Univerzitet u Beogradu

Matematički fakultet

Master rad

ANALIZA PREŽIVLJAVANJA

Mentor :

Prof. dr Vesna Jevremović

Student :

Jasmina Miščević

br. indeksa : 1024/2010

Beograd, septembar 2012.

Sadržaj

Uvod

Analiza preživljavanja je kolekcija statističkih procedura za analizu podataka gde je slučajna promenljiva od interesa vreme do pojave određenog događaja. Pod događajem podrazumevamo smrt, bolest, odgovor na tretman, oporavak (npr. povratak na posao) ili bilo koje iskustvo od interesa za posmatranje, koje se može dogoditi posmatranom subjektu. Pod vremenom podrazumevamo godine, mesece, nedelje ili dane od početka posmatranja subjekta do pojave događaja; na primer, vreme se može odnositi na godine individue u trenutku pojavljivanja događaja. U analizi preživljavanja, vremenska promenljiva se obično opisuje kao vreme preživljavanja, jer predstavlja vreme koje je subjekat „preživeo“ tokom nekog perioda posmatranja. Na primer, vreme preživljavanja može biti vreme bez tumora, vreme od početka tretmana pa do odgovora na isti, dužina remisije, kao i vreme do smrti. Sam događaj se uglavnom opisuje kao neuspeh, jer je događaj od interesa često smrt ili neko drugo negativno individualno iskustvo. Međutim, vreme preživljavanja može biti i vreme do povratka na posao nakon neke hiruške intervencije, tada je neuspeh pozitivan događaj.

Ova analiza se prvo razvila u medicini i biologiji, gde su predmet posmatranja živa bića, a događaj je najčešće smrt, oboljenje ili povratak neke bolesti. Podaci o preživljavanju mogu da uključe vreme preživljavanja, odgovor na dati tretman, karakteristike pacijenata bitne za taj odgovor, kao i za opstanak i razvoj bolesti. Analiza tih podataka je usmerena na predviđanje verovatnoće preživljavanja, ili prosečne dužine života, poredeći raspodele preživljavanja eksperimentalnih životinjskih ili ljudskih pacijenata i identifikovanjem rizika i/ili prognostičkih faktora vezanih za odgovor, opstanak i razvoj bolesti.

Naravno, sve metode su pogodne i za primene u ekonomiji, industriji, društvenim naukama. U slučaju kada su mašine posmatrani subjekti, tada je događaj uglavnom njihov kvar. U industriji se može posmatrati životni vek elektronskih uređaja, komponenti ili sistema. U analizi društva postoje veoma interesantni primeri, kao što su vreme „preživljavanja“ brakova, vreme do napuštanja škole ili vreme do izvršavanja zločina. U ekonomiji se može posmatrati „preživljavanje“ neke delatnosti ili vreme „preživljavanja“ nekog proizvoda; dužina pretplate za novine ili magazine (marketing); naknada štete za radnike (osiguranje), kao i rizik i prognostički faktori koji na to utiču.

Ključni analitički problem u analizi preživljavanja je takozvano cenzurisanje. U suštini, cenzurisanje se javlja kada postoji neka informacija o vremenu preživljavanja subjekta, ali ne postoji tačno vreme preživljavanja. Razmotrimo tri tipa cenzurisanja.

Cenzurisanje I tipa se javlja kada je istraživanje dizajnirano tako da se završi nakon tačnog određenog vremenskog perioda. U tom slučaju, subjekti koji nisu ostvarili događaj za vreme perioda trajanja istraživanja su cenzurisani u trenutku njegovog završetka. Ispitivanja na životinjama obično počinju sa fiksiranim brojem životinja kojima se daju određeni tretmani. Zbog vremenskih i/ili ograničenja troškova, istraživač često ne može da čeka na smrt svih

životinja. Jedna je opcija je da se istraživanje vremenski ograniči, nakon čega preživele životinje bivaju žrtvovane. Vremena preživljavanja zabeležena za životinje koje su uginule za vreme istraživačkog perioda su vremena od početka eksperimenta do njihove smrti. Ta vremena se nazivaju tačna ili necenzurisana. Vremena preživljavanja žrtvovanih životinja nisu tačno poznata, ali se vode makar kao dužina trajanja eksperimenta. Ta vremena se nazivaju cenzurisana. Neke životinje se mogu izgubiti ili slučajno uginuti. Njihova vremena preživljavanja, od početka eksperimenta do gubitka ili smrti, su takođe cenzurisana zapažanja. Kod cenzurisanja I tipa, ukoliko nema slučajnih gubitaka, sva cenzurisana zapažanja su jednaka dužini trajanja eksperimenta. Na primer, pretpostavimo da se kancerogene ćelije ubrizgavaju u 6 pacova. Posmatra se vreme razvijanja tumora određene veličine. Istraživač odlučuje da prekine eksperiment nakon 30 nedelja. Kod tri pacova se tumor razvija nakon 10, 15 i 25 nedelje, respektivno. Kod dva se tumor ne razvija do kraja studije, pa su njihova vremena preživljavanja 30+ nedelja. I jedan pacov je slučajno uginuo bez tumora nakon 19 nedelje posmatranja. Vremena preživljavanja su 10, 15, 30+, 25, 30+ i 19+ nedelja , gde + označava cenzurisane podatke.

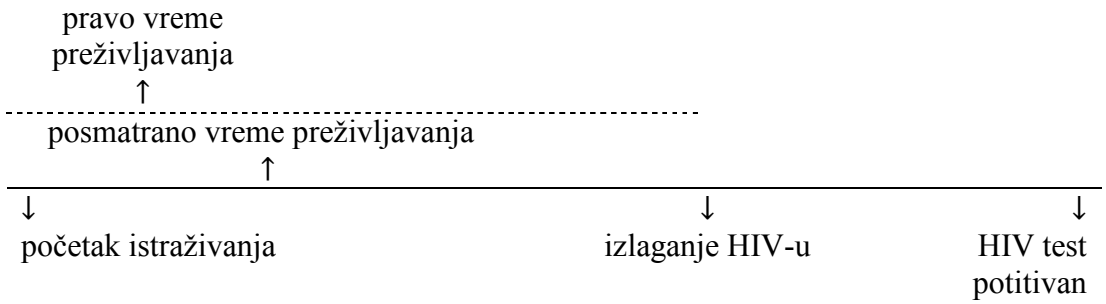
Cenzurisanje II tipa se javlja kada je istraživanje dizajnirano tako da se završi kada se ostvari prethodno određen broj događaja. Druga opcija kod ispitivanja na životinjama jeste čekati da fiksiran deo životinja ugine, nakon čega se žrtvuju preživele životinje. U slučaju cenzurisanja II tipa, ukoliko nema slučajnih gubitaka, cenzurisana vremena su jednaka najvećem necenzurisanom vremenu. Na primer, u eksperimentu sa 6 pacova istraživač može da odluči da prekida eksperiment kada se kod 4 od 6 pacova razviju tumori. Tada su vremena preživljavanja 10, 15, 35+, 25, 35 i 19+ nedelja.

Kod cenzurisanje III tipa, istraživanje je dizajnirano tako da se završi nakon tačnog određenog vremenskog perioda, ali cenzurisani subjekti nemaju svi isto vreme cenzurisanja. Kod većine kliničkih i epidemioloških studija period istraživanja je fiksiran i pacijenti se uključuju u studiju u različitim vremenima tokom tog perioda. Neki mogu umreti pre kraja studije; njihova tačna vremena preživljavanja su poznata. Drugi se mogu povući pre kraja studije i izgubiti za praćenje. Drugi, opet, mogu biti živi na kraju studije. Za „izgubljene“ pacijente vremena preživljavanja su najmanje jednaka vremenu od ulaska u studiju do poslednjeg kontakta. Za pacijente koje su i dalje živi, vremena preživljavanja su najmanje jednaka vremenu od ulaska do kraja studije. Poslednja dva zapažanja nazivamo cenzurisanim. Kako vremena ulaska u studiju nisu istovremena, i cenzurisana vremena su različita. Na primer, pretpostavimo da se šest pacijenata sa akutnom leukemijom uključuje u studiju tokom njenog istraživačkog perioda od godinu dana. Pretpostavimo, takođe, da se kod svih pacijenata javlja odgovor na terapiju i postiže remisija. Prvi, treći i šesti pacijent postiže remisiju na početku drugog, četvrtog i devetog meseca i recidiv nakon četiri, šest i tri meseca, respektivno. Drugi pacijent postiže remisiju početkom trećeg meseca, ali se izgubio četiri meseca kasnije; stoga je trajanje remisije najmanje 4 meseca. Četvrti i peti pacijent postižu remisiju na početku petog i desetog meseca, respektivno, i još uvek su u remisiji na kraju studije; njihova vremena u remisiju su stoga najmanje 8 i 3 meseca.

Odgovarajuća vremena preživljavanja (u remisiju) ovih šest pacijenata su 4, 4+, 6, 8+, 3 i 3+ meseca.

Cenzurisana zapažanja I i II tipa se često nazivaju pojedinačno cenzurisanim podacima, a tipa III postepeno cenzurisanim podacima. Čest naziv za cenzurisanje III tipa je slučajno cenzurisanje. Svi navedeni tipovi cenzurisanja su desna cenzurisanja ili cenzurisanja sa desne strane. Podaci su cenzurisani sa desne strane ako je potpun interval vremena preživljavanja, koji je u stvari nepoznat, presečen (tj. cenzurisan) sa desne strane, što se dešava kada se subjektu ne desi događaj do kraja istraživanja, ili kada je subjekat izgubljen za praćenje tokom perioda istraživanja, ili ispadne iz istraživanja zbog nekog događaja koji nije od interesa istraživanja.

Levo cenzurisanje se javlja kada je poznato da se događaj od interesa desio pre izvesnog vremena t, ali tačno vreme nastanka je nepoznato, odnosno kada je tačno vreme preživljavanja subjekta manje ili jednako od posmatranog vremena preživljavanja subjekta. Na primer, epidemiolog želi da zna starost pacijenta prilikom dijagnoze u studiji praćenja dijabetske retinopatije[1]. Tokom pregleda, kod 50-godišnjeg ispitanika je ustanovljeno da ima već razvijenu retinopatiju, ali ne postoji evidencija o tačnom vremenskom trenutku kad se bolest razvila. Tako da su godine ispitanika tokom pregleda levo cenzurisano zapažanje. To znači da je starost ovog pacijenta prilikom dijagnoze najviše 50 godina. Razmotrimo još jedan primer, u kome posmatramo osobe dok ne postanu HIV pozitivne, možemo da beležimo neuspeh kada su testovi individue pozitvni na virus. Najčešće nećemo znati vreme prvog izlaganja virusu, i samim tim, ni tačno vreme kada se neuspeh dogodio. Stoga, vreme preživljavanja je cenzurisano sa leve strane, jer je pravo vreme preživljavanja, koje se završava u trenutku izlaganja virusu, manje od posmatranog vremena, koje se završava kada su testovi subjekta pozitivni.



Intervalno cenzurisanje se javlja kada je poznato da se događaj od interesa dogodio između vremena *a* i *b*. Na primer, ako medicinska evidencija upućuje na to da pacijent iz prethodnog primera sa 45 godina nije imao retinopatiju, njegove godine prilikom dijagnoze su između 45 i 50.

Desno cenzurisanje se najčešće javlja među podacima preživljavanja, tako da ćemo u nastavku rada podrazumevati da su cenzurisani podaci cenzurisani sa desne strane.

[1] Dijabetska retinopatija je oštećenje malih krvnih sudova mrežnjače oka usled neregulisane šećerne bolesti.

Funkcije vremena preživljavanja

U analizi preživljavanja posmatramo vreme do pojave određenog događaja, i to vreme nazivamo vreme preživljavanja. Raspodelu vremena preživljavanja obično opisujemo pomoću tri funkcije: (1) funkcije preživljavanja, (2) funkcije gustine i (3) funkcije rizika (ili hazardne funkcije). Ove tri funkcije su matematički ekvivalentne, odnosno ako je jedna od njih data, druge dve mogu biti izvedene.

U praksi, ove tri funkcije se mogu koristi za ilustrovanje različitih aspekata podataka. Osnovni problem u analizi preživljavanja jeste ocenjivanje na osnovu uzorka jedne ili više od ove tri funkcije i izvođenje zaključaka o obrascu preživljavanja u populaciji.

Sa T označavamo slučajnu promenljivu koja predstavlja vreme preživljavanja subjekta. Moguće vrednosti od T su svi nenegativni brojevi ili $T \geq 0$. Raspodela od T se može opisati pomoću tri ekvivalentne funkcije.

Funkcija preživljavanja

Funkcija preživljavanja, označena sa $S(t)$, je definisana kao verovatnoća da subjekat preživi duže od nekog određenog vremena :

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

Na osnovu definicije funkcije raspodele $F(t)$ od T,

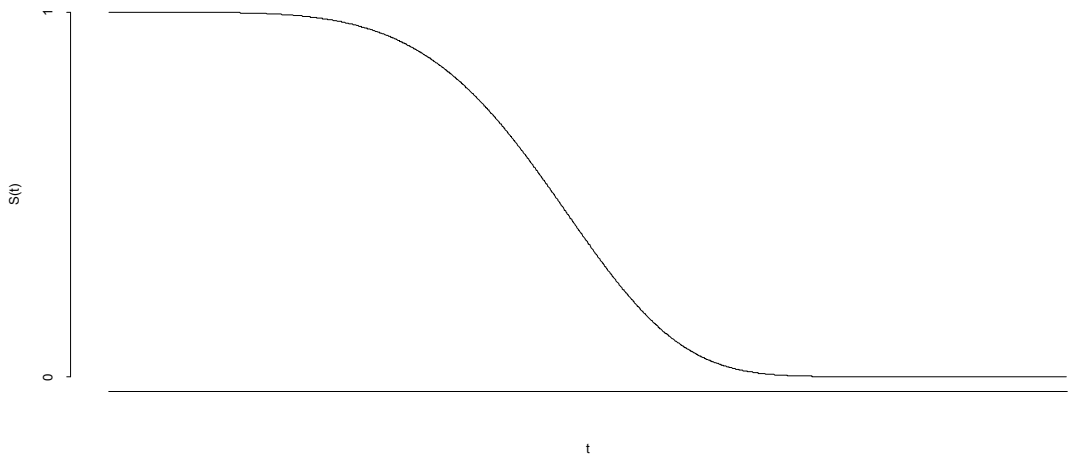
$$S(t) = 1 - P(T \leq t) = 1 - F(t) \qquad (1)$$

Teoretski gledano, $S(t)$ je nerastuća funkcija po vremenu t sa osobinama

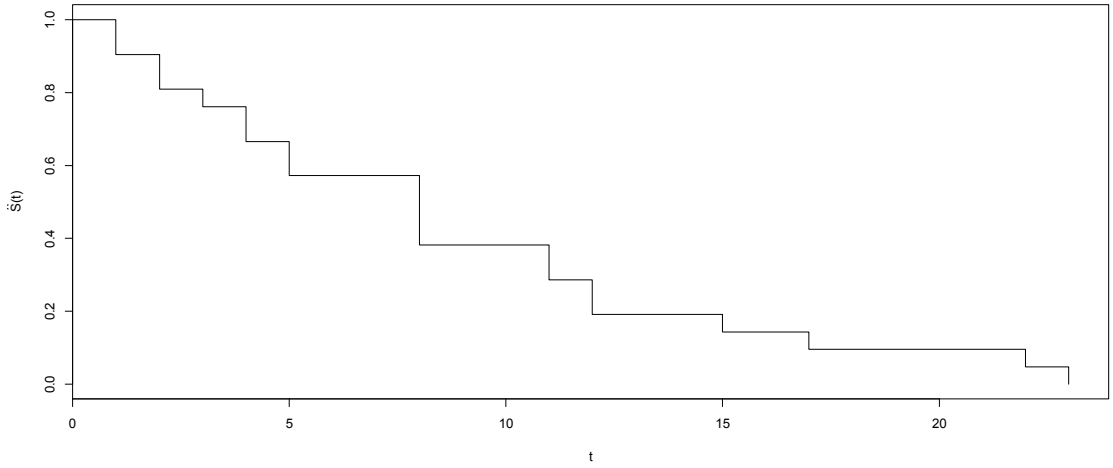
$$S(t) = \begin{cases} 1 & za\ t = 0 \\ 0 & za\ t = \infty \end{cases}$$

Odnosno, na početku istraživanja, kada još uvek nijedan subjekat ne ostvaruje događaj, verovatnoća preživljavanja vremena 0 je jedan. Takođe, ako se period trajanja istraživanja, teoretski gledano, bezgranično poveća, na kraju niko neće preživeti, pa kriva preživljavanja mora na kraju pasti na nulu.

Funkcija $S(t)$, koja direktno opisuje preživljavanje posmatrane studijske grupe, je poznata i kao kumulativna stopa preživljavanja. Grafik funkcije $S(t)$ se naziva kriva preživljavanja. Strma kriva preživljavanja ukazuje na nisku stopu preživljavanja ili kratko vreme preživljavanja. Dok postepeno opadajuća ili ravna kriva preživljavanja ukazuje na visoku stopu preživljavanja ili duže preživljavanje.



U praksi, kada se koriste stvarni podaci, obično su dobijeni grafici funkcije preživljavanja stepenaste funkcije, pre nego glatke krive. Pre svega, jer periodi istraživanja nikada nisu beskonačno dugi, pa je moguće da neki od posmatranih subjekata ne ostvare događaj.



Funkcija preživljavanja se koristi za nalaženje medijane i drugih kvantila vremena preživljavanja, kao i za poređenje raspodela preživljavanja dve ili više grupa. Srednja vrednost se obično koristi za opisivanje centralne tendencije raspodele, ali je kod raspodela preživljavanja medijana često bolja jer mali broj subjekata sa izuzetno dugim ili kratkim vremenima preživljavanja može prouzrokovati da srednja vrednost vremena preživljavanja bude nesrazmerno velika ili mala. U praksi, ukoliko ne postoje cenzurisana zapažanja, funkcija preživljavanja se ocenjuje kao proporcija subjekata čija su vremena preživljavanja veća od t :

$$\hat{S}(t) = \frac{broj\ subjekata\ koji\ su\ preživeli\ duže\ od\ t}{ukupan\ broj\ subjekata}$$

Kada postoje cenzurisana zapažanja, brojilac se ne može uvek odrediti. Na primer, posmatrajmo sledeći skup vremena preživljavanja: $\{4, 6, 6+, 10+, 15, 20\}$. Koristeći gornju jednačinu možemo izračunati $\hat{S}(5) = \frac{5}{6} = 0.833$. Međutim, ne možemo izračunati $\hat{S}(11)$ jer je tačan broj subjekata koji su preživeli duže od 11 nepoznat. I treći i četvrti subjekat su mogli da prežive duže ili kraće od 11. Stoga, kada postoje cenzurisana zapažanja, gornja jednačina nije više prikladna za ocenjivanje funkcije preživljavanja. Neparametarske metode za ocenjivanje funkcije preživljavanja za cenzurisane podatke biće objašnjene kasnije.

Funkcija gustine verovatnoće

Kao i svaka druga neprekidna slučajna promenljiva, vreme preživljavanja T ima gustinu raspodele definisanu kao limes verovatnoće da vreme preživljavanja subjekta upadne u mali interval od t do $t + \Delta t$ po jedinici širine Δt, ili jednostavnije, verovatnoće da se neuspeh desi u malom vremenskom interval po jedinici vremena. Može se izraziti na sledeći način :

$$f(t) = \frac{\lim_{\Delta t \to 0} P[da\ subjekat\ ostvari\ događaj\ u\ intervalu\ (t, t + \Delta t)]}{\Delta t}$$

Grafik funkcije $f(t)$ naziva se kriva gustine. Funkcija gistina ima sledeće dve karakteristike:

1. $f(t)$ je nenegativna funkcija:
$$f(t) \geq 0 \quad \text{za sve } t \geq 0$$
$$f(t) = 0 \quad \text{za sve } t = 0$$

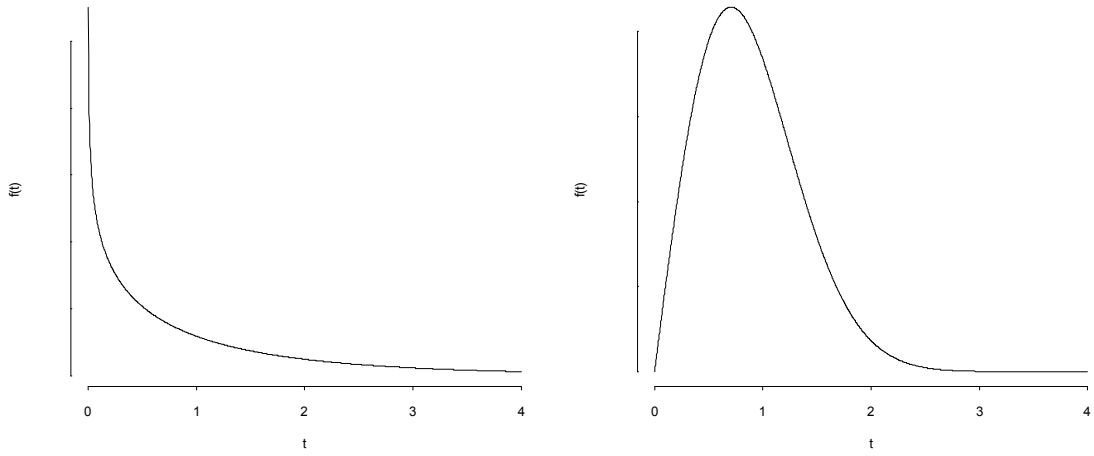
2. Površina između krive gustine i t ose jednaka je 1.

U praksi, ukoliko ne postoje cenzurisana opažanja, funkcija gustine raspodele se ocenjuje kao proporcija subjekata koji ostvaruju događaj u intervalu po jedinici vremena :

$$\hat{f}(t) = \frac{broj\ subjekata\ koji\ ostvaruju\ događaj\ u\ intervalu\ počevši\ od\ trenutka\ t}{(ukupan\ broj\ subjekata) \times (dužina\ intervala)}$$

Slično kao i kod ocenjivanja funkcije preživljavanja, kada postoje cenzurisana opažanja, gornja jednačina nije primenjiva. O pogodnijim metodama će biti više reči kasnije.

Funkcija gustina nam daje proporciju subjekata koji ostvaruju događaj u bilo kom vremenskom intervalu i vrhove visoke frekvetnosti neuspeha (ostvarivanja događaja). Gustina raspodele je takođe poznata pod nazivom bezuslovna stopa promašaja.



Funkcija rizika

Funkcija rizika $h(t)$ vremena preživljavanja T predstavlja trenutni potencijal po jedinici vremena da se događaj pojavi, ako se zna da se nije pojavio do trenutka t (tj. subjekat je preživeo do trenutka t). Ona je definisana sa :

$$h(t) = \frac{\lim_{\Delta t \to 0} P\left[\begin{matrix} subjekat\ ostvari\ događaj\ u\ intervalu\ (t, t+\Delta t) \\ pod\ uslovom\ da\ je\ preživeo\ do\ t \end{matrix}\right]}{\Delta t} \qquad (2)$$

Suprotno od funkcije preživljavanja koja se fokusira na to da se događaj ne pojavi, funkcija rizika se fokusira na neuspeh, odnosno da se događaj pojavi. Drugim rečima, kada funkcija preživljavanja raste onda funkcija rizika opada, i obrnuto. Rizik je stopa, a ne verovatnoća i funkcija rizika se ponekad naziva i uslovna stopa preživljavanja (uslovna zbog toga što je brojilac razlomka u formuli kojom je zadata funkcija rizika uslovna verovatnoća da vreme preživljavanja bude između t i $t + \Delta t$ ako je vreme preživljavanja veće ili jednako od t ; stopa jer se deli sa Δt). Funkcija rizika $h(t)$ je nenegativna funkcija i nema gornja ograničenja. Ona igra jako važnu ulogu u analizi preživljavanja, jer se, u matematičkom smislu, modeli preživljavanja često opisuju pomoću funkcije rizika.

Funkcija rizika može biti definisana i uz pomoć funkcije raspodele $F(t)$ i funkcije gustine raspodele $f(t)$:

$$h(t) = \frac{f(t)}{1 - F(t)} \qquad (3)$$

Funkcija rizika se naziva još i trenutna stopa neuspeha, snaga mortaliteta, uslovna stopa mortaliteta, i starosno-specifična stopa neuspeha. Ukoliko t u jednačini (2) predstavlja godine, to je mera sklonosti ka neuspehu kao funkcija godina subjekta u smislu da je kvantitet $\Delta t h(t)$

očekivana proporcija subjekata starosti t koji će doživeti neuspeh u malom vremenskom intervalu $t + \Delta t$. Takva funkcija rizika govori o riziku neuspeha po jedinici vremena tokom procesa starenja.

U praksi, kada ne postoje cenzurisana opažanja, funkcija rizika se ocenjuje kao proporcija subjekata koji ostvaruju događaj u intervalu po jedinici vremena, pod pretpostavkom da su ti subjekti preživeli do početka intervala:

$$\hat{h}(t) = \frac{broj\ subjekata\ koji\ ostvaruje\ događaj\ u\ intervalu\ sa\ početkom\ u\ trenutku\ t}{(broj\ subjekata\ koji\ preživljava\ do\ t) \times (dužina\ intervala)} =$$

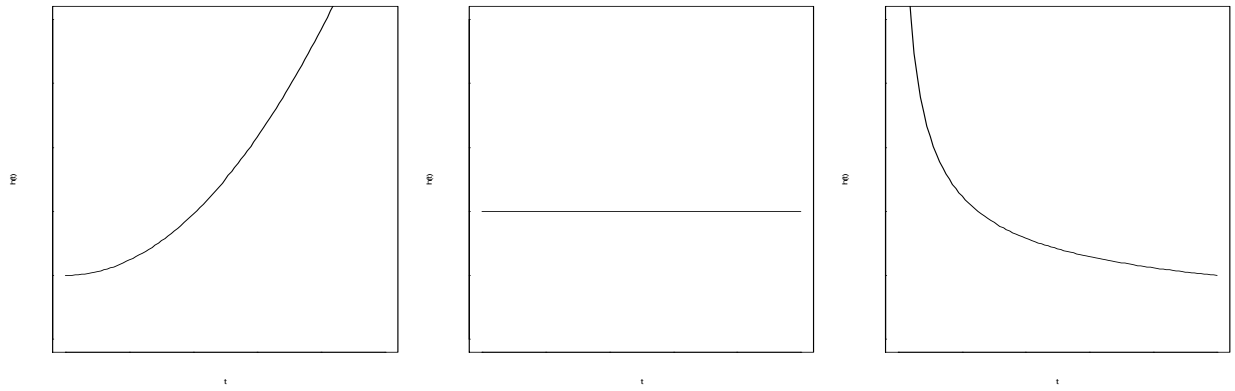
$$= \frac{broj\ subjekata\ koji\ ostvaruje\ događaj\ po\ jedinici\ vremena\ u\ intervalu}{broj\ subjekata\ koji\ preživljava\ do\ t}$$

Aktuari uglavnom koriste prosečnu stopu rizika intervala u kome je broj subjekata koji ostvaruje događaj po jedinici vremena u intervalu podeljen sa prosečnim brojem preživelih do sredine intervala:

$$\hat{h}(t) = \frac{broj\ subjekata\ koji\ ostvaruje\ događaj\ po\ jedinici\ vremena\ u\ intervalu}{(broj\ subjekata\ koji\ preživljava\ do\ t) - (broj\ smrti\ u\ intervalu)/2}$$

Ova aktuarska procena daje veću stopu rizika nego prethodna, pa je samim tim konzervativnija procena.

Funkcija rizika može rasti, opadati, biti konstantna ili ukazivati na neki složeniji proces. Na primer, pacijenti sa akutnom leukemijom koji ne reaguju na tretman imaju rastuću stopu rizika. Opadajuća funkcija rizika, na primer, ukazuje na rizik vojnika sa ranama od vatrenog oružja koji se podvrguju operaciji. Glavna opasnost je sama operacija i ona se smanjuje ako operacija prođe uspešno. Primer konstantne hazardne funkcije je rizik zdravih osoba između 18 do 40 godina starosti čije su glavne opasnosti od smrti nesreće.



Kumulativna funkcija rizika je opisana kao

$$H(t) = \int_0^t h(x)dx.$$

Biće pokazano kasnije da je

$$H(t) = -\log S(t)$$

Samim tim, za $t = 0$, $S(t) = 1$, $H(t) = 0$, i za $t = \infty$, $S(t) = 0$, $H(t) = \infty$. Kumulativna funkcija rizika uzima bilo koje vrednosti između 0 i ∞.

Odnosi funkcija preživljavanja

Tri prethodno definisane funkcije su matematički ekvivalentne.

1. Iz (1) i (3) dobijamo

$$h(t) = \frac{f(t)}{S(t)}$$

   Ovaj odnos možemo dobiti i iz (2) koristeći osnovne definicije uslovne verovatnoće.

2. Kako je funkcija gustine raspodele derivat funkcije raspodele,

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t)$$

3. Zamenom 2. u 1. dobijamo

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t)$$

4. Integraljenjem 3. od nula do t i korišćenjem $S(0) = 1$, dobijamo

$$-\int_0^t h(x)dx = \log S(t)$$

$$H(t) = -\log S(t)$$

$$S(t) = exp[-H(t)] = exp\left[-\int_0^t h(x)dx\right]$$

5. Iz 1. i 4. dobijamo

$$f(t) = h(t)exp[-H(t)]$$

Dakle, ako je $f(t)$ poznata, funkcija preživljavanja se može dobiti iz osnovnog odnosa $f(t)$, $F(t)$ i 2. Funkcija rizika se onda može odrediti na osnovu 1. Ako je $S(t)$ poznata, $f(t)$ i $h(t)$ se

mogu odrediti iz 2. i 1., respektivno, ili se prvo $h(t)$ može izvesti iz 3. i $f(t)$ onda iz 1. Ako je $h(t)$ data, $S(t)$ i $f(t)$ se mogu odrediti, respektivno, iz 4. i 5. Dakle, ako je data bilo koja od tri funkcije preživljavanja, druge dve se lako mogu izvesti.

Prikaz podataka

Razmotrimo dva tipa prikaza podataka u analizi preživljavanja.

Pretpostavimo da je dat skup podataka preživljavanja koji se sastoji od n subjekata.

Posmatrani subjekti	t	δ	x_1	...	x_p
1	t_1	δ_1	x_{11}	...	x_{p1}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	t_n	δ_n	x_{n1}	...	x_{pn}

Prva kolona predstavlja posmatrane subjekte, od 1 do n. Druga kolona daje informacije o posmatranom vremenu preživljavanja svakog subjekta, bez obzira na to da li je subjekat ostvario događaj ili je cenzurisan. Treća kolona je slučajna promenljiva $\delta = (0,1)$ koja se odnosi na neuspeh ili cenzuru. Odnosno, $\delta = 1$ za neuspeh, ako se događaj pojavi tokom perioda istraživanja, $\delta = 0$ ako je vreme preživljavanja cenzurisano. Ovu promenljivu nazivamo i statusna promenljiva. Primetimo da kada sumiramo sve δ_j u ovoj koloni dobićemo ukupan broj neuspeha subjekata. I ovaj broj će biti manji od ili jednak n, jer neki od subjekata možda neće ostvariti događaj. Ostale informacije u tabeli predstavljaju vrednosti za objašnjavajuće promenljive od interesa (npr. godine, pol, rasa...). Ako na primer želimo da ispitamo da li se vremena preživljavanja pacijenata lečenih od leukemije lekovima dve različite farmaceutske komapanije razlikuju, onda imamo samo jednu promenljivu od interesa i to je farmaceutska kompanija čije su lekove pili pacijenti, ili na primer ako nas interesuje razlika u vremenima preživljavanja između muškaraca i žena onda bi promenljiva od interesa bila pol.

Drugi (uređeni) prikaz podataka preživljavanja je podesan za razumevanje kako se sprovodi analiza preživljavanja, posebno kako se dobijaju krive preživljavanja.

$t_{(j)}$	d	q_j	$R(t_{(j)})$
$t_{(0)} = 0$	$d_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	d_1	q_1	$R(t_{(1)})$
.	.	.	.
.	.	.	.
.	.	.	.
$t_{(k)}$	d_k	q_k	$R(t_{(k)})$

Prva kolona, $t_{(j)}$, opisuje različita vremena neuspeha po redu, od najkraćeg do najdužeg. Dakle, k je broj različitih vremena u kojima su subjekti ostvarili događaj, $k \leq n$. Druga kolona, d_j, predstavlja broj neuspeha u trenutku $t_{(j)}$. Kada nema istih vremena neuspeha $d_j = 1$. U trećoj koloni, q_j, su frekvencije pojavljivanja cenzurisanih subjekata u vremenskom intervalu $[t_{(j)}, t_{(j+1)})$. Primetimo da se broje subjekti čije je vreme cenzurisanja $t_{(j)}$ i ne broje oni čije je vreme cenzurisanja $t_{(j+1)}$. Napomenimo i da nam ubačeni red na početku tabele sa odgovarajućim vremenom 0 dozvoljava mogućnost da subjekat bude cenzurisan nakon početka studije, ali pre prvog vremena neuspeha. Poslednja kolona predstavlja skup rizika $R(t_{(j)})$, koji je definisan kao kolekcija subjekata koji su preživeli najmanje do trenutka $t_{(j)}$. Odnosno, svaki subjekat u $R(t_{(j)})$ ima vreme preživljavanja $t_{(j)}$ ili duže, bez obzira na status cenzurisanja.

Značaj ovakvog prikaza podataka jeste u tome što nam omogućava da uključimo cenzurisane opservacije u analizu preživljavanja. Iako su cenzurisani podaci nepotpuni, u smislu da ne postoji tačno vreme preživljavanja, jako je bitno za analizu da iskoristimo informacije o cenzurisanom subjektu do trenutka kada gubimo trag o njemu. Neuključivanje cenzurisanih podataka prilikom analize može dovesti do ozbiljne pristrasnosti u ocenama raspodele vremena preživljavanja i njenim kvantilima.

Da bi se izračunala verovatnoća preživljavanja u datom trenutku, koristi se skup rizika kako bi se uključile informacije koje imamo i o cenzurisanim subjektima do trenutka cenzurisanja. Za izračunavanje takve verovatnoće koristimo Kaplan-Meier-ovu metodu sa kojom ćemo se upoznati u nastavku rada.

Kaplan-Meier ocene funkcije preživljavanja

Kaplan-Meier (KM) ocena funkcije preživljavanja, koja se, takođe, naziva ocena granične vrednosti proizvoda, je neparametarska metoda za ocenjivanje $S(t)$ za necenzurisane i desno cenzurisane podatke. Jako je popularna, jer zahteva samo jako slabe pretpostavke i koristi informacije iz svih raspoloživih opservacija (necenzurisanih i cenzurisanih).

Pretpostavimo da su $t_{(1)} < \cdots < t_{(k)}$ uređena vremena neuspeha i da su podaci preživljavanja prikazani gornjom tabelom. S tim što, uvodimo i oznaku n_j, koja predstavlja broj subjekata u skupu rizika $R(t_{(j)})$.

Ideja KM ocene jeste da se $\hat{S}(t)$ održava konstantnom između dva vremena neuspeha i umanjuje u vremenima neuspeha sa verovatnoćom preživljavanja u tim vremenima. U trenutku $t_{(j)}$, postoji n_j subjekata u skupu rizika i njih d_j je ostvarilo događaj. Prirodna ocena verovatnoće neuspeha u tom trenutku je d_j/n_j , a ocena verovatnoće preživljavanja je $1 - d_j/n_j$, odnosno

$$\hat{P}\left(T > t_{(j)} \middle|\, T \geq t_{(j)}\right) = \frac{n_j - d_j}{n_j}$$

Za $t_{(i)} \leq t < t_{(i+1)}$ verovatnoća preživljavanja nakon vremena t je

$$S(t) = P(T > t) = P\left(T > t, T > t_{(i)}\right) = P\left(T > t \middle| T > t_{(i)}\right) P\left(T > t_{(i)}\right) =$$
$$= P\left(T > t \middle| T > t_{(i)}\right) P\left(T > t_{(i)} \middle| T > t_{(i-1)}\right) P\left(T > t_{(i-1)}\right) =$$
$$= P\left(T > t \middle| T > t_{(i)}\right) P\left(T > t_{(i)} \middle| T > t_{(i-1)}\right) P\left(T > t_{(i-1)} \middle| T > t_{(i-2)}\right) \ldots P\left(T > t_{(1)} \middle| T > t_{(0)}\right) P\left(T > t_{(0)}\right)$$

$$S(t) \approx \prod_{t_{(i)} \leq t} P\left(T > t_{(i)} \middle| T \geq t_{(i)}\right)$$

gde je $t_{(0)} = 0$ i $t_{(k+1)} = +\infty$.

KM ocena funkcije preživljavanja u trenutku t je data sa

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

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

Šta se dešava sa $\hat{S}(t)$ ako je t veće od maksimalnog posmatranog vremena događaja, t_{max} ? Efron (1967.) je preložio da $\hat{S}(t) = 0$ za $t > t_{max}$, Gill (1980.) je predložio $\hat{S}(t) = \hat{S}(t_{max})$, i

Brown (1974.) je predložio da $\hat{S}(t) = exp\left\{\log\left(\hat{S}(t_{max})\right) t/t_{max}\right\}$. Međutim, možda je najbolja odluka ne pokušavati oceniti, jer se valjanost ocene ne može proceniti. Bolje je zaustaviti se u poslednjem posmatranom vremenu događaja.

KM ocene su ocene maksimalne verodostojnosti.

Disperzija Kaplan-Meier ocene

Kao kod svih statističkih ocena, važno je proceniti preciznost KM ocene $\hat{S}(t)$, odnosno odrediti standardnu grešku te ocene. Kako je standardna greška kvadratni koren od disperzije, problem se svodi na određivanje disperzije od KM ocene $\hat{S}(t)$.

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

Suma disperzija nezavisnih događaja se lako nalazi, ali to nije slučaj kada imamo proizvod disperzija. Da bi olakšali sebi zadatak radićemo sa logaritmom ocene funkcije preživljavanja,

$$D\left(\log \hat{S}(t)\right) = D\left(\sum_{i:\, t_{(i)} \le t} \log \hat{p}_i\right) = \sum_{i:\, t_{(i)} \le t} D(\log \hat{p}_i)$$

pod pretpostavkom da se događaji ostvaruju nezavisno među populacijom.

Uočimo da smo na ovaj način stvorili drugi problem, određivanje odnosa disperzija $D\left(\hat{S}(t)\right)$ i $D\left(\log \hat{S}(t)\right)$, i u čijem rešavanju će nam pomoći sledeća metoda.

Delta metoda

Statističari često koriste proceduru pod nazivom delta metoda za dobijanje ocene disperzije kada ocena, čiju disperziju ocenjujemo, nije prosta suma opservacija. Delta metoda koristi Tejlorov razvoj prvog reda funkcije f (koja mora biti glatka kriva da bi mogli da primenimo Tejlorov razvoj) slučajne promenljive X oko $\mu = E(X)$ za aproksimativno određivanje disperzije od $f(X)$:

$$f(X) \cong f(\mu) + f'(\mu)(X - \mu) \qquad (4)$$

gde je

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

$$D\left(f(X)\right) \cong D(f(\mu) + f'(\mu)(X - \mu)) = f'^2(\mu) D(X - \mu) = f'^2(\mu) D(X)$$

Ocena disperzije delta metodom je

$$\widehat{D}\big(f(X)\big) \cong f'^2(\hat{\mu})\hat{\sigma}^2 \qquad (5)$$

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

Na primer, posmatrajmo funkciju $\log X$. Razvoj iz (4) je

$$\log X \cong \log \mu + (X - \mu)\frac{1}{\mu}$$

Na osnovu (5), ocena disperzije delta metodom je

$$\widehat{D}(\log X) \cong \frac{1}{\hat{\mu}^2}\hat{\sigma}^2$$

Vratimo se na traženje disperzije KM ocene funkcije preživljavanja.

$$D\big(\log \hat{S}(t)\big) = \sum_{i:\, t_{(i)} \le t} D(\log \hat{p}_i)$$

gde je $\hat{p}_i = \frac{n_i - d_i}{n_i}$.

Druga pretpostavka za dobijanje disperzije ocene jeste da su opservacije u skupu rizika u trenutku $t_{(i)}$ nezavisne Bernulijeve opservacije sa konstantnom verovatnoćom p_i. Pod ovom pretpostavkom, ocena ove verovatnoće je $\hat{p}_i$ sa disperzijom $\hat{p}_i(1 - \hat{p}_i)/n_i$. Ocena disperzije $\log \hat{p}_i$ delta metodom je

$$\widehat{D}(\log \hat{p}_i) \cong \frac{1}{\hat{p}_i^2}\frac{\hat{p}_i(1 - \hat{p}_i)}{n_i} \cong \frac{d_i}{n_i(n_i - d_i)}$$

Dakle, ocena disperzije log KM ocene delta metodom je

$$\widehat{D}\big(\log \hat{S}(t)\big) \cong \sum_{i:\, t_{(i)} \le t} \widehat{D}(\log \hat{p}_i) \cong \sum_{i:\, t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

Ocena disperzije KM ocene funkcije preživljavanja se dobija ponovnom primenom delta metode. U ovoj primeni funkcija je

$$f(X) = \exp(X), \text{ odnosno } \hat{S}(t) = exp\big(\log \hat{S}(t)\big)$$

Na osnovu (4) sledi da je razvoj funkcije

$$exp(X) \cong exp(\mu) + exp(\mu)(X - \mu)$$

i iz (5) da je aproksimativna ocena disperzije

$$\widehat{D}\big(exp(X)\big) \cong [exp(\hat{\mu})]^2 \hat{\sigma}^2$$

Odatle sledi da je ocena disperzije KM ocene

$$\widehat{D}\left(\hat{S}(t)\right) \cong \left[\hat{S}(t)\right]^2 \sum_{i:\, t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

Ova ocena je poznata pod nazivom Greenwood-ova formula za ocenu disperzije KM ocene funkcije preživljavanja.

Intervali poverenja

Za velike uzorke KM ocena ima aproksimativno normalnu raspodelu[2], pa možemo izračunati krajeve $(1-\alpha)100\%$ interval poverenja za funkciju preživljavanja u trenutku t :

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

gde je $Z_{1-\alpha/2}$ $1-\alpha/2$-i kvantil standardne normalne raspodele. Međutim, problem ovog pristupa jeste da su krajevi intervala često izvan 0 i 1, odnosno $\hat{S}(t) \in [0,1]$, a $\hat{S}(t) \pm Z_{1-\alpha/2}\sqrt{\widehat{D}\left(\hat{S}(t)\right)} \in (-\infty, +\infty)$, kao i normalnost KM ocene ako obim uzorka nije veliki.

Da bi se rešili ovi problemi, Kalbfleisch i Prentice su predložili da se ocenjivanje intervala poverenja zasniva na funkciji $\log\big[-\log \hat{S}(t)\big]$. Prednost ove funkcije je u tome što ona uzima vrednosti u intervalu $(-\infty, +\infty)$. Ocena disperzije log-log KM ocene funkcije preživljavanja, koju dobijamo primenom delta metode na $X = \log \hat{S}(t)$, je

$$\widehat{D}\big(\log\big[-\log \hat{S}(t)\big]\big) \cong \big[\log \hat{S}(t)\big]^2 \sum_{i:\, t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

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

$$\log\big[-\log \hat{S}(t)\big] \pm Z_{1-\alpha/2}\sqrt{\widehat{D}\big(\log\big[-\log \hat{S}(t)\big]\big)}$$

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

[2] Da KM ocena ima aproksimativno normalnu raspodelu dokazujemo pomoću teorije procesa brojanja, koja je izvan domašaja ovog master rada.

$$(exp[-e^{c_2}], exp[-e^{c_1}])$$

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

Borgan i Listol (1990.) su pokazali da je log-log interval poverenja za funkciju preživljavanja bolji od običnog, linearnog intervala. Log-log interval poverenja daje skoro ispravnu verovatnoću pokrivanja za $(1-\alpha)100\%$ interval za uzorke malog obima kao 25, sa čak 50% cenzurisanja. Verovatnoća pokrivanja za linearne interavale u ovim slučajevima je dosta manja od $(1-\alpha)$. Za jako velike uzorke, ove dve metode su ekvivalentne. Log-log intervali poverenja, za razliku od linearnih, nisu simetrični oko tačkaste ocene funkcije preživljavanja. To je prikladno za male uzorke u kojima su tačkaste ocene pristrasne i raspodela ocena pomerena.

Prethodno razmatrani intervali poverenja funkcije preživljavanja su važeći samo za fiksirano vreme t. Oni su konstruisani da obezbede da, sa datim nivoem poverenja $(1-\alpha)$, prava vrednost funkcije preživljavanja u prethodno određenom vremenu t upada u konstruisani interval. Česta netačna primena ovih intervala, koje nazivamo i tačka po tačka intervali, jeste njihovo prikazivanje za sve vrednosti t i interpretacija dobijenih krivih kao interval poverenja za celokupnu funkciju preživljavanja.

Hall i Wellner su uveli intervale poverenja koji garantuju da, sa datim nivoem poverenja, funkcija preživljavanja upada unutar tog intervala za svako t. Za njihovu konstrukciju potrebna nam je sledeća tabela kvantila[3] :

$$\hat{a} = n\hat{\sigma}^2(t_{(k)})/[1 + n\hat{\sigma}^2(t_{(k)})]$$

$(1-\alpha)$	0.1	0.25	0.40	0.50	0.60	0.75	0.90	1.0
0.90	0.599	0.894	1.062	1.133	1.181	1.217	1.224	1.224
0.95	0.682	1.014	1.198	1.273	1.321	1.354	1.358	1.358
0.99	0.851	1.256	1.470	1.552	1.600	1.626	1.628	1.628

$$\hat{\sigma}^2(t) = \sum_{i:\, t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

Uz pomoć date tabele, nije teško izračunati intervale poverenja zasnovane na ocenjenoj funkciji preživljavanja ili na njenoj log-log transformaciji. Preporuka je da se ovi intervali poverenja ograniče na vrednosti vremena manje ili jednake od najvećeg posmatranog vremena neuspeha, označenog sa $t_{(k)}$. Krajevi $(1-\alpha)100\%$ intervala poverenja za log-log transformaciju na domenu $[0, t_{(k)}]$ su

[3] Tabela kvantila se sastoji od α-kvantila Braunovog mosta.

$$\log\left[-\log \hat{S}(t)\right] \pm H_{\hat{a},\alpha} \frac{1 + n\hat{\sigma}^2(t)}{\sqrt{n}\left|\log \hat{S}(t)\right|}$$

gde je

$$\hat{\sigma}^2(t) = \sum_{i:\, t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}$$

ocena disperzije logaritmovane KM ocene funckije preživljavanja i $H_{\hat{a},\alpha}$ kvantili iz tabele, gde je

$$\hat{a} = n\hat{\sigma}^2\left(t_{(k)}\right)/\left[1 + n\hat{\sigma}^2\left(t_{(k)}\right)\right]$$

Ako označimo donje i gornje krajeve intervala sa $\hat{b}_1$ i $\hat{b}_2$, tada je Hall i Wellner interval poverenja za funkciju preživljavanja na domenu $\left[0, t_{(k)}\right]$

$$\left(exp\left[-\exp\left(\hat{b}_2\right)\right], exp\left[-\exp\left(\hat{b}_1\right)\right]\right)$$

U slučaju kada je $\hat{a} < 0.9$, linearna interpolacija između dve tabelirane vrednosti može biti potrebna za izračunavanje najtačnije vrednosti. Da bi dobili interval poverenja, izračunavamo prethodno definisane krajeve intervala za svako vreme neuspeha. Ignorišemo cenzurisanje s obzirom na to da su ocenjena funkcija preživljavanja i njena disperizija konstantne između posmatranih vremena neuspeha. Povećena širina u odnosu na tačka po tačka interval poverenja, je potrebna da bi se obezbedilo da sa verovatnoćom $(1 - \alpha)$ svaki istovremeno ocenjeni $(1 - \alpha)100\%$ interval poverenja pokriva njemu odgovarajući parametar.

Ocene kvantila

Ocenjena funkcija preživljavanja i njeni intervali poverenja nam pružaju korisnu deskriptivnu meru opšteg obrasca vremena preživljavanja. Međutim, često je korisno upoznati se sa tačkastim i intervalnim ocenama ključnih kvantila. Ocenjena funkcija preživljavanja se može koristiti za ocenu kvantila raspodele vremena preživljavanja. Ocena p-tog kvantila vremena preživljavanja je

$$\hat{t}_p = min\{t|\, \hat{S}(t) \le p\} \qquad (6)$$

Ova ocena je ekvivalentna grafičkoj oceni, koju dobijamo povlačenjem horizontalne linije od $\hat{S}(t) = p$ tačke y-ose, odnosno ose na kojoj su vrednosti KM ocene funkcije preživljavanja, do prvog preseka sa ocenjenom funkcijom preživljavanja. Povlačenjem vertikalne linije ka vremenskoj osi dobijamo ocenjeni kvantil. Da bi ocena bila konačna, horizontalna linija mora preseći ocenjenu funkciju preživljavanja. Dakle, najmanji mogući ocenjeni kvantil koji ima konačnu vrednost je posmatrani minimum funkcije preživljavanja i samo se mogu oceniti kvantili unutar domena ocenjene funkcije preživljavanja. Grafička metoda je laka za korišćenje, ali nije naročito precizna. Formula (6) daje preciznije numeričke vrednosti.

Kako, za velike uzorke, ocena kvantila ima normalnu raspodelu sa očekivanjem jednakim ocenjenom kvantilu, mogu se izvesti aproksimativni intervali poverenja za kvantile. Predložena ocena za disperziju ocene p-tog kvantila, koja se može dobiti primenom delta metode, je

$$\widehat{D}(\hat{t}_p) \cong \frac{\widehat{D}(\hat{S}(\hat{t}_p))}{[\hat{f}(\hat{t}_p)]^2}$$

gde je brojilac Greenwood-ova ocena, a imenilac ocena funkcije gustine raspodele vremena preživljavanja. Ocena funkcije gustine, koju koriste mnogi softverski paketi, je

$$\hat{f}(\hat{t}_p) = \frac{\hat{S}(\hat{u}_p) - \hat{S}(\hat{l}_p)}{\hat{l}_p - \hat{u}_p}$$

Vrednosti $\hat{l}_p$ i $\hat{u}_p$ su izabrane tako da $\hat{u}_p < \hat{t}_p < \hat{l}_p$ i najčešće se dobijaju iz jednačina:

$$\hat{u}_p = max\{t|\, \hat{S}(t) \geq p + 0.05\} \text{ i } \hat{l}_p = min\{t|\, \hat{S}(t) \leq p - 0.05\}$$

Krajevi $(1-\alpha)100\%$ interval poverenja za p-ti kvantil su

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

gde je $Z_{1-\alpha/2}$ $1-\alpha/2$-i kvantil standardne normalne raspodele.

Prilikom ocenjivanja očekivanja vremena preživljavanja javlja se problem kada je najveća opservacija cenzurisana. Postoje dva pristupa u zavisnosti od toga koja se vremena koriste za ocenjivanje i ne postoji jedinstveno mišljenje o tome koji je najbolji. Po prvom koristimo samo vremena neuspeha (ocena je pristrasna), a po drugom se koriste sve opservacije (pretvaramo se da je najveća opservacija vreme neuspeha, ali se ocena tumači uslovno na posmatranom opsegu).

Za sve pozitivne neprekidne promenljive očekivanje je jednako površini ispod funkcije preživljavanja

$$\mu = \int_0^{+\infty} S(u)du$$

Ako ograničimo promenljivu na interval $[0, t^*]$, očekivanje promenljive je

$$\mu = \int_0^{t^*} S(u)du$$

Ocena očekivanja se dobija korišćenjem KM ocene za funkciju preživljavanja. Razlog za ograničavanje domena na kojem se računa očekivanje jeste što je KM ocena nedefinisana nakon najveće opservacije. Vrednost t^* zavisi od izbora pristupa za dobijanje očekivanja. Označimo uređena vremena neuspeha sa $t_{(i)},\ i = 1, \dots, k$, a najveću opservaciju vremena sa $t_{(n)}$. Po prvom pristupu $t^* = t_{(k)}$, a po drugom $t^* = t_{(n)}$. U slučaju kada je najveća opservacija vreme neuspeha, ova dva pristupa daju identičnu ocenu. Vrednost ocene očekivanja vremena preživljavanja je površina ispod stepenaste funkcije, definisana pomoću KM ocene i izabranog intervala vremena.

Ocena očekivanja zasnovana na posmatranom opsegu vremena neuspeha je

$$\hat{\mu}\left(t_{(k)}\right) = \sum_{i=1}^{k} \hat{S}\left(t_{(i-1)}\right)\left(t_{(i)} - t_{(i-1)}\right) \qquad (7)$$

gde je $\hat{S}\left(t_{(0)}\right) = 1$ i $t_{(0)} = 0$. Ocena očekivanja zasnovana na svim opservacijama je

$$\hat{\mu}\left(t_{(n)}\right) = \hat{\mu}\left(t_{(k)}\right) + \left(1 - c_{(n)}\right)\hat{S}\left(t_{(n)}\right)\left(t_{(n)} - t_{(k)}\right) \qquad (8)$$

gde sa $c_{(n)}$ označavamo status cenzurisanja, $(0,1)$, te opservacije.

Preporučuje se ocena zasnovana na celokupnom domenu opservacija, jer ocena zasnovana samo na opsegu vremena neuspeha ne koristi informacije o preživljavanju date vremenima većim od najvećeg vremena neuspeha. Međutim, mogu da postoje situacije (npr. kada postoji značajna nesigurnost u merenju najvećeg cenzursanog vremena), kada je bolja ocena zasnovana samo na vremenima neuspeha.

Ocena disperzije uzoračkog očekivanja izračunatog koristeći (7) je data sa

$$\widehat{D}\left(\hat{\mu}\left(t_{(k)}\right)\right) = \frac{n_d}{n_d - 1}\sum_{i=1}^{k-1} \frac{A_i^2 d_i}{n_i(n_i - d_i)}$$

gde je $n_d = \sum_i^k d_i$ ukupan broj subjekata koji su ostvarili događaj i

$$A_i = \sum_{j=i}^{k-1} \hat{S}\left(t_{(j)}\right)\left(t_{(j+1)} - t_{(j)}\right)$$

Ocena disperzije uzoračkog očekivanja izračunatog koristeći (8) se dobija tako što se pretvaramo da je najveće posmatrano vreme vreme neuspeha, ali vrednost n_d se ne menja.

Aproksimativni $(1 - \alpha)100\%$ interval poverenja za očekivanje je

$$\left(\hat{\mu} - Z_{1-\alpha/2}\sqrt{\widehat{D}(\hat{\mu})}, \hat{\mu} + Z_{1-\alpha/2}\sqrt{\widehat{D}(\hat{\mu})}\right)$$

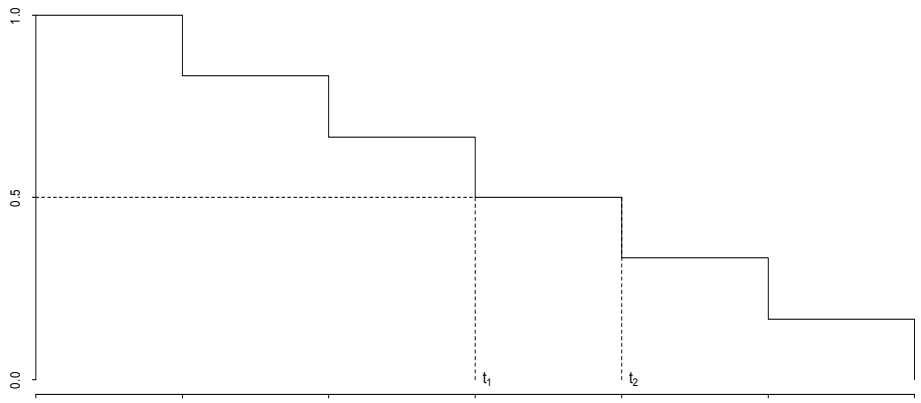
gde je $Z_{1-\alpha/2}$ $1-\alpha/2$-i kvantil standardne normalne raspodele.

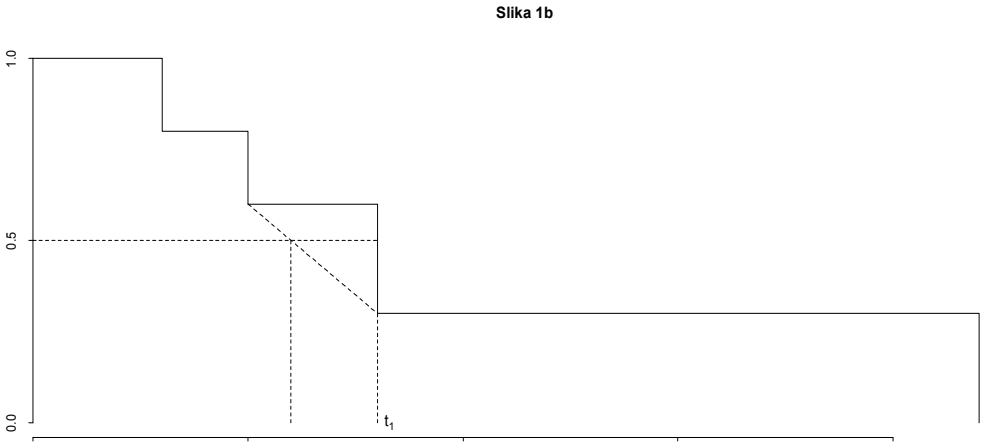
Mane Kaplan-Meier ocene

1. Kaplan-Majer ocene su ograničene na vremenski interval u koji opservacije upadaju. Ako je najveća opservacija necenzurisana, KM ocena u tom trenutku je jednaka nuli. Iako ova ocena možda neće biti pozdravljena od strane istraživača, ona je tačna s obzirom na to da niko iz uzorka ne živi duže. Ako je najveća opservacija cenzurisana, KM ocena nikada nije jednaka nuli i nedefinisana je iza najveće vrednosti.

2. Najčešće korišćena opisna statistika u analizi preživljavanja je medijana vremena preživljavnja. Do ocene medijane se jednostavno može doći na osnovu ocenjene krive preživljavanja pomoću KM metode kao vreme t za koje je $\hat{S}(t) = 0.5$. Međutim, rešenje ne mora biti jedinstveno. Posmatrajmo sliku 1a, gde je kriva preživljavanja presečena horizontalnom linijom u $\hat{S}(t) = 0.5$; bilo koja vrednost t iz intervala t_1 do t_2 je razumna ocena medijane. Praktično rešenje jeste da se uzme sredina intervala za KM ocenu medijane. Slika 1b predstavlja drugačiji slučaj u kome ocena (t_1) ima tendenciju da preceni medijanu. Praktičan način za rešenje ovog problema jeste da se povežu tačke krive i pronađe medijana.

Slika 1a





3. Ako je manje od 50% opservacija necenzurisano i najveća opservacija cenzurisana, medijana vremena preživljavanja se ne može oceniti. Praktičan način za rešenje ovog problema jeste da se koriste verovatnoće preživljavanja datih dužina vremena, recimo 1, 3 ili 5 godina ili srednje vreme preživljavanja ograničeno na dato vreme t.

4. KM metoda pretpostavlja da su cenzurisana vremena nezavisna od vremena preživljavanja. Odnosno, razlog zašto je neka opservacija cenzurisana nije povezan sa uzrokom pojave događaja. Ova pretpostavka je tačna ako je subjekat još uvek živ, nije ostvario događaj, na kraju istraživanja. Međutim, ova pretpostavka je prekršena ako se, na primer u nekom medicinskom istraživanju, gde je događaj od interesa smrt, kod pacijenta pojave ozbiljni neželjeni efekti lečenja pa je primoran da napusti istraživanje pre smrti ili ako je smrt pacijenta prouzrokovana nečim drugim od predmeta istraživanja (na primer: smrt kao posledica saobraćajne nesreće u istraživanju preživljavanja od raka). Kada postoji neprimerena cenzura, KM metoda nije odgovarajuća. U praksi, jedini način za ublažavanje ovog problema jeste da se izbegne ili da se smanji na minimum.

Neparametarske metode za poređenje raspodela preživljavanja

Nakon razmatranja ukupnog iskustva preživljavanja subjekata u istraživanju, obično našu pažnju usmeravamo na poređenje funkcija preživljavanja ključnih podgrupa podataka. Ove grupe su definisane pomoću kategorija kovarijanti koje mislimo da utiču na vremena preživljavanja.

Kovarijante se nazivaju i prateće promenljive, nezavisne promenljive, promenljive koje objašnjavaju, prognostičke faktori ili faktori rizika. One mogu biti numeričke i nenumeričke promenljive. Numeričke prognostičke promenljive mogu biti diskretne, kao što je broj prethodnih šlogova, ili neprekidne, kao što su godine ili krvni pritisak. Neprekidne promenljive mogu postati diskretne grupisanjem subjekata u potkategorije, na primer promenljiva godine može biti grupisana u 4 podgrupe: $< 20,\ 20-39,\ 40-59,\ \geq 60$. Nenumeričke prognostičke promenljive mogu biti neordinalne, kao što su rasa, pol, pušački status, ili ordinalne, na primer ozbiljnost bolesti može biti primarna, lokalna ili metastaza. Promenljive mogu biti dihotomne, na primer jetra je uvećena ili ne.

U ovom radu bavićemo se isključivo vremenski nezavisnim kovarijantama. Vremenski nezavisna promenljiva je definisana kao promenljiva čija se vrednost za datog subjekta ne menja kroz vreme. Primeri takvih promenljivih su pol i pušački status. Iako se pušački status može menjati kroz vreme, za ciljeve naše analize je uzeto da se jednom utvrđen status neće menjati. Takođe, primećujemo da se promenljive kao što su godine i težina menjaju kroz vreme, ali može biti veoma zgodno tretirati takve promenljive kao vremenski nezavisne, ukoliko se njihova vrednost ne menja drastično tokom vremena ili ako efekat takvih promenljivih na rizik preživljavanja u biti zavisi od jednom utvrđene vrednosti tih promenljivih.

Kada poredimo grupe subjekata, prvo počinjemo sa grafičkim prikazom podataka za svaku grupu. Prikazujemo na istom grafiku KM ocene funkcija preživljavanja za svaku grupu. Uopšteno, obrazac kada se jedna kriva preživljavanja nalazi iznad druge, znači da grupa definisana pomoću gornje krive duže živi, odnosno da ima pogodnije iskustvo opstanka, nego grupa definisana pomoću donje krive.

Razlike između vremena preživljavanja različitih grupa se mogu uočiti na osnovu grafika ocenjenih funkcija preživljavanja, ali to daje samo grubu predstavu o razlici između raspodela. Ne otkriva da li se radi o značajnim razlikama ili samo o slučajnim varijacijama. Statistički test je neophodan.

Poređenje dve funkcije preživljavanja

Procedura svakog od testova je zasnovana na tabeli kontingencije grupa u odnosu na status u svakom posmatranom vremenu preživljavanja. Neka je broj subjekata u skupu rizika u trenutku $t_{(i)}$ označen sa n_{1i} za grupu 1 i n_{2i} za grupu 2; broj posmatranih neuspeha u svakoj od ove dve

grupe označen sa d_{1i} i d_{2i}, respektivno; ukupan broj subjekata u riziku označen sa n_i i ukupan broj neuspeha sa d_i.

Tabela koja se koristi za poređenje funkcija preživljavanja dve grupe u posmatranom vremenu preživljavanja $t_{(i)}$:

Događaj/Grupa	1	2	Totali
Ostvaruje se	d_{1i}	d_{2i}	d_i
Ne ostvaruje se	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
U skupu riziku	n_{1i}	n_{2i}	n_i

Pretpostavimo da su opservacije u grupi 1 uzorci iz raspodele sa funkcijom preživljavanja $S_1(t)$ i opservacije u grupi 2 uzorci iz raspodele sa funkcijom preživljavanja $S_2(t)$. Nulta hipoteza koju razmatramo je

$$H_0 : \ S_1(t) = S_2(t)$$

protiv alternative

$$H_1 : \ S_1(t) > S_2(t) \quad \text{ili} \quad H_1 : \ S_1(t) < S_2(t) \quad \text{ili} \quad H_1 : \ S_1(t) \neq S_2(t)$$

Kada ne postoje cenzurisane opservacije, standardni neparametarski testovi se mogu koristiti za poređenje dve funkcije preživljavanja. Na primer, Wilcoxon testom ili Mann-Whitney testom se može testirati jednakost dve nezavisne populacije ili se test znakova može koristiti za zavisne uzorke.

U slučaju da postoje cenzurisane opservacije, razmotrićemo nekoliko neparametraskih testova pogodnih za upoređivanje funkcija preživljavanja koji se mogu definisati pomoću uopštene test statistike na sledeći način.

Pretpostavimo da su funkcije preživljavanja iste za ove dve grupe, odnosno da je H_0 tačna, tada je ocena očekivanog broja neuspeha u grupi 1 i grupi 2 u posmatranom vremenu neuspeha $t_{(i)}$:

$$\hat{e}_{1i} = \frac{n_{1i} d_i}{n_i} \qquad \hat{e}_{2(i)} = \frac{n_{2i} d_i}{n_i}$$

Ocena disperzije od broja neuspeha, na primer d_{1i}, zasnovana na hipergeometrijskoj raspodeli, definisana je sa

$$\widehat{D}_{1i} = \frac{n_{1i} n_{2i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$$

Uopštena test statistika je definisana sa

$$Q = \frac{\left[\sum_i^k w_i (d_i - \hat{e}_{1i})\right]^2}{\sum_i^k w_i^2 \widehat{D}_{1i}}$$

za neke težine w_i, a gde je k broj različitih vremena neuspeha. Primetimo da se test statistika zasniva samo na jednom uzorku vremena preživljavanja.

Pretpostavke testova su :

1. Cenzurisana iskustva su nezavisna u odnosu na grupu
2. Ukupan broj neuspeha d_i je veliki
3. Suma očekivanog broja neuspeha je velika

Pod datim pretpostavkama i pretpostavkom da je H_0 tačna, test statistika ima asimptotski hi-kvadrat raspodelu sa jednim stepenom slobode.

Testovi se razlikuju po tome koje se težine koriste. Na izbor težina utiče vrsta razlike između funkcija preživljavanja koje je test u mogućnosti najbolje da uoči. Najčešće se korsiti test koji se naziva log-rank test (Peto i Peto 1972.), čije su težine $w_i = 1$. Uopšteni Wilcoxon test koristi težine $w_i = n_i$ (Gehan 1965., Breslow 1970.). Uopšteni Wilcoxon test daje više značaja ranijim neuspesima nego kasnijim, dok log-rank test daje više značaja kasnijim neuspesima. Samim tim, uopšteni Wilcoxon test je moćniji u otkrivanju ranih razlika između dve funkcije preživljavanja, dok je log-rank test osetljiviji na razlike u desnim repovima. Log-rank test je moćniji od Wilcoxon testa u otkrivanju odstupanja kada su dve hazardne funkcije paralelne (rizici proporcionalni) ili kada postoji slučajno, ali jednako cenzurisanje i kada ne postoji cenzurisanje u uzorcima. Uopšteni Wilcoxon test se pokazuje moćniji u otkrivanju mnogih drugih razlika, na primer kada funkcije rizika nisu paralelene i kada ne postoji cenzurisanje i kada logaritmovana vremena preživljavanja prate normalnu raspodelu sa jednakim disperzijama, ali moguće različitim očekivanjima.

Kada koristimo težine $w_i = \sqrt{n_i}$ u pitanju je Tarone-Ware test (1977.), koji stavlja akcenat na razlike funkcija u srednjim vrednostima vremena.

Peto i Peto (1972.) i Prentice (1977.) su predložili korišćenje funkcije težine koja izrazitije zavisi od pomatranog iskustva preživljavanja kombinovanog uzorka. Funkcija težine je modifikacija KM ocene i definisana je na takav način da je njena vrednost poznata baš pre posmatranog vremena neuspeha. Vrednost bilo koje ocenjene funkcije preživljavanja u određenom trenutku neuspeha je poznata samo nakon što se opservacija dogodi. Osobina poznavanja vrednosti prave

opservacije neuspeha unapred naziva se predvidljivost u terminologiji procesa brojanja. Modifikovana ocena funkcije preživljavanja je

$$\tilde{S}(t) = \prod_{t_{(j)} \le t} \frac{n_j + 1 - d_j}{n_j + 1}$$

i težina koja se koristi je

$$w_i = \tilde{S}\left(t_{(i-1)}\right) \frac{n_i}{n_i + 1}$$

Harrington i Fleming (1982.) su predložili klasu testova koja za težinu koristi

$$w_i = \left[\hat{S}\left(t_{(i-1)}\right)\right]^p \left[1 - \hat{S}\left(t_{(i-1)}\right)\right]^q, \quad p, q \ge 0$$

gde je $\hat{S}$ KM ocena funkcije preživljavanja kombinovanog uzorka. Za $p = q = 0$ svodi se na log-rank test. Za $p > 0,\ q = 0$ test više težine stavlja na ranija vremena neuspeha, za $p = 0,\ q > 0$ više težine stavlja na kasnija vremena neuspeha.

Glavna prednost ovih testova u odnosu na uopšteni Wilcoxen test jeste da se težine odnose na celokupno iskustvo preživljavanja. Uopšten Wilcoxen test koristi veličinu skupa rizika i stoga težine zavise kako od cenzurisanja tako i od iskustva preživljavanja. Ako je obrazac cenzurisanja značajno različit u svakoj od grupa, onda ovaj test može odbaciti ili prihvatiti nultu hipotezu ne na osnovu razlika ili sličnosti funkcija preživljavanja, već na osnovu obrasca cenzurisanja. Iz tog razloga, većina softverskih paketa pruža informaciju o obrascima cenzurisanja u svakoj od ove dve grupe. Ovu informaciju treba proveriti, pogotovo kada su rezultati nekoliko testova značajno razlikuju.

Samo ćemo napomenuti da se stratifikovan pristup može primeniti na bilo koju vrstu težina. Međutim, ograničenje ovog pristupa jeste u smanjenju obima uzorka unutar svakog stratuma.

Uopšteno, različito ponderisanje treba da da slične rezultate i uglavnom će dovesti do istih zaključaka. Jedini razlog za izbor jednog testa pre drugog jeste ukoliko je moćniji, odnosno ukoliko je verovatnije odbacivanje netačne hipoteze. Treba doneti a priori odluku o tome koji test koristiti, umesto traženja željene p-vrednosti. Traženje željenih rezultata može dovesti do pristrasnosti.

Poređenje $G > 2$ funkcija preživljavanja

Ako postoji $G > 2$ populacijskih grupa od interesa, slični testovi se mogu izgraditi uopštenjem oznaka. Test statistika se zasniva na tabelama kontingencije za svako pomatrano vreme neuspeha $t_{(i)}$, koje imaju sledeći oblik :

Događaj/Grupa	1	2	...	G	Totali
Ostvaruje se	d_{1i}	d_{2i}		d_{Gi}	d_i
Ne ostvaruje se	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	...	$n_{Gi} - d_{Gi}$	$n_i - d_i$
U skupu riziku	n_{1i}	n_{2i}	...	n_{Gi}	n_i

Nulta hipoteza koju testiramo je

$$H_0 : \ S_1(t) = S_2(t) = \cdots = S_G(t)$$

protiv alternative

$$H_1 : \text{Postoji najmanje jedan par grupa } j \text{ i } l \text{ takav da je } S_j(t) \neq S_l(t)$$

Pod pretpostavkom da su sve funkcije preživljavanja jednake, ocenjujemo očekivani broj neuspeha u trenutku neuspeha $t_{(i)}$ za svaku grupu sa :

$$\hat{e}_{ji} = \frac{n_{ji} d_i}{n_i}, \quad j = 1, \dots, G$$

Poredimo posmatrane sa očekivane brojevima događaja za $G - 1$ od G grupa.

Neka su

$$\mathbb{d}_i' = (d_{1i}, d_{2i}, \dots, d_{G-1i}) \quad \text{i} \quad \hat{\mathbb{e}}_i' = (\hat{e}_{1i}, \hat{e}_{2i}, \dots, \hat{e}_{G-1i})$$

vektorski zapisi posmatranog broja neuspeha i ocena očekivanog broja neuspeha. Razlika ova dva vektora je

$$(\mathbb{d}_i - \hat{\mathbb{e}}_i)' = (d_{1i} - \hat{e}_{1i}, d_{2i} - \hat{e}_{2i}, \dots, d_{G-1i} - \hat{e}_{G-1i})$$

Iz praktičnih razloga koristimo prvih $G - 1$ grupa, ali bilo koja kolekcija od $G - 1$ grupa se može podjednako dobro koristiti.

Da bi izračunali test statistiku, potrebna nam je ocena kovarijansne matrica od $\mathbb{d}_i$. Elementi ove matrice se izračunavaju pretpostavljajući da posmatrani broj neuspeha ima višedimenzionalnu centralnu hipergeometrijsku raspodelu. Dijagonalni elementi $(G-1) \times (G-1)$ matrice, označene sa $\hat{\mathbb{D}}_i$, su

$$\hat{D}_{jji} = \frac{n_{ji}(n_i - n_{ji}) d_i (n_i - d_i)}{n_i^2 (n_i - 1)}, \quad j = 1, \dots, G-1$$

i nedijagonalni elementi su

$$\widehat{D}_{jli} = -\frac{n_{ji}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad j, l = 1, \dots, G-1, \quad j \neq l$$

Definišemo $(G-1) \times (G-1)$ dijagonalnu matricu težina, $\mathbb{W}_i = diag(w_i)$, za svako različito vreme neuspeha $t_{(i)}$. Težine w_i mogu biti bilo koje od prethodno razmatranih težina kod testova za dve grupe. Test statistika kojom se porede iskustva preživljavanja G grupa je

$$Q = \left[\sum_{i=1}^{k} \mathbb{W}_i(\mathbb{d}_i - \widehat{\mathbb{e}}_i)\right]' \left[\sum_{i=1}^{k} \mathbb{W}_i \widehat{\mathbb{D}}_i \mathbb{W}_i\right]^{-1} \left[\sum_{i=1}^{k} \mathbb{W}_i(\mathbb{d}_i - \widehat{\mathbb{e}}_i)\right]$$

Razlog zašto koristimo $G-1$ od G mogućih posmatrano-očekivanih poređenja jeste da bi sprečili da matrica u sredini desne strane gornje jednakosti bude singularna. Vrednost test statistike je ista, bez obzira koja se kolekcija $G-1$ grupa koristi. Za $G = 2$ ova test statistika se svodi na test statistiku testova za dve grupe. Pod pretpostavkom da je nulta hipoteza tačna i da je suma ocenjenih očekivanih brojeva neuspeha velika, test statistika asimptotski ima hi-kvadrat raspodelu sa $G-1$ stepenom slobode.

Napomenimo samo da, i u slučaju višestrukog poređenja, važe prethodna zapažanja o tome kako izbor težina utiče na sposobnost testa da otkrije određene razlike funkcija preživljavanja.

Cox-ov model sa proporcionalnim rizicima

Modele preživljavanja analiziramo posmatrajući dve fundamentalne stavke, a to su osnovna funcija rizika koja opisuje kako se menja rizik tokom vremena i efekat parametara koji opisuju kako rizik varira u odnosu na nezavisne promenljive. David Cox je uočio da ukoliko pretpostavimo da je rizik proporcionalan moguće je oceniti efekat parametara bez određivanja same funkcionalne forme rizika. Ovaj pristup analizi podataka preživljavanja se zove primena Cox-ovog modela sa proporcionalnim rizikom (Cox-ov PH model).

Cox (1972.) je unapredio predviđanje vremena preživljavanja subjekta bez pretpostavki o osnovnoj funkciji rizika subjekta, ali pretpostavljajući da su funkcije rizika različitih subjekata proporcionalne. Dakle, Cox-ov PH model ima osobinu da je $[h(t,x_1)/h(t,x_2)]$ količnik funkcija rizika dva subjekta sa prognostičkim faktorima ili kovarijantama $x_1 = (x_{11}, x_{21}, \dots, x_{p1})'$ i $x_2 = (x_{12}, x_{22}, \dots, x_{p2})'$ konstantan (ne menja se sa vremenom). Ovo znači da je odnos rizika od neuspeha za dva subjekta isti bez obzira na to koliko su dugo živeli. Ova osobina ukazuje na to da se funkcija rizika za dat skup kovarijanti $\mathbb{x} = (x_1, x_2, \dots, x_p)'$ može napisati kao funkcija osnovne funkcije rizika i funkcije, $g(x_1, x_2, \dots, x_p)$, koja zavisi samo od kovarijanti, odnosno

$$h(t|x_1, x_2, \dots, x_p) = h_0(t)g(x_1, x_2, \dots, x_p) \text{ ili } h(t|\mathbb{x}) = h_0(t)g(\mathbb{x})$$

Osnovna funkcija rizika, $h_0(t)$, predstavlja promenu rizika sa vremenom, a $g(\mathbb{x})$ predstavlja efekte kovarijanti. $h_0(t)$ se može interpretirati kao funkcija rizika kada se ignorišu sve kovarijante ili kada je $g(\mathbb{x}) = 1$, i takođe se naziva polazna funkcija rizika. Hazardni odnos (količnik) dva subjekta sa različitim kovarijantama $\mathbb{x}_1$ i $\mathbb{x}_2$ je

$$\frac{h(t|\mathbb{x}_1)}{h(t|\mathbb{x}_2)} = \frac{h_0(t)g(\mathbb{x}_1)}{h_0(t)g(\mathbb{x}_2)} = \frac{g(\mathbb{x}_1)}{g(\mathbb{x}_2)}$$

konstantan, nezavisan od vremena.

Cox-ov PH model pretpostavlja da je $g(\mathbb{x})$ eksponencijalna funkcija kovarijanti, odnosno

$$g(\mathbb{x}) = exp\left(\sum_{j=1}^{p} b_j x_j\right) = \exp(\mathbb{b}'\mathbb{x})$$

i funkcija rizika je

$$h(t|\mathbb{x}) = h_0(t)exp\left(\sum_{j=1}^{p} b_j x_j\right) = h_0(t)\exp(\mathbb{b}'\mathbb{x}) \qquad (9)$$

gde sa $\mathbb{b} = (b_1, \dots, b_p)$ označavamo koeficijente kovarijanti. Ovi koeficijenti, koji se mogu oceniti iz posmatranih podataka, označavaju jačinu uticaja njihovih odgovarajućih kovarijanti. Na primer, ukoliko imamo samo jednu kovarijantu, neka $x_1 = 0$ ako osoba dobija placebo, a $x_1 = 1$ ako osoba dobija eksperimentalni lek. Hazardni odnos pacijenta koji dobija lek i pacijenta koji ne dobija lek je

$$\frac{h(t|x_1 = 1)}{h(t|x_2 = 0)} = \exp(b_1)$$

Prema tome, dva tretmana su podjednako efektna ako je $b_1 = 0$ i eksperimentalni lek predstavlja manji (veći) rizik za preživljavanje od placeba ako je $b_1 < 0$ $(b_1 > 0)$.

Primetimo da Cox-ova formula (9) ima osobinu da ukoliko su sve nezavisne promenljive jednake nuli da se ona svodi na funkciju osnovnog rizika, jer je $e^0 = 1$. Ova osobina Cox-ovog modela jeste i razlog zašto se $h_0(t)$ zove osnovna funkcija rizika. Drugačije rečeno, Cox-ov model se svodi na osnovnu funkciju rizika kada u modelu nema nezavisnih promenljivih.

Druga važna osobina Cox-ovog modela je ta da je osnovna funkcija rizika, $h_0(t)$, neodređena funkcija. Ova osobina čini Cox-ov model neparametarskim modelom.

Deljenjem obe strane jednačine (9) sa $h_0(t)$ i logaritmovanjem dobijamo

$$log\frac{h_1(t)}{h_0(t)} = b_1x_{1i} + b_2x_{2i} + \cdots + b_px_{pi} = \sum_{j=1}^{p} b_jx_{ji} = \mathbb{b}'\mathbb{x}_i$$

gde su x_{ji} kovarijante za i-ti subjekat. Leva strana gornje jednačine je funkcija hazardnog količnika (ili rezlativni rizik), a desna strana je linearna funkcija kovarijanti i njihovih odgovarajućih koeficijenata.

Bavimo se primenom gornje jednačine i glavni cilj nam je nalaženje značajnih prognostičkih faktora. Drugim rečima, želimo da identifikujemo među p kovarijanti podskup promenljivih koje značajnije utiču na rizik, kao i na dužinu preživljavanja subjekata. Zanimaju nas regresioni koeficijenti. Ako je b_i nula, odgovarajuća kovarijanta ne utiče na opstanak. Ako je b_i nije nula, ono predstavlja jačinu uticaja x_i na rizik kada se istovremeno razmatraju druge kovarijante.

Može se pokazati da je funkcija preživljavanja Cox-ovog PH modela ekvivalentna sa

$$S(t|\mathbb{x}) = [S_0(t)]^{exp\left(\sum_{j=1}^{p} b_jx_j\right)} = [S_0(t)]^{exp(\mathbb{b}\prime\mathbb{x})},$$

gde sa $S_0(t)$ označavamo osnovnu funkciju preživljavanja koja odgovara osnovnoj funkciji rizika $h_0(t)$.

Ocenjivanje koeficijenata u Cox-ovom PH modelu

Kako parametre obično ocenjujemo uz pomoć funkcije verodostojnosti, razmotrimo oblik te funkcije u slučaju Cox-ovog PH modela.

Neka $f(t|\mathbb{x},\mathbb{b})$ predstavlja gustinu verovatnoće neuspeha u trenutku t ako su date vrednosti kovarijanti $\mathbb{x}$ i njihovi regresioni koeficijenti $\mathbb{b}$. Tada subjekat i koji je ostvario događaj u trenutku t_i doprinosi funkciji verodostojnosti sa $f(t_i\,|\mathbb{x}_i,\mathbb{b})$. Dok za subjekat i koji je (desno) cenzurisan u trenutku t_i, sve što znamo jeste da je preživeo do tog trenutka, i stoga doprinosi funkciji verodostojnosti sa $S(t_i\,|\mathbb{x}_i,\mathbb{b})$.

Tada se funkcija verodostojnosti za n podataka preživljavanja definiše sa

$$L(\mathbb{b}) = \prod_{i=1}^{n}\left(f(t_i\,|\mathbb{x}_i,\mathbb{b})\right)^{\delta_i}\left(S(t_i\,|\mathbb{x}_i,\mathbb{b})\right)^{1-\delta_i}$$

gde je δ_i statusna promenljiva.

Koristeći dobro poznati odnos funkcija preživljavanja $h(t) = \frac{f(t)}{S(t)}$, dobijamo da je

$$L(\mathbb{b}) = \prod_{i=1}^{n}\left(h(t_i\,|\mathbb{x}_i,\mathbb{b})\right)^{\delta_i} S(t_i\,|\mathbb{x}_i,\mathbb{b})$$

Prema Cox-ovom PH modelu

$$h(t_i\,|\mathbb{x}_i,\mathbb{b}) = h_0(t_i)\exp\left(\mathbb{b}'\mathbb{x}_i\right) \quad \text{i} \quad S(t_i\,|\mathbb{x}_i,\mathbb{b}) = [S_0(t_i)]^{exp(\mathbb{b}\prime\mathbb{x}_i)}$$

Samim tim, funkcija verodostojnosti ima sledeći oblik

$$L(\mathbb{b}) = \prod_{i=1}^{n} h_0(t_i)^{\delta_i}\exp\left(\mathbb{b}'\mathbb{x}_i\right)^{\delta_i} S_0(t_i)^{exp(\mathbb{b}\prime\mathbb{x}_i)}$$

$$\log L(\mathbb{b}) = \sum_{i=1}^{n} \delta_i \log h_0(t_i) + \delta_i \mathbb{b}'\mathbb{x}_i + exp(\mathbb{b}'\mathbb{x}_i) S_0(t_i)$$

Uočavamo da ne možemo naći ocenu maksimalne verodostojnosti parametara $\mathbb{b}$ bez određivanja oblika osnovne funkcije rizika $h_0(t_i)$.

Za ocenu koeficijenata , $b_1, \ldots, b_p$, Cox (1972.) predlaže parcijalnu funkciju verodostojnosti zasnovanu na uslovnoj verovatnoći neuspeha, pretpostavljajući da nema istih vrednosti među vremenima preživljavanja. Međutim, pošto se u praksi često sreću ista vremena preživljavanja, Cox-ova parcijalna funkcija verodostojnosti je modifikovana da se izbori sa istim vrednostima.

U nastavku opisujemo postupak ocenjivanja koeficijenta sa i bez istih vrednosti vremena neuspeha.

Parcijalna funkcija verodostojnosti za jedinstvena vremena neuspeha

Pretpostavimo da je k vremena preživljavanja od n subjekata necenzurisano i različito, i $n-k$ cenzurisano sa desne strane. Neka su $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ k različitih vremena neuspeha poređanih od najkraćeg ka najdužem sa odgovarajućim kovarijantama $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. Neka je $R(t_{(i)})$ skup rizika u trenutku $t_{(i)}$. $R(t_{(i)})$ se sastoji od svih subjekata čija su vremena preživljavanja najmanje $t_{(i)}$. Uslovna verovatnoća da subjekat i ostvari događaj u trenutku $t_{(i)}$, ako je dat skup rizika $R(t_{(i)})$ i ako znamo da samo jedan subjekat doživljava neuspeh u tom trenutku, je

$$\frac{exp\left(\sum_{j=1}^{p} b_j x_{j(i)}\right)}{\sum_{l \in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)} \quad \left(= \frac{\exp\left(\mathbb{b}'\mathbb{x}_{(i)}\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\mathbb{b}'\mathbb{x}_{l}\right)}\right)$$

Prema Cox-ovom PH modelu, rizik subjekta i koji je ostvario događaj u trenutku $t_{(i)}$ je proporcionalan $\exp\left(\mathbb{b}'\mathbb{x}_{(i)}\right)$. Samim tim, ovaj odnos, takođe, izražava rizik subjekta i u odnosu na kumulativni rizik svih subjekata koji su u rizičnom skupu u trenutku kada je ostvaren događaj subjekta i.

Dakle, parcijalna funkcija verodostojnosti je

$$L(\mathbb{b}) = \prod_{i=1}^{n} \left(\frac{exp\left(\sum_{j=1}^{p} b_j x_{j(i)}\right)}{\sum_{l \in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)} \right)^{\delta_i}$$

Kako statusna promenljiva δ_i uzima vrednosti 0 ili 1 u zavisnosti od toga da li postoji cezurisanje ili ne, parcijalna funkcija verodostojnosti za jedinstvena vremena neuspeha se svodi na

$$L(\mathbb{b}) = \prod_{i=1}^{k} \frac{exp\left(\sum_{j=1}^{p} b_j x_{j(i)}\right)}{\sum_{l \in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)} \qquad \left(= \prod_{i=1}^{k} \frac{\exp\left(\mathbb{b}'\mathbb{x}_{(i)}\right)}{\sum_{l \in R(t_{(i)})} \exp\left(\mathbb{b}'\mathbb{x}_{l}\right)}\right) \qquad (10)$$

Želimo one vrednosti koeficijenata koji će predvideti da je rizik bio veliki za subjekte u trenucima u kojima su se događaji ostvarili.

Kada logaritmujemo dobijamo

$$l(\mathbb{b}) = logL(\mathbb{b}) = \sum_{i=1}^{k}\sum_{j=1}^{p} b_j x_{j(i)} - \sum_{i=1}^{k} log\left[\sum_{l\in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)\right]$$
$$= \sum_{i=1}^{k}\left\{\mathbb{b}'\mathbb{x}_{(i)} - log\left[\sum_{l\in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)\right]\right\}$$

Ocena maksimalne parcijalne verodostojnosti (OMPV) $\widehat{\mathbb{b}}$ od $\mathbb{b}$ je skup $\hat{b}_1, \dots, \hat{b}_p$ koji maksimizuje $l(\mathbb{b})$:

$$l(\widehat{\mathbb{b}}) = \max_{\text{svim } \mathbb{b}} (l(\mathbb{b}))$$

Jasno je da je $\widehat{\mathbb{b}}$ rešenje sledećih jednačina

$$\frac{\partial(l(\mathbb{b}))}{\partial \mathbb{b}} = 0$$

Ili

$$\frac{\partial(l(\mathbb{b}))}{\partial b_u} = \sum_{i=1}^{k}\left[x_{u(i)} - A_{ui}(\mathbb{b})\right] = 0, \quad u = \overline{1,p}$$

gde je

$$A_{ui}(\mathbb{b}) = \frac{\sum_{l\in R(t_{(i)})} x_{ul} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)}{\sum_{l\in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)} = \frac{\sum_{l\in R(t_{(i)})} x_{ul} \exp(\mathbb{b}'\mathbb{x}_l)}{\sum_{l\in R(t_{(i)})} \exp(\mathbb{b}'\mathbb{x}_l)}$$

Često se za izračunavanje OMPV $\widehat{\mathbb{b}}$ koristi numerička metoda pod nazivom Newton-Raphson iterativna procedura. Sastoji se od niza iteracija, pri čemu se pretpostavlja da ocena postaje sve bolja sa svakom iteracijom. Ova metoda je i deterministička, jer ne postoji element slučajnosti u potrazi za optimalnim vrednostima. Ona često, ali ne uvek, konvergira ka željenim ocenama maksimalne verodostojnosti. Napomenula bih samo da i R koristi ovu metodu za ocenjivanje parametara.

Newton-Raphson metoda se može predstaviti na sledeći način :

1. Neka su inicijalne vrednosti od $b_1, \dots, b_p$ nula, odnosno neka je $\mathbb{b}^0 = 0$.
2. Promene $\mathbb{b}$ u svakom sledećem koraku, koje označavamo sa $\Delta^{(j)}$, izračunavamo preko drugog izvoda log-parcijalno verovatnosne funkcije :

$$\Delta^{(j)} = \left[-\frac{\partial^2 l(\mathbb{b}^{(j-1)})}{\partial \mathbb{b} \partial \mathbb{b}'} \right]^{-1} \frac{\partial l(\mathbb{b}^{(j-1)})}{\partial \mathbb{b}}$$

3. Koristeći $\Delta^{(j)}$, vrednost $\mathbb{b}^{(j)}$ u j-tom koraku je

$$\mathbb{b}^{(j)} = \mathbb{b}^{(j-1)} + \Delta^{(j)}, \quad j = 1,2, \dots$$

Iteracija se završava, u recimo, m-tom koraku ako je $\left\|\Delta^{(m)}\right\| < \delta$, gde je δ zadata preciznost, često jako mali broj, 10^{-4} ili 10^{-5}. Onda se OMPV $\widehat{\mathbb{b}}$ definiše kao

$$\widehat{\mathbb{b}} = \mathbb{b}^{(m-1)}$$

Drugi parcijalni derivat od $l(\mathbb{b})$ u odnosu na b_u i b_v , $u, v = \overline{1, p}$ u Newton-Raphson itrerativnom postupku je

$$I_{uv}(\mathbb{b}) = \frac{\partial^2 l(\mathbb{b})}{\partial b_u \partial b_v} = -\sum_{i=1}^{k} C_{uvi}(b_1, \dots, b_p) = -\sum_{i=1}^{k} C_{uvi}(\mathbb{b}) \qquad u, v = \overline{1, p} \qquad (3)$$

gde je

$$C_{uvi}(\mathbb{b}) = \frac{\sum_{l \in R(t_{(i)})} x_{ul} x_{vl} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)}{\sum_{l \in R(t_{(i)})} exp\left(\sum_{j=1}^{p} b_j x_{jl}\right)} - A_{ui}(\mathbb{b}) A_{vi}(\mathbb{b})$$

Ova metoda je dobra ukoliko su početne vrednosti $\mathbb{b}^0$ dovoljno blizu ciljanih vrednosti $\widehat{\mathbb{b}}$. Ukoliko nisu, mogu dovesti do velikih skokova daleko od ciljanih vrednosti. Da bi smanjili rizik da se to dogodi, možemo promeniti treći korak u

$$\mathbb{b}^{(j)} = \mathbb{b}^{(j-1)} + \xi \Delta^{(j)}, \quad j = 1,2, \dots$$

gde se $\xi < 1$ ponaša kao kočnica, ograničavajući veličinu skoka, što povećava broj iteracija za dostizanje željenih vrednosti $\widehat{\mathbb{b}}$.

Iako nisu u potpunosti efikasne, OMPV imaju druge opšte osobine ocena maksimalne verodostojnosti.

Asimptotska ocena kovarijanse matrice od OMPV $\widehat{\mathbb{b}}$ je data sa

$$\widehat{D}\left(\widehat{\mathbb{b}}\right) = \widehat{Cov}\left(\widehat{\mathbb{b}}\right) = \left[-\frac{\partial^2 l\left(\widehat{\mathbb{b}}\right)}{\partial \mathbb{b} \partial \mathbb{b}'} \right]^{-1}$$

gde se $-\partial^2 l\left(\widehat{\mathbb{b}}\right)/\partial \mathbb{b} \partial \mathbb{b}'$ naziva posmatrana informaciona matrica sa $-I_{uv}(\widehat{\mathbb{b}})$ kao njenim (u, v) elementom, gde je $I_{uv}(\mathbb{b})$ definisano jednačinom (3). Označimo sa v_{ij} (i, j) element $\widehat{D}\left(\widehat{\mathbb{b}}\right)$, tada je $100(1-\alpha)\%$ interval poverenja za b_i

$$\left(\hat{b}_i - Z_{1-\alpha/2}\sqrt{v_{ii}},\ \hat{b}_i + Z_{1-\alpha/2}\sqrt{v_{ii}}\right)$$

Parcijalna funkcija verodostojnosti svoj naziv duguje tome da ona ne koristi podatke preživljavanja u potpunosti: tačna vremena neuspeha nisu bitna, samo njihovo rangiranje. Kalbfleisch i Prentice su pokazali da je parcijalna funkcija verodostojnosti za jedinstvena vremena neuspeha marginalna verodostojnost rangova opservacija, dobijena posmatrajući samo redosled po kom su subjekti ostvarivali događaj, ne i sama vremena neuspeha.

Parcijalna funkcija verodostojnosti za ponovljena vremena neuspeha

Pretpostavimo da među n posmatranih vremena preživljavanja imamo k različitih necenzurisanih vremena $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$. Neka d_i označava broj subjekata koji ostvaruju događaj u trenutku $t_{(i)}$ ili mnogostrukost $t_{(i)}$. Označimo sa $R(t_{(i)})$ skup subjekata u riziku u trenutku $t_{(i)}$ i sa n_i broj takvih subjekata. Na primer, u sledećem skupu vremena preživljavanja $\{15, 16+, 20, 20, 20, 21, 24, 24\}$, $n = 8,\ k = 4,\ t_{(1)} = 15,\ t_{(2)} = 20,\ t_{(3)} = 21,\ t_{(4)} = 24,\ d_1 = 1,\ d_2 = 3,\ d_3 = 1$ i $d_4 = 2$. Onda $R(t_{(1)})$ obuhvata svih osam subjekata. U $R(t_{(2)})$ uključeni su subjekti sa vremenima preživljavanja 20, 21, 24, u $R(t_{(3)})$ subjekti sa vremenima preživljavanja 21 i 24 i $R(t_{(4)})$ oni sa vremenom preživljavanja 24; prema tome $n_1 = 8,\ n_2 = 6,\ n_3 = 3$ i $n_4 = 2$.

Da bi razmotrili metode ocenjivanja kada postoje iste vrednosti, uvodimo par novih oznaka. Iz svakog skupa $R(t_{(i)})$ možemo izabrati na slučajan način d_i subjekata. Označimo svaki od ovih d_i izbora sa $u_{(j)}$. Postoji $n_i!/[d_i!\,(n_i - d_i)!]$ mogućih $u_{(j)}$-ova. Sa U_i označimo skup koji sadrži sve $u_{(j)}$-ove. Na primer, iz $R(t_{(2)})$ možemo na slučajan način izabrati $d_2 = 3$ od mogućih $n_2 = 6$ subjekata. Ukupno imamo 20 takvih izbora (ili podskupova) i jedan od $u_{(j)}$-ova je, na primer, { $tri\ subjekta\ sa\ vremenima\ preživljavanja\ 20, 20\ i\ 24$ }. $U_2 = \{u_{(1)}, u_{(2)}, \ldots, u_{(20)}\}$ sadrži svih 20 podskupova. Fokusirajmo se sada na opservacije sa istim vrednostima. Sa $\mathbb{x}_k = (x_{1k}, \ldots, x_{pk})'$ označavamo kovarijante k -tog subjekta, $z_{u_{(j)}} = \sum_{k \in u_{(j)}} \mathbb{x}_k = (z_{1u_{(j)}}, \ldots, z_{pu_{(j)}})'$, gde je $z_{lu_{(j)}}$ suma l-tih kovarijanti od d_i subjekata koji su u $u_{(j)}$. Sa $u^*_{(i)}$ označavamo skup od d_i subjekata koji su ostvarili događaj u trenutku $t_{(i)}$ i $z_{u^*_{(i)}} = \sum_{k \in u^*_{(i)}} \mathbb{x}_k = (z^*_{1u^*_{(i)}}, \ldots, z^*_{pu^*_{(i)}})'$, gde je $z^*_{lu^*_{(i)}}$ suma l-tih kovarijanti od d_i subjekata koji su u $u^*_{(i)}$. Na primer, za skup $R(t_{(2)})$, $z^*_{1u^*_{(2)}}$ je jednak sumi vrednosti prve kovarijante od tri subjekta čije je vreme preživljavanja 20.

Neprekidna vremenska skala

U slučaju neprekidne vremenske skale, za d_i subjekata koji ostvaruju događaj u trenutku $t_{(i)}$, razumno je smatrati da vremena preživljavanja d_i subjekata nisu identična jer su najverovatnije iste vrednosti rezultat nepreciznog merenja. Ukoliko je moguće izvršiti tačna merenja, ova d_i

vremena preživljavanja možemo poređati u niz i možemo koristiti parcijalnu funkciju verodostojnosti (10). U odsustvu poznavanja pravog poretka (realan slučaj), moramo razmotriti sve moguće poretke ovih posmatranih d_i vremena preživljavanja (istih). Za svako $t_{(i)}$, d_i posmatranih istih vremena preživljavanja mogu biti poređani na $d_i!$ mogućih različitih načina. Za svaki od ovih mogućih poredaka imaćemo proizvod kao u (10) za odgovarajućih d_i vremena preživljavanja. Stoga, kad je vreme preživljavanja mereno na neprekidnoj vremenskoj skali, izgradnja i računanje tačne parcijalne funkcije verodostojnosti je veoma kompleksan zadatak ako je d_i veliko.

Kao aproksimacija tačne parcijalne funkcije verodostojnosti, mogu se koristi dve sledeće funkcije verodostojnosti kada je svaki d_i mali u odnosu na n_i .

Breslow (1974.) je uveo sledeću aproksimaciju :

$$L_B(\mathbb{b}) = \prod_{i=1}^{k} \frac{\exp\left(z'_{u^*_{(i)}}\mathbb{b}\right)}{\left[\sum_{l\in R(t_{(i)})} exp(x'_l\mathbb{b})\right]^{d_i}}$$

Alternativnu (precizniju) aproksimaciju je uveo Efron (1977.) :

$$L_E(\mathbb{b}) = \prod_{i=1}^{k} \frac{\exp\left(z'_{u^*_{(i)}}\mathbb{b}\right)}{\prod_{j=1}^{d_i}\left[\sum_{l\in R(t_{(i)})} exp(x'_l\mathbb{b}) - [(j-1)/d_i]\sum_{l\in u^*_{(i)}} exp(x'_l\mathbb{b})\right]}$$

Diskretna vremenska skala

U slučaju diskretne vremenske skale, iste opservacije su zaista iste, odnosno ovi događaji se stvarno dešavaju u istom trenutku. Cox (1972.) je predložio sledeći model :

$$\frac{h_i(t)dt}{1-h_i(t)dt} = \frac{h_0(t)dt}{1-h_0(t)dt} exp\left(\sum_{j=1}^{p} b_j x_{ji}\right) = \frac{h_0(t)dt}{1-h_0(t)dt}\exp\left(\mathbb{b}'\mathbb{x}_i\right)$$

Ovaj model se svodi na (9) kada imamo neprekidnu vremensku skalu. Koristeći model i zamenom i-tog člana u (10) sledećim izrazom u slučaju istih opservacija u $t_{(i)}$:

$$\frac{\exp\left(z'_{u^*_{(i)}}\mathbb{b}\right)}{\sum_{u_{(j)}\in U_i}\exp\left(z'_{u_{(j)}}\mathbb{b}\right)}$$

dobijamo parcijalnu funkciju verodostojnosti za iste opservacije u diskretnoj vremenskoj skali

$$L_d(\mathbb{b}) = \prod_{i=1}^{k} \frac{\exp\left(z'_{u^*_{(i)}}\mathbb{b}\right)}{\sum_{u_{(j)} \in U_i} \exp\left(z'_{u_{(j)}}\mathbb{b}\right)}$$

i -ti član u ovom izrazu predstavlja uslovnu verovatnoću posmatranja d_i neuspeha pod pretpostavkom da postoji d_i neuspeha u trenutku $t_{(i)}$ i skup rizika $R(t_{(i)})$ u $t_{(i)}$. Broj članova u imeniocu i-tog člana je $n_i!/[d_i!\,(n_i - d_i)!]$ i biće veoma veliki ako je d_i veliko. Srećom, rekurzivni algoritam koji je predložio Gail (1981.) čini račun izvodljivim. Gornja jednačina se takođe može smatrati i aproksimacijom parcijalne funkcije verodostojnosti za neprekidna vremena preživljavanja sa istim vrednostima.

U praksi, gornje tri parcijalne funkcije verodostojnosti se pokazuju kao dobra aproksimacija tačne parcijalne funkcije verodostojnosti za neprekidna vremena preživljavanja sa istim vrednostima. Ukoliko nema istih vrednosti te jednačine se svode na (10). OMPV od $\mathbb{b}$ se dobija koristeći slične procedure koje smo već opisali.

Identifikacija značajnih kovarijanti

U praksi obično postoji veliki broj mogućih kovarijanti povezanih sa ishodima. Jedan od načina da se smanji broj kovarijanti jeste ispitivanje odnosa između svake pojedinačne kovarijante i zavisne promenljive, vremena preživljavanja. Kovarijante koje imaju mali ili nemaju uticaj na zavisnu promenljivu mogu biti isključeni iz modela. Stoga, jedan od glavnih ciljeva analize preživljavanja jeste identifikacija značajnih kovarijanti. Ovo uključuje testiranje hipoteza i procedure selekcija kovarijanti.

Testiranje hipoteza

Za testiranje hipoteze da neka kovarijanta nema efekta na rizik, odnosno testiranje hipoteze $H_0: b_i = 0$ koristimo Wald test statistiku :

$$X_W = \frac{\hat{b}_i^{\,2}}{\widehat{D}_i(\widehat{\mathbb{b}})}$$

gde je $\hat{b}_i$ OMPV od b_i , $\widehat{\mathbb{b}} = (\hat{b}_1, \dots, \hat{b}_i, \dots, \hat{b}_p)$ je OMPV od $\mathbb{b}$, vektora parametara kovarijanti, i $\widehat{D}_i(\widehat{\mathbb{b}})$ je podmatrica ocenjene kovarijansne matrice koja odgovara b_i .

Za dati prag značajnosti α , H_0 se odbacuje ako je je $X_W > \chi^2_{1,\alpha/2}$ ili $X_W < \chi^2_{1,1-\alpha/2}$ (dvostrani test) ili $X_W > \chi^2_{1,\alpha}$ (jednostrani test), gde su $\chi^2_{1,\alpha}$, $\chi^2_{1,\alpha/2}$ i $\chi^2_{1,1-\alpha/2}$ odgovarajući kvantili hi-kvadrat raspodele sa jednim stepenom slobode.

Uopštimo nultu hipotezu. Bez gubitka opštosti želimo da testiramo hipotezu da je prvih $1 \le q \le p$ elemenata vektora parametara $\mathbb{b}$, gde je p dimenzija od $\mathbb{b}$, jednako nekim određenim vrednostima b_j^* , $j = 1, \dots, q$. Ostalih $p - q$ elemenata su slobodni parametri. Alternativna hipoteza glasi: bar jedan od tih q parametara nije jednak pretpostavljenoj vrednosti.

Detaljnije ćemo opisati test log verovatnosnog količnika, koji je, po Hosmer-u (2008.), najpoželjnije koristiti. Test izvodimo pomoću dva ugneždena modela. Uopšteni model nam dozvoljava da svih p parametara budu ocenjeni sa OMPV. Specifičan model fiksira prvih q parametara na njihove pretpostavljene vrednosti, a ostalih $p - q$ parametara ocenjuje sa OMPV. Parametarski prostor uopštenog modela sadrži manje dimenzionalan parametarski prostor specifičnog modela, tj. specifičan model je ugnežden u uopšten model.

Neka je $l(\widehat{\mathbb{b}})$ vrednost logaritmovane parcijalne funkcije verodostojnosti u OMPV $\widehat{\mathbb{b}}$ uopštenog modela.

Neka je $l(b^*_{1:q}, \hat{b}_{q+1:p})$ vrednost logaritmovane parcijalne funkcije verodostojnosti u OMPV od $p - q$ slobodnih parametara, pod uslovom q fiksiranih parametara u specifičnom modelu.

Log verovatnosna količnik statistika (test statistika ovog testa) je

$$X_L = 2\left[l\left(\widehat{\mathbb{b}}\right) - l\left(b^*_{1:q}, \hat{b}_{q+1:p}\right)\right]$$

Za testiranje ovako definisane hipoteze mogu se koristiti i sledeće dve test statistike.

Wald statistika :

$$X_W = \left(\widehat{\mathbb{b}}_q - \mathbb{b}^*_q\right)' \frac{\partial^2 l\left(b^*_{1:q}, \hat{b}_{q+1:p}\right)}{\partial \mathbb{b} \partial \mathbb{b}'} \left(\widehat{\mathbb{b}}_q - \mathbb{b}^*_q\right)$$

Score statistika :

$$X_S = \left[\frac{\partial l\left(b^*_{1:q}, \hat{b}_{q+1:p}\right)}{\partial \mathbb{b}}\right]' \left[-\frac{\partial^2 l\left(b^*_{1:q}, \hat{b}_{q+1:p}\right)}{\partial \mathbb{b} \partial \mathbb{b}'}\right]^{-1} \frac{\partial l\left(b^*_{1:q}, \hat{b}_{q+1:p}\right)}{\partial \mathbb{b}}$$

Pod pretpostavkom da je H_0 tačna i da $\widehat{\mathbb{b}}$ aproksimativno ima višedimenzionu normalnu raspodelu, svaka od ove tri statistike asimptotski ima hi-kvadrat raspodelu sa q stepeni slobode.

Za dati prag značajnosti α , H_0 se odbacuje ako je $X_L > \chi^2_{q,\alpha}$ kada koristimo log verovatnosnu količnik statistiku ; ili ako je $X_W > \chi^2_{q,\alpha/2}$ ili $X_W < \chi^2_{q,1-\alpha/2}$, kada koristimo Wald statistiku ; ili kada se koristi score statistika ako je $X_S > \chi^2_{q,\alpha/2}$ ili $X_S > \chi^2_{q,\alpha/2}$.

Obično koristimo ovakvo testiranje prilikom biranja nekog od dva modela, gde jedan ima dodatni parametar, možda interakciju kovarijanti, za čiji koeficijent se pitamo da li je jednak nuli.

Da bi testirali da sve kovarijante nemaju efekta na rizik, nulta hipoteza je da su svi koeficijenti $\mathbb{b}$ kovarijanti jednaki nuli, H_0: $\mathbb{b} = 0$. Napomenula bih samo da se u R-u testira hipoteza da su svi koeficijenti jednaki nekim pretpostavljenim vrednostima, standardno nulama. Ova tri testa obično daju različite p-vrednosti.

Procedure selekcija kovarijanti

Sledeće metode se mogu koristiti za izbor optimalnog podskupa kovarijanti u smislu da izabrani podskup ima statistički najznačajnije efekte na vreme preživljavanja među svim podskupovima kovarijanti.

Postupak izbora unapred

Postupak izbora unapred je proces u kome se jedna kovarijanta bira i dodaje u model pri svakom koraku. Prvo se moraju oceniti koeficijenti odgovarajućih kovarijanti koje su već u modelu.

Zatim izračunavamo odgovarajuće hi-kvadrat statistike svake promenljive koja nije u modelu i uočavamo najveću od njih. Ako je najveća hi-kvadrat statistika značajna za dati prag značajnosti α (obično je $\alpha = 0.15$), odgovarajuća kovarijanta se dodaje u model.

Neka je $\mathbb{b}_1$ vektor koeficijenata kovarijanti koje su već u modelu i neka je $l(.)$ logaritmovana parcijalna funkcija verodostojnosti. Postupak napred izbora će izabrati kovarijantu x_j, koja nije još u modelu, da uđe u model ako je razlika između vrednosti $l(.)$ sa x_j i bez x_j najveća među svim x_k-ovima koji nisu u modelu. Odnosno, ako koeficijent b_j od x_j zadovoljava

$$X_L = 2\left[l\left(\hat{b}_j, \hat{\mathbb{b}}_1\right) - l(\hat{\mathbb{b}}_{1j}(0))\right] = \max_k \left\{2\left[l\left(\hat{b}_k, \hat{\mathbb{b}}_1\right) - l(\hat{\mathbb{b}}_{1k}(0))\right], za\ bilo\ koje\ x_k\ koje\ nije\ u\ modelu\right\}$$

i $X_L > \chi^2_{1,\alpha}$, gde je b_k koeficijent od x_k koji još nije u modelu, $\left(\hat{b}_k, \hat{\mathbb{b}}_1\right)$ je OMPV od $(b_k, \mathbb{b}_1)$, $\hat{\mathbb{b}}_{1k}(0)$ je OMPV od $\mathbb{b}_1$ ako je dato da je $b_k = 0$ i $\chi^2_{1,\alpha}$ je α-nivoa kritična tačka hi-kvadrat raspodele sa jednim stepenom slobode. U ovoj proceduri, jednom kad kovarijanta uđe u model, nikada se ne uklanja. Proces se ponavlja dok nijedna od preostalih kovarijanti ne zadovoljava određeni prag za ulazak ili dok unapred određeni broj kovarijanti ne uđe u model.

Postupak izbora unazad

Postupak izbora unazad je proces u kome su sve kovarijante u početku uključene u model, a zatim se uklanjaju jedna po jedna prema kriterijumu značajnosti. Prvo se ocenjuju koeficijenti svih kovarijanti. Zatim se Wald testom ispituje svaka kovarijanta. Najmanje značajna kovarijanta koja ne ispunjava prag značajnosti α (obično, $\alpha = 0.15$) za opstanak u modelu se uklanja. Odnosno, kovarijanta x_j će biti uklonjena iz modela ako

$$X_W = \frac{\hat{b}_j^2}{v_{jj}^2} = \min_k \left\{ \frac{\hat{b}_k^2}{v_{kk}^2}\ za\ bilo\ koju\ x_k\ koja\ je\ u\ modelu \right\}$$

i $X_W \leq \chi^2_{1,\alpha}$, gde je b_j odgovarajući koeficijent od x_j i v_{jj} ocenjena disperzija od $\hat{b}_j$. U ovoj proceduri, jednom uklonjena kovarijanta iz modela, ostaje isključena. Proces se ponavlja sve dok sve preostale kovarijante u modelu ne zadovolje određeni prag značajnosti za opstanak ili dok prethodno određeni broj kovarijanti ne ostane u modelu.

Postupak postupne selekcije

Postupak postupne selekcije je kombinacija procedura izbora unapred i unazad. U početku ona je slična postupku selekcije unapred; međutim, kovarijante koje su već u modelu ne moraju obavezno i ostati. Kovarijante koje su već u modelu mogu biti uklonjene kasnije ako nisu više značajne. Korak po korak proces selekcije prestaje ako nema značajnih kovarijanti koje bi se dodale u model ili ukoliko je kovarijanta koja je upravo ušla u model uklonjena i više nijedna kovarijanta se ne može dodati.

Opisaćemo detaljnije algoritam postupne selekcije, pre svega zbog kasnije primene u R-u.

Nulti korak. Pretpostavimo da postoji p mogućih kovarijanti označenih sa x_j, $j = \overline{1,p}$. Primenjujemo test log verovatnosnog količnika na svaku kovarijantu poredeći model koji sadrži x_j sa model koji nema kovarijante. Nalazimo sledeće test statistike i p-vrednosti testa:

$$X_L^{(0)}(j) = 2\left[l^{(0)}(j) - l(0)\right] \quad i \quad p^{(0)}(j) = P\left\{\chi^2(1) \geq X_L^{(0)}(j)\right\} \quad j = \overline{1,p}$$

gde je $l^{(0)}(j)$ logaritam parcijalne funkcije verodostojnosti modela sa kovarijantom x_j , $l(0)$ logaritam parcijalne funkcije verodostojnosti modela bez kovarijanti. Kandidat za ulazak u model u prvom koraku je najznačajnija kovarijanta i označena je sa x_{e_1}, gde je $p^{(0)}(e_1) = \min_j\{p^{(0)}(j)\}$.

Kovarijanta x_{e_1} ulazi u model ako je njena p-vrednost manja od nekog prethodno izabranog praga značajnosti, u oznaci p_E. Ako je kovarijanta izabrana za ulazak značajna ($p^{(0)}(e_1) < p_E$), tada procedura prelazi na prvi korak, u suprotnom se završava.

Prvi korak. Ovaj korak počinje sa promenljivom x_{e_1} u modelu. Zatim se prave novi PH modeli (svaki model uključuje po jednu preostalu kovarijantu zajedno sa x_{e_1}) , koji se koriste za izračunavanje log verovatnosne test statistike i p-vrednosti :

$$X_L^{(1)}(j) = 2\left[l^{(1)}(j) - l\left(x_{e_1}\right)\right] \quad i \quad p^{(1)}(j) = P\left\{\chi^2(1) \geq X_L^{(1)}(j)\right\} \quad j = \overline{1,p}, j \neq e_1$$

Promenljiva izabrana za kandidata koji može da uđe u model u drugom koraku je x_{e_2} , gde je $p^{(1)}(e_2) = \min_{j \neq e_1}\{p^{(1)}(j)\}$. Ako je izabrana kovarijanta značajna ($p^{(1)}(e_2) < p_E$), tada procedura prelazi na drugi korak, u suprotnom se završava.

Drugi korak. Ovaj korak počinje sa x_{e_1} i x_{e_2} u modelu. Tokom ovog koraka dve različite provere se dešavaju. Prvo proveravamo da li kovarijanta x_{e_1} još uvek doprinosi modelu nakon dodavanja x_{e_2} u model. Biramo prag značajnosti za tu proveru, u oznaci p_R, tako da je $p_R > p_E$ da bi eliminisali mogućnost ulaska i izlaska iste promenljive u beskrajnom broju uzastopnih koraka. Pretpostavimo da je promenljiva koja je uključena u model u prvom koraku još uvek značajna.

Sada ponovo pravimo $p - 2$ novih PH modela (svaki uključuje po jednu novu kovarijantu zajedno sa x_{e_1} i x_{e_2}), i izračunavamo log verovatnosne test statistike i p-vrednosti :

$$X_L^{(2)}(j) = 2\left[l^{(2)}(j) - l\left(x_{e_1}, x_{e_2}\right)\right] \quad i \quad p^{(2)}(j) = P\left\{\chi^2(1) \geq X_L^{(2)}(j)\right\} \quad j = \overline{1,p}, j \neq e_1, e_2$$

Kovarijanta x_{e_3} izabrana za ulazak u model u trećem koraku je ona sa najmanjom p-vrednošću. Procedura prelazi na treći korak ako je $p^{(2)}(e_3) < p_E$, u suprotnom se završava.

Treći korak. Ako je dostignut ovaj korak, procesom eliminacije prvo proveravamo da li su kovarijante, koje su ušle u model u ranijim koracima, još uvek značajne. Zatim, primenjujemo postupak izbora kovarijante identičan onom iz ranijih koraka. Ova procedura se nastavlja do poslednjeg K-tog koraka.

K-ti korak. U ovom koraku su sve kovarijante u modelu i nijedna se ne može isključiti iz njega, ili svaka kovarijanta koja nije u modelu ima $p^{(K)}(j) > p_E$.

Ocene funkcije preživljavanja sa kovarijantama

U Cox-ovom PH modelu funkcija preživljavanja sa kovarijantama x_j je data sa

$$S(t|\mathbb{x}) = [S_0(t)]^{exp\left(\sum_{j=1}^{p} b_j x_j\right)}$$

Jednom kada ocenimo regresione koeficijente b_j, ostaje nam samo da ocenimo osnovnu funkciju preživljavanja, $S_0(t)$. Na osnovu ocenjene osnovne funkcije preživljavanja možemo jednostavno da ocenimo verovatnoću da subjekat živi duže od datog vremenskog trenutka ako je dat skup kovarijanti $x_1, \dots, x_p$.

Uz pretpostavku da je osnovna funkcija rizika konstantna između svakog para posmatranih uzastopnih vremena neuspeha, Breslow je predložio sledeću ocenu osnovne kumulativne funkcije rizika :

$$\widehat{H}_0(t) = \sum_{t_{(i)} \le t} \frac{d_i}{\sum_{l \in R(t_{(i)})} \exp\left(\mathbb{x}_l' \widehat{\mathbb{b}}\right)}$$

Na osnovu poznatog odnosa, osnovna funkcija preživljavanja se može oceniti sa

$$\hat{S}_0(t) = exp\left[-\widehat{H}_0(t)\right] = \prod_{t_{(i)} \le t} \left\{ exp\left[- \frac{d_i}{\sum_{l \in R(t_{(i)})} \exp\left(\mathbb{x}_l' \widehat{\mathbb{b}}\right)} \right] \right\}$$

i funkcija preživljavanja subjekta sa skupom kovarijanti $\mathbb{x} = \left(x_1, \dots, x_p\right)$ je

$$\hat{S}(t|\mathbb{x}) = \left[\hat{S}_0(t)\right]^{exp\left(\sum_{j=1}^{p} \hat{b}_j x_j\right)} = \left[\hat{S}_0(t)\right]^{exp\left(\widehat{\mathbb{b}}' \mathbb{x}\right)}$$

Uz slabe pretpostavke, $\hat{S}(t|\mathbb{x})$ ima asimptotski normalnu raspodelu sa očekivanjem $S(t|\mathbb{x})$. Kako je $S(t|\mathbb{x}) = exp[-H(t|\mathbb{x})]$ ocena disperzija od $\hat{S}(t|\mathbb{x})$ je

$$\widehat{D}\left(\hat{S}(t|\mathbb{x})\right) \simeq \left[\hat{S}(t|\mathbb{x})\right]^2 \widehat{D}\left(\widehat{H}(t|\mathbb{x})\right)$$

Asimptotski interval poverenja za funkciju preživljavanja je

$$\left(\hat{S}(t|\mathbb{x}) - Z_{1-\alpha/2} \sqrt{\widehat{D}\left(\hat{S}(t|\mathbb{x})\right)}, \quad \hat{S}(t|\mathbb{x}) + Z_{1-\alpha/2} \sqrt{\widehat{D}\left(\hat{S}(t|\mathbb{x})\right)} \right)$$

gde je $Z_{1-\alpha/2}$ $1-\alpha/2$-i kvantil standardne normalne raspodele.

Kalbfleisch i Prentice su predložili alternativnu ocenu po kojoj je osnovna funkcija preživljavanja ocenjena tako da bude stepenasta funkcija i

$$\hat{S}_0(t) = \prod_{j=0}^{i-1} \hat{\alpha}_j \qquad t_{(i-1)} < t \leq t_{(i)}, \quad i = 1, \dots, k+1$$

gde je $\hat{\alpha}_0 = 1$ i $\hat{\alpha}_1, \dots, \hat{\alpha}_k$ su rešenja sledećih k jednačina :

$$\sum_{j \in u^*_{(i)}} \frac{\exp(\mathbb{x}'_j \hat{\mathbb{b}})}{1 - \hat{\alpha}_i^{\exp(\mathbb{x}'_j \hat{\mathbb{b}})}} = \sum_{l \epsilon R(t_{(i)})} \exp(\mathbb{x}'_l \hat{\mathbb{b}}) \qquad i = 1, \dots, k$$

Kada nema istih vrednosti,

$$\hat{\alpha}_i = \left[1 - \frac{\exp(\mathbb{x}'_{(i)} \hat{\mathbb{b}})}{\sum_{l \epsilon R(t_{(i)})} \exp(\mathbb{x}'_l \hat{\mathbb{b}})}\right]^{\exp(-\mathbb{x}'_{(i)} \hat{\mathbb{b}})} \qquad i = 1, \dots, k$$

i

$$\hat{S}_0(t) = \prod_{j=0}^{i-1} \left[1 - \frac{\exp(\mathbb{x}'_{(j)} \hat{\mathbb{b}})}{\sum_{l \epsilon R(t_{(j)})} \exp(\mathbb{x}'_l \hat{\mathbb{b}})}\right]^{\exp(-\mathbb{x}'_{(j)} \hat{\mathbb{b}})} \qquad t_{(i-1)} < t \leq t_{(i)}, \quad i = 1, \dots, k+1$$

Prema tome,

$$\hat{S}(t|\mathbb{x}) = \left[\hat{S}_0(t)\right]^{exp(\hat{\mathbb{b}}' \mathbb{x})}$$

Uz slabe pretpostavke, Kalbfleisch i Prentice ocena $\hat{S}(t|\mathbb{x})$ takođe ima asimptotski normalnu raspodelu sa očekivanjem $S(t|\mathbb{x})$ i disperzijom koja se može oceniti.

Procena adekvatnosti Cox-ovog PH modela

Opravdanost statističkih zaključaka koji dovode do identifikacije bitnih prognostičkih faktora u velikoj meri zavisi od adekvatnosti izabranog modela. Potrebno je proceniti adekvatnost Cox-ovog PH modela, odnosno pretpostavku o proporcionalnosti rizika i goodness-of-fit (dobrotu pristajanja). U narednom delu objasnićemo nekoliko metoda za datu procenu.

Grafičke metode za procenu opravdanosti pretpostavke o proporcionalnim rizicima

Postoje dva grafička pristupa za proveru PH pretpostavke: upoređivanje log-log krivih preživljavanja i upoređivanje posmatranih i očekivanih krivih preživljavanja.

Log-log kriva preživljavanja je transformacija ocenjene krive preživljavanja koja podrazumeva dvostruko prirodno logaritmovanje ocenjene funkcije preživljavanja. Matematički, log-log krive pišemo kao $-\log\left(-\log \hat{S}(t)\right)$. Primetimo da je logaritam verovatnoće, kao što je $\hat{S}(t)$, uvek negativan broj, pa zato pre drugog logaritmovanja negiramo prvi rezultat. Kao i da $-\log\left(-\log \hat{S}(t)\right)$ kriva uzima vrednosti između $-\infty$ i $+\infty$. Drugi minus postoji kako bi krive opadale sa vremenom. Pokazaćemo da PH pretpostavka može biti ocenjena pomoću procene da li su log-log krive paralelne.

Cox-ova PH funkcija rizika je

$$h(t|\mathbb{x}) = h_0(t)exp\left(\sum_{j=1}^{p} b_j x_j\right)$$

i njegova odgovarajuća funkcija preživljavanja

$$S(t|\mathbb{x}) = [S_0(t)]^{exp\left(\sum_{j=1}^{p} b_j x_j\right)}$$

gde je $S_0(t)$ osnovna funkcija preživljavanja koja odgovara osnovnoj funkciji rizika $h_0(t)$.

$$\log S(t|\mathbb{x}) = exp\left(\sum_{j=1}^{p} b_j x_j\right)\log S_0(t)$$

$$\log(-\log S(t|\mathbb{x})) = \log\left[-exp\left(\sum_{j=1}^{p} b_j x_j\right)\log S_0(t)\right] = \sum_{j=1}^{p} b_j x_j + \log[-\log S_0(t)]$$

$$-\log(-\log S(t|\mathbb{x})) = -\sum_{j=1}^{p} b_j x_j - \log[-\log S_0(t)]$$

Pretpostavimo da su $\mathbb{x}_1 = (x_{11}, \ldots, x_{1p})$ i $\mathbb{x}_2 = (x_{21}, \ldots, x_{2p})$ odgovarajuće kovarijante dva subjekta. Odgovarajuće log-log krive ova dva subjekta su

$$-\log(-\log S(t|\mathbb{x}_1)) - (-\log(-\log S(t|\mathbb{x}_2))) = \sum_{j=1}^{p} b_j(x_{2j} - x_{1j})$$

odnosno

$$-\log(-\log S(t|\mathbb{x}_1)) = -\log(-\log S(t|\mathbb{x}_2)) + \sum_{j=1}^{p} b_j(x_{2j} - x_{1j})$$

Na osnovu gornjeg izraza dolazimo do zaključka da, ukoliko se Cox-ov PH model koristi, grafici ocenjenih log-log krivih preživljavanja će biti aproksimativno paralelni. Razlika između ove dve krive je linearni izraz koji uključuje razlike u vrednostima prognostičkih promenljivih, koje ne zavise od vremena. Uopšteno, ukoliko je vertikalna razlika između dve krive konstantna, krive su paralelne. Ukoliko je PH model odgovarajući za dati skup prognostičkih faktora, treba očekivati da empirijski grafici log-log krivih za različite subjekte budu aproksimativno paralelni. Pod empirijskim graficima podrazumevamo prikaz log-log krivih preživljavanja zasnovan na KM ocenama, ne na Cox-ovom PH modelu. Mogu se prikazati i prilagođene log-log krive prognostičkim faktorima za koje se već pretpostavlja da zadovoljavaju PH pretpostavku, ali se ne uključuju oni faktori koji se ocenjuju.

Glavni problem kod ove grafičke metode jeste kako odlučiti „koliko paralelno je paralelno“. Ova odluka može biti jako subjektivna, pogotovo ukoliko je obim uzorka relativno mali. Preporuka je da se u odlučivanju koristi konzervativna strategija, odnosno da pretpostavljamo da je PH pretpostavka zadovoljena, osim ako ne postoji jak dokaz neparalelnosti log-log krivih. Drugi problem se tiče kategorizacije neprekidne promenljive. Ukoliko postoji puno kategorija, podaci se „stanjuju“ u svakoj, što otežava upoređivanje različitih krivih. Takođe, različite kategorizacije u isti broj grupa mogu dati različiti grafički prikaz. Kod kategorizacije neprekidnih promenljivih preporučuje se da broj kategorija bude razumno mali (npr. dve ili tri) ako je moguće, i da izbor kategorija bude što je moguće smisleniji i da obezbeđuje razuman balans brojeva.

Javlja se i problem kako proceniti PH pretpostavku za nekoliko promenljivih istovremeno. Jedna strategija je kategorizacija svih promenljivih posebno, formiranje kombinacija kategorija, i zatim poređenje log-log krivih za sve kombinacije na istom grafiku. Mana ove strategije jeste da će se podaci sve više smanjivati kako broj kombinacija postaje čak i umereno veliki. Takođe, čak i ako imamo dovoljan broj podataka za svaku od kombinovanih kategorija, često je teško odrediti koja

promenljiva je odgovorna za neparalelnost koju možemo uočiti. Alternativna strategija za posmatranje nekoliko prognostičkih faktora istovremeno jeste procena PH pretpostavke za jedan prognostički faktor, a prilagođen za druge za koje pretpostavljamo da zadovoljavaju pretpostavku. Ovo podrazumeva poređenje prilagođenih log-log krivih. U izračunavanju prilagođenih krivih preživljavanja, stratifikujemo podatke po kategorijama prognističkog faktora za koji želimo da proverimo pretpostavku, podešavamo PH model za svaki stratum i zatim izračunavamo prilagođene verovatnoće preživljavanja koristeći ukupnu prosečnu vrednost ostalih prognostičkih faktora za koje smo pretpostavili da je pretpostavka zadovoljena.

Druga grafička metoda za proveru PH pretpostavke jeste upoređivanje posmatrane sa očekivanom krivom preživljavanja. Ovu metodu ćemo primeniti procenjujući PH pretpostavku za jednu po jednu kovarijantu. Ukoliko je kovarijanta za koju želimo da procenimo PH pretpostavku kategorička, prvo stratifikujemo podatke pomoću kategorija te kovarijante. Zatim, dobijamo posmatrane krive preživljavanja primenjujući KM ocene posebno za svaku kategoriju. Očekivane krive preživljavanja dobijamo pod pretpostavkom da je Cox-ov PH model ispravan, tj. da su rizici zaista proporcionalni, podešavanjem Cox-ovog PH modela za tu kovarijantu. Zamenjujemo vrednost svake kategorije promenljive u formuli za procenu krive preživljavanja (u Cox-ovom PH modelu) , i tako dobijamo ocenjene krive preživljavanja za svaku kategoriju posebno.

Da bi uporedili ove krive smeštamo ih na isti grafik. Ako su za svaku kategoriju kovarijante za koju procenjujemo PH pretpostavku, posmatrane i očekivane krive „blizu“ jedna drugoj, možemo zaključiti da je PH pretpostavka zadovoljena. Ako su, međutim, za jednu ili više kategorija posmatrane i očekivane krive protivrečne, zaključujemo da je narušena PH pretpostavka. Očigledna mana ove grafičke metode leži u odluci „koliko blizu je blizu“ prilikom upoređivanja krivih za date kategorije kovarijante. Preporuka je da se smatra da je PH pretpostavka narušena samo kada su posmatrane i očekivane krive jako protivrečne.

Kada koristimo ovu grafičku metodu za procenu PH pretpostavke za neprekidne promenljive, posmatrane krive dobijamo na isti način kao kod kategoričkih promenljivih, formiranjem stratuma na osnovu kategorija neprekidne promenljive i izvođenjem KM krivih za svaku kategoriju. Za dobijanje očekivanih krivih moguće su dve opcije. Jedna opcije jeste korišćenje Cox-ovog PH modela koji sadrži $k-1$ indikator promenljivih koje ukazuju na k kategorija kovarijante. Očekivana kriva za datu kategoriju se dobija kao prilagođena kriva preživljavanja zamenom vrednosti indikator promenljivih koje definišu datu kategoriju u formulu za ocenjivanje krive preživljavanja.

$$h(t|\mathbb{x}) = h_0(t)exp\left(\sum_{i=1}^{k-1} b_i X_i\right)$$

predstavlja Cox-ov PH model sa $k-1$ indikator promenljivih X_i za k kategorija. Dok je dobijena prilagođena kriva preživljavanja za kategoriju c

$$\hat{S}(t|\mathbb{X}_c) = \left[\hat{S}_0(t)\right]^{exp\left(\sum_{i=1}^{k-1} \hat{b}_i X_{ci}\right)}$$

gde $\mathbb{X}_c = (X_{c1}, X_{c2}, \dots, X_{c2})$ predstavlja vektor vrednosti indikator promenljivih za kategoriju c.

Druga opcija je korišćenje Cox-ovog PH modela koji sadrži datu neprekidnu kovarijantu. Očekivane krive se dobijaju kao prilagođene krive preživljavanja određivanjem vrednosti kovarijante koje se razlikuju po kategorijama, na primer, uključivanjem srednjih vrednosti kovarijante svake kategorije u podešen model. Odnosno, koristimo sledeći PH model

$$h(t|X) = h_0(t)exp(bX)$$

Očekivana kriva preživljavanja za kategoriju c

$$\hat{S}(t|\bar{X}_c) = \left[\hat{S}_0(t)\right]^{exp(\hat{b}\bar{X}_c)}$$

gde je $\bar{X}_c$ srednja vrednost promenljive X u kategoriji c.

Primetimo da se i kod ove grafičke metode se javlja problem kategorizacije neprekidnih promenljivih. Preporuke za rešavanje ovog problema su iste kao kod log-log krivih. Takođe, kada želimo da procenimo PH pretpostavku za nekoliko promenljivih istovremeno, postupak je sličan onom kod log-log krivih.

Reziduali

Postoje grafičke metode, zasnovane na rezidualima, koje se često koriste kao dijagnostički alat. U nastavku uvešćemo sledeće tipove reziduala : Cox-Snell, martingalne reziduale, reziduale odstupanja i Schoenfeld reziduale. Oni se mogu grafički prikazati u odnosu na vreme preživljavanja ili kovarijantu. Obrazac na grafiku pruža informacije o adekvatnosti Cox-ovog PH modela. Grafik takođe pruža informacije o autlajerima i drugim obrascima. Slično drugim grafičkim metodama, tumačenje grafičkih prikaza reziduala može biti subjektivno.

Cox-Snell rezidual, R_i , za i -ti subjekat sa posmatranim vremenom preživljavanja t_i i vrednostima kovarijanti $\mathbb{x}_i$ je definisan sa $R_i = -\log\left(\hat{S}(t_i, \mathbb{x}_i)\right)$, što predstavlja ocenjen akumulirani rizik zasnovan na Cox-ovom PH modelu. Ako je posmatrano t_i cenzurisano, onda je odgovarajući R_i takođe cenzurisan. Ako je Cox-ov PH model odgovarajući, grafički prikaz R_i i njegove KM ocene funkcije preživljavanja $\left(\hat{S}(R)\right)$ treba da izgleda kao prava linija pod uglom od 45^0.

R_{M_i} martingalni rezidual (Fleming i Harrington, 1991.) za i-ti subjekat su definisani sa

$$R_{M_i} = \delta_i - R_i \qquad i = 1, \dots, n$$

gde je δ_i statusna promenljiva, a R_i Cox-Snell rezidual i-tog subjekata.

Martingalni reziduali se koriste za ispitivanje najbolje funkcionalne forme neprekidne kovarijante od interesa koristeći pretpostavljeni Cox-ov PH model za ostale kovarijante. Neka je vektor kovarijanti $\mathbb{x}$ podeljen na vektor $\mathbb{x}_*$ kovarijanti čiji funkcionalni oblik znamo i jednu neprekidnu kovarijantu x za koju nismo sigurni koju funkcionalnu formu da koristimo. Pretpostavljamo da je x nezavisna od $\mathbb{x}_*$. Sa $g(.)$ označimo najbolju funkciju od x za objašnjavanje njenog uticaja na preživljavanje. Tada je Cox-ov PH model

$$h(t|\mathbb{x}) = h_0(t)\exp(\mathbb{b}_*{}'\mathbb{x}_*)\exp(g(x))$$

gde je $\mathbb{b}_*$ $p-1$ dimezioni vektor koeficijenata. Da bi našli $g(.)$, primenjujemo Cox-ov PH model na podatke zasnovane na $\mathbb{x}_*$ i izračunavamo martingalne reziduale R_{M_i}, $i = 1, \dots, n$. Ovi reziduali se zatim grafički predstavljaju u odnosu na vrednosti kovarijante x za sve subjekte ($x_i, i = 1, \dots, n$). Obično se glatka kriva podešava na rasejani grafik, i ukazuje na funkciju $g(.)$. Ako je ovaj grafik linearan, nije potrebna trasformacija kovarijante x.

Suma martingalnih reziduala jednaka je nuli. Za velike uzorke, martingalni reziduali imaju pomerenu raspodelu sa očekivanjem nula. R_{M_i} uzima vrednosti od $-\infty$ do 1, pa iz toga sledi da nisu simetrični oko nule.

Reziduali odstupanja (1990.) su definisani sa

$$R_{D_i} = sgn\big(R_{M_i}\big)\sqrt{2\big[-R_{M_i} - \delta_i \log\big(\delta_i - R_{M_i}\big)\big]} \qquad i = 1,2, \dots, n$$

gde je $sgn(.)$ znakovna funkcija, koja uzima vrednost 1 ako je argument pozitivan, 0 ako je nula, -1 ako je negativan, a R_{M_i} je martingalni rezidual i-tog subjekta.

Reziduali odstupanja imaju očekivanje nula za velike uzorke i simetrično su raspoređeni oko nule ako je model adekvatan. Reziduali odstupanja su pozitivni za subjekte koji žive kraće od očekivanog i negativni za one koji žive duže. Reziduali odstupanja se često koriste u proceni goodness-of-fit Cox-ovog PH modela. Potencijalni autlajeri odgovaraju velikim apsolutnim vrednostima reziduala odstupanja.

Schoenfeld rezidual za j-tu kovarijantu i-tog subjekta sa posmatranim vremenom preživljavanja t_i je

$$R_{ji} = \delta_i \left[x_{ji} - \frac{\sum_{l \in R(t_i)} x_{jl} \exp\left(\widehat{\mathbb{b}}'\mathbb{x}_l\right)}{\sum_{l \in R(t_i)} \exp\left(\widehat{\mathbb{b}}'\mathbb{x}_l\right)} \right], \quad j = \overline{1, p} \quad , \quad i = \overline{1, n}$$

gde je $\widehat{\mathbb{b}}$ ocena maksimalne parcijalne verodostojnosti od $\mathbb{b}$ i δ_i statusna promenljiva subjekta i. Schoenfeld reziduali su definisani samo u necenzurisanim vremenima preživljavanja; cenzurisane opservacije se posmatraju kao nestale. Uočavamo da, ako pretpostavimo da subjekat i postiže događaj u trenutku t_i, Schoenfeld reziduali tog subjekata su jednaki razlici posmatrane

vrednosti kovarijante od interesa i ponderisane srednje vrednosti kovarijante od interesa za druge subjekte koji su još uvek u rizičnom skupu u trenutku t_i. Težine su rizici svakog subjekta.

Suma Schoenfeld reziduala za kovarijantu je nula (ako je PH pretpostavka tačna, $\widehat{\mathbb{b}}$ OMPV). Tako asimptotski, Schoenfeld reziduali imaju očekivanje nula. Takođe, ako je PH pretpostavka tačna ovi reziduali su nezavisni od vremena neuspeha.

Grambsch i Therneau (1994.) su predložili da Schoenfeld reziduali budu ponderisani pomoću inverzne ocenjene kovarijansne matrice od $R_i = \left(R_{1i}, \dots, R_{pi}\right)'$ označene sa $\widehat{D}(R_i)$, odnosno

$$R_i^* = \left[\widehat{D}(R_i)\right]^{-1} R_i$$

Ponderisani Schoenfeld reziduali imaju bolju dijagnostičku moć i češće se koriste od neponderisanih u proceni pretpostavke o proporcionalnosti rizika. Da bi se pojednostavilo izračunavanje, Grambsch i Therneau su predložili aproksimaciju $\left[\widehat{D}(R_i)\right]^{-1}$:

$$\left[\widehat{D}(R_i)\right]^{-1} \simeq r\widehat{D}(\widehat{\mathbb{b}})$$

gde je r broj događaja ili broj posmatranih necenzurisanih vremena preživljavanja i $\widehat{D}(\widehat{\mathbb{b}})$ ocenjena kovarijansna matrica od $\widehat{\mathbb{b}}$. Sa ovom aproksimacijom, ponderisani Schoenfeld reziduali se mogu aproksimirati pomoću

$$R_i^* = r\widehat{D}(\widehat{\mathbb{b}})R_i$$

Grafički prikazi (ponderisanih) Schoenfeld reziduala u odnosu na vreme preživljavanja ili kovarijantu se mogu koristiti za proveru adekvatnosti Cox-ovog PH modela. Prisustvo određenih obrazaca u ovim graficima može ukazivati na narušavanje pretpostavke o proporcionalnosti rizika, dok ekstremna odstupanja od glavne grupe ukazuje na moguće autlajere i potencijalne probleme stabilnosti modela.

Jako je bitno što i formalno možemo proveriti adekvatnost Cox-ovog PH modela pomoću (ponderisanih) Schoenfeld reziduala. Ideja iza ovog statističkog testa je da ako PH pretpostavka važi za određenu kovarijantu, onda Schoenfeld reziduali za tu kovarijantu neće biti povezani sa vremenima neuspeha. Kleinbaum i Klein (2005) su predložili korišćenje rangova vremena, pre nego samih vremena neuspeha.

Nultu hipotezu je da su reziduali nezavisni od vremena neuspeha. Što povlači da su nekolerisani. Dok obratno ne mora da važi. Grambsch i Therneau (1994.) su predložili hi-kvadrat test za testiranje ove hipoteze, koji nećemo navoditi. Odbacivanje nulte hipoteze dovodi do zaključka da je narušena PH pretpostavka.

Važno je naglasiti da se nulta hipoteza nikad ne dokazuje ovim testom. Najviše što možemo reći jeste da ne postoji dovoljno dokaza za njeno odbacivanje. P-vrednost testa zavisi od obima uzorka. Veliko narušavanje nulte hipoteze ne može biti statistički značajno ako je uzorak veoma

mali. Obratno, malo narušavanje nulte hipoteze može biti jako značajno ako je uzorak jako veliki.
Statistički test nudi objektivni pristup za procenu PH pretpostavke u odnosu na subjektivnost grafičkog pristupa. Međutim, grafički pristup omogućava istraživaču da otkrije specifična mimoilaženja sa PH pretpostavkom. Preporuka je da kada se procenjuje opravdanost PH pretpostavke istraživač koristi i grafički pristup i statistički test pre donošenja konačne odluke.

Procena PH pretpostavke pomoću kovarijanti koje zavise od vremena

Cox-ov PH model pretpostavlja da je hazardni odnos dva subjekta nezavisan od vremena. To zahteva da kovarijante ne budu vremenski zavisne promenljive. Ako se bilo koja kovarijanta menja tokom vremena, pretpostavka o proporcionalnosti rizika je narušena. Ova činjenica se može iskoristiti za testiranje pretpostavke uključujući vremenski zavisnu kovarijantu u model i testirajući da li je koeficijent te kovarijante značajno različit od nule. Oblik proširenog Cox-ovog modela je

$$h(t|\mathbb{x}) = h_0(t)exp\left(\sum_{i=1}^{p} b_i x_i + c_i x_i g_i(t)\right)$$

gde je $g_i(t)$ funkcija koja zavisi od vremena za i-tu kovarijantu i c_i parametar i-te interakcije (proizvoda $x_i g_i(t)$) . Možemo testirati pretpostavku o proporcionalnosti rizika testiranjem $H_0 : c_i = 0\,,\ i = \overline{1,p}$. Procedure koje se mogu koristiti za testiranje ove hipoteze su slične onima objašnjenim ranije. Na primer, možemo koristiti sledeću log verovatnosnu količnik statistiku

$$X_L = 2\left[\log L_{proširenog\ CM} - \log L_{\ CPHM}\ \right]$$

koja ima hi-kvadrat raspodelu sa p stepeni slobode ako je H_0 tačna. Sa dodatim članovima, parcijalna funkcija verodostojnosti uopštenog Cox-ovog modela postaje komplikovanija. Srećom, dostupni su kompjuterski softveri za izvršavanje kalkulacija. Ako ne odbacimo H_0 , model se svodi na običan Cox-ov PH model, odnosno zaključujemo da je pretpostavka o proporcionalnosti rizika zadovoljena. Ako odbacimo H_0 , zaključujemo da je narušena pretpostavka o proporcionalnosti rizika za najmanje jednu kovarijanti iz modela. Da bi odredili koja kovarijanta ne zadovoljava PH pretpostavku nastavljamo proceduru eliminacijom unazad neznačajnih interakcija sve dok ne dođemo do konačnog modela.

Različite kovarijante mogu zahtevati različite funkcije $g_i(t)$. Neki od mogućih izbora za $g_i(t)$ su

- $g_i(t) = t$
- $g_i(t) = \log t$
- $g_i(t) = \begin{cases} 1 & t \geq t_0 \\ 0 & t < t_0 \end{cases}$

Primetimo da je izbor $g_i(t)$ subjektivan. Na odluku o izboru funkcije može da utiče grafička procena PH pretpostavke, ali ona se i dalje oslanja na subjektivnu procenu istraživača. Upravo je to nedostatak ovog pristupa u odnosu na Schoenfeld reziduale.

Provera skale neprekidnih kovarijanti

Ne postoji razlog zašto bi neprekidna kovarijanta bila modelirana sa exp (bx), a ne sa, na primer, exp $(b \log x)$, exp $(b\sqrt{x})$ ili exp (bx^2). Ove vrste modela se lako grade, na primer u R-u, ali zato donose nove probleme: ako razmatramo različite modele za neku neprekidnu kovarijantu, koji treba da koristimo?

Želeli bismo da procenimo koji od konkurentnih modela za „skalu" kovarijante najbolje odgovara podacima. Kada bismo samo razmatrali modele koji su ugneždeni jedan u drugi, kao $h_0(t)$exp (bx) koji je ugnežden u $h_0(t)$exp $(b_1x + b_2x^2)$, mogli bi da koristimo test log verovatnosnog količnika, objašnjen ranije. Međutim, ovaj način testiranja nije odgovarajući za neugneždene modele, kao što je slučaj kada upoređujemo $h_0(t)$exp (bx) i $h_0(t)$exp $(b \log x)$. Potrebna nam je alternativa.

Akaike-ov informacioni kriterijum (AIC) je metoda za biranje između dva ili više konkurentnih modela, moguće sa različitim brojem parametara. On balansira štedljivost sa goodness-of-fit kažnjavajući modele koji imaju veći broj parametara.

AIC za model m sa p_m dimenzionim vektorom parametara θ_m je definisan sa

$$\text{AIC}(m) = 2p_m - 2 \log L(\hat{\theta}_m)$$

gde je L funkcija verodostojnosti. Ona se, u slučaju Cox-ovog PH modela, može zameniti parcijalnom funkcijom verodostojnosti.

Preporučuje se onaj model čija je vrednost AIC najmanja. „Standarno" pravilo: ako je razlika vrednosti AIC dva modela manja od 2, onda kao i da ne postoji izbor ; ako ta razlika ide do 7 (po nekima i do 10), onda je jedan mnogo bolji od drugog, a ako je veća od 7 (10), onda je gori model toliko lošiji da ga nije trebalo ni uzeti za razmatranje.

Postoji beskonačan broj funkcija koje se mogu koristiti za skaliranje. Preporuka je da se prvo pokuša sa kvadratom; ako se model pokaže boljim, pokušati sa trećim stepenom. Zatim probati par osnovnih funkcija od x, kao što su $\log x$ ili $\sqrt{x}$.

Popularnost Cox-ovog PH modela

Ključni razlog za popularnost Cox-ovog modela leži u činjenici da iako je osnovna funkcija rizika neodređena, dobre ocene koeficijenata regresije, hazard količnici i prilagođene krive preživljavanja se mogu izvesti za širok spektar podataka. Drugim rečima, Cox-ov model je „čvrst“ model. Rezultati dobijeni upotrebom Cox-ovog modela su veoma približni rezultatima tačnog parametarskog modela.

U principu, uvek koristimo parametarski model ukoliko smo sigurni u pravilnost modela. Postoje različite metode za procenu prednosti korišćenja parametarskog modela, ali nikada ne možemo biti potpuno sigurni da je dati parametarski model prikladan. Baš iz tog razloga što često dolazimo u nedoumicu, biramo Cox-ov model jer on daje dovoljno pouzdane rezultate i možemo ga smatrati sigurnim izborom.

Uopšteno gledano „čvrstina“ Cox-ovog modela i njegov specifičan oblik je atraktivan iz nekoliko razloga.

Kao što znamo formula za Cox-ov model je proizvod osnovne funkcije rizika koja zavisi od vremena i eksponencijalnog izraza koji zavisi od prognostičkih promenljivih, ali ne i od vremena. Eksponencijalan deo ove formule je privlačan jer obezbeđuje nenegativne ocene prilagođenog modela. Pošto se po definiciji vrednost bilo koje funkcije rizika mora kretati između 0 i $+\infty$ želimo da i drugi deo formule bude nenegativan. Ukoliko bi umesto eksponencijalnog dela imali lineranu funkciju, mogli bi dobiti negativne ocene rizika što nije dozvoljeno.

Još jedna bitna osobina Cox-ovog modela je to što iako osnovna funkcija rizika nije određena možemo oceniti parametre u eksponencijalnom delu modela, koji su nam potrebni da bi procenili efekat promenljivih od interesa. Mera efekta, koja se zove hazard količnik, se takođe računa bez ocene osnovne funkcije rizika. Primetimo da funkcija rizika i odgovarajuća funkcija preživljavanja mogu biti ocenjene za Cox-ov model, čak iako osnovna funkcija rizika nije određena. To znači da sa Cox-ovim modelom uz minimum pretpostavki možemo dobiti primarne informacije iz analize preživljavanja, a to su hazard količnik i kriva preživljavanja.

Analiza preživljavanja pacijenata obolelih od leukemije

U ovom radu proučavaćemo podatke iz jedne kliničke studije koja posmatra 42 pacijenta koja boluju od leukemije, i prati njihovo vreme u remisiji, u nedeljama, do povratka bolesti. Baza[4] se sastoji od pet promenljivih. Prva promenljiva predstavlja vreme preživljavanja, u nedeljama. Druga je statusna promenljiva. Treća promenljiva je kovarijanta koja predstavlja pol pacijenta (0-muški, 1-ženski). Četvrta je neprekidna kovarijanta i predstavlja broj belih krvnih zrnaca na početku studije. Peta promenljiva je nominalna kovarijanta kodirana tako da 0 označava da su pacijenti primili tretman, a 1 placebo. Analiza ovih podataka preživljavanja teći će analogno teorijskim postavkama u dosadašnjem radu.

Učitavamo bazu i paket *survival* u R-u, i zatim kreiramo objekat preživljavanja, kojim definišemo vremena preživljavanja. Sa + su označena cenzurisana vremena.

```
Surv(Time, Censor)

[1] 35+ 34+ 32+ 32+ 25+ 23  22  20+ 19+ 17+ 16  13  11+ 10+ 10   9+  7   6+  6
[20]  6   6  23  22  17  15  12  12  11  11   8   8   8   8   5   5   4   4   3
[39]  2   2   1   1
```

Sledeća funkcija nam daje Kaplan-Meier ocene i standardne greške te ocene na osnovu Greenwood-ove formule :

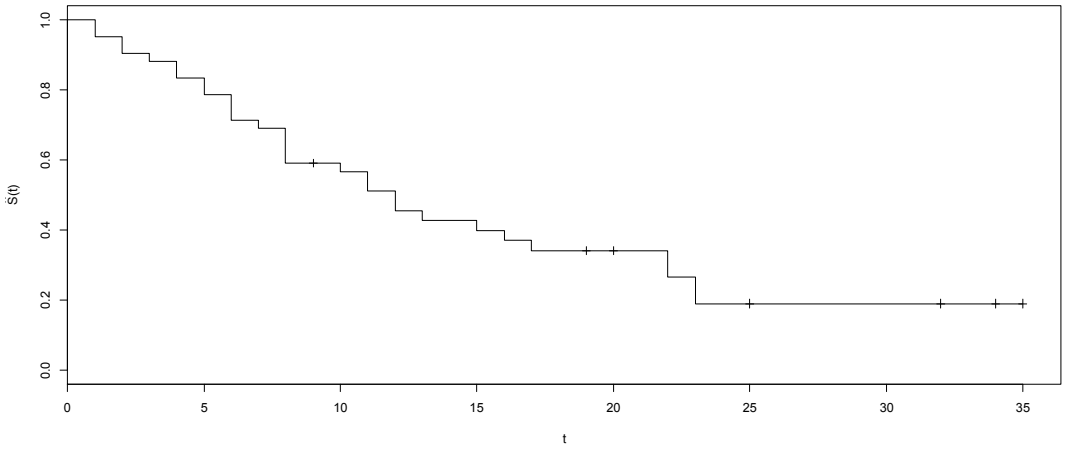
```
leukemia.km<-survfit(Surv(Time, Censor)~1,conf.type = "none", type ="kaplan-meier")
summary(leukemia.km)

Call: survfit(formula = Surv(Time, Censor) ~ 1, conf.type = "none",
    type = "kaplan-meier")
 time n.risk n.event survival std.err
    1     42       2    0.952  0.0329
    2     40       2    0.905  0.0453
    3     38       1    0.881  0.0500
    4     37       2    0.833  0.0575
    5     35       2    0.786  0.0633
    6     33       3    0.714  0.0697
    7     29       1    0.690  0.0715
    8     28       4    0.591  0.0764
   10     23       1    0.565  0.0773
   11     21       2    0.512  0.0788
   12     18       2    0.455  0.0796
   13     16       1    0.426  0.0795
   15     15       1    0.398  0.0791
   16     14       1    0.369  0.0784
```

[4] http://www-stat.stanford.edu/~olshen/hrp262spring01/spring01Assignments/anderson.txt

17	13	1	0.341	0.0774
22	9	2	0.265	0.0765
23	7	2	0.189	0.0710

Pomoću funkcije *plot* dajemo prikaz KM ocena funkcije preživljavanja na osnovu svih vremena preživljavanja, uz napomenu da su na grafiku znakom „+“ označena cenzurisana vremena.

plot(leukemia.km,xlab="t",ylab=expression(hat(S)*"(t)"))



Takođe, možemo izračunati i intervale poverenja za KM ocenu funkcije preživljavanja i grafički ih predstaviti, uz napomenu da su ovde predstavljeni log-log intervali.

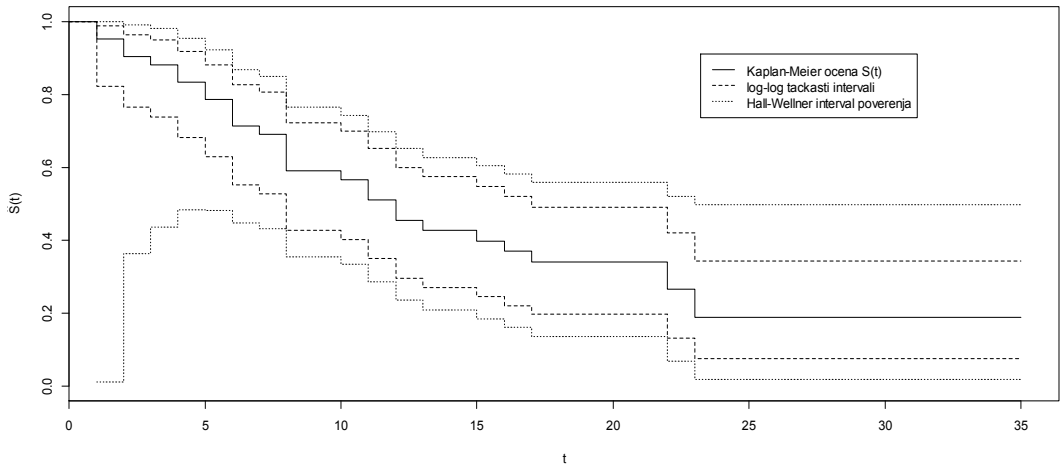
leukemia2.km<-survfit(Surv(Time, Censor)~1, conf.type = "log-log", type ="kaplan-meier")
confBands(Surv(Time, Censor),confType="log-log",confLevel=0.95,type="hall")[5]

Tačka po tačka IP		IP za svako t	
Donja	Gornja	Donja	Gornja
0.8227	0.988	0.01186638	0.9994633
0.7658	0.963	0.36334363	0.9901548
0.7373	0.949	0.43699291	0.9807799
0.6819	0.917	0.48475569	0.9551315
0.6286	0.882	0.48251111	0.9232949
0.5521	0.826	0.44725068	0.8687485
0.5262	0.807	0.43061491	0.8488602
0.4269	0.723	0.35489927	0.7658324
0.4017	0.700	0.33331983	0.7438632
0.3495	0.652	0.28631177	0.6982390
0.2958	0.601	0.23513833	0.6511676
0.2700	0.574	0.20983993	0.6278047

[5] Da bi se koristila ova funkcija neophodno je učitati u R paket OIsurv.

0.2449	0.547	0.18487756	0.6046475
0.2204	0.519	0.16037625	0.5817821
0.1966	0.491	0.13648406	0.5593183
0.1311	0.420	0.06729409	0.5207200
0.0753	0.343	0.01896927	0.4976157
		0.01896927	0.4976157

```
plot(leukemia2.km,xlab="t",ylab=expression(hat(S)*"(t)"),mark.time=F)
lines(confBands(Surv(Time, Censor),confType="log-log",confLevel=0.95,type="hall"), lty=3)
legend(locator(n=1), legend=c("Kaplan-Meier ocena S(t)","log-log tackasti intervali","Hall-Wellner interval poverenja"), lty=1:3)
```



Da bi izračunali tačkaste ocene kvantila date formulom (6), i njihove intervalne ocene, definišemo u R-u sledeću funkciju u skladu sa teorijskom postavkom. Argumenti date funkcije su traženi kvantil u procentima*, survfit* objekat i željeni nivo poverenja intervalnih ocena.

```
squantile = function(qn, y, alpha)
  {
  temp<-data.frame(time=y$time, surv=y$surv, std.err=y$std.err)
  attach(temp)

q.lp<-temp[surv<= qn/100 -.05,][1,]
  q<-temp[surv<=qn/100,][1,]
  q.u<-temp[surv>=qn/100+.05,]
  rnm<-nrow(q.u)
  q.up<-q.u[rnm, ]
  fp = (q.up$surv - q.lp$surv)/( q.lp$time - q.up$time)
  std = (q$std.err)/fp
  lower = q$time - qnorm(1-alpha/2)*std
  upper = q$time + qnorm(1-alpha/2)*std
 print(rbind(c(quantile=qn, time=q$time, std.err=std, cie.lower=lower, cie.upper=upper))) }
```

Ocenu medijanu dobijamo pozivom funkcije *squantile(50,leukemia2.km,0.95)*. Tačkasta ocena medijane $\hat{t}_{0.05} = 12$, a dobijeni 95% interval poverenja za medijanu je (11.76329 , 12.23671).

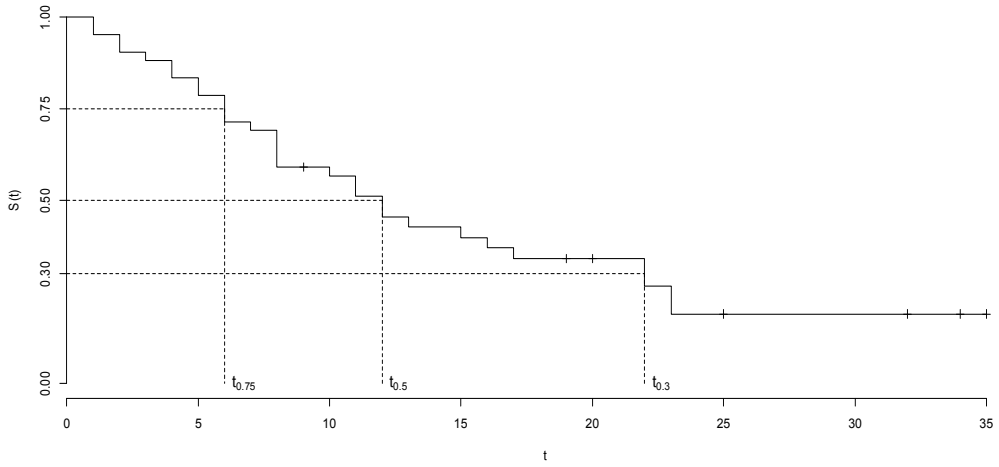
Rezultati za $\hat{t}_{0.75}$ i $\hat{t}_{0.3}$ ocenjene kvantile su sledeći:

```
squantile(75,leukemia2.km,0.95)

   quantile  time   std.err     cie.lower    cie.upper
[1,]      75    6  2.037679   5.872224    6.127776
```

```
squantile(30,leukemia2.km,0.95)

   quantile  time   std.err     cie.lower    cie.upper
[1,]      30   22    11.2186   21.29652    22.70348
```

Grafičku ocenu kvantila funkcije preživljavanja nalazimo na sledeći način:

```
plot(leukemia.km,xlab="t",ylab=expression(hat(S)*"(t)"), axes=F)
axis(1)
axis(2,at=c(0,0.3,0.5,0.75,1))
lines(x=c(0,12),y=c(.5,.5), lty=2)
lines(x=c(12,12),y=c(.5,0), lty=2)
text(12,0,expression(t[0.5]),pos=4,adj=c(0.5,1),cex=1.2)
lines(x=c(0,6),y=c(.75,.75), lty=2)
lines(x=c(6,6),y=c(.75,0), lty=2)
text(6,0,expression(t[0.75]),pos=4,adj=c(0.5,1),cex=1.2)
lines(x=c(0,22),y=c(.3,.3), lty=2)
lines(x=c(22,22),y=c(.3,0), lty=2)
text(22,0,expression(t[0.3]),pos=4,adj=c(0.5,1),cex=1.2)
```



Ocenu očekivanja nalazimo na sledeći način:

```
print(leukemia2.km, print.rmean=TRUE)
```

```
Call: survfit(formula = Surv(Time, Censor) ~ 1, conf.int = 0.95, conf.type = "log-log",
    type = "kaplan-meier")

  records     n.max   n.start    events    *rmean *se(rmean)    median
    42.00     42.00     42.00     30.00     15.34      1.86     12.00

  0.95LCL    0.95UCL
     8.00      17.00
   * restricted mean with upper limit =  35
```

Za gornju granicu ocene očekivanja R postavlja najveće vreme preživljavanja, koje može biti i cenzurisano. U konkretnom primeru 35 je cenzurisano vreme.

Na ovaj način smo dobili ocene celokupnog iskustva preživljavanja u posmatranoj kliničkoj studiji zasnovane na vremenima preživljavanja svih pacijenata u studiji, ne obazirući se na bitne podgrupe pacijenata definisane pomoću kovarijanti.

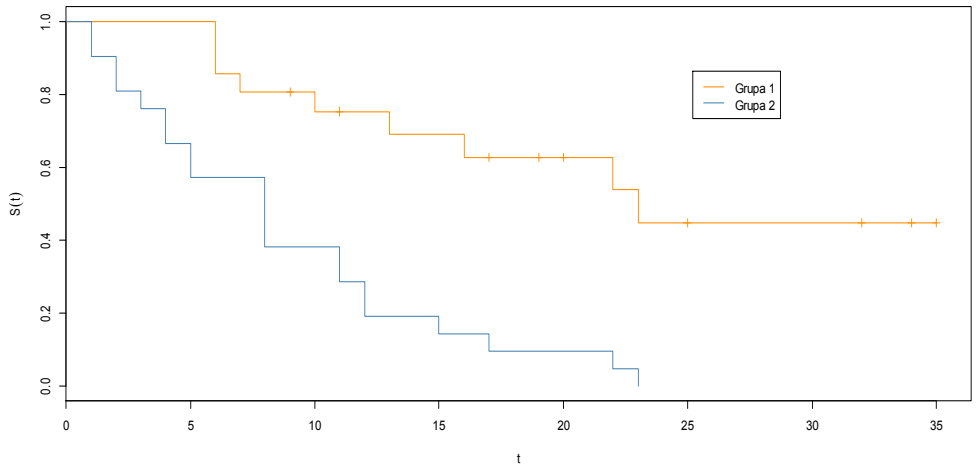
Osnovni cilj dalje analize biće poređenje iskustva preživljavanja pacijenata u odnosu na to da li su primili tretman ili placebo. Prvo ćemo razlike u vremenima preživljavanja pokušati da uočimo pomoću grafika odgovarajućih KM krivih.

```
fit1 = survfit(Surv(Time,Censor)~Rx, conf.type = "log-log")

summary(fit1)

Call: survfit(formula = Surv(Time, Censor) ~ Rx, conf.type = "log-log")

                Rx=0
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   6     21       3    0.857  0.0764       0.620       0.952
   7     17       1    0.807  0.0869       0.563       0.923
  10     15       1    0.753  0.0963       0.503       0.889
  13     12       1    0.690  0.1068       0.432       0.849
  16     11       1    0.627  0.1141       0.368       0.805
  22      7       1    0.538  0.1282       0.268       0.747
  23      6       1    0.448  0.1346       0.188       0.680

                Rx=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.67005       0.975
    2     19       2   0.8095  0.0857      0.56891       0.924
    3     17       1   0.7619  0.0929      0.51939       0.893
    4     16       2   0.6667  0.1029      0.42535       0.825
    5     14       2   0.5714  0.1080      0.33798       0.749
    8     12       4   0.3810  0.1060      0.18307       0.578
   11      8       2   0.2857  0.0986      0.11656       0.482
```

12	6	2	0.1905	0.0857	0.05948	0.377
15	4	1	0.1429	0.0764	0.03566	0.321
17	3	1	0.0952	0.0641	0.01626	0.261
22	2	1	0.0476	0.0465	0.00332	0.197
23	1	1	0.0000	NaN	NA	NA

Kreiranjem datog *survfit* objekta, dobijamo KM ocene funkcija preživljavanja pacijenata koji su primili tretman, i pacijenata koji su primili placebo, kao i tačka po tačka log-log 95% intervale poverenja funkcija preživljavanja za odgovarajuća vremena neuspeha.

```
plot(fit1,xlab="t",ylab=expression(hat(S)(t)), col=c("darkorange","steelblue"))
legend(locator(n=1),legend=c("Grupa 1", "Grupa 2"),col=c("darkorange","steelblue"),lty=c(1,1))
```

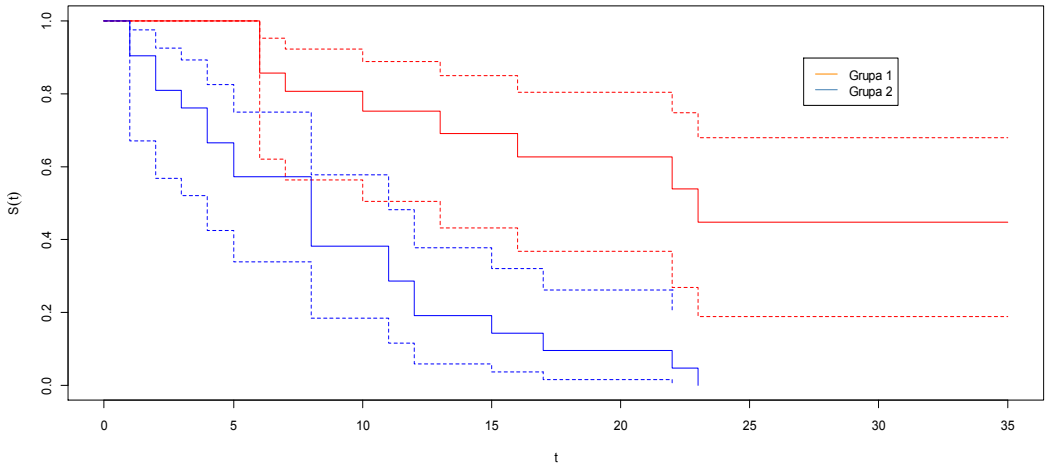


Uočavamo da Grupa 1, odnosno grupa pacijenata koja je primila tretman, ima bolju prognozu preživljavanja od Grupe 2, grupe pacijenata koji su dobili placebo. Štaviše, kako se povećava broj nedelja u remisiji, tako se ovde dve KM krive sve više udaljavaju. To ukazuje na to da su pozitivni efekti tretmana bolji u odnosu na placebo što duže pacijent ostaje u remisiji.

Možemo predstaviti ove krive i sa njihovim log-log intervalima poverenja na sledeći način:

```
fit1$label=c(rep(1,fit1$strata[1]),rep(2,fit1$strata[2]))
 t1=c(0,subset(fit1$time,fit1$label==1))
 t2=c(0,subset(fit1$time,fit1$label==2))
 St1=c(1,subset(fit1$surv,fit1$label==1))
 St2=c(1,subset(fit1$surv,fit1$label==2))
 uSt1=c(1,subset(fit1$upper,fit1$label==1))
 uSt2=c(1,subset(fit1$upper,fit1$label==2))
 lSt1=c(1,subset(fit1$lower,fit1$label==1))
 lSt2=c(1,subset(fit1$lower,fit1$label==2))
 plot(0,0,pch=" ",ylim=0:1,xlim=range(t1,t2),xlab="t",ylab=expression(hat(S)(t)))
 lines(t1,uSt1,lty=2,type='s',col="red")
```

```
 lines(t1,lSt1,lty=2,type='s',col="red")
 lines(t2,uSt2,lty=2,type='s',col="blue")
 lines(t2,lSt2,lty=2,type='s',col="blue")
 lines(t1,St1,type='s',col="red")
 lines(t2,St2,type='s',col="blue")
legend(locator(n=1), legend=c("Grupa 1", "Grupa2"),col=c("darkorange","steelblue"),lty=c(1,1))
```



U R za testiranje značajnosti razlika između iskustava preživljavanja više grupa, koristimo *survdiff* testove. Funkcija *survdiff* predstavlja familiju testova Harrington-a i Fleming-a definisanih pomoću parametra rho. Test statistika ovih testova ima težine $\hat{S}(t)^{\rho}$ koje zavisi od parametra ρ. Sa $\rho = 0$ (standarno u R) definišemo log-rank test, sa $\rho = 1$ uopšteni Wilcoxen test, i sa $\rho = 1/2$ definišemo Tarone-Ware test.
Log-rank test stavlja akcenat na veće vrednosti vremena preživljavanja, gde se po grafiku KM krivih uočava najveća razlika u preživljavanju. Stoga, primenjujemo log-rank test da bi ispitali značajnost razlike iskustva preživljavanja između ove dve grupe pacijenta.

```
survdiff(Surv(Time,Censor)~Rx)

Call:
survdiff(formula = Surv(Time, Censor) ~ Rx)

      N Observed Expected (O-E)^2/E (O-E)^2/V
Rx=0 21        9     19.3      5.46      16.8
Rx=1 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Na osnovu p-vrednosti ovog testa odbacujemo nultu hipotezu, odnosno zaključujemo da su iskustva preživljavanja pacijenata koji su primili tretman i dobili placebo značajno različita.

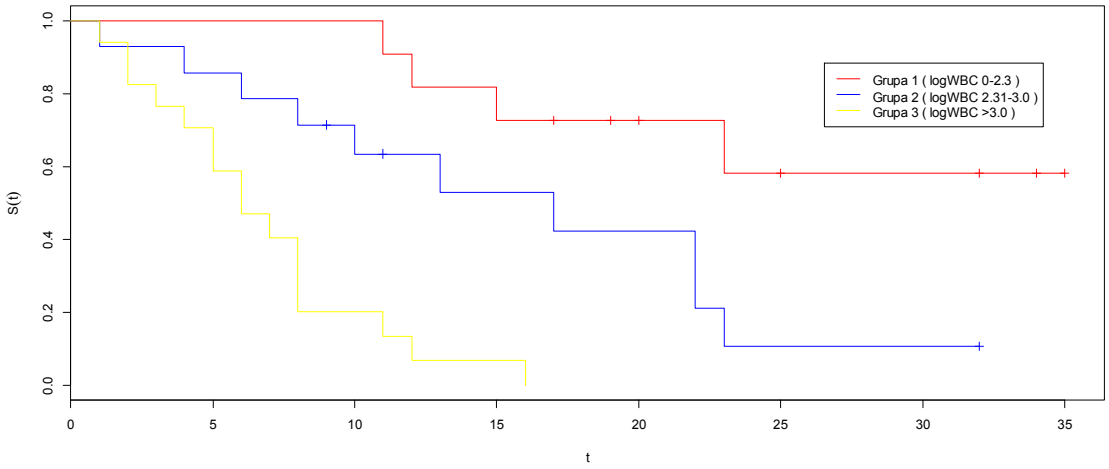
Smatra se da broj belih krvnih zrnaca, logWBC, ima jak uticaj na preživljavanje, u smislu da pacijenti sa lošijom prognozom za leukemiju imaju povećani broj belih krvnih zrnaca. Samim tim, možemo da testiramo razlike iskustava preživljavanja pacijenata u odnosu na podgrupe kovarijante logWBC, za koje znamo da su klinički bitne. Ako je $\text{logWBC} \leq 2.30$ smatra se da pacijent ima nizak broj belih krvih zrnaca, ako je $2.30 < \text{logWBC} \leq 3$ smatra se da imaju srednji broj, a ako je $\text{logWBC} > 3$ broj krvnih zrnaca se smatra visokim.

```
wbc<-cut(logWBC,c(0,2.30,3.00,5.00))
levels(wbc)<-c("nizak","srednji","visok")

survfit(Surv(Time,Censor)~wbc, conf.type="log-log")

Call: survfit(formula = Surv(Time, Censor) ~ wbc, conf.type = "log-log")

          records n.max n.start events median 0.95LCL 0.95UCL
wbc=nizak      11    11      11      4     NA      12      NA
wbc=srednji    14    14      14     10     17       6      22
wbc=visok      17    17      17     16      6       3       8

summary(survfit(Surv(Time,Censor)~wbc, conf.type="log-log"))

Call: survfit(formula = Surv(Time, Censor) ~ wbc, conf.type = "log-log")

                wbc=nizak
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   11     11       1    0.909  0.0867       0.508       0.987
   12     10       1    0.818  0.1163       0.447       0.951
   15      9       1    0.727  0.1343       0.371       0.903
   23      5       1    0.582  0.1687       0.213       0.827

                wbc=srednji
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     14       1    0.929  0.0688      0.5908       0.990
    4     13       1    0.857  0.0935      0.5394       0.962
    6     12       1    0.786  0.1097      0.4725       0.925
    8     11       1    0.714  0.1207      0.4063       0.882
   10      9       1    0.635  0.1308      0.3312       0.830
   13      6       1    0.529  0.1457      0.2263       0.761
   17      5       1    0.423  0.1501      0.1452       0.682
   22      4       2    0.212  0.1297      0.0345       0.489
   23      2       1    0.106  0.0990      0.0062       0.371

                wbc=visok
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     17       1   0.9412  0.0571     0.65018       0.991
    2     16       2   0.8235  0.0925     0.54713       0.939
```

3	14	1	0.7647	0.1029	0.48828	0.904
4	13	1	0.7059	0.1105	0.43148	0.866
5	12	2	0.5882	0.1194	0.32537	0.778
6	10	2	0.4706	0.1211	0.22960	0.680
7	7	1	0.4034	0.1210	0.17641	0.622
8	6	3	0.2017	0.1022	0.05096	0.423
11	3	1	0.1345	0.0875	0.02263	0.345
12	2	1	0.0672	0.0646	0.00441	0.261
16	1	1	0.0000	NaN	NA	NA

Opet, kreiranjem *survfit* objekta dobijamo tačkaste i intervalne KM ocene funkcija preživljavanja datih podgrupa kovarijante logWBC.

```
plot(survfit(Surv(Time,Censor)~wbc),xlab="t",ylab=expression(hat(S)(t)),col=c("red","blue",
"yellow"))
legend(locator(n=1), legend=c("Grupa 1 ( logWBC 0-2.3 )", "Grupa 2 ( logWBC 2.31-3.0
)","Grupa 3 ( logWBC >3.0 )"), lty=c(1,1,1),col=c("red","blue","yellow"))
```



Uočavamo da su KM krive prilične različite. Grupa 1, grupa pacijenata sa niskim brojem belih krvnih zrnaca, ima bolju prognozu od Grupe 2, grupe pacijenata sa srednjim brojem, koja opet ima bolju prognozu od Grupe 3, grupe pacijenata sa velikim brojem belih krvnih zrnaca. Primetimo, takođe, da je razlika između Grupe 1 i 2 otprilike ista tokom vremena, dok se krive Grupa 2 i 3 udaljavaju kako vreme prolazi.

Da bi utvrdili da li su razlike statistički značajne, koristimo log-rank test.

```
survdiff(Surv(Time,Censor)~wbc)
```

```
Call:
survdiff(formula = Surv(Time, Censor) ~ wbc)

        N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
wbc=nizak   11       4    13.06    6.2880   12.7695
wbc=srednji 14      10    10.72    0.0489    0.0809
wbc=visok   17      16     6.21   15.4173   23.1040
```

Chisq= 26.4 on 2 degrees of freedom, p= 1.86e-06

Log-rank statistika (26.4) je jako značajna sa malom p-vrednošću. Ovi rezultati ukazuju da postoji neka opšta razlika u preživljavanju između ove tri grupe pacijenata.

Rezultati prethodne analize, pre svega testova, su veoma korisni u proceni da li kovarijante utiču na preživljavanje. Međutim, ne pružaju nam odgovor na pitanje u koliko većem riziku je jedna grupa subjekata u odnosu na drugu. Želimo da istražimo funkcionalnu vezu između kovarijanti i opstanka, da analiziramo efekte, a ne samo prisustvo, kovarijanti na preživljavanje. Da bismo dobili takve odgovore, na datu bazu podataka primenjujemo Cox-ov model sa proporcionalnim rizicima.
Prvo vršimo identifikaciju značajnih kovarijanti primenjujući algoritam postupne selekcije kovarijanti. Želimo da proverimo koja od tri kovarijante, Rx (tretman ili placebo), logWBC i Sex (pol), iz baze će ući u model.

Definišemo p-vrednosti za ulazak i izlazak iz modela (po Hosmer-u) :

```
pE<-0.15;        logpE<-log(pE)
pR<-0.2;         logpR<-log(pR)
```

$\text{logpE} = -1.89712$ predstavlja logaritmovanu p-vrednost za ulazak u model
$\text{logpR} = -1.609438$ predstavlja logaritmovanu p-vrednost za izlazak iz model

Nulti korak algoritma:

Pomoću funkcije *coxph(S~1)* definišemo početni model bez kovarijanti, gde je sa *S<-Surv(Time, Censor)* definisan objekat preživljavanja.

```
m0<-coxph(S~1);            ll0<-m0$logl[1]
```

Zatim definišemo i niz sledećih PH modela u koje je uključena po jedna kovarijanta od interesa:

```
m1<-coxph(S~Rx);          ll1<-m1$logl[2]
m2<-coxph(S~logWBC);  ll2<-m2$logl[2]
m3<-coxph(S~Sex);         ll3<-m3$logl[2]
```

```
lp1<-pchisq(2*(ll1-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
lp2<-pchisq(2*(ll2-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
lp3<-pchisq(2*(ll3-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
```

lp1	lp2	lp3
-9.852619	-19.44887	-0.8218939

Najmanju logaritmovanu (jer su p-vrednosti jako male) p-vrednost ima kovarijanta logWBC i ona postaje kandidat za ulazak u model. Kako je ova vrednost manja od $logpE$, logWBC je značajna kovarijanta za ulazak u model.

Prvi korak:

```
m0<-coxph(S~logWBC);          ll0<-m0$logl[2]
m2<-coxph(S~logWBC+Rx);       ll2<-m2$logl[2]
m3<-coxph(S~logWBC+Sex);      ll3<-m3$logl[2]

lp2<-pchisq(2*(ll2-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
lp3<-pchisq(2*(ll3-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
```

lp2	lp3
-7.4694828	-0.5637398

Model sa kovarijantama logWBC i Rx ima najmanju log p-vrednost. Opet, to je dovoljno malo da uključi Rx u model.

Moramo da proverimo da dodavanje kovarijante Rx ne dovodi do gubitka značajnosti kovarijante logWBC.

```
m0=coxph(S~logWBC+Rx) ;ll0=m0$logl[2]
m1=coxph(S~Rx) ;ll1=m1$logl[2]

pchisq(2*(ll0-ll1),lower.tail=FALSE,log.p=TRUE,df=1)

[1] -17.14327
```

Ova vrednost je manja od $logpR$, samim tim ne izbacujemo iz modela prvu dodatu kovarijantu.

U drugom koraku proveravamo da li se kovarijanta Sex može uključiti u postojeći model:

```
m0<-coxph(S~logWBC+Rx);  ll0<-m0$logl[2]
m1<-coxph(S~logWBC+Rx+Sex);  ll1<-m1$logl[2]

lp1<-pchisq(2*(ll1-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
```

lp1	logpE
-0.712151	-1.897120

Ova logaritmovana p-vrednost nije manja od $logpE$, pa kovarijanta Sex nije značajna za ulazak u model i algoritam se prekida jer ne postoji više kovarijanti čiju značajnost za ulazak možemo da ispitamo.

Na osnovu dosadašnje procedure izbora kovarijanti, u Cox-ov PH model ulaze Rx i logWBC. Moramo proveriti i mogući efekat interakcije ove dve promenljive.
Uopšteno, ako imamo p kovarijanti u modelu, postoji $p(p-1)/2$ mogućih interakcija. U našem slučaju, proveravamo uticaj samo jedne interakcije, Rx:logWBC, koju posmatramo kao treću promenljivu, i ponavljamo proceduru postupnog izbora.

```
m0<-coxph(S~Rx+logWBC);  ll0<-m0$logl[2]
m1<-coxph(S~Rx+logWBC+Rx*logWBC);  ll1<-m1$logl[2]

lp1<-pchisq(2*(ll1-ll0),lower.tail=FALSE,log.p=TRUE,df=1)
```

lp1	logpE
-0.599979	-1.897120

Nedovoljno mala p-vrednost ukazuje na to da ne postoji značajna interakcija tretmana /placeba sa brojem belih krvnih zrnaca.

Odlučujemo se za model koji sadrži dve kovarijante Rx i logWBC, odnosno

$$h(t|\text{Rx}, \text{logWBC}) = h_0(t)\exp\left(\hat{b}_1\text{Rx} + \hat{b}_2\text{logWBC}\right)$$

Tačkaste ocene parametara dobijamo pozivom funkcije *coxph* ili *summary(coxph)*, pri tom možemo birati metodu za ocenjivanje promenom argumenta *method* ("*breslow*", "*efron*" ,"*exact*"). Ocena Efron-ovom aproksimacijom parcijalne funkcije preživljavanja je standardna u R-u.

```
model_b<-coxph(Surv(Time,Censor)~Rx+logWBC,method="breslow")
model_b
Call:
coxph(formula = Surv(Time, Censor) ~ Rx + logWBC, method = "breslow")

           coef exp(coef) se(coef)    z       p
Rx         1.29      3.65    0.422    3.07   2.2e-03
logWBC 1.60      4.97    0.329    4.87   1.1e-06
Likelihood ratio test=43.4  on 2 df, p=3.74e-10  n= 42, number of events= 30

model<-coxph(Surv(Time,Censor)~Rx+logWBC,method="efron")
model
Call:
coxph(formula = Surv(Time, Censor) ~ Rx + logWBC, method = "efron")

           coef    exp(coef) se(coef)    z       p
Rx         1.39      4.00       0.425   3.26   1.1e-03
logWBC 1.69      5.42       0.336   5.03   4.8e-07

Likelihood ratio test=46.7  on 2 df, p=7.19e-11  n= 42, number of events= 30
```

```
model_e<-coxph(Surv(Time,Censor)~Rx+logWBC,method="exact")
model_e
Call:
coxph(formula = Surv(Time, Censor) ~ Rx + logWBC, method = "exact")

          coef  exp(coef) se(coef)    z       p
Rx        1.44      4.24     0.455  3.18  1.5e-03
logWBC 1.76      5.83     0.359   4.91  9.2e-07

Likelihood ratio test=46.6  on 2 df, p=7.72e-11  n= 42, number of events= 30
```

Možemo primetiti da p-vrednosti pojednačnih kovarijanti (Wald test statistike) ukazuju na značajan uticaj datih kovarijanti na preživljavanje bez obzira na to koja se aproksimacija parcijalne funkcije verodostojnosti koristila. I p-vrednost testa log verovatnosnog količnika, koji provera da li su svi koeficijenti jednaki nuli, je jako mala, što potvrđuje značajnost uticaja kovarijanti.

Kako se Efon-ova aproksimacija smatra najboljom, ocene koeficijenata dobijene pomoću nje ćemo koristiti. Funkcija rizika Cox-ovog modela sa proporcionalnim rizicima je

$$h(t|\text{Rx}, \text{logWBC}) = h_0(t)\exp\left(1.39\text{Rx} + 1.69\text{logWBC}\right)$$

Korišćenjem funkcije *summary* dobijamo još detaljniji izlaz.

```
summary(model)
Call:
coxph(formula = Surv(Time, Censor) ~ Rx + logWBC)

  n= 42, number of events= 30

             coef    exp(coef) se(coef)     z      Pr(>|z|)
Rx        1.3861    3.9991    0.4248   3.263   0.0011 **
logWBC 1.6909    5.4243    0.3359   5.034  4.8e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

          exp(coef) exp(-coef) lower .95 upper .95
Rx            3.999     0.2501     1.739     9.195
logWBC     5.424     0.1844     2.808    10.478

Concordance= 0.852  (se = 0.062 )
Rsquare= 0.671   (max possible= 0.988 )
Likelihood ratio test= 46.71  on 2 df,   p=7.187e-11
Wald test            = 33.6  on 2 df,   p=5.061e-08
Score (logrank) test = 46.07  on 2 df,   p=9.921e-11
```

Kak R daje standardnu grešku ocene koeficijenata, možemo odrediti i intervalu ocenu koeficijenata. Na primer, 95% interval poverenja koeficijenta uz Rx je

(1.3861-qnorm(1-0.025)*0.4248, 1.3861+qnorm(1-0.025)*0.4248)=(0.56, 2.22)

Dat je i ocenjeni hazardni količnik pacijenata koji su dobili placebo i pacijenata koji su dobili tretman $\widehat{\mathrm{HR}} = \exp(\hat{b}_1) = 3.9991$. Na osnovu čega zaključujemo da je rizik da se ponovo pojavi leukemija kod pacijenata koji su dobili placebo četiri puta veći nego rizik kod pacijenata koji su primili tretman. Odnosno, lečenje značajno smanjuje rizik od ponovne pojave bolesti.

Kod neprekidnih kovarijanti, kao što je logWBC, da bi protumačili ocenjeni hazardni količnik moramo odrediti klinički značajnu jedinicu c za kovarijantu.

$$\widehat{\mathrm{HR}}(\mathrm{c}) = \frac{h(t|\mathrm{logWBC} + \mathrm{c})}{h(t|logWBC)} = \exp\,(c\hat{b}_2)$$

R nam daje $\widehat{\mathrm{HR}}(1) = 5.424$ za logWBC. Na osnovu toga zaključujemo da je opstanak negativno povezan sa brojem belih krvnih zrnaca. Povećanje belih krvnih zrnaca za jedan na log skali, dovodi do 5.4 puta većeg rizika za povratak leukemije kod pacijenta.

Da bismo detaljnije opisali uticaj logWBC na preživljavanje, ovu neprekidnu promenljivu možemo pretvoriti u kategoričku grupisanjem podataka na način ranije opisan pomoću faktora wbc. Međutim, ako nominalna promenljiva ima više od dve kategorije, označene sa K, moramo je, pre primene Cox-ov modela, modelirati koristeći kolekciju od K-1 dizajniranih indikator promenljivih. Najčešće korišćena metoda za kodiranje ovih dizajniranih promenljivih u Cox-ovom modelu se naziva kodiranje u odnosu na referentnu ćeliju. Sa ovom metodom, biramo jedan nivo promenljive za referentni nivo, u odnosu na koji se svi drugi nivoi porede. Dobijeni hazard količnici porede stopu rizika svake grupe sa referentnom grupom.

wbc	logWBC_2	logWBC_3
nizak	0	0
srednji	1	0
visok	0	1

Ocenjeni hazard količnici obezbeđuju zgodnu meru za poređenje iskustva preživljavanja ove tri grupe kovarijante logWBC. R automatski uzima prvu kategoriju za referentnu, što nama i odgovara jer pretpostavljamo da će rizik za tu grupu biti najmanji.

$$\widehat{\mathrm{HR}}(\mathrm{srednji, nizak}) = \frac{h(t|\mathrm{srednji})}{h(t|\mathrm{nizak})} = \frac{\exp\,(\hat{b}_2\mathrm{logWBC\_2[srednji]} + \hat{b}_3\mathrm{ogWBC\_3[srednji]})}{\exp\,(\hat{b}_2\mathrm{logWBC\_2[nizak]} + \hat{b}_3\mathrm{ogWBC\_3[nizak]})}$$

$$\widehat{\mathrm{HR}}(\mathrm{srednji, nizak}) = \frac{h(t|\mathrm{srednji})}{h(t|\mathrm{nizak})} = \exp\,(\hat{b}_2)$$

Analogno,

$$\widehat{\mathrm{HR}}(\text{visok}, \text{nizak}) = \frac{h(t|\text{visok})}{h(t|\text{nizak})} = \exp(\hat{b}_3)$$

coxph(Surv(Time,Censor)~Rx+wbc)

Call:
coxph(formula = Surv(Time, Censor) ~ Rx + wbc)

	coef	exp(coef)	se(coef)	z	p
Rx	1.33	3.78	0.437	3.04	0.00230
wbcSrednji	1.33	3.79	0.605	2.20	0.02800
wbcVisok	2.49	12.08	0.653	3.82	0.00014

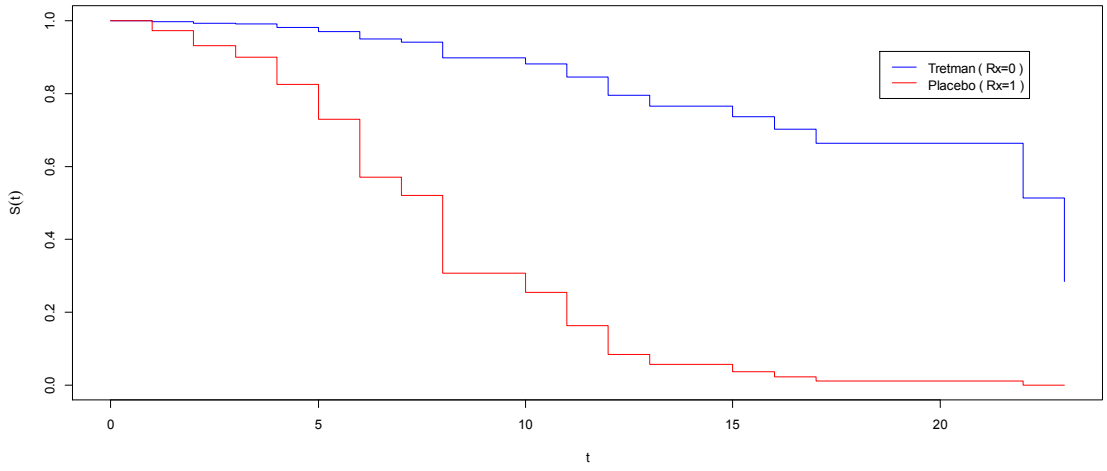
Likelihood ratio test=35.6 on 3 df, p=9.19e-08 n= 42, number of events= 30

Na osnovu izlaza u R-u koji uključuje sada nominalnu kovarijantu wbc, zaključujemo da je rizik za povratak leukemije pacijenata koji imaju srednji broj belih krvnih zrnaca 1.33 puta veći nego kod pacijenata sa niskim brojem; kao i da je ovaj rizik čak 12 puta veći za pacijente koji imaju visok broj belih krvnih zrnaca u odnosu na one koji imaju mali. Sve nam ovo ukazuje da je logWBC bitan prognostički indikator za preživljavanje pacijenata koji boluju od leukemije, i da se povećanje broja belih krvnih zrnaca negativno odražava na opstanak pacijenata.
Primetimo da se i ocena koeficijenta kovarijante Rx promenila, tačnije malo se pogoršala preciznost same ocene, interval poverenja je sada (0.47, 2.19). U nastavku rada se vraćamo na model sa nestratifokovanom kovarijantom logWBC.

Da bi grafički prikazali ocenjene funkcije preživljavanja, na osnovu Cox-ovog modela, u odnosu na neku kovarijantu od interesa moramo prilagoditi ostale vrednosti kovarijanti iz modela. To prilagođavanje se obično vrši uzimanjem srednjih vrednosti ostalih kovarijanti. Na sledeći način u R-u možemo prikazati prilagođene krive preživljavanja za tretman i placebo grupu.

```
d.phm = coxph.detail(model)
times = c(0,d.phm$t)
h0 = c(0,d.phm$hazard)
S0 = exp(-cumsum(h0))
b = model2$coef
meanx = c(mean(Rx),mean(logWBC))
x_1 = c(0,mean(subset(logWBC,Rx==0)))-meanx
Sx_1 = S0 ^ exp(t(b) %*% x_1)
x_2 = c(1,mean(subset(logWBC,Rx==1)))-meanx
Sx_2 = S0 ^ exp(t(b) %*% x_2)

plot(times,Sx_1,xlab="t",ylab=expression(hat(S)(t)),ylim=0:1,type="s",col="blue")
lines(times,Sx_2,col=2,type="s")
```

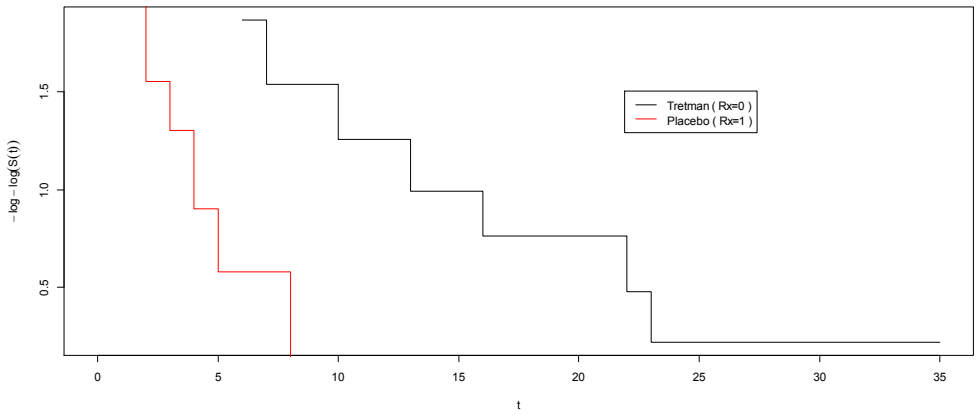
legend(locator(n=1),legend=c("Tretman(Rx=0)","Placebo(Rx=1)"),col=c("blue","red"),lty=c(1,1
))



Sada ćemo poceniti adekvatnost izabranog modela, primenjujući niz metoda ranije objašnjenih. Prvo ćemo proceniti pretpostavku o proporcionalnim rizicima uz pomoć dve grafičke metode.

Upoređivanjem log-log krivih preživljavanja procenićemo PH pretpostavku za kovarijantu Rx :

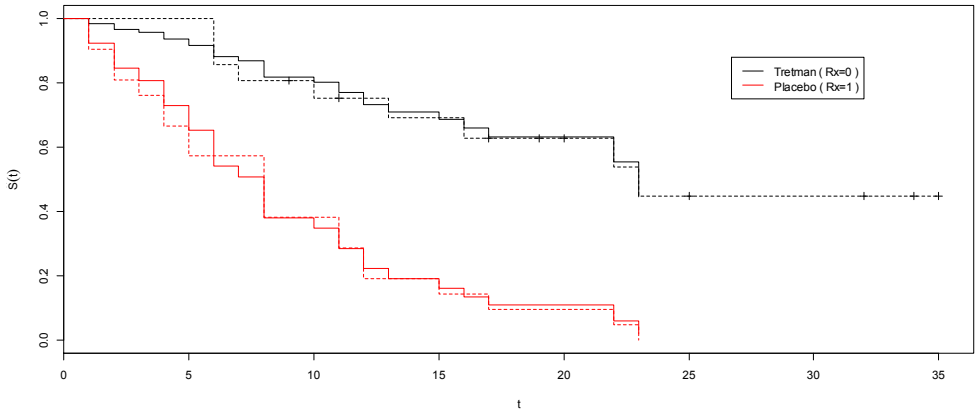
```
k1<-survfit(Surv(Time[Rx==0],Censor[Rx==0])~1)
t1=c(0,k1$time)
St1=c(1,k1$surv)
k2<-survfit(Surv(Time[Rx==1],Censor[Rx==1])~1)
t2=c(0,k2$time)
St2=c(1,k2$surv)

plot(t1,-log(-log(St1)),col=1,type="s",xlab="t",ylab=expression(-log-log(hat(S)(t))))
lines(t2,-log(-log(St2)),col=2,type="s")
legend(locator(n=1), legend=c("Tretman ( Rx=0 )", "Placebo ( Rx=1 )"), col=c(1,2),lty=1)
```



Na osnovu grafiku uočavamo „grubu“ paralelnost log-log krivih u odnosu na nivoe kovarijante Rx. Procenjujemo da je za kovarijantu Rx zadovoljena PH pretpostavka.

Proverimo da li je ova pretpostavka zadovoljena za kovarijantu Rx i pomoću druge grafičke metode, koja poredi posmatrane sa očekivanim krivama preživljavanja.

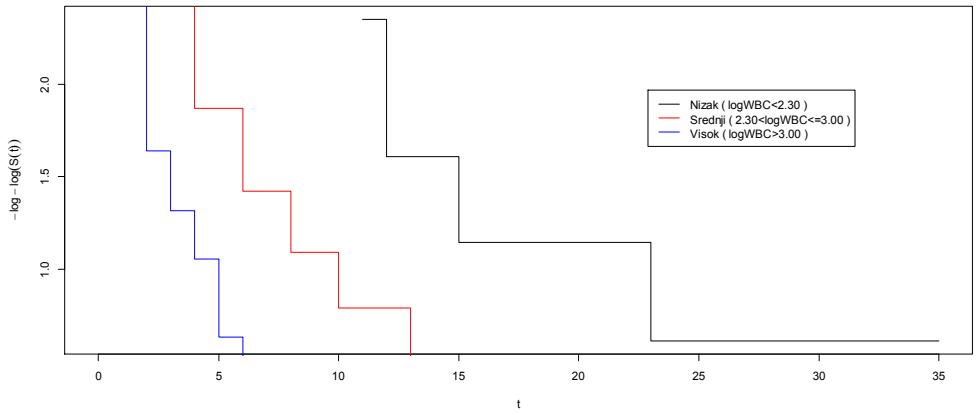
```
model1<-coxph(Surv(Time,Censor)~Rx)
d.model 1= coxph.detail(model1)
times =c(0,d.model1$t)
h0 = c(0,d.model1$hazard)
S0=exp(-cumsum(h0))
beta = c(model1$coef)
x1=c(0)-mean(Rx)
Sx1 = S0 ^ exp(t(beta) %*% x1)
x2=c(1)-mean(Rx)
Sx2 = S0 ^ exp(t(beta) %*% x2)
k=survfit(Surv(Time,Censor)~Rx)
plot(k,col=1:2,lty=2,xlab="t",ylab=expression(hat(S)(t)))
lines(times,Sx1,col=1,type="s")
lines(times,Sx2,col=2,type="s")
legend(locator(n=1), legend=c("Tretman ( Rx=0 )", "Placebo ( Rx=1 )"), col=c(1,2),lty=1)
```



Uočavamo da su i za grupu pacijenta koji su primili tretman, i za grupu koja je dobila placebo, posmatrane i očekivane krive preživljavanja blizu, pa i na osnovu ove grafičke metode zaključujemo da je zadovoljena PH pretpostavka za kovarijantu Rx.

Procenimo PH pretpostavku za kovarijantu logWBC iz modela primenjujući grafičku metodu koja procenjuje paralelnost log-log krivih:

```
k1<-survfit(Surv(Time[wbc=="Nizak"],Censor[wbc=="Nizak"])~1)
t1=c(0,k1$time)
St1=c(1,k1$surv)
k2<-survfit(Surv(Time[wbc=="Srednji"],Censor[wbc=="Srednji"])~1)
t2=c(0,k2$time)
St2=c(1,k2$surv)
k3<-survfit(Surv(Time[wbc=="Visok"],Censor[wbc=="Visok"])~1)
t3=c(0,k3$time)
St3=c(1,k3$surv)

plot(t1,-log(-log(St1)),col=1,type="s",xlab="t",ylab=expression(-log-log(hat(S)(t))))
lines(t2,-log(-log(St2)),col=2,type="s")
lines(t3,-log(-log(St3)),col=4,type="s")
legend(locator(n=1), legend=c("Nizak ( logWBC<2.30 )","Srednji ( 2.30<=logWBC<=3.00 )","Visok ( logWBC>3.00 )"), col=c(1,2,4),lty=1)
```



Zaključujemo da je zadovoljena PH pretpostavka za kovarijantu logWBC.

Sada ćemo proveriti skalu neprekidne kovarijante logWBC pomoću AIC kriterijuma i martingalnih reziduala.

```
phm.WBC1 =coxph(Surv(Time,Censor)~logWBC)
phm.WBC2 =coxph(Surv(Time,Censor)~logWBC+I(logWBC^2))
phm.WBC3 =coxph(Surv(Time,Censor)~log(logWBC))
phm.WBC4 =coxph(Surv(Time,Censor)~sqrt(logWBC))
phm.WBC5 =coxph(Surv(Time,Censor)~logWBC+I(logWBC^2)+I(logWBC^3))
```

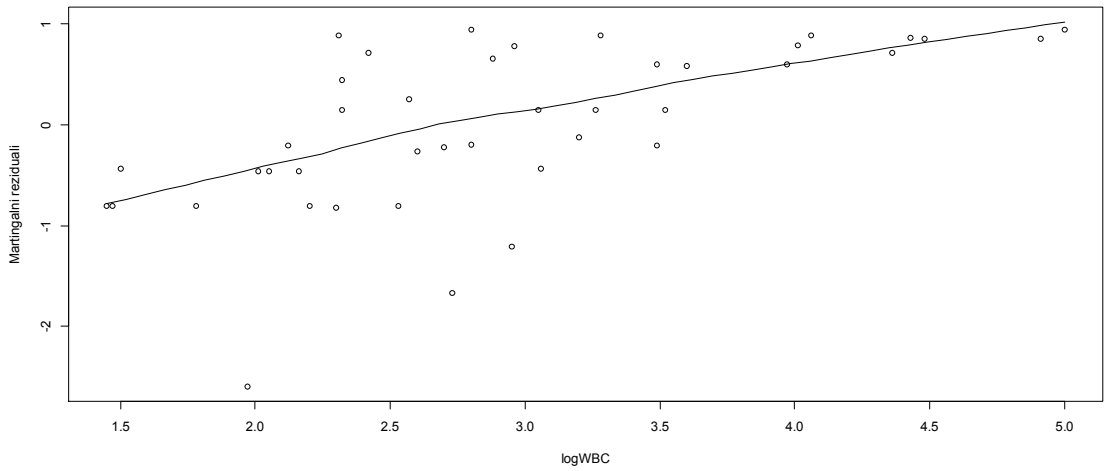
```
aic1=extractAIC(phm.WBC1)[2]
aic2=extractAIC(phm.WBC2)[2]
aic3=extractAIC(phm.WBC3)[2]
aic4=extractAIC(phm.WBC4)[2]
aic5=extractAIC(phm.WBC5)[2]
```

Model		AIC	AIC - minAIC
1	linearni	153.53	0
2	kvadratni	155.17	1.64
3	logaritamski	155.61	2.08
4	kvadratni koren	154.33	0.80
5	kubni	156.67	3.14

Na osnovu Akaike-ovog informacionog kriterijuma zaključujemo da je linearni model najbolji.

Martingalne reziduale grafički predstavljamo radi provere funkcionalne forme neprekidne kovarijante u modelu. Funkcija *resid(fit)*, gde je *fit coxph* objekat daje martingalne reziduale.

```
scatter.smooth(logWBC,resid(model1),type="p",xlab="logWBC",ylab="Martingalni reziduali")
```

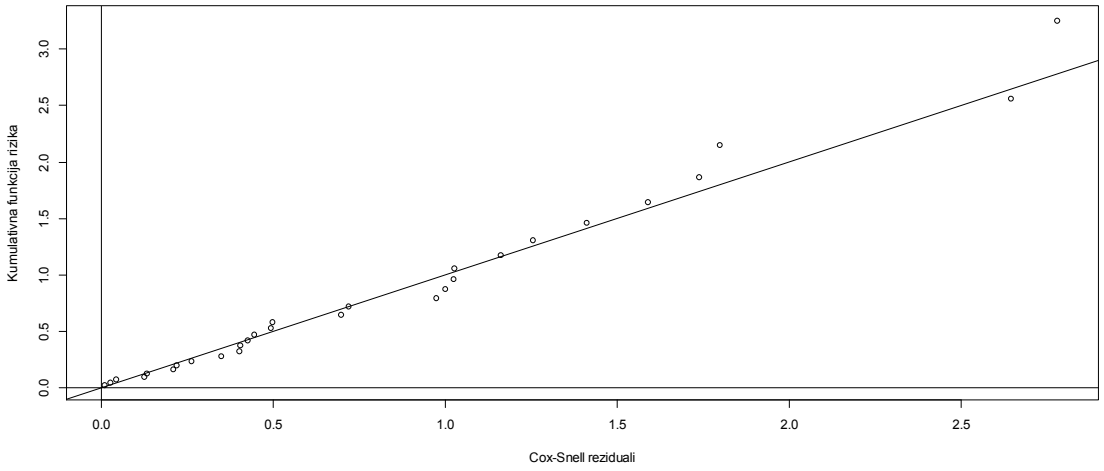


I na osnovu grafika možemo da zaključimo da je linearna forma kovarijante logWBC odgovarajuća.

Da bi proverili goodnes-of-fit modela na grafiku predstavljamo Cox-Snell reziduale, koje izvodimo iz martingalnih reziduala, jer nisu direktno obezbeđeni u R-u.

```
rc <- abs(Censor - model$residuals)      # Cox-Snell reziduali
km.rc <- survfit(Surv(rc,Censor) ~ 1)
summary.km.rc <- summary(km.rc)
rcu <- summary.km.rc$time                # Cox-Snell za necezurisana vremena
surv.rc <- summary.km.rc$surv
```

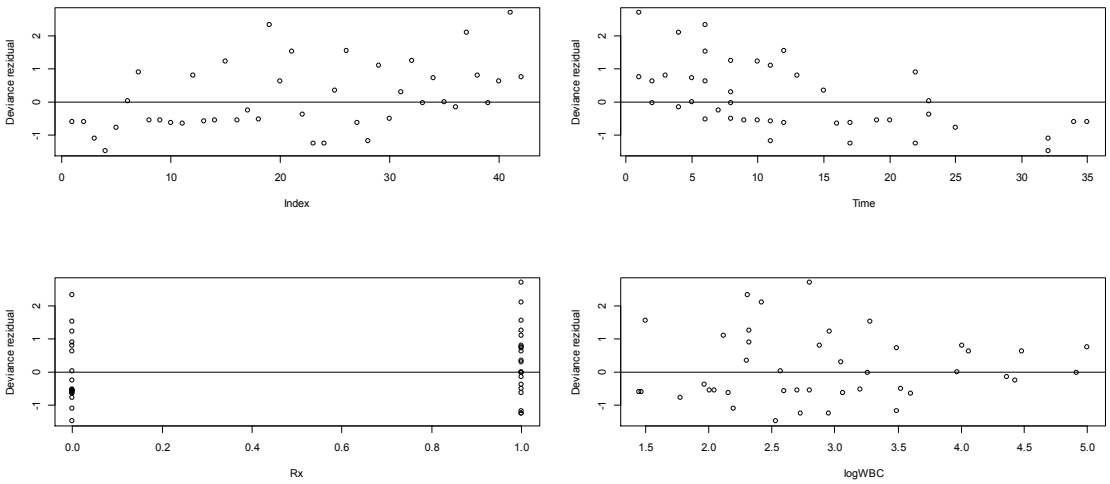
```
plot(rcu,-log(surv.rc),type="p",xlab="Cox-Snell reziduali",ylab="Kumulativna funkcija rizika")
abline(a=0,b=1); abline(v=0); abline(h=0)
```



Na osnovu grafika Cox-Snell reziduala uočavamo da konačni model razumno pristaje podacima. Uopšteno reziduali padaju na pravu liniju sa odsečkom nula i nagibom jedan. Takođe, ne postaje velika odstupanja od prave linije.

Grafičko predstavljanje reziduala odstupanja (deviance) nam pruža informaciju o mogućim autlajerima.

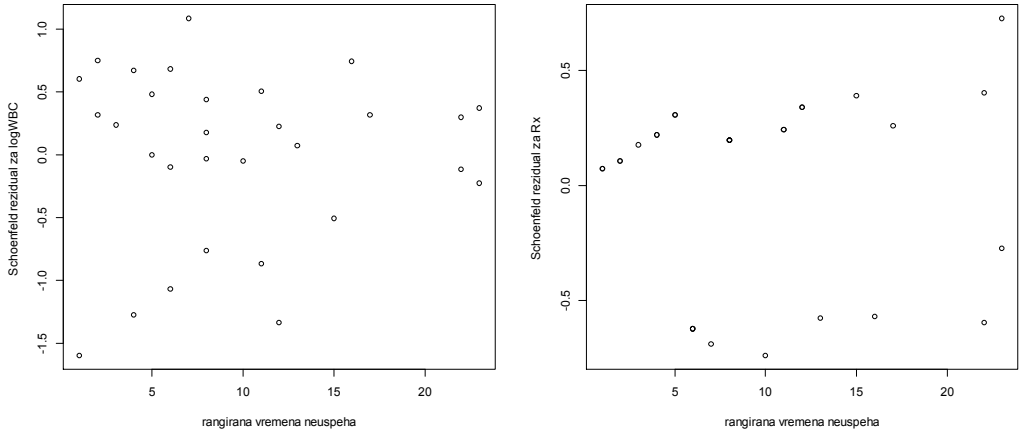
```
dresid <- resid(model,type="deviance")   # deviance reziduali

opar<-par(mfrow=c(2,2))
plot(dresid,type="p",ylab="Deviance rezidual")
abline(h=0)
plot(Time,dresid,type="p",ylab="Deviance rezidual")
abline(h=0)
plot(Rx,dresid,type="p",ylab="Deviance rezidual")
abline(h=0)
plot(logWBC,dresid,type="p",ylab="Deviance rezidual")
abline(h=0)
par(opar)
```



Grafik reziduala odstupanja pokazuje blagi trend da veća vremena preživljavanja imaju negativne reziduale. Ovo sugeriše da model precenjuje šansu umiranja u velikim vremenima. Međutim, postoji samo jedan mogući autlajer u najranijem vremenu i to ne mora izazvati zabrinutost o adekvatnosti modela. Svi ostali grafici pokazuju da su reziduali simetrični oko nule i da postoji najviše jedan mogući autlajer.

Schoenfeld rezidualima proveravamo pristajanje modela podacima i detektujemo udaljene vrednosti kovarijanti.

```
d.model <- coxph.detail(model)
time <- d.model$y[,2]              # rangirana vremena preživljavanja uključujući i cenzurisana
status <- d.model$y[,3]            # status cenzurisanja
sch <- resid(model2,type="schoenfeld")   # Schoenfeld reziduali

par(mfrow=c(1,2));
plot(time[status==1],sch[,2],xlab="rangirana vremena neuspeha",ylab="Schoenfeld rezidual za logWBC ")
plot(time[status==1],sch[,1],xlab="rangirana vremena neuspeha",ylab="Schoenfeld rezidual za Rx ")
```



Na oba grafika ne uočavamo velike vrednosti reziduala. Dakle, PH pretpostavka se čini odgovarajuća.

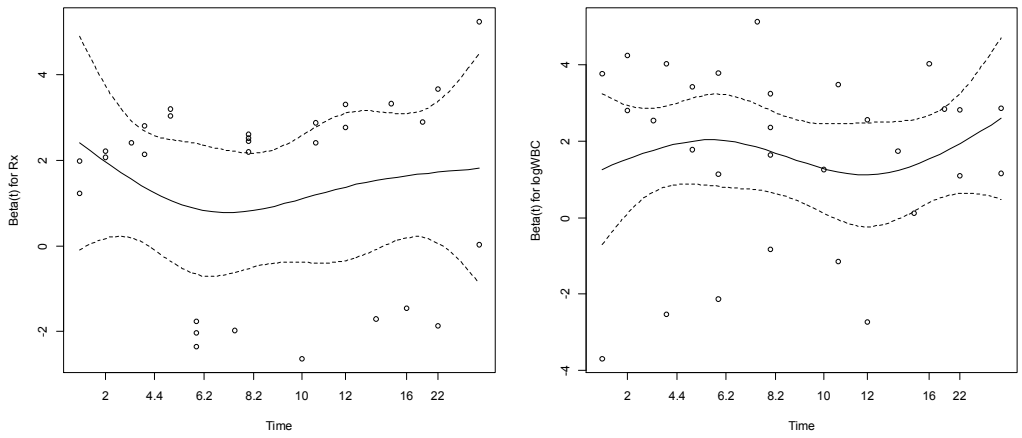
Procenimo i formalno PH pretpostavku pomoću Grambsch i Therneau testa koji je zasnovan na ponderisanim Schoenfeld rezidualima.

(PH.test <- cox.zph(model))

	rho	chisq	p
Rx	0.00451	0.000542	0.981
logWBC	0.02764	0.034455	0.853
GLOBAL	NA	0.034529	0.983

Rezultati testa ukazuju da je PH pretpostavka zadovoljena za obe kovarijante, što podržavaju i sledeći grafici. Na taj način smo i potvrdili da je izabrani model za datu bazu podataka najbolji.

par(mfrow=c(1,2)); plot(PH.test)



Literatura

[1] Hosmer, D.W. , Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd ed. Wiley, New York

[2] Lee, H.T., Wang, J.W. (2003), *Statistical Methods for Survival Data Analysis*, 3rd ed. Wiley, New York

[3] Kleinbaum, D. G. , Klein, M. (1996), *Survival Analysis*: *A Self-Learning Text*, 2nd ed. Springer-Verlag, New York

[4] Klein, J. P., Moeschberger, M. L. (1997), *Survival Analysis: Techniques for Censored and Truncated Data* , 2nd ed. Springer-Verlag, New York

[5] Tableman, M., Kim, J.S. (2004), *Survival Analysis using S: Analysis of Time To Event Data*, Champan and Hall, London

[6] Stevenson M. (2009), *An Introduction to Survival Analysis*, EpiCentre, Massey University, dostupno na http://epicentre.massey.ac.nz/resources/acvsc_grp/docs/Stevenson_survival_analysis_195.721.pdf

[7] Ibrahim, J.G. (2005), *Applied Survival Analysis*, University of North Carolina at Chapel Hill, dostupno na http://www.amstat.org/chapters/northeasternillinois/pastevents/presentations/summer05_Ibrahim_J.pdf

[8] Diez, D. M. *Survival Analysis in R*, dostupno na http://www.stat.ucdavis.edu/~hiwang/teaching/10fall/R_tutorial%201.pdf

[9] Rodriguez, G. (2005), *Non-parametric Estimation in Survival Models*, dostupno na http://data.princeton.edu/pop509/NonParametricSurvival.pdf

[10] Cook, A. (2008) , *An Introduction to Survival Analysis*, dostupno na http://courses.nus.edu.sg/course/stacar/internet/st3242/st3242.html

[11] R textbook examples: *Applied Survival Analysis* , by Hosmer and Lemeshow, dostupno na http://www.ats.ucla.edu/stat/R/examples/asa/