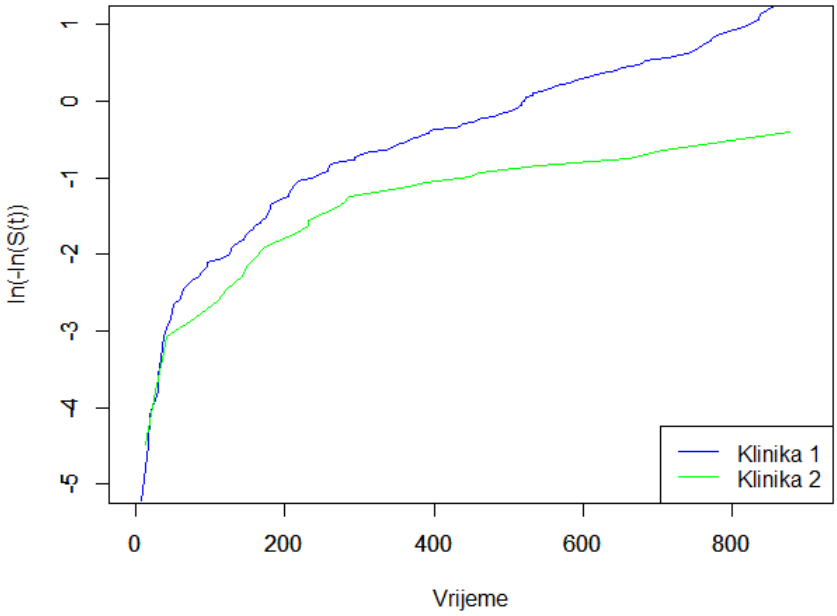
Usporedimo li log-log krivulje po kategorijama varijable Klinika, pri čemu u procjeni pripadnih funkcija doživljenja sudjeluju varijable Doza i Ustanova, dobivamo dvije neparalelne krivulje, koje su prikazane na slici 3.3. Grafički prikaz nam sugerira da pretpostavka proporcionalnog hazarda nije zadovoljena za varijablu Klinika jer razlika funkcijskih vrijednosti sve više raste kroz vrijeme. Dakle, možemo zaključiti da omjer hazarda između klinika nije konstantan kroz vrijeme. Zato bi nam uključivanje varijable Klinika u Coxov model dalo nepouzdane rezultate o omjeru hazarda. Posljedično, prilagodbom Coxovog modela ne možemo dati odgovor na pitanje koliko je puta rizik odlaska osobe iz klinike 1 veći ili manji u odnosu na kliniku 2.



Slika 3.2: Usporedba log-log krivulja po kategorijama varijable Ustanova

Napomenimo da se log-log krivulje mogu odrediti i na način da funkciju doživljenja procjenjujemo Kaplan-Meierovom metodom. U tom slučaju za svaku varijablu posebno promatramo log-log krivulje po kategorijama te varijable kako bismo ispitali zadovoljavaju li pretpostavku proporcionalnog hazarda, ne uzimajući pritom u obzir ostale varijable.

Prirodno pitanje koje se nameće je pitanje granice između paralelnosti i neparalelnosti dobivenih krivulja. Naime, log-log krivulje su u praksi uvijek približno paralelne krivulje, a problem se javlja u subjektivnom zaključivanju o paralelnosti. Stoga je neformalni dogovor da se krivulje smatraju paralelnima dokle god ne postoji čvrst dokaz da krivulje nisu paralelne. Ipak, poželjno je potvrditi zaključke nekim statističkim testom.



Slika 3.3: Usporedba log-log krivulja po kategorijama varijable Klinika

3.2 Schoenfeldovi reziduali i *zph* test

Kako se zaključak o zadovoljenosti pretpostavke proporcionalnog hazarda ne bi temeljio samo na grafičkoj analizi, poželjno je navedenu pretpostavku ispitati statističkim testom koji daje objektivniji zaključak. Jedan od najpoznatijih testova za provjeru te pretpostavke bazira se na rezidualima koje je predložio D. Schoenfeld, po kojem su i dobili naziv. Ukoliko promatramo Coxov model sa p kovarijata $X_1, \ldots, X_p$, za i-tu osobu te kovarijatu X_j definira se **Schoenfeldov rezidual** sa

$$r_j^i := \begin{cases} \frac{\partial}{\partial \beta_j} \ln L_i\left(\widehat{\boldsymbol{\beta}}\right) & \text{, ako se događaj dogodio za osobu } i \\ 0, & \text{ako je podatak o događaju za osobu } i \text{ cenzuriran.} \end{cases}$$

Odredimo čemu je jednak izraz

$$\frac{\partial}{\partial \beta_j} \ln L_i\left(\widehat{\boldsymbol{\beta}}\right).$$

Iz (2.10) imamo:

$$\ln L_i(\boldsymbol{\beta}) = \ln\left(\frac{\exp\left(\boldsymbol{\beta}' \cdot \mathbf{x}^i\right)}{\sum_{r\in R(t_i)} \exp\left(\boldsymbol{\beta}' \cdot \mathbf{x}^r\right)}\right) = \boldsymbol{\beta}' \cdot \mathbf{x}^i - \ln\left(\sum_{r\in R(t_i)} \exp\left(\boldsymbol{\beta}' \cdot \mathbf{x}^r\right)\right), \tag{3.8}$$

pa parcijanim deriviranjem jednadžbe (3.8) po varijabli β_j dobivamo

$$\frac{\partial}{\partial \beta_j} \ln L_i(\boldsymbol{\beta}) = x_j^i - \frac{\sum_{r\in R(t_i)} x_j^r \cdot \exp\left(\boldsymbol{\beta}' \cdot \mathbf{x}^r\right)}{\sum_{r\in R(t_i)} \exp\left(\boldsymbol{\beta}' \cdot \mathbf{x}^r\right)}.$$

Stoga je Schoenfeldov rezidual koji pripada osobi i te kovarijati X_j dan sa

$$r_j^i = \begin{cases} x_j^i - \frac{\sum_{r\in R(t_i)} x_j^r \cdot \exp\left(\widehat{\boldsymbol{\beta}}' \cdot \mathbf{x}^r\right)}{\sum_{r\in R(t_i)} \exp\left(\widehat{\boldsymbol{\beta}}' \cdot \mathbf{x}^r\right)}, & \text{ako se događaj dogodio za osobu } i \\ 0, & \text{ako je podatak o događaju za osobu } i \text{ cenzuriran.} \end{cases} \tag{3.9}$$

Dakle, u modelu sa n osoba i p kovarijata definiramo ukupno $n \cdot p$ Schoenfeldovih reziduala.

Schoenfeldovi reziduali su koristan alat za provjeru pretpostavke proporcionalnog hazarda nakon prilagodbe Coxovog modela na podacima. Grambsch i Therneau (vidi [1]) pokazuju da vrijedi

$$\mathbb{E}\left(r_j^{i*}\right) \approx \beta_j(t_i) - \widehat{\beta}_j, \tag{3.10}$$

pri čemu je r_j^{i*} skalirani Schoenfeldov rezidual za osobu i te kovarijatu X_j, a $\beta_j(t_i)$ predstavlja parametar koji odgovara kovarijati X_j te koji ovisi o vremenu događaja t_i za i-tu osobu. Pritom je skalirani Schoenfeldovi rezidual r_j^{i*} jednak j-toj komponenti vektora Schoenfeldovih reziduala $\mathbf{r}^{i*} = \left(r_1^{i*}, \ldots, r_p^{i*}\right)$ koji je definiran sa

$$\mathbf{r}^{i*} := J \cdot \mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}\right) \cdot r^i,$$

gdje J kao i ranije označava broj osoba za koje se dogodio događaj, $r^i = \left(r_1^i, \ldots, r_p^i\right)$ je vektor neskaliranih Schoenfeldovih reziduala, dok $\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}\right)$ označava kovarijacijsku matricu.

Rezultat (3.10) daje nam ideju kako grafički provjeriti je li zadovoljena pretpostavka proporcionalnog hazarda za kovarijatu X_j. Naime, ona nam sugerira da ukoliko prikažemo skalirane Schoenfeldove reziduale r_j^{i*} u odnosu na vremena događaja t_i, za $i = 1, \ldots, n$, morali bismo dobiti informaciju o funkcijskoj formi vremensko-ovisnog koeficijenta $\beta_j(t_i)$ kovarijate X_j. Kako bismo lakše grafički interpretirali rezultate, osim odnosa skaliranih Schoenfeldovih reziduala i vremena događaja, crtamo regresijsku krivulju skaliranih Schoenfeldovih reziduala. Ukoliko je dobivena krivulja približno jednaka horizontalnom pravcu $y = 0$, možemo zaključiti da je zadovoljena pretpostavka proporcionalnog hazarda. Osim

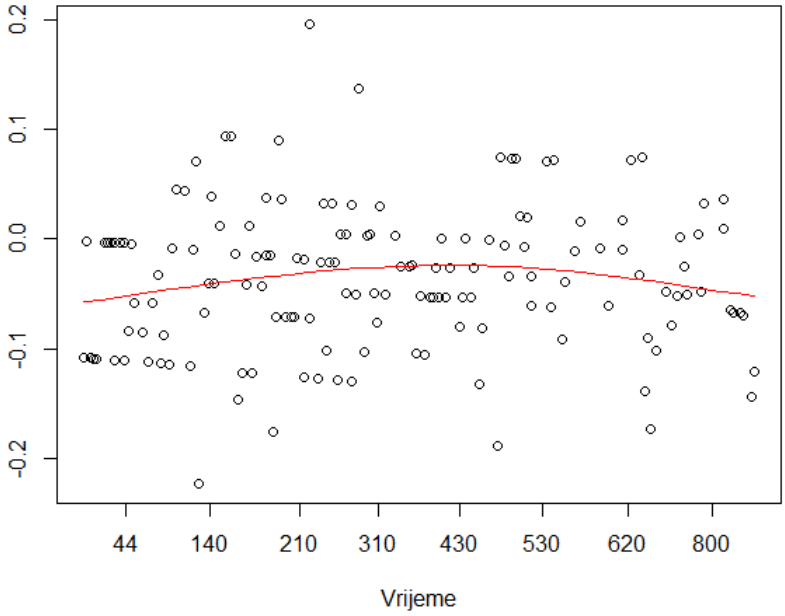
toga, Grambsch i Therneau pokazali su (vidi [1]) da se na temelju Schoenfeldovih reziduala može konstruirati statistički test kojim se provjerava pretpostavka proporcionalnog hazarda za svaku kovarijatu posebno. Ideja statističkog testa leži u činjenici da ukoliko je pretpostavka proporcionalnog hazarda zadovoljena za kovarijatu X_j, da tada ne postoji linearna veza između $\mathbb{E}\left(r_j^{i*}\right)$ i vremena t. Navedeni test naziva se **Grambsch-Therneau test proporcionalnog hazarda** ili ***zph* test**. Osim toga, konstruirali su test kojim se testira zadovoljavalju li sve kovarijate u modelu istovremeno pretpostavku proporcionalnog hazarda što je ujedno i jedan od pokazatelja da je odabir Coxovog modela razuman. Dakle, na temelju p-vrijednosti tog testa, lako možemo zaključiti ima li smisla prilagođavati Coxov model sa odabranim kovarijatama. Takav test često se naziva **globalni test**.

Analizom Schoenfeldovih reziduala i *zph* testom provjerit ćemo zadovoljavaju li varijable Ustanova, Doza i Klinika pretpostavku proporcionalnog hazarda. Ukoliko ustanovimo da neka varijabla zadovoljava pretpostavku, dalje ćemo je koristiti u modelu kojeg promatramo. Provjerimo najprije zadovoljava li variabla Doza pretpostavku proporcionalnog hazarda. *Zph* test i globalni test rezultiraju s p-vrijednostima koje su dane u tablici 3.2. P-vrijednost *zph* testa jednaka je 0.62, što ukazuje da varijabla Doza zadovoljava pretpos-

Model s varijablom Doza	
	P-vrijednost
Doza	0.62
Globalni test	0.62

Tablica 3.2: Rezultati testova

tavku proporcionalnog hazarda. Budući da promatramo sada samo model s varijablom Doza, i globalni test rezultira istom p-vrijednosti što zapravo znači da je opravdano prilagoditi Coxov model sa varijablom Doza na podatke. Nadalje, na slici 3.4 vidimo graf Schoenfeldovih reziduala u ovisnosti o vremenu. Budući da je crvena krivulja približno jednaka horizontalnom pravcu $y = 0$, možemo zaključiti da varijabla Doza doista zadovoljava pretpostavku proporcionalnog hazarda. Ovaj zaključak u skladu je sa prethodnim zaključkom dobivenim na temelju analize log-log krivulja.



Slika 3.4: Schoenfeldovi reziduali za model s varijablom Doza

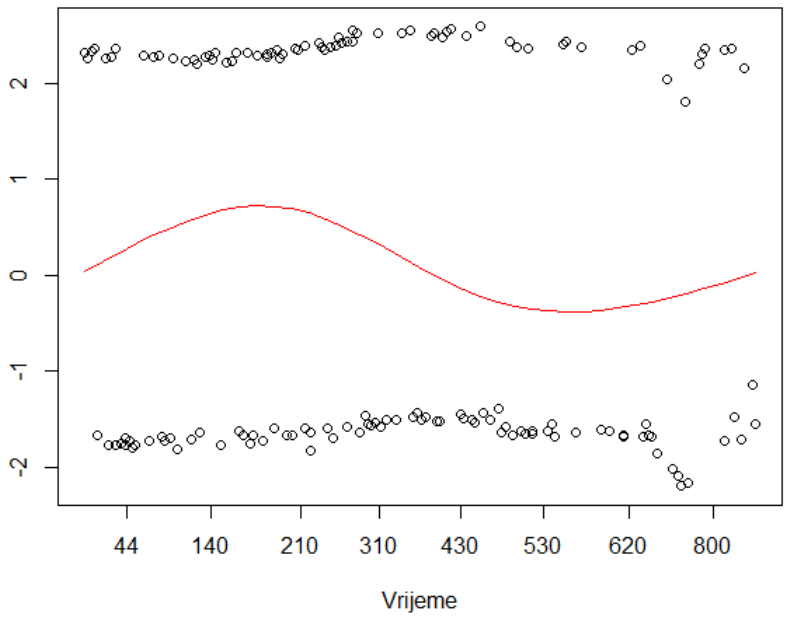
Koristeći činjenicu da varijabla Doza zadovoljava pretpostavku proporcionalnog hazarda, provjerimo zadovoljava li ju i varijabla Ustanova. Provodeći testove, dobivamo p-vrijednosti dane u tablici 3.3. Dobivena je p-vrijednost jednaka 0.095, pa na razini

Model s varijablama Doza i Ustanova	
	P-vrijednost
Ustanova	0.095
Globalni test	0.203

Tablica 3.3: Rezultati testova

značajnosti od 5% nećemo odbaciti nultu hipotezu o nekoreliranosti Schoenfeldovih reziduala i vremena. Prema tome, možemo zaključiti da varijabla Ustanova zadovoljava pretpostavku proporcionalnog hazarda. Na slici 3.5 možemo vidjeti prikaz Schoenfeldovih

reziduala za varijablu Ustanova. Na njoj crvena krivulja ne prati pravac $y = 0$ što ukazuje na nezadovoljavanje pretpostavke proporcionalnog hazarda. Međutim, kako nam grafički prikaz ne daje siguran zaključak, uzimajući u obzir rezultat testa, možemo zaključiti da varijabla Ustanova zadovoljava pretpostavku proporcionalnog hazarda. Globalni test sa p-vrijednošću 0.203 daje nam zaključak da je Coxov model koji uključuje varijable Ustanova i Doza dobar.



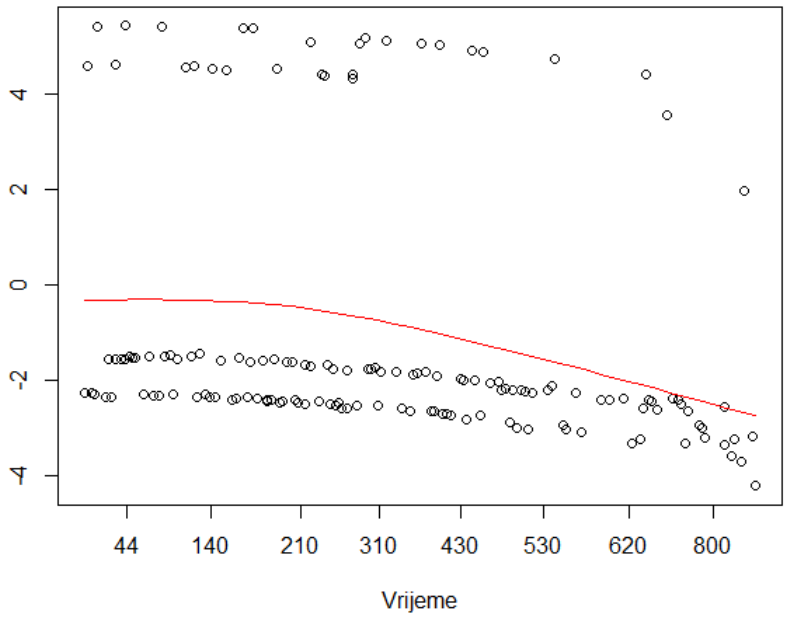
Slika 3.5: Schoenfeldovi reziduali za model s varijablama Doza i Ustanova

Provjerimo još zadovoljava li varijabla Klinika navedenu pretpostavku, pri čemu ćemo koristiti činjenicu da ju varijable Ustanova i Doza zadovoljavaju. Prikaz p-vrijednosti testova dan je u tablici 3.4.

Naime, dobivena p-vrijednost jednaka 0.00079 ukazuje na činjenicu da pretpostavka proporcionalnog hazarda nije zadovoljena za varijablu Klinika. Osim toga, globalni test daje zaključak da prilagodba Coxovog modela s varijablama Doza, Ustanova i Klinika nije opravdana. Na slici 3.6 skaliranih Schoenfeldovih reziduala možemo primijetiti trend pada crvene krivulje, što nam sugerira da Schoenfeldovi reziduali nisu nezavisni sa vremenom.

Model s varijablama Doza, Ustanova i Klinika	
	P-vrijednost
Klinika	0.00079
Globalni test	0.00616

Tablica 3.4: Rezultati testova



Slika 3.6: Schoenfeldovi reziduali za model s varijablama Doza, Ustanova i Klinika

Na kraju, možemo primijetiti da ovaj pristup provjere pretpostavke proporcionalnog hazarda daje iste zaključke kao i kod provjere putem log-log krivulja. Prema tome, zaključak o proporcionalnosti hazarda za varijable Ustanova i Doza je razuman, dok za varijablu Klinika zaključujemo da narušava pretpostavku proporcionalnog hazarda.

Osim putem navedena dva pristupa, zadovoljenost pretpostavke proporcionalnog hazarda možemo provjeriti uključivanjem vremenski-ovisne kovarijate u model. Kako bismo opisali takav pristup, najprije moramo uvesti pojam proširenog Coxovog modela.

Poglavlje 4

Prošireni Coxov model

4.1 Osnovna formula

Česta pojava vremensko-ovisnih kovarijata u praksi potiče nas na definiranje modela koji će ih uključivati. U kontekstu Coxovog modela (2.3) nameće se njegovo proširenje uvođenjem vremensko-ovisnih kovarijata. Zato definiramo **prošireni Coxov model** sa

$$h(t, \mathbf{X}(t)) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p_1} \beta_i \cdot X_i + \sum_{j=1}^{p_2} \gamma_j \cdot X_j(t)\right), \tag{4.1}$$

pri čemu $\mathbf{X}(t) = (X_1, \ldots, X_{p_1}, X_1(t), \ldots, X_{p_2}(t))'$ predstavlja vektor kovarijata. Pritom je očito da su $X_1, \ldots, X_{p_1}$ vremenski nezavisne, a $X_1(t), \ldots, X_{p_2}(t)$ vremenski zavisne varijable. Općenito, prošireni Coxov model može uključivati i varijable sa vremenskim odmakom, odnosno varijable čiji utjecaj na hazard u trenutku t može ovisiti o vremenima iz skupa $\{\ldots, t-1, t, t+1, \ldots\}$, a ne nužno samo o vremenu t. U tom slučaju, model koji uključuje i takve varijable dan je sa

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p_1} \beta_i \cdot X_i + \sum_{j=1}^{p_2} \gamma_j \cdot X_j\left(t - N_j\right)\right), \tag{4.2}$$

pri čemu su $N_1, \ldots, N_{p_2} \in \mathbb{N}_0$. Model (4.2) je alternativni način zapisa proširenog Coxovog modela koji, dakle, dopušta kovarijate sa vremenskim odmakom. Međutim, u ovom radu ograničit ćemo se na prošireni Coxov model (4.1).

Uvođenje vremensko-ovisnih kovarijata u osnovni Coxov model (2.3) očigledno rezultira drugačijom formulom za omjer rizika. Konkretno, procjena omjera rizika za dva subjekta, s vektorom kovarijata $\mathbf{X}^1 = (X_1^1, \ldots, X_{p_1}^1, X_1^1(t), \ldots, X_{p_2}^1(t))'$ za prvu osobu te vek-

torom kovarijata $\mathbf{X}^2 = (X_1^2, \ldots, X_{p_1}^2, X_1^2(t), \ldots, X_{p_2}^2(t))'$ za drugu osobu, dana je sa

$$\widehat{OR} = \frac{\widehat{h_1}\left(t|\mathbf{X}^1\right)}{\widehat{h_2}\left(t|\mathbf{X}^2\right)} = \exp\left(\sum_{i=1}^{p_1}\widehat{\beta_i}\cdot\left(X_i^1 - X_i^2\right) + \sum_{j=1}^{p_2}\widehat{\gamma_j}\cdot\left(X_j^1(t) - X_j^2(t)\right)\right). \tag{4.3}$$

U formuli (4.3) očito je vidljivo da omjer rizika ovisi o vremenu, odnosno da nije konstantna vrijednost, kao u slučaju modela (2.3). Zato prilikom prilagodbe proširenog Coxovog modela moramo uzeti u obzir činjenicu da narušavamo pretpostavku proporcionalnog hazarda. Pored toga, s obzirom da je omjer rizika funkcija vremena, procjena omjera rizika jednom vrijednošću ne daje nam neki vjerodostojan zaključak. Smislenu informaciju o omjeru rizika može nam dati upravo procjena koeficijenata $\gamma_1, \ldots, \gamma_{p_2}$, u smislu da $\gamma_i > 0$, za $i \in \{1, \ldots, p_2\}$, daje zaključak o povećanju omjera rizika kroz vrijeme, dok $\gamma_i < 0$ sugerira smanjenje omjera rizika kroz vrijeme.

Nadalje, uočimo da iako se vrijednost kovarijate $X_i(t)$ ($i \in \{1, \ldots, p_2\}$) mijenja kroz vrijeme, pripadni parametar γ_i ne ovisi o vremenu, nego je fiksan. Drugim riječima, utjecaj varijable $X_i(t)$ na vjerojatnost doživljenja procijenit ćemo samo jednim parametrom γ_i, sa kojim ćemo dati generalni zaključak o njenom utjecaju na vjerojatnost doživljenja, pritom uzimajući u obzir sva vremena u kojima su izmjerene vrijednosti te varijable.

Osim omjera rizika, i funkcija vjerodostojnosti je drugačijeg oblika nego kod osnovnog Coxovog modela. Doduše, ideja za konstrukciju funkcije vjerodostojnosti je analogna kao i kod osnovnog modela, a razliku očigledno čine vremensko-ovisne kovarijate. Preciznije, uz oznake kao i ranije, funkcija vjerodostojnosti proširenog Coxovog modela dana je sa

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{J}\frac{\exp(\boldsymbol{\beta}'\cdot\mathbf{x}^i + \boldsymbol{\gamma}'\cdot\mathbf{x}^i(t))}{\sum_{r\in R(t_i)}\exp(\boldsymbol{\beta}'\cdot\mathbf{x}^r + \boldsymbol{\gamma}'\cdot\mathbf{x}^r(t))}, \tag{4.4}$$

pri čemu je $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_{p_1})$ i $\boldsymbol{\gamma}' = (\gamma_1, \ldots, \gamma_{p_2})$. Iz formule (4.4) jasno je da funkcija vjerodostojnosti ovisi o vremenu, što nije bio slučaj kod osnovnog Coxovog modela.

4.2 Provjera pretpostavke proporcionalnog hazarda

Kao što je već spomenuto, pretpostavku proporcionalnog hazarda možemo provjeriti uključivanjem vremensko-ovisnih kovarijata u model. Ideja ovog pristupa je proširivanje Coxovog modela (2.3) dodavanjem produkta kovarijate X_i sa nekom funkcijom vremena $g_i(t)$, za sve $i = 1, \ldots, p$. Time dobivamo prošireni Coxov model oblika

$$h\left(t, \mathbf{X}(t)\right) = h_0(t)\cdot\exp\left(\sum_{i=1}^{p}\beta_i\cdot X_i + \sum_{i=1}^{p}\gamma_i\cdot(X_i\cdot g_i(t))\right). \tag{4.5}$$

Kako bismo donijeli zaključke o zadovoljenosti pretpostavke proporcionalnog hazarda, uobičajeno testiramo sljedeće dvije hipoteze statističkim testovima:

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0^1 \tag{4.6}$$

i

$$H_0 : \gamma_L = 0, \tag{4.7}$$

za $L \in \{1, \ldots, p\}$. Testiranjem hipoteze (4.6) dobivamo informaciju o tome je li u Coxovom modelu koji uključuje kovarijate $X_1, \ldots, X_p$ narušena pretpostavka proporcionalnog hazarda. Preciznije, odbacivanje hipoteze H_0 daje nam zaključak da postoji kovarijata X_i, $i \in \{1, \ldots, p\}$ za koju vrijedi $\gamma_i \neq 0$, odnosno, koja ne zadovoljava pretpostavku proporcionalnog hazarda. Zaključak o odbacivanju H_0 donosimo provođenjem **testa omjera vjerodostojnosti** ili **LR testa** (engl. *Likelihood Ratio Test*) s pripadnom testnom statistikom

$$LR := -2 \ln L_R - (-2 \ln L_P) \overset{H_0}{\sim} \chi^2(p), \tag{4.8}$$

pri čemu L_P označava funkciju vjerodostojnosti proširenog Coxovog modela (4.5), dok L_R označava funkciju vjerodostojnosti osnovnog (u ovom kontekstu još kažemo i reduciranog) Coxovog modela (2.3). Uočimo da nam test omjera vjerodostojnosti ne daje informaciju o tome koja varijabla narušava pretpostavku porporcionalnog hazarda. Stoga, kako bismo ju doznali, testiramo hipotezu (4.7). Naime, ukoliko nas zanima zadovoljava li varijabla X_L za neki $L \in \{1, \ldots, p\}$ pretpostavku proporcionalnog hazarda, u modelu (4.5) stavljamo

$$g_i(t) = \begin{cases} g_L(t), & \text{za } i = L \\ 0, & \text{za } i \neq L, \end{cases}$$

čime dobivamo model

$$h(t, \mathbf{X}(t)) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i \cdot X_i + \gamma_L \cdot (X_L \cdot g_L(t)) \right). \tag{4.9}$$

Zaključak o zadovoljenju pretpostavke proporcionalnog hazarda za kovarijatu X_L donosimo provođenjem ranije opisanog Waldovog testa s nultom hipotezom (4.7). Ukoliko odbacimo hipotezu H_0, zaključujemo da kovarijata X_L ne zadovoljava pretpostavku proporcionalnog hazarda jer je u tom slučaju parametar γ_L statistički značajno različit od nule. Pritom je važno napomenuti da je za različit odabir oblika funkcije $g_L(t)$ moguće dobiti različit zaključak o odbacivanju H_0.

Ukoliko kovarijata X_L ne zadovoljava pretpostavku proporcionalnog hazarda, znamo da omjer rizika nije konstantan kroz vrijeme, zbog čega ga nema smisla procjenjivati jednom vrijednošću, kao što je to bio slučaj kod osnovnog Coxovog modela. Zato se postavlja pitanje procjene omjera rizika u slučaju kada kovarijata X_L narušava pretpostavku

[1] Uočimo da ukoliko vrijedi H_0, model (4.5) je reduciran na osnovni Coxov model (2.3).

proporcionalnog hazarda. Odgovor leži u particioniranju vremenskog intervala kojeg promatramo na q ekvidistantnih podintervala, te zatim u procjeni omjera rizika na svakom od njih posebno. U tu svrhu definiramo **odskočne ili Heavisideove funkcije** (*engl. Heaviside Function*) $g_1, \ldots, g_q$ koje su za $k \in \{1, \ldots q\}$ oblika

$$g_k(t) = \begin{cases} 1, & \text{za } t \in [t_{k-1}, t_k\rangle \\ 0, & \text{inače,} \end{cases} \tag{4.10}$$

za vremena $t_0 = 0, t_1, t_2, \ldots, t_q \in \mathbb{R}$. Uvođenjem ovako definiranih funkcija u osnovni Coxov model (2.3) dobivamo prošireni Coxov model oblika

$$h(t, \mathbf{X}(t)) = h_0(t) \cdot \exp\left(\sum_{\substack{i=1 \\ i \neq L}}^{p} \beta_i \cdot X_i + \left(\gamma_1 \cdot X_L(t) \cdot g_1(t) + \ldots + \gamma_q \cdot X_L(t) \cdot g_q(t) \right) \right).$$

Preciznije, za $t \in [t_{k-1}, t_k\rangle$ dobivamo model oblika

$$h(t, \mathbf{X}(t)) = h_0(t) \cdot \exp\left(\sum_{\substack{i=1 \\ i \neq L}}^{p} \beta_i \cdot X_i + \gamma_k \cdot X_L(t) \right). \tag{4.11}$$

S obzirom da se omjer rizika mijenja kroz vrijeme, ovaj pristup daje nam mogućnost njegove procjene na svakom od intervala $[t_{k-1}, t_k\rangle$, za $k = 1, \ldots, q$, što u konačnici rezultira preciznijim zaključcima o omjerima rizika. Prema tome, dobivamo q različitih omjera rizika koji sljedećeg oblika

$$\widehat{OR}_{t \in [t_{k-1}, t_k\rangle} := \frac{\widehat{h}_1(t \mid t \in [t_{k-1}, t_k\rangle)}{\widehat{h}_2(t \mid t \in [t_{k-1}, t_k\rangle)},$$

pri čemu su $\widehat{h}_1$ i $\widehat{h}_2$ funkcije hazarda oblika (4.11), a koje pripadaju dvama subjektima.

4.3 Primjena na podacima

Koristeći prethodno opisanu teorijsku pozadinu, najprije želimo provjeriti je li zadovoljena pretpostavka proporcionalnog hazarda u modelu

$$h(t) = h_0(t) \cdot \exp\left(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \beta_3 \cdot \text{Klinika}\right). \tag{4.12}$$

Zato proširujemo model (4.12) do modela oblika (4.5), pri čemu stavljamo $g_1(t) = g_2(t) = g_3(t) = t$. Time dobivamo model

$$h(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \beta_3 \cdot \text{Klinika} + \gamma_1 \cdot \text{Ustanova} \cdot t + \gamma_2 \cdot \text{Doza} \cdot t + \gamma_3 \cdot \text{Klinika} \cdot t),$$

pa testom omjera vjerodostojnosti testiramo hipotezu

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

Realizacija testne statistike jednaka je

$$LR = -2 \ln L_R - (-2 \ln L_P) = -2 \cdot (-673.4024) - (-2 \cdot (-667.1910)) = 12.42283,$$

a osim toga, znamo da vrijedi

$$LR \overset{H_0}{\sim} \chi^2(3),$$

pa je pripadna p-vrijednost jednaka 0.006, zbog čega odbacujemo H_0 na svim standardnim razinama značajnosti. Stoga možemo zaključiti da je u modelu (4.12) narušena pretpostavka proporcionalnog hazarda.

Kako bismo identificirali koje varijable narušavaju tu pretpostavku, promatramo sljedeća tri modela s ciljem testiranja zadovoljenja pretpostavke proporcionalnog hazarda za varijable Ustanova, Doza i Klinika, redom.

$$h(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \beta_3 \cdot \text{Klinika}) + \gamma \cdot \text{Ustanova} \cdot t, \quad (4.13)$$

$$h(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \beta_3 \cdot \text{Klinika}) + \gamma \cdot \text{Doza} \cdot t, \quad (4.14)$$

$$h(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \beta_3 \cdot \text{Klinika}) + \gamma \cdot \text{Klinika} \cdot t. \quad (4.15)$$

Sada u svakom od modela (4.13), (4.14) i (4.15) Waldovim testom testiramo hipotezu

$$H_0 : \gamma = 0.$$

Dobivene p-vrijednosti testova prikazane su u tablici 4.1 koja slijedi.

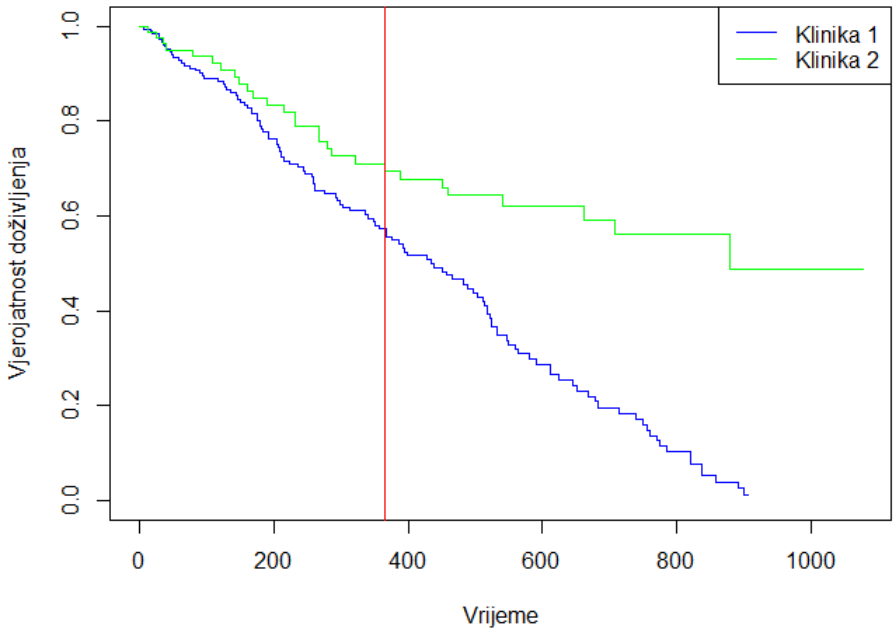
Model	Varijabla	Koeficijent γ	P-vrijednost
(4.13)	Ustanova $\cdot t$	-0.0006536	0.3214
(4.14)	Doza $\cdot t$	$1.647 \cdot 10^{-5}$	0.485536
(4.15)	Klinika $\cdot t$	-0.0030207	0.000551

Tablica 4.1: Hipoteze i pripadne p-vrijednosti Waldovog testa

Na temelju tablice 4.1 zaključujemo da varijabla Klinika narušava pretpostavku proporcionalnog hazarda, dok ju varijable Ustanova i Doza ne narušavaju, što je u skladu s prethodnim zaključcima. Uočimo da je pripadni koeficijent γ za varijablu Klinika $\cdot t$ negativan, što nam sugerira da se omjer rizika između klinika ekponencijalno smanjuje kroz vrijeme.

Gornja analiza daje nam zaključak da, ukoliko polazimo od Coxovog modela koji uključuje varijable Ustanova i Doza, dodavanjem varijable Klinika u taj model narušavamo

pretpostavku proporcionalnog hazarda. Osim toga, zaključili smo da omjer rizika između klinika ovisi o vremenu. Zato nam je cilj prilagoditi prošireni Coxov model uz korištenje odskočnih funkcija, kako bismo dali pouzdanije procjene omjera rizika. U našem primjeru, odskočne funkcije bit će određene grafičkom analizom funkcija doživljenja po kategorijama varijable Klinika. Navedeni grafički prikaz dan je na slici 4.1.



Slika 4.1: Funkcije doživljenja po klinikama

Prikazane krivulje doživljenja po klinikama sugeriraju da je vjerojatnost zadržavanja pacijenata u klinici 2 malo veća nego u klinici 1 u prvih godinu dana, dok se nakon tog razdoblja vjerojatnost zadržavanja u klinikama sve više razlikuje s vremenom. Prema tome, možemo zaključiti da će se omjer rizika klinike 2 u odnosu na kliniku 1 smanjivati kroz vrijeme, što nam potvrđuje već doneseni zaključak da omjer rizika po klinikama ovisi o vremenu.

Kako bismo dali procjenu omjera rizika za razdoblje prve godine dana, a zatim za razdoblje nakon godine dana, prilagodit ćemo prošireni Coxov model na podatke uvođenjem odskočnih funkcija na način koji slijedi. Promatramo model

$$h(t) = h_0(t) \cdot \exp\left(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \gamma_1 \cdot \text{Klinika} \cdot g_1(t) + \gamma_2 \cdot \text{Klinika} \cdot g_2(t)\right), \tag{4.16}$$

pri čemu su

$$g_1(t) = \begin{cases} 1, & \text{ako je } t < 365 \\ 0, & \text{ako je } t \geq 365 \end{cases}$$

$$g_2(t) = \begin{cases} 1, & \text{ako je } t \geq 365 \\ 0, & \text{ako je } t < 365. \end{cases}$$

Uočimo da je sada Coxov model (4.16) za $t < 365$ jednak

$$h(t) = h_0(t) \cdot \exp\left(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \gamma_1 \cdot \text{Klinika}\right),$$

dok je za $t \geq 365$ jednak

$$h(t) = h_0(t) \cdot \exp\left(\beta_1 \cdot \text{Ustanova} + \beta_2 \cdot \text{Doza} + \gamma_2 \cdot \text{Klinika}\right).$$

Prema tome, koeficijent γ_1 dat će nam informaciju o omjeru rizika klinika za prvih godinu dana, dok će nam γ_2 dati informaciju o omjeru rizika za vremensko razdoblje nakon godine dana.

Kako bismo odgovorili na pitanje koliko je puta vjerojatnost odlaska iz klinike 2 veća u odnosu na kliniku 1, trebamo procijeniti omjer hazarda

$$\frac{h(t, (\text{Ustanova}, \text{Doza}, \text{Klinika} = 2))}{h(t, (\text{Ustanova}, \text{Doza}, \text{Klinika} = 1))},$$

pri čemu varijable Ustanova i Doza imaju konstantne vrijednosti. Prema tome, vrijedi

$$\begin{aligned} \widehat{OR}_{t<365} &= \frac{h(t, (\text{Ustanova}, \text{Doza}, \text{Klinika} = 2) \mid t < 365)}{h(t, (\text{Ustanova}, \text{Doza}, \text{Klinika} = 1) \mid t < 365)} \\ &= \exp\left(\widehat{\gamma_1}\right) \end{aligned}$$

te

$$\begin{aligned} \widehat{OR}_{t\geq365} &= \frac{h(t, (\text{Ustanova}, \text{Doza}, \text{Klinika} = 2) \mid t \geq 365)}{h(t, (\text{Ustanova}, \text{Doza}, \text{Klinika} = 1) \mid t \geq 365)} \\ &= \exp\left(\widehat{\gamma_2}\right). \end{aligned}$$

Procjene koeficijenata γ_1 i γ_2, dane su u tablici 4.2 koja slijedi.

Varijabla	Koeficijent	$e^{\text{Koeficijent}}$	Standardna greška koeficijenta	P-vrijednost
Klinika $\cdot g_1(t)$	-0.4596	0.632	0.25529	0.0769
Klinika $\cdot g_2(t)$	-1.8282	0.1607	0.38595	$4.39 \cdot 10^{-6}$

Tablica 4.2: Rezultati prilagodbe proširenog Coxovog modela

Dakle, traženi omjeri su jednaki

$$\widehat{OR}_{t<365} = 0.632$$

$$\widehat{OR}_{t\geq 365} = 0.1607.$$

Možemo primijetiti da je omjer rizika nakon godine dana gotovo 4 puta manji nego unutar prve godine, što nam ukazuje na vremensku ovisnost omjera hazarda. Stoga, kao generalni zaključak možemo reći da je rizik odlaska iz klinike 2 manji nego iz klinike 1, no omjer rizika se smanjuje s prolaskom vremena. Osim toga, p-vrijednosti testova značajnosti koeficijenata γ_1 i γ_2 iz tablice 4.2 ukazuju da su značajno različiti od nule. To nam sugerira da je pretpostavka proporcionalnog hazarda narušena za varijablu Klinika, što potvrđuje već ranije donesene zaključke.

Dodatak A

Kod u programskom paketu R

```
##### 1. Analiza dozivljenja

# Ucitavanje podataka.

podaci<-read.table("podaci.txt")

# Zadavanje imena stupaca.

colnames(podaci)=c("id","Klinika","Status","Vrijeme","Ustanova","Doza")

# Slika 1.1: Primjer distribucije Vremena dozivljenja T.

hist(podaci$Vrijeme, main="", col="lightblue", xlab="Vrijeme",
probability = T,ylab="Relativna frekvencija",xlim=c(0,1000),
ylim=c(0,0.002))

# Slika 1.2: Kaplan-Meierova procjena funkcije dozivljenja S(t).

library(survival)
Vremena=c(2,5,5,5,10,17,22)
d=c(1,1,1,1,1,0,1)
Z=Surv(Vremena,d)
kmfit=survfit(Z~1)
plot(kmfit,conf.int=F,col="blue", xlab="Vrijeme",
ylab="Procjena funkcije dozivljenja",xaxt = "n",yaxt="n")
axis(1,at = c(0,2,5,10,17,22),labels = c(0,expression("t"[1]),
expression("t"[2]),expression("t"[3]),expression("t"[4]),
expression("t"[5])))
```

```
axis(2,at = c(0,2/7,3/7,6/7,1),labels = c(0,expression(frac(2,7)),
expression(frac(3,7)),expression(frac(6,7),1)), las=2)

##### 2.5 Primjena na podacima

# Spremanje podataka o vremenu u varijablu Y.

Y=Surv(podaci$Vrijeme,podaci$Status==1)

# Odredivanje ukupnog broja pacijenata te broja pacijenata po Klinikama.

length(podaci$id[podaci$Klinika==1]) # 163
length(podaci$id[podaci$Klinika==2]) # 75
length(podaci$id) # 238

# Slika 2.1: Kaplan-Meierova procjena funkcije dozivljenja s medijalnim
# vremenom dozivljenja.

# Odredivanje medijalnog vremena dozivljenja i intervala pouzdanosti 95%.
median(Y)

# Kaplan-Meierova procjena s medijalnim vremenom dozivljenja.
KM1=survfit(Y~1)
plot(KM1, col=c("black","orchid1","orchid1"), xlab="Vrijeme",
ylab="Vjerojatnost␣dozivljenja",lty=c("solid","solid","solid"))
abline(v=504,col="red3",lty=2)
abline(h=0.5, col="red3",lty=2)
axis(side=1,at=504, las=1, col="darkred",lty=2)
axis(side=2,at=0.5, las=3, col="darkred",lty=2)
legend("topright",c("Procjena␣S(t)","95%␣pouzdani␣interval"),
col=c("black","orchid1"),lty=c("solid","solid"),text.font=1)

# Slika 2.2: Usporedba funkcija dozivljenja po klinikama.

KM2=survfit(Y~podaci$Klinika)
plot(KM2, col=c("blue","green"), xlab="Vrijeme",
ylab="Vjerojatnost␣dozivljenja",main="")
legend("topright", c("Klinika␣1", "Klinika␣2"),col=c("blue","green"),
lty=c("solid","solid"))

# Test log-rangova.
survdiff(Surv(podaci$Vrijeme,podaci$Status)~podaci$Klinika)
```

```
# Slika 2.3: Usporedba funkcija dozivljenja po varijabli Ustanova.

KM3=survfit(Y~podaci$Ustanova)
plot(KM3, col=c("purple","orange"), xlab="Vrijeme",
ylab="Vjerojatnost␣dozivljenja")
legend("topright", c("Ustanova␣=␣0", "Ustanova␣=␣1"),
col=c("purple","orange"), lty=c("solid","solid"))

# Test log-rangova.
survdiff(Surv(podaci$Vrijeme,podaci$Status)~podaci$Ustanova)

# Tablica 2.3: Rezultati prilagodbe Coxovog modela.

MODEL1=coxph(Y~Ustanova+Doza,data=podaci,method="breslow")
MODEL1

# Slika 2.4: Usporedba funkcija dozivljenja.

plot(survfit(MODEL1),conf.int=F,col="red", xlab="Vrijeme",
ylab="Vjerojatnost␣dozivljenja")
par(new=T)
plot(KM1, col="purple", xlab="Vrijeme",axes=F,conf.int=F)
legend("topright",c("Kaplan-Meierova␣krivulja␣dozivljenja",
"Coxova␣krivulja␣dozivljenja"),col=c("purple","red"),
lty=c("solid","solid"))
par(new=F)

##### 3.1: Log-log krivulje

# Slika 3.1: Usporedba log-log krivulja po kategorijama varijable Doza.

# Mala doza.
A=podaci$Doza>=20&podaci$Doza<=55
FIT1=survfit(coxph(Y[A]~1),method="breslow")
Vrijeme_A=summary(FIT1)$time
Procjena_A=summary(FIT1)$surv
Procjena_A=log(-log(Procjena_A))
plot(Vrijeme_A,Procjena_A,col="red",type="l",xlim=c(0,900),
xlab="Vrijeme", ylab="ln(-ln(S(t))",ylim=c(-5,1))

# Srednja doza.
```

```
par(new=T)
B=podaci$Doza>55&podaci$Doza<=65
FIT2=survfit(coxph(Y[B]~1),method="breslow")
Vrijeme_B=summary(FIT2)$time
Procjena_B=summary(FIT2)$surv
Procjena_B=log(-log(Procjena_B))
plot(Vrijeme_B,Procjena_B,col="purple",type="l",xlim=c(0,900),xlab="",
ylab="",ylim=c(-5,1))

# Velika doza.
par(new=T)
C=podaci$Doza>65&podaci$Doza<=110
FIT3=survfit(coxph(Y[C]~1),method="breslow")
Vrijeme_C=summary(FIT3)$time
Procjena_C=summary(FIT3)$surv
Procjena_C=log(-log(Procjena_C))
plot(Vrijeme_C,Procjena_C,col="orange",type="l",xlim=c(0,900),xlab="",
ylab="",ylim=c(-5,1))
legend("bottomright",c("Mala␣Doza","Srednja␣Doza","Velika␣Doza"),
col=c("red","purple","orange"),lty=c("solid","solid","solid"), cex=1)
par(new=F)

# Slika 3.2: Usporedba log-log krivulja po kategorijama varijable
# Ustanova.

# Ustanova=0
D=podaci$Ustanova==0
FIT4=survfit(coxph(Y[D]~podaci$Doza[D]),method="breslow")
Vrijeme_D=summary(FIT4)$time
Procjena_D=summary(FIT4)$surv
Procjena_D=log(-log(Procjena_D))
plot(Vrijeme_D,Procjena_D,col="purple",type="l",xlim=c(0,900),
xlab="Vrijeme",ylab="ln(-ln(S(t))",ylim=c(-5,1))

# Ustanova=1
par(new=T)
E=podaci$Ustanova==1
FIT5=survfit(coxph(Y[E]~podaci$Doza[E]),method="breslow")
Vrijeme_E=summary(FIT5)$time
Procjena_E=summary(FIT5)$surv
Procjena_E=log(-log(Procjena_E))
plot(Vrijeme_E,Procjena_E,col="orange",type="l",xlim=c(0,900),xlab="",
```

```
ylab="",ylim=c(-5,1))
legend("bottomright",c("Ustanova = 0", "Ustanova = 1"),
col=c("purple","orange"),lty=c("solid","solid"))
par(new=F)

# Slika 3.3: Usporedba log-log krivulja po kategorijama varijable
# Klinika.

# Klinika 1
G=podaci$Klinika==1
FIT6=survfit(coxph(Y[G]~podaci$Doza[G]+podaci$Ustanova[G]),
method="breslow")
Vrijeme_G=summary(FIT6)$time
Procjena_G=summary(FIT6)$surv
Procjena_G=log(-log(Procjena_G))
plot(Vrijeme_G,Procjena_G,col="blue",type="l",xlim=c(0,900),
xlab="Vrijeme",ylab="ln(-ln(S(t))",ylim=c(-5,1))

# Klinika 2
par(new=T)
H=podaci$Klinika==2
FIT7=survfit(coxph(Y[H]~podaci$Doza[H]+podaci$Ustanova[H]),
method="breslow")
Vrijeme_H=summary(FIT7)$time
Procjena_H=summary(FIT7)$surv
Procjena_H=log(-log(Procjena_H))
plot(Vrijeme_H,Procjena_H,col="green",type="l",xlim=c(0,900),xlab="",
ylab="",ylim=c(-5,1))
legend("bottomright",c("Klinika 1","Klinika 2"),col=c("blue","green"),
lty=c("solid","solid"))

##### 3.2: Schoenfeldovi reziduali i zph test

# Tablica 3.2: Rezultati testova.

MODEL2=coxph(Y~Doza,data=podaci,method="breslow")
cox.zph(MODEL2,transform=rank)

# Slika 3.4: Schoenfeldovi reziduali za model s varijablom Doza.

plot(cox.zph(MODEL2,transform=rank), var='Doza',col="red",se=F,
xlab="Vrijeme", ylab="")
```

```
# Tablica 3.3: Rezultati testova.

MODEL3=coxph(Y~Doza+Ustanova,data=podaci,method="breslow")
cox.zph(MODEL3,transform=rank)

# Slika 3.5: Schoenfeldovi reziduali za model s varijablama
# Doza i Ustanova.

plot(cox.zph(MODEL3,transform=rank), var='Ustanova',col="red",se=F,
xlab="Vrijeme",ylab="") #Ustanova

# Tablica 3.4: Rezultati testova.

MODEL4=coxph(Y~Ustanova+Doza+Klinika,data=podaci,method="breslow")
cox.zph(MODEL4,transform=rank)

# Slika 3.5: Schoenfeldovi reziduali za model s varijablama
# Doza i Ustanova.

plot(cox.zph( MODEL4,transform=rank),var='Klinika', se=F,col="red",
xlab="Vrijeme",ylab="")

##### 4.3: Primjena na podacima

# Priprema podataka.

podaci.cp=survSplit(podaci,cut=podaci$Vrijeme[podaci$Status==1],
end="Vrijeme",event="Status",start="start")

# Tablica 4.1: Hipoteze i pripadne p-vrijednosti Waldovog testa.

# Model (4.13)
podaci.cp$t_Ustanova=podaci.cp$Ustanova*podaci.cp$Vrijeme
coxph(Surv(podaci.cp$start,podaci.cp$Vrijeme,podaci.cp$Status)~Ustanova
+Doza+Klinika+t_Ustanova+cluster(id),data=podaci.cp,method="breslow")

# Model (4.14)
podaci.cp$t_Doza=podaci.cp$Doza*podaci.cp$Vrijeme
coxph(Surv(podaci.cp$start,podaci.cp$Vrijeme,podaci.cp$Status)~Ustanova
+Doza+Klinika+t_Doza+cluster(id),data=podaci.cp,method="breslow")
```

```
# Model (4.15)
podaci.cp$t_Klinika=podaci.cp$Klinika*podaci.cp$Vrijeme
coxph(Surv(podaci.cp$start,podaci.cp$Vrijeme,podaci.cp$Status)~Ustanova
+Doza+Klinika+t_Klinika+cluster(id),data=podaci.cp,method="breslow")

# Test omjera vjerodostojnosti.

lnL_R=coxph(Y~Ustanova+Doza+Klinika,data=podaci,method="breslow")$loglik
lnL_P=coxph(Surv(podaci.cp$start,podaci.cp$Vrijeme,podaci.cp$Status)
~Ustanova+Doza+Klinika+t_Klinika+t_Doza+t_Ustanova+cluster(id),
data=podaci.cp, method="breslow")$loglik
LR=-2*lnL_R+2*lnL_P
p_vrijednost=1-pchisq(LR,3)
p_vrijednost

# Slika 4.1: Funkcije dozivljenja po klinikama.

MODEL5=coxph(Y~Ustanova+Doza+strata(Klinika),data=podaci,
method="breslow")
plot(survfit(MODEL5),conf.int=F,lty=c("solid","solid"),
col=c("blue","green"),xlab="Vrijeme",ylab="Vjerojatnost␣dozivljenja")
abline(v=365, col="red")
legend("topright",c("Klinika␣1","Klinika␣2"),lty=c("solid","solid"),
col=c("blue","green"))

# Tablica 4.2: Rezultati prilagodbe prosirenog Coxovog modela.

# Spremanje podataka.

podaci.cp365=survSplit(podaci,cut=365,end="Vrijeme",
event="Status",start="start")
Y365=Surv(podaci.cp365$start,podaci.cp365$Vrijeme,podaci.cp365$Status)

# Definiranje odskocne funkcije.

podaci.cp365$hv1=podaci.cp365$Klinika*(podaci.cp365$start<365)
podaci.cp365$hv2=podaci.cp365$Klinika*(podaci.cp365$start>=365)

# Prilagodba prosirenog Coxovog modela.

coxph(Y365~Ustanova+Doza+hv1+hv2+cluster(id), data=podaci.cp365,
method="breslow")
```

Bibliografija

[1] D. Collett, *Modelling survival data in medical research, third edition*, Chapman & Hall, New York, 2015.

[2] M. Huzak, *Vjerojatnost i matematička statistika*, http://aktuari.math.pmf.unizg.hr/docs/vms.pdf, (listopad, 2022.).

[3] D. G. Kleinbaum i M. Klein, *Survival Analysis: A Self-Learning Text*, Springer, New York, 2005.

[4] P. McCullagh i J.A. Nelder, *Generalized Linear Models, Second Edition*, Chapman & Hall, New York, 1989.

[5] M. Nikulin i H. Wu, *The Cox Model and Its Applications*, Springer, New York, 2016.

[6] N. Sarapa, *Teorija Vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[7] T. M. Therneau i P. M. Grambsch, *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York, 2000.

[8] Ministarstvo zdravstva i socijalne skrbi Republike Hrvatske, *Smjernice za farmakoterapiju opijatskih ovisnika metadonom*, https://drogeiovisnosti.gov.hr/UserDocsImages/dokumenti/Smjernice/smjernice_metadon.pdf, (listopad, 2022.).

Sažetak

U kontekstu analize doživljenja, kvantitativna analiza podataka podrazumijeva promatranje dvije fundamentalne funkcije: funkciju doživljenja i funkciju hazarda. Cilj analize doživljenja je procijeniti navedene dvije funkcije, koje nam daju sažete i bitne informacije iz podataka. Funkcija doživljenja daje nam informaciju o vjerojatnosti da se neki događaj od interesa nije dogodio do nekog fiksnog trenutka, dok funkciju hazarda uvodimo iz razloga što ona predstavlja glavni alat za matematički prikaz modela u analizi doživljenja. U ovom radu je funkcijom hazarda

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

gdje je $\mathbf{X} = (X_1, \ldots, X_p)'$ vektor kovarijata, predstavljen jedan od najpoznatijih matematičkih modela u analizi doživljenja: **Coxov model proporcionalnih hazarda**. Njegova najbitnija pretpostavka je proporcionalnost hazarda, ili ekvivalentno, vremenska nezavisnost omjera rizika. Kako bi prilagodba Coxovog modela bila valjana, bitno je provjeriti zadovoljavaju li kovarijate $X_1, \ldots, X_p$ navedenu pretpostavku, koju uobičajeno testiramo korištenjem tri pristupa. Prvi se temelji na analizi log-log krivulja, drugi objedinjuje analizu Schoenfeldovih reziduala i *zph* test, a treći počiva na uvođenju vremensko-ovisnih kovarijata u osnovni Coxov model. Ukoliko neka od kovarijata ne zadovoljava pretpostavku proporcionalnog hazarda, proširujemo osnovni Coxov model sa vremensko-ovisnim kovarijatama, čime dobivamo prošireni Coxov model

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

gdje je $\mathbf{X}(t) = (X_1, \ldots, X_{p_1}, X_1(t), \ldots, X_{p_2}(t))'$ vektor kovarijata. U principu, prošireni Coxov model koristan je kao alat za provjeru zadovoljenosti pretpostavke proporcionalnog hazarda te za procjenu omjera rizika na particiji vremenskog intervala kojeg promatramo. U ovom radu dan je detaljan prikaz Coxovog modela i metoda provjere pretpostavke proporcionalnog hazarda, a također je dan i uvid u njegovo proširenje. Kako bi se postiglo bolje razumijevanje same teorije, kroz cijeli rad proteže se problem kliničke studije modeliran Coxovim modelom.

Summary

In the context of survival analysis, quantitative data analysis considers two fundamental functions: the survivor function and the hazard function. The goal of the survival analysis is to estimate these two functions, which provide us crucial and summarized information from the data. The survivor function gives us the probability that an event of interest did not occur until a certain fixed moment, while we introduce the hazard function because it is the main tool for the mathematical representation of the model in the survival analysis. In this paper, the Cox proportional hazard model , which is one of the most famous mathematical models in the survival analysis, is represented by the hazard function

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

where $\mathbf{X} = (X_1, \ldots, X_p)'$ denotes the vector of covariates. The most important assumption of that model is the proportional hazard assumption, or equivalently, the hazard ratio independence of time. In order to achieve the valid adjustment of the Cox model, it is essential to check whether the covariates $X_1, \ldots, X_p$ satisfy the proportional hazard assumption. We usually asses it using three approaches. The first one involves comparing log-log survivor curves, the second one combines the analysis of Schoenfeld residuals and the *zph* test, and the third one is based on using time-dependent covariates in order to extend the basic Cox model. If one of the covariates does not satisfy the proportional hazard assumption, we extend the basic Cox model with time-dependent covariates, which gives us an extended Cox model

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

where $\mathbf{X}(t) = (X_1, \ldots, X_{p_1}, X_1(t), \ldots, X_{p_2}(t))'$ denotes the vector of covariates. In general, the extended Cox model is usually used for checking the proportional hazard assumption, and also for estimating the hazard ratio over the partition of the time interval. In this paper, a detailed presentation of the Cox model and its extension is given, as well as methods for checking the proportional hazard assumption. In order to achieve a better understanding of the theory itself, in this paper the medical research problem is modeled by the Cox model.

Životopis

Zovem se Eva Kolarek i rođena sam 20.07.1997. godine u Čakovcu. Nakon završene Gimnazije Josipa Slavenskog u Čakovcu s odličnim uspjehom, upisala sam preddiplomski studij Matematike, nastavničkog smjera, na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta u Zagrebu. Tijekom preddiplomskog studija bila sam demonstrator iz kolegija Računarski praktikum 1 i 2, a istog sam završila 2020. godine. Te godine nastavila sam obrazovanje na istom fakultetu, upisavši diplomski studij Matematičke statistike. Tijekom njegovog pohađanja uvidjela sam da područja mog interesa uključuju biostatistiku, strojno učenje i modeliranje rizika, a pri samom kraju studija zaposlila sam se u Addiko banci u Zagrebu, u odjelu Kontrole kreditnog rizika kao član tima kvantitativnih istraživanja.