Analiza ponovnog pojavljivanja raka dojke analizom doživljenja

Končić, Dominik

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Dominik Končić

ANALIZA PONOVNOG POJAVLJIVANJA RAKA DOJKE ANALIZOM DOŽIVLJENJA

Diplomski rad

Voditelj rada:
prof. dr. sc.
Anamarija Jazbec

Zagreb, rujan 2023.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Kod analize događaja koji se mogu dogoditi ili ne (npr. smrt), osim ostvarenja tog događaja, bitnu ulogu igra vrijeme koje je prošlo do događaja. Analiza doživljenja je skup statističkih metoda koje analiziraju vrijeme koje je prošlo do nekog događaja.
U samim počecima (17. stoljeće) analiza doživljenja bila je isključivo korištena za analiziranje događaja smrti, što objašnjava ime. Kasnije se, osim u svrhu istraživanja mortaliteta, počinje koristi u i drugim medicinskim istraživanjima, ali i aktuarstvu, inžinjerstvu, psihologiji, ekonomiji...
Najzaslužniji za razvoj ovog skupa statističkih metoda su američki statističari Edward Lynn Kaplan i Paul Meier koji su osmislili neparametarski procjenitelj za funkciju doživljenja te britanski statističar David Cox koji je razvio Coxov regresijski model.
Detaljnije o analizi doživljenja i Coxovom regresijskom modelu reći ćemo u prvom i drugom poglavlju koja su temeljena na knjigama [2] i [3]. Tim statističkim metodama analizirat ćemo podatke vezane uz ponovno pojavljivanje raka dojke koji se proučavaju u članku [4]. Za grafove i statističke testove koristit ćemo programski jezik SAS uz pomoć knjige [1].

Poglavlje 1

Analiza doživljenja

1.1 Osnove analize doživljenja

Analiza doživljenja (engl. *survival analysis*) je skup statističkih metoda za analizu podataka kod kojih su varijable od interesa neki promatrani događaj (dihotomna varijabla) te vrijeme do događaja. Varijabla koja označava događaj govori dogodio li se događaj ili ne dok vrijeme do događaja (vrijeme doživljenja) označava vrijeme od početka studije do ranijeg od događaja ili kraja studije. Neki primjeri događaja koji se promatraju su smrt, greška u mehaničkom stroju, pojavljivanje neke bolesti, prestanak pušenja, remisija nakon tretmana...
Da bi se analiza doživljenja mogla provoditi podaci moraju sadržavati vrijeme do nekog događaja koji se prati. Dakle, ključno je da vrijeme praćenja bude dobro definirano, tj. mora se znati početak i kraj analiziranog razdoblja.
Ovaj skup statističkih metoda stekao je popularnost zbog svojih pozitivnih karakteristika:

- mogućnost procjene vjerojatnosti doživljenja nekog vremena umjesto očekivanog vremena doživljenja;
- može dobro rješavati problem cenzuriranja jer model sadrži informaciju dogodio li se promatrani događaj ili ne;
- dobro koristi varijable koje ovise o vremenu.

Primjer 1.1.1. *Promatramo osobe koje su izašle iz zatvora godinu dana nakon izlaska. Događaj od interesa je prvo uhićenje. Cilj je odrediti kako razne varijable (kovarijate) utječu na pojavljivanje događaja i vrijeme pojavljivanja.*
Ako podatke pokušamo analizirati logističkom regresijom, ignoriramo vrijeme uhićenja. Osobe koje su uhićene prvi tjedan nakon puštanja iz zatvora imaju veću sklonost uhićenju od osoba koje su uhićene u 52. tjednu.

Idući pokušaj bi bio da u model uključimo vrijeme koje je prošlo do uhićenja i promatramo linearnu regresiju. Sada imamo problem s osobama koje nisu uhićene u prvoj godini nakon puštanja iz zatvora. Takve slučajeve zovemo cenzuriranim. Kada bi izbacivali iz modela takve slučajeve moglo bi se dogoditi da izbacimo previše podataka.
*Također, postavlja se pitanje kako koristiti kovarijate koje ovise o vremenu (npr. zaposlila li se osoba). Mogli bismo za svaki tjedan u godini uvesti indikatorsku varijablu koja bi označavala status zaposlenja (*dummy *varijabla). Time bismo zakomplicirali model koji bi tada bio neefikasan.*
S druge strane, analiza doživljenja dopušta cenzuriranje i kovarijate koje ovise o vremenu.

Problemom cenzuriranja bavimo se u sljedećem poglavlju.

1.2 Cenzuriranje

Cenzuriranje predstavlja nepoznato vrijeme kada se neki događaj dogodio. Preciznije, imamo neku informaciju o vremenu doživljenja, ali ne znamo točno vrijeme.
Kao jednostavan primjer cenzuriranja, gledamo pacijente koji boluju od leukemije i pratimo ih sve dok ne izađu van remisije. Ako istraživanje završi, a neki pacijent je još uvijek u remisiji (kod njega se nije dogodio praćeni događaj), za njegovo vrijeme doživljenja kažemo da je cenzurirano. Razlikujemo tri vrste cenzuriranja:

- desno cenzuriranje;
- lijevo cenzuriranje;
- intervalno cenzuriranje.

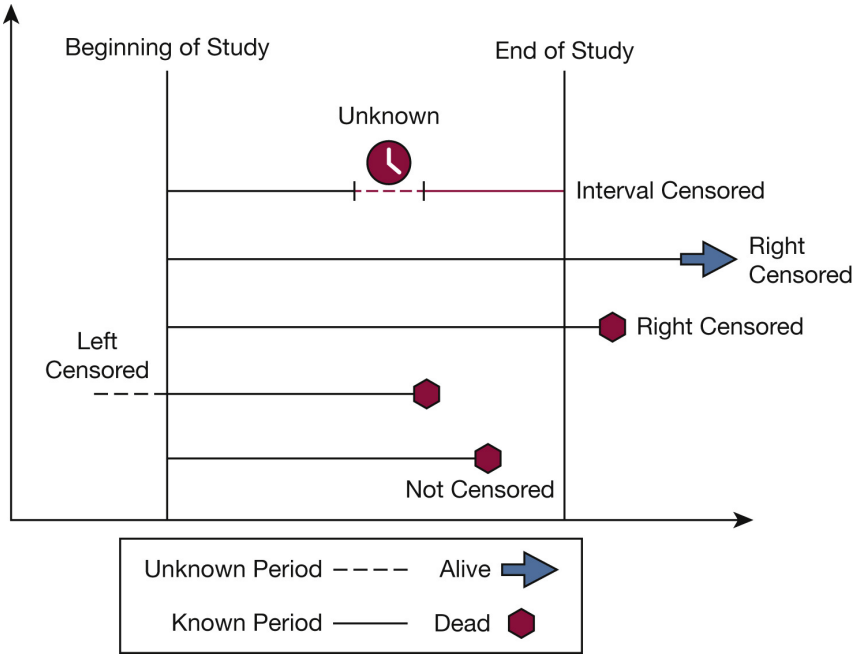
Kod *desno cenzuriranih* subjekata pravo vrijeme doživljenja veće je ili jednako od promatranog razdoblja, odnosno događaj se dogodio nakon vremena do kojeg se prati populacija ili se još uvijek nije dogodio. Za subjekte koji su povućeni iz istraživanja dok ono traje ili se izgubi kontakt s njima također kažemo da su desno cenzurirani.

Za subjekte kažemo da su *lijevo cenzurirani* ako je pravo vrijeme doživljenja manje ili jednako od promatranog vremena doživljenja, odnosno događaj se dogodio prije nekog poznatog vremena, ali ne znamo točan trenutak kada se dogodio. Naprimjer, možemo pratiti u kojoj dobi je neko dijete ispunilo određeni zadatak. Kažemo da je vrijeme doživljenja lijevo cenzurirano za dijete koje je već prije ulaska u studiju taj zadatak ispunilo.

Intervalno cenzurirani subjekti su oni kod kojih je pravo vrijeme doživljenja unutar promatranog intervala, ali ne znamo kada se točno dogodio događaj. Kao primjer pogledajmo istraživanje u kojem je događaj prvo pojavljivanje simptoma neke bolesti. Ako u trenutku

t_1 neka osoba nema simptome i u trenutku t_2 ($t_1 < t_2$) osoba ima simptome, događaj se dogodio negdje u intervalu (t_1, t_2). Tada za vrijeme doživljenja te osobe kažemo da je intervalno cenzurirano. Desno cenzurirani i lijevo cenzurirani subjekti su zapravo poseban slučaj intervalno cenzuriranih subjekata. Kod lijevo cenzuriranih subjekata imamo da je $t_1 = 0$ te je t_2 jednak vremenu cenzuriranja. U slučaju desnog cenzuriranja t_1 je jednak vremenu nakon kojeg osoba doživi događaj dok je $t_2 = +\infty$.

Različite vrste cenzuriranja možemo vidjeti na slici 1.1.



Slika 1.1: Vrste cenzuriranja. *Beginning of Study* i *End of Study* označavaju početak i kraj studije. *Left Censored, Right Censored*, *Interval Censored* i *Not Censored* označavaju redom lijevo, desno, intervalno cenzuriranje te podatak koji nije cenzuriran. *Unknown Period* i *Known Period* označavaju nepoznat, odnosno poznat period. *Alive* i *Dead* označavaju živog (događaj se nije dogodio) i neživog (događaj se dogodio) subjekta. (Izvor: `https://www.sciencedirect.com/science/article/pii/S0012369220304700`, 30.08.2023.)

1.3 Funkcija doživljenja

Neka je T nenegativna slučajna varijabla koja označava vrijeme do nekog događaja.

Definicija 1.3.1. Funkcija doživljenja *za nenegativnu slučajnu varijablu T je* $S : [0, +\infty) \to [0, 1]$ *definirana formulom*

$$S(t) = \mathbb{P}(T > t) = 1 - F_T(t), \tag{1.1}$$

gdje je F_T *funkcija distribucije slučajne varijable* T.

U slučaju da je T nenegativna i neprekidna slučajna varijabla postoji nenegativna funkcija f koju zovemo *funkcija gustoće slučajne varijable* T za koju vrijedi

$$F_T(x) = \int_{-\infty}^{x} f(t)dt = \int_{0}^{x} f(t)dt,\ x \geq 0.$$

Tada je funkcija doživljenja slučajne varijable T

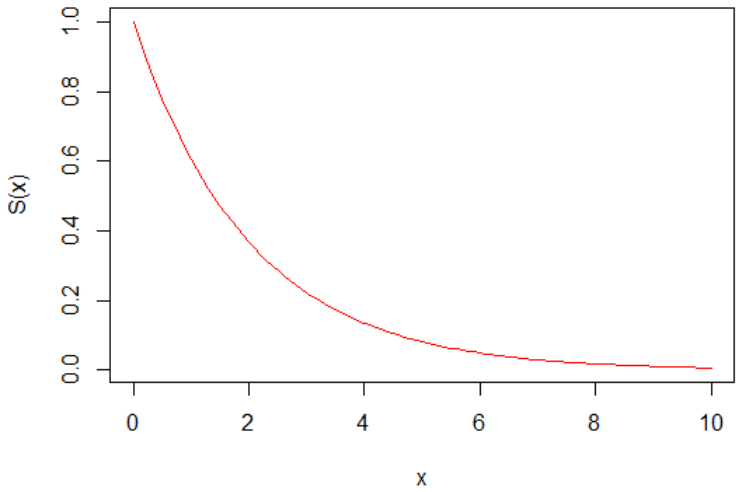
$$S(t) = \int_{t}^{\infty} f(x)dx,\ t \geq 0.$$

Funkcija doživljenja u točki t predstavlja vjerojatnost da se promatrani događaj dogodi nakon vremena t ili da se još uvijek nije dogodio. Tu vjerojatnost zovemo *vjerojatnost doživljenja vremena* t. Ona je fundamentalna u analizi doživljenja.
Funkcija doživljenja S ima sljedeća svojstva:

- nerastuća je;
- na početku studije (za $t = 0$) $S(t) = S(0) = 1$, tj. vjerojatnost doživljenja početnog vremena studije je 1;
- za vrijeme $t = +\infty$, $S(t) = S(+\infty) = 0$, tj. kada bi studija trajala beskonačno dugo za sve subjekte događaj bi se dogodio.

Primjer 1.3.2. *Pogledajmo nekoliko primjera funkcija doživljenja:*

1. *Neka je T slučajna varijabla koja ima eksponencijalnu distribuciju s parametrom* $\lambda = 0.5$. *Tada je funkcija doživljenja za varijablu T prikazana na slici 1.2.*

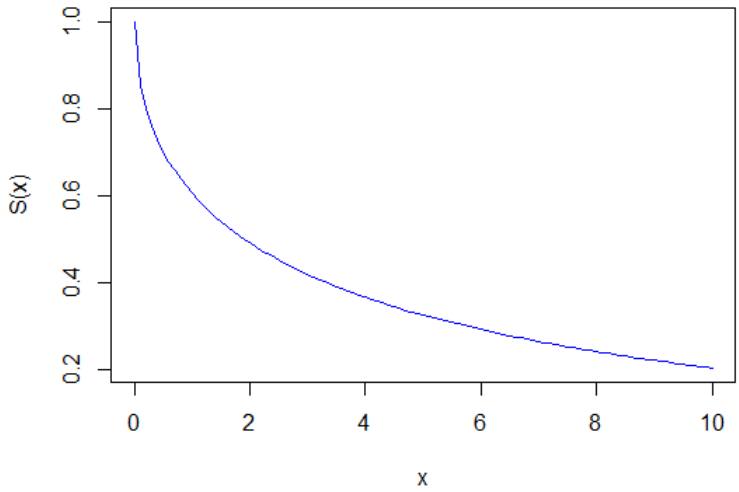


Slika 1.2: Eksponencijalna funkcija doživljenja ($\lambda = 0.5$)

2. *Neka je T slučajna varijabla koja ima Weibullovu distribuciju s parametrima* $\lambda = 0.5$ *i* $\gamma = 0.5$. *Tada je funkcija doživljenja od T*

$$S(t) = e^{-\lambda t^{\gamma}}, \ za\ t \geq 0,$$

prikazana na slici 1.3.

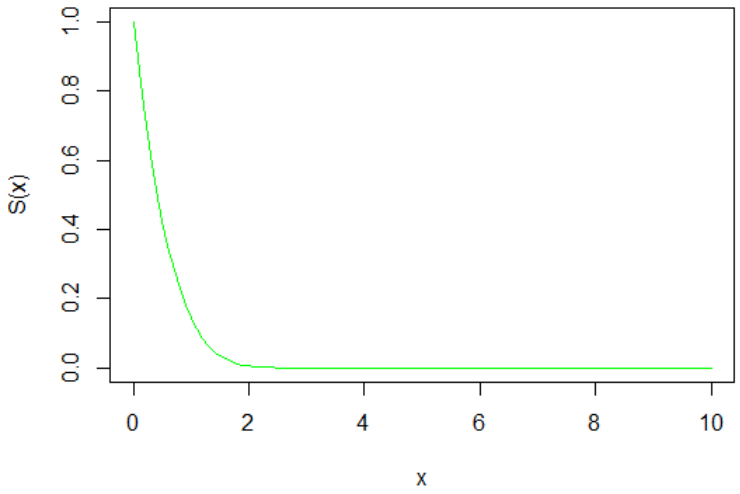


Slika 1.3: Weibullova funkcija doživljenja ($\lambda = 0.5$, $\gamma = 0.5$)

3. *Neka je T slučajna varijabla koja ima Gompertzovu distribuciju s parametrima $\gamma = 1.5$ i $\theta = 0.5$. Tada je funkcija doživljenja od T*

$$S(t) = e^{-\frac{\lambda}{\theta}(e^{\theta t}-1)}, \ za\ t \geq 0,$$

prikazana na slici 1.4.



Slika 1.4: Gompertzova funkcija doživljenja ($\gamma = 1.5$, $\theta = 0.5$)

1.4 Funkcija hazarda

Definicija 1.4.1. *Neka je T nenegativna slučajna varijabla.* Funkcija hazarda slučajne varijable T *je* $h : [0, +\infty) \to [0, +\infty)$ *zadana formulom*

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

Funkciju hazarda interpretiramo kao potencijal, po jedinici vremena, da se događaj dogodi neposredno nakon nekog trenutka, uz uvjet da je subjekt doživio taj trenutak. Funkcija hazarda u svakoj točki t direktno odgovara intuitivnom shvaćanju rizika da se neki događaj dogodio baš u vremenu t. Bitno je primijetiti da funkcija hazarda nije vjerojatnost. Zbog nazivnika u limesu iz definicije vrijednosti funkcije hazarda su u intervalu $[0, +\infty)$.

Neka je T neprekidna slučajna varijabla, f pripadna funkcija gustoće, F_T pripadna funkcija distribucije, S pripadna funkcija doživljenja te h pripadna funkcija hazarda. Tada

vrijedi:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t)}{\Delta t \cdot \mathbb{P}(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{F_T(t + \Delta t) - F_T(t)}{S(t)\Delta t} \\ &= \frac{\partial F_T(t)}{\partial t} \cdot \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{1.3}$$

$$\frac{\partial \ln S(t)}{\partial t} = \frac{\partial S(t)}{\partial t} \cdot \frac{1}{S(t)} = \frac{\partial (1 - F_T(t))}{\partial t} \cdot \frac{1}{S(t)} = -\frac{f(t)}{S(t)}. \tag{1.4}$$

Iz raspisa 1.3 i 1.4 dobivamo da vrijedi

$$h(t) = -\frac{\partial \ln S(t)}{\partial t}. \tag{1.5}$$

Definicija 1.4.2. Kumulativna funkcija hazarda za nenegativnu slučajnu varijablu T *je* $H : [0, +\infty) \to [0, +\infty)$ *zadana formulom*

$$H(t) = \int_0^t h(x)dx = -\ln S(t), \text{ gdje druga jednakost slijedi iz 1.5.}$$

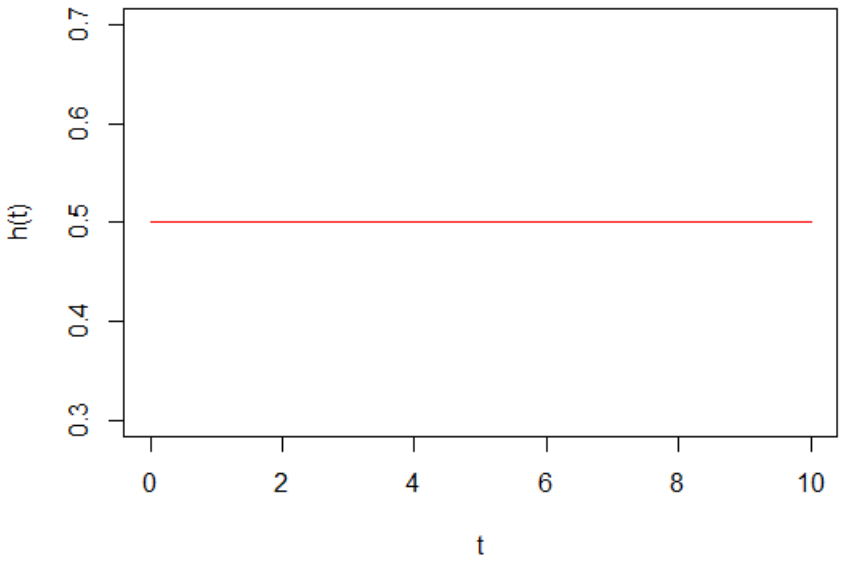
Sada iz definicije kumulativne funkcije hazarda slijedi da je

$$S(t) = e^{-H(t)}.$$

Dakle, da bi za neku neprekidnu slučajnu varijablu T odredili njezinu funkciju doživljenja, funkciju hazarda i kumulativnu funkciju hazarda, dovoljno je znati jednu od tih triju funkcija.

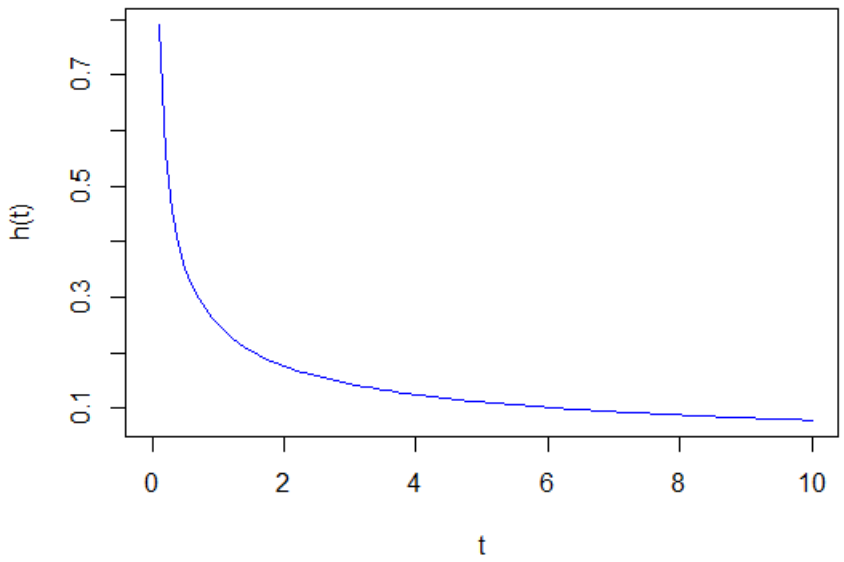
Primjer 1.4.3. *Pogledajmo nekoliko primjera funkcija hazarda:*

1. *Konstantna funkcija hazarda* $h(t) = \lambda$, *za* $\lambda = 0.5$ *i* $t \geq 0$ *(slika 1.5).*
   Možemo izvesti formulu za funkciju doživljenja $S(t) = e^{-\lambda t}$, *što je funkcija doživljenja za eksponencijalnu funkciju distribucije već viđenu u primjeru 1.3.2.*



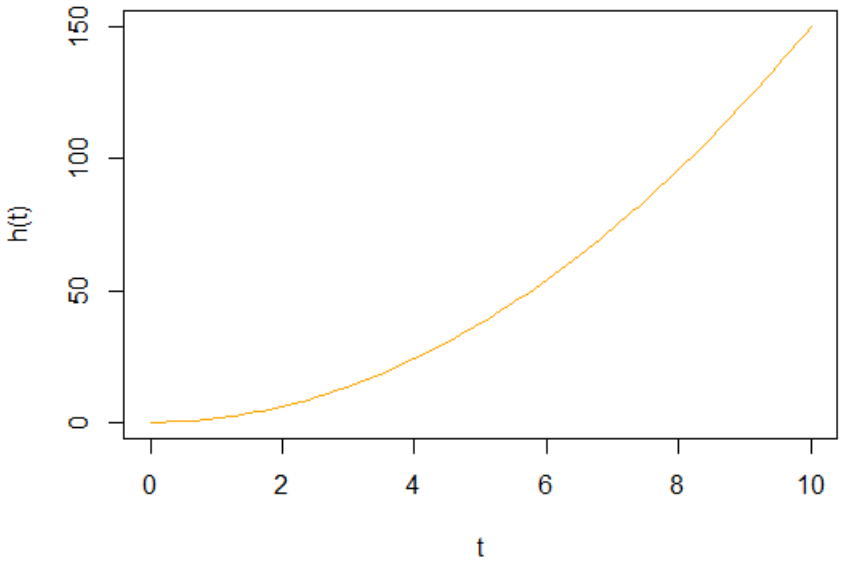
Slika 1.5: Konstantna funkcija hazarda ($\lambda = 0.5$)

2. *Padajuća funkcija hazarda* $h(t) = \gamma\lambda t^{\gamma-1}$, *za* $\gamma = 0.5$, $\lambda = 0.5$ *i* $t \geq 0$ *(slika 1.6). Tada dobivamo Weibullovu funkciju doživljenja* $S(t) = e^{-\lambda t^{\gamma}}$, *za* $t \geq 0$, *koju smo također vidjeli u primjeru 1.3.2. Padajuće funkcije hazarda u primjeni možemo pronaći u slučajevima kad postoji velika vjerojatnost da se događaj dogodi vrlo rano, npr. kvar u mehaničkim strojevima koji imaju neku grešku, smrt kod pacijenata koji primaju transplantant.*



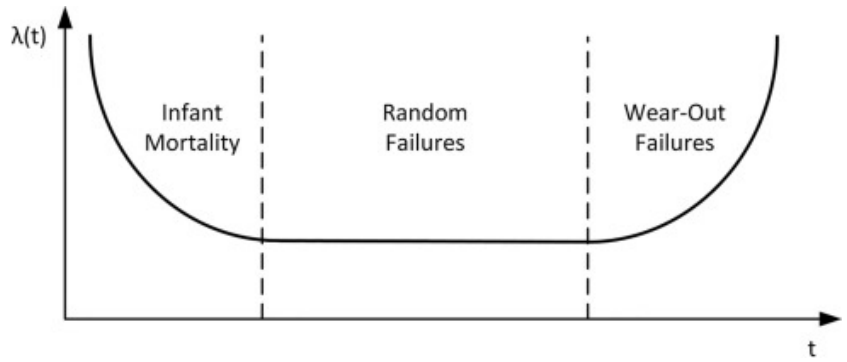
Slika 1.6: Padajuća funkcija hazarda ($\gamma = 0.5,\ \lambda = 0.5$)

3. *Rastuća funkcija hazarda (slika 1.7). Slično kao u prošlom primjeru, ali uzimamo drugačije parametre.*
   $h(t) = \gamma\lambda t^{\gamma-1}$, *za* $\gamma = 3,\ \lambda = 0.5$ *i* $t \geq 0$.
   Rastuća funkcija hazarda u primjeni se javlja kod populacija s prirodnim starenjem ili istrošenošću.



Slika 1.7: Rastuća funkcija hazarda ($\gamma = 3,\ \lambda = 0.5$)

4. *Funkcija hazarda u obliku olova* U *prikazana je na slici 1.8. Najčešća je u primjeni jer se javlja kod mortaliteta ljudske populacije koja se prati od rođenja. U najranijem razdoblju do smrti dolazi zbog bolesti dojenčadi, nakon toga dolazi do stabilizacije pa onda do povećane smrtnosti zbog prirodnog procesa starenja.*



Slika 1.8: Funkcija hazarda u obliku slova *U*. Apscisa i ordinata su vrijeme t i funkcija hazarda u vremenu t, dok su *Infant Mortality, Random Failures, Wear-Out Failures* redom smrtnost dojenčadi, slučajni događaji te događaji do kojih dovodi proces istrošenosti (starenja). (Izvor: `https://www.sciencedirect.com/topics/engineering/hazard-rate`, 30.08.2023.)

Neka je T diskretna slučajna varijabla koja poprima vrijednosti u skupu $\{t_1, t_2, ... \mid t_1 < t_2 < ...\}$. Tada je funkcija hazarda dana formulom

$$h(t_j) = \mathbb{P}(T = t_j | T \geq t_j) = \frac{\mathbb{P}(T = t_j)}{\mathbb{P}(T \geq t_j)} = \frac{\mathbb{P}(T > t_{j-1}) - \mathbb{P}(T > t_j)}{S(t_{j-1})} = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})},$$

za $j = 2, 3, ...$
S druge strane, funkciju doživljenja za diskretnu slučajnu varijablu možemo zapisati kao

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

Dakle, funkciju doživljenja možemo izraziti pomoću funkcije hazarda

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

Uočimo da je kod diskretnih slučajnih varijabli funkcija hazarda jednaka nuli osim u točkama iz domene slučajne varijable (u tim vremenima se događaj može dogoditi).

1.5 Kaplan-Meierov procjenitelj funkcije doživljenja

Kada želimo provoditi analizu doživljenja nad dobivenim podacima najčešće nemamo nikakve informacije o distribuciji vremena do događaja. Tada je potrebno procijeniti funkciju

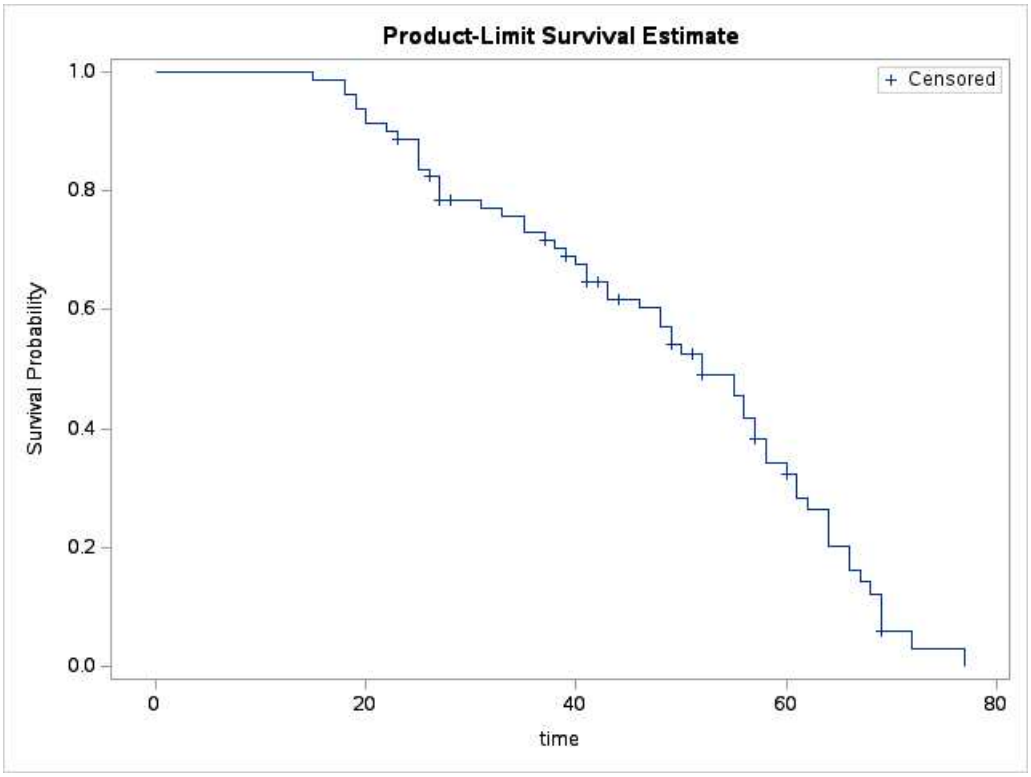
doživljenja.
Najpoznatiji neparametarski procjenitelj za funkciju doživljenja je Kaplan-Meierov procjenitelj. Edward L. Kaplan i Paul Meier predložili su, svaki zasebno, slične ideje, a zatim su zajedno 1958. godine predložili procjenitelj funkcije doživljenja danas poznat kao Kaplan-Meierov procjenitelj.
Neka je T nenegativna slučajna varijabla koja označava vrijeme do nekog događaja. Neka su $t_1, t_2, ..., t_D$, $t_1 < t_2 < ... < t_D$, vrijednosti koje varijabla može poprimiti (moguća vremena do događaja). Pretpostavimo da se u trenutku t_i dogodi d_i događaja, $i = 1, 2, ..., D$. Neka je y_i broj jedinki koje su "rizične" u trenutku t_i. "Rizične" su one jedinke kod kojih se događaj još nije dogodio i nisu cenzurirane prije trenutka t_i. Ako je neka jedinka cenzurirana baš u trenutku t_i, nju također smatramo "rizičnom" u trenutku t_i. Faktor $\frac{d_i}{y_i}$ predstavlja uvjetnu vjerojatnost da jedinka doživi događaj u trenutku t_i, uz uvjet da se za nju događaj nije dogodio do trenutka t_i.
Kaplan-Meierov procjenitelj funkcije doživljenja definira se kao

$$\hat{S}(t) = \begin{cases} 1, & t < t_1 \\ \prod\limits_{\substack{i=1 \\ t_i \leq t}}^{D} \left(1 - \frac{d_i}{y_i}\right), & t \geq t_1. \end{cases} \tag{1.6}$$

Ako je najveće opaženo vrijeme doživljenja (t_D) ujedno i vrijeme do događaja (nije cenzurirano) onda je Kaplan-Meierov procjenitelj jednak nuli za vremena veća od t_D. Ako je pak t_D cenzurirano onda Kaplan-Meierov procjenitelj nije definiran za vremena veća od t_D jer ne znamo kada bi jedinka s najvećim vremenom doživljenja doživjela događaj da nije cenzurirana. Tom problemu pokušalo se doskočiti tako da se definiralo $\hat{S}(t) = \hat{S}(t_D)$, za $t > t_D$.
Varijancu Kaplan-Meierovog procjenitelja možemo procjeniti Greenwoodovom formulom:

$$\hat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \cdot \sum_{\substack{i=1 \\ t_i \leq t}}^{D} \frac{d_i}{y_i(y_i - d_i)}.$$

Primjer 1.5.1. *Kaplan-Meierova procjena funckije doživljenja (slika 1.9). Kaplan-Meierov procjenitelj je funkcija u obliku stepenica (engl.* step function*) koja ima skokove u trenucima u kojima su opaženi događaji. Veličina skoka u nekom trenutku ovisi o broju opaženih događaja u tom trenutku i o broju cenzuriranih podataka do tog trenutka. Cenzurirani podaci su označeni plusom.*



Slika 1.9: Grafički prikaz Kaplan-Meierove procjene funkcije doživljenja. Prikazne su procijenjene vjerojatnosti doživljenja (*Survival Probability*) s obzirom na vrijeme (*time*).

Kaplan-Meierov procjenitelj možemo koristiti i za procjenu kumulativne funkcije hazarda $H(t) = -\ln S(t)$. Procijenjena kumulativna funkcija hazarda je

$$\hat{H}(t) = -\ln \hat{S}(t).$$

Alternativno, za procjenu kumulativne funkcije hazarda možemo koristiti Nelson-Aalenov procjenitelj definiran kao

$$\hat{H}(t) = \begin{cases} 0, & t < t_1 \\ \sum\limits_{\substack{i=1 \\ t_i \leq t}}^{D} \frac{d_i}{y_i}, & t_1 \leq t. \end{cases}, \tag{1.7}$$

gdje su $d_1, d_2, ..., d_D$ i $t_1, t_2, ..., t_D$ kao u 1.6. Nelson-Aalenov procjenitelj definiran je samo za trenutke manje od najvećeg opaženog trenutka. U praksi se pokazalo da je Nelson-Aalenov procjenitelj za kumulativnu funkciju hazarda bolji od Kaplan-Meierovog za male

uzorke. Procjena varijance kumulativne funkcije hazarda pomoću Nelson-Aalenovog procjenitelja dana je s

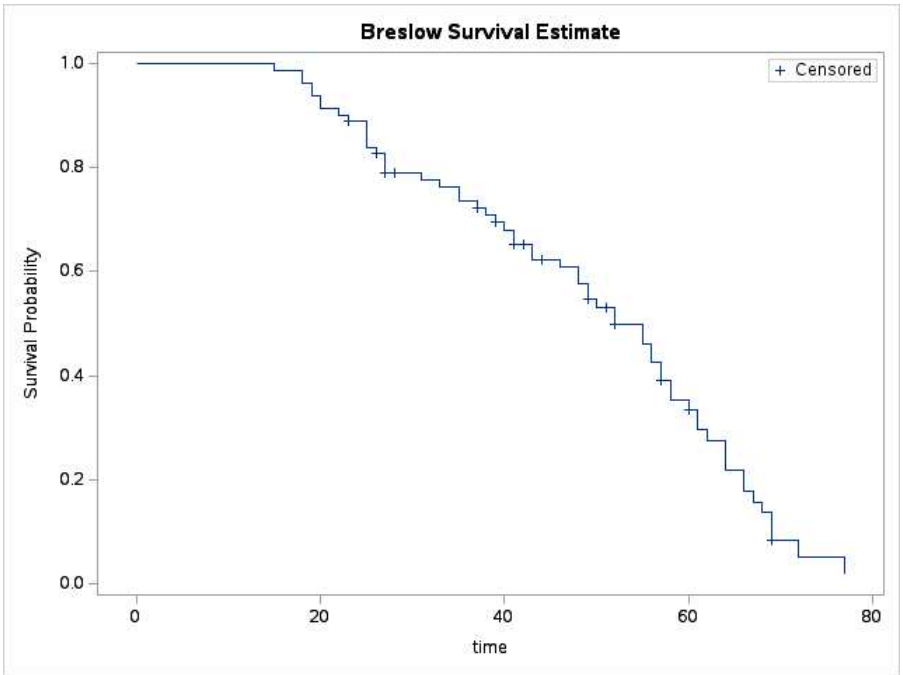
$$\sigma^2_H(t) = \sum_{\substack{i=1 \\ t_i \leq t}}^{D} \frac{d_i}{y_i^2}.$$

Pomoću Nelson-Aalenovog procjenitelja kumulativne funkcije hazarda možemo dobiti alternativni procjenitelj funkcije doživljenja

$$\hat{S}(t) = e^{-\hat{H}(t)}.$$

Nelson-Aalenov procjenitelj koristi se za odabir parametarskog modela za vrijeme do događaja. Naprimjer, graf funkcije $\hat{H}(t)$ će biti otprilike linearan ako eksponencijalna funkcija dobro opisuje vremena do događaja. Također, koristi se i za grubu procjenu funkcije hazarda.

Primjer 1.5.2. *Nelson-Aalenov procjenitelj funkcije doživljenja (slika 1.10). Koristimo iste podatke kao i u primjeru 1.5.1.*



Slika 1.10: Grafički prikaz Nelson-Aalenove procjene funkcije doživljenja. Prikazne su procijenjene vjerojatnosti doživljenja (*Survival Probability*) s obzirom na vrijeme (*time*).

1.6 Testiranje hipoteza za dva ili više uzoraka

Pretpostavimo da imamo K ($K \geq 2$) populacija. Želimo usporediti funkcije doživljenja K populacija. Neka je $S_i(t)$ vrijednost funkcije doživljenja i-te populacije u trenutku t, $t \geq 0$, $i = 1, ..., K$. Testiramo sljedeće hipoteze:
$H_0 : S_1(t) = S_2(t) = ... = S_K(t)$, za sve $t \leq \tau$;
H_1 : barem neki od $S_i(t)$, $i \in 1, 2, ..., K$, su međusobno različiti za neki $t \leq \tau$.
Ovdje je τ najveće opaženo vrijeme za koje sve grupe imaju barem jednu "rizičnu" jedinku.
Alternativno, možemo testirati sljedeće hipoteze:
$H_0 : h_1(t) = h_2(t) = ... = h_K(t)$, za sve $t \leq \tau$;
H_1 : barem neki od $h_j(t)$, $j \in 1, 2, ..., K$, su međusobno različiti za neki $t \leq \tau$.
Neka su $t_1, ..., t_D$, $t_1 < ... < t_D$, vremena događaja u ujedinjenom uzorku svih K nezavisnih uzoraka. Neka je d_{ij} broj događaja s vremenom t_i u j-tom uzorku i neka je y_{ij} broj jedinki koje su "rizične" u vremenu t_i u j-tom uzorku, za $j = 1, ..., K$, $i = 1, ..., D$. Definiramo $d_i = \sum_{j=1}^{K} d_{ij}$ i $y_i = \sum_{j=1}^{K} y_{ij}$ kao broj događaja i broj "rizičnih" jedinki u ujedinjenom uzorku u vremenu t_i, $i = 1, ..., D$.
Test hipoteza baziran je na usporedbi težinskih razlika između procijenjene funkcije hazarda za j-ti uzorak pod nultom i alternativnom hipotezom, koristeći ranije viđene Nelson-Aalenove procjenitelje. Neka je $W_j(t)$ pozitivna težinska funkcija takva da je $W_j(t_i) = 0$ kada je $y_{ij} = 0$. Testna statistika, uz uvjet da vrijedi H_0, je

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i)\left(\frac{d_{ij}}{y_{ij}} - \frac{d_i}{y_i}\right), \ j = 1, ..., K.$$

U većini testova koji se upotrebljavaju u primjeni uzima se da je $W_j(t_i) = y_{ij}W(t_i)$, $j = 1, ..., K$, $i = 1, ..., D$. Tada je $W(t_i)$ zajednička težina u točki t_i koju dijele svi uzorci. Sada dobivamo novu formulu za testnu statistiku

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i)\left(d_{ij} - y_{ij}\left(\frac{d_i}{y_i}\right)\right), \ j = 1, ..., K.$$

Primjetimo da je tada testna statistika suma težinskih razlika između opaženih brojeva događaja koji su se dogodili i očekivanog broja događaja pod pretpostavkom nulte hipoteze u j-tom uzorku. Očekivani broj događaja u j-tom uzorku u vremenu t_i je proporcija "rizičnih" jedinki $\frac{y_{ij}}{y_i}$ u j-tom uzorku u vremenu t_i pomnožena s brojem događaja u vremenu t_i.
Varijanca od $Z_j(\tau)$ dana je formulom

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{y_{ij}}{y_i}\left(1 - \frac{y_{ij}}{y_i}\right)\left(\frac{y_i - d_i}{y_i - 1}\right)d_i, \ j = 1, ..., K$$

dok je kovarijanca od $Z_j(\tau)$ i $Z_g(\tau)$ dana s

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{y_{ij}}{y_i}\frac{y_{ig}}{y_i}\left(\frac{y_i - d_i}{y_i - 1}\right) d_i,\ g \neq j.$$

Komponente vektora $(Z_1(\tau), ..., Z_K(\tau))$ su linearno zavisne jer je $\sum_{j=1}^{K} Z_j(\tau) = 0$. Testna statistika se konstruira tako da se izabere proizvoljnih $K-1$ komponenta vektora $(Z_1(\tau), ..., Z_K(\tau))$. Procijenjena kovarijacijska matrica tih statistika je $\Sigma \in M_{K-1,K-1}(\mathbb{R})$ formirana pomoću pripadnih $\hat{\sigma}_{jg}$. Testna statistika, uz pretpostavku da je nulta hipoteza istinita, dana je s

$$(Z_1(\tau), ..., Z_{K-1}(\tau))\Sigma^{-1}(Z_1(\tau), ..., Z_{K-1}(\tau))^T \ \sim\ \chi^2(K-1).$$

Kada je $K = 2$, testna statistika može se zapisati kao

$$Z = \frac{\sum_{i=1}^{D} W(t_i)\left(d_{i1} - y_{i1}\left(\frac{d_i}{y_i}\right)\right)}{\sqrt{\sum_{i=1}^{D} W(t_i)^2 \left(\frac{y_{i1}}{y_i}\right)\left(1 - \frac{y_{i1}}{y_i}\right)\left(\frac{y_i - d_i}{y_i - 1}\right) d_i}}.$$

Prethodna statistika ima standardnu normalnu distribuciju za velike uzorke kada je nulta hipoteza istinita.
Najpoznatija težinska funkcija je $W(t) = 1$, za sve t. Takav izborom funkcije dolazimo do *log-rank testa*. Za težinsku funkciju možemo uzeti i $W(t_i) = y_i$. Takva funkcija daje generalizaciju Mann-Whitney-Wilcoxonovog testa za dva uzorka. Ako želimo dati više težine razlikama između opaženih i očekivanih brojeva događaja koji su se dogodili u j-tom uzorku u vremenima gdje imamo najviše podataka, onda možemo uzeti da je $W(t_i) = y_i^{1/2}$. Fleming i Harrington (1981.) predložili su široku klasu testova koji kao specijalne slučajeve uključuju *log-rang test* i verziju Mann-Whitney-Wilcoxonovog testa. Neka je $\hat{S}(t)$ Kaplan-Meierov procjenitelj funkcije doživljenja definiran kao i ranije. Tada je

$$W_{p,q}(t_i) = \hat{S}(t_{i-1})^p \left(1 - \hat{S}(t_{i-1})\right)^q,\ p \geq 0,\ q \geq 0,\ i = 1, ..., D.$$

Ovdje je funkcija doživljenja u prethodnom vremenu događaja korištena kako bi se osiguralo da su težine poznate samo prethodno vremenu u kojem se radi usporedba. Uzimamo da je $S(t_0) = 1$.
Kada je $p = q = 0$ imamo *log-rank test*, a kada je $p = 1$, $q = 0$ imamo verziju Mann-Whitney-Wilcoxonovog testa.

Poglavlje 2

Coxov regresijski model

2.1 Osnove Coxovog regresijskog modela

Do sada smo se bavili s varijablama koje su opisivale dogodio li se neki događaj ili ne i kad se dogodio (ako se je dogodio). Te varijable zvati ćemo zavisnim varijablama. U ovom poglavlju uz zavisne varijable zanimat će nas i nezavisne varijable. To su neke karakteristike subjekata za koje ćemo ispitivati utječu li na zavisne varijable. Naprimjer, svaki subjekt u nekom istraživanju ima spol, godinu rođenja, visinu, težinu...
Također, nezavisne varijable možemo koristiti da bi predvidjeli ishod i vrijeme događaja za subjekte s određenim karakteristikama. Nezavisne varijable još nazivamo kovarijatama. Pretpostavimo da imamo uzorak $(T_j, \delta_j, \mathbb{X}_j)$, $j = 1, ..., n$, gdje je T_j vrijeme događaja ili cenzuriranja j-tog subjekta, δ_j je indikatorska varijabla koja označava dogodio li se za j-tog subjekta događaj ili ne te $\mathbb{X}_j = (X_{j1}, ..., X_{jp})$ vektor kovarijata za j-tog subjekta. Neka je $h(t|\mathbb{X})$ funkcija hazarda u trenutku t za subjekta s vektorom kovarijata $\mathbb{X}$. Osnovni Coxov model glasi:

$$h(t|\mathbb{X}) = h_0(t)c(\boldsymbol{\beta}^T\mathbb{X}),$$

gdje je h_0 bazna funkcija hazarda, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ su koeficijenti (parametri modela) i c je poznata funkcija. Ovakav model naziva se semiparametarski jer distribucija bazne funkcije nije poznata te se ona tretira neparametarski, dok je parametarski model pretpostavljen za efekt kovarijata.
Coxov regresijski model pretpostavlja da je

$$c(\boldsymbol{\beta}^T\mathbb{X}) = e^{\sum_{i=1}^{p}\beta_i X_i}$$

pa je onda

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

Logaritmiranjem Coxovog regresijskog modela dobivamo:

$$\ln h(t|\mathbb{X}) = \ln h_0(t) + \sum_{i=1}^{p} \beta_i X_i.$$

Dakle, ako povećamo X_i za 1 u bilo kojem trenutku t i sve ostale kovarijate držimo konstantnima dobivamo promjenu u log-hazardu koja iznosi upravo β_i.

Coxov regresijski model naziva se još *model proporcionalnog hazarda*. Taj naziv je dobio jer je omjer funkcija hazarda za subjekte s kovarijatama $\mathbb{X}$ i $\mathbb{X}^*$ konstantan:

$$HR := \frac{h(t|\mathbb{X})}{h(t|\mathbb{X}^*)} = \frac{h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}}{h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i^*}} = e^{\sum_{i=1}^{p} \beta_i (X_i - X_i^*)}.$$

Omjer funkcija hazarda označavamo s *HR* (engl. *hazard ratio*). Uzmimo za primjer prvog pacijenta koji je primio prvi tretman, tj. $X_1 = 1$ (njegove kovarijate označavamo s $\mathbb{X}$) te drugog pacijenta koji je primio drugi tretman, tj. $X_1^* = 2$ (njegove kovarijate označavamo s $\mathbb{X}^*$). Pretpostavimo da sve ostale kovarijate (osim X_1 i X_1^*) imaju iste vrijednosti. Tada je $\frac{h(t|\mathbb{X})}{h(t|\mathbb{X}^*)} = e^{\beta_1}$, što označava rizik pojavljivanja događaja za subjekte koji primaju prvi tretman relativno s obzirom na rizik pojavljivanja događaja za subjekte koji primaju drugi tretman.

Napomenimo još da je Coxov model moguće definirati za kovarijate $\mathbb{X}$ koje ovise o vremenu t. Tada vektor kovarijata označavamo s $\mathbb{X}(t) = (X_1(t), ..., X_p(t))$. Za takve kovarijate kažemo da su vremenski promjenjive. Takav model zovemo *prošireni Coxov model*:

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

Naravno, takav model više nije model proporcionalnog hazarda jer omjer funkcija hazarda više nije konstantan, tj. ovisi o vremenu.

2.2 Procjena parametara

Parametri Coxovog regresijskog modela su koeficijenti $\beta_1, ..., \beta_p$ uz kovarijate $X_1, ..., X_p$. Procjenjujemo ih metodom temeljenoj na metodi maksimalne vjerodostojnosti (engl. *maximum likelihood estimation*). Formula za Coxovu funkciju vjerodostojnosti zapravo je "parcijalna" jer uzima u obzir samo vjerojatnosti onih subjekata kod kojih se dogodio događaj, a zanemaruje one koji su cenzurirani.
Kao i ranije, imamo uzorak $(T_j, \delta_j, \mathbb{X}_j)$, $j = 1, ..., n,$. Pretpostavimo da su $t_{(1)} < t_{(2)} < ... <$

$t_{(D)}$ uređena vremena događaja te da su ona nezavisna s vremenima cenzuriranja za svaki subjekt. Neka je $X_{(i)k}$ k-ta kovarijata subjekta čije vrijeme događaja je $t_{(i)}$. Označimo s $R(t_i)$ skup svih subjekata koji su rizični u vremenu $t_{(i)}$, tj. još uvijek sudjeluju u studiji malo prije vremena $t_{(i)}$. Coxova funkcija parcijalne vjerodostojnosti dana je formulom

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{e^{\sum_{k=1}^{p} \beta_k X_{(i)k}}}{\sum_{j \in R(t_i)} e^{\sum_{k=1}^{p} \beta_k X_{jk}}}.$$

Iako ovo nije "prava" funkcija vjerodostojnosti, s njoj postupamo kako bi i inaće postupali s funkcijom vjerodostojnosti. Numerički je jednostavnije naći maksimum logaritma funkcije vjerodostojnosti.

$$LL(\boldsymbol{\beta}) := \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k X_{(i)k} - \sum_{i=1}^{D} \ln \left(\sum_{j \in R(t_{(i)})} e^{\sum_{k=1}^{p} \beta_k X_{jk}} \right).$$

Procjenu od $\boldsymbol{\beta}$ dobivamo tražeći maksimum funkcije LL pa želimo naći stacionarne točke.

$$U_l(\boldsymbol{\beta}) := \frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_l} = \sum_{i=1}^{D} X_{(i)l} - \sum_{i=1}^{D} \frac{\sum_{j \in R(t_{(i)})} X_{jl} e^{\sum_{k=1}^{p} \beta_k X_{jk}}}{\sum_{j \in R(t_{(i)})} e^{\sum_{k=1}^{p} \beta_k X_{jk}}}. \tag{2.1}$$

Sustav $U_l = 0$, za $l = 1, ..., p$ rješavamo nekom iterativnom metodom. Najpopularnija je Newton-Raphsonova metoda kod koje $\boldsymbol{\beta}$ procjenjujemo na sljedeći način:

$$\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + I^{-1}(\hat{\boldsymbol{\beta}}) U(\hat{\boldsymbol{\beta}}),\ r = 0, 1, 2, ...$$

$I(\boldsymbol{\beta})$ je $p \times p$ dimenzionalna matrica nenegativnih drugih derivacija od $LL(\boldsymbol{\beta})$ koju nazivamo matrica informacija. Algoritam kreće od $\hat{\boldsymbol{\beta}}_0 = \mathbf{0}$ te se zaustavlja kad je razlika funkcija log-vjerodostojnosti dovoljno mala ili kad je najveća relativna promjena u vrijednosti procijenjenog parametra zadovoljavajuće mala.

2.3 Testovi za testiranje hipoteza o koeficijentu $\boldsymbol{\beta}$

Tri su glavna testa za testiranje hipoteza o koeficijentu $\boldsymbol{\beta}$. Neka je $\hat{\boldsymbol{\beta}}$ (parcijalna) procjena od $\boldsymbol{\beta}$ metodom maksimalne vjerodostojnosti, $I(\boldsymbol{\beta})$ pripadna matrica informacija i $\boldsymbol{\beta}_0$ p-dimenzionalan vektor.

- Prvi test temelji se na asimptotskoj normalnosti procjenitelja (parcijalne) funkcije maksimalne vjerodostojnosti, poznat kao Waldov test. Za veliki uzorak, $\hat{\boldsymbol{\beta}}$ ima p-dimenzionalnu normalnu distribuciju s očekivanjem $\boldsymbol{\beta}$ i kovarijacijskom matricom procjenjenom s $I^{-1}(\hat{\boldsymbol{\beta}})$.
  Testiramo početnu hipotezu $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ i koristimo testnu statistiku

  $$X_W^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \ \sim \ \chi^2(p), \text{ uz pretpostavku da vrijedi nulta hipoteza.}$$

- Sljedeći test je test omjera vjerodostojnosti (engl. *likelihood ratio*) koji se jos zove - *2log likelihood*. Opet imamo nultu hipotezu $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ te koristimo testnu statistiku

  $$X_{LR}^2 = 2(LL(\hat{\boldsymbol{\beta}}) - LL(\boldsymbol{\beta}_0)) \ \sim \ \chi^2(p), \text{ uz pretpostavku da vrijedi nulta hipoteza za veliki uzorak.}$$

- Posljednji test je test skorova (engl. *score test*). Neka je $U(\boldsymbol{\beta}) = (U_1(\boldsymbol{\beta}), ..., U_p(\boldsymbol{\beta}))^T$, gdje je $U_l(\boldsymbol{\beta})$, $l = 1, ..., p$, definiran u 2.1. Za velike uzorke, uzima se da $U(\boldsymbol{\beta})$ ima p-dimenzionalnu normalnu distribuciju s očekivanjem 0 i kovarijacijskom matricom $I(\boldsymbol{\beta})$, kada je istinita nulta hipoteza $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. Testna statistika je tada

  $$X_{SC}^2 = U(\boldsymbol{\beta}_0)^T I^{-1}(\boldsymbol{\beta}_0) U(\boldsymbol{\beta}_0) \ \sim \ \chi^2(p).$$

2.4 Parcijalna vjerodostojnost kada postoje podudaranja u vremenima događaja

Do sada smo promatrali slučaj kod kojeg su vremena događaja bila različita za sve subjekte. Zbog načina na koji su vremena bilježena u istraživanjima (npr. umjesto starosti u danima promatramo starost u godinama) često imamo slučaj gdje postoji više subjekata s istim vremenom događaja.
Neka su $t_1 < t_2 < ... < t_D$ različita, uređena vremena događaja te neka je d_i broj događaja u vremenu t_i. Označimo s $\boldsymbol{A}_i$ skup svih subjekata kod kojih se dogodio događaj u vremenu t_i. Neka je $\boldsymbol{s}_i = \sum_{j \in \boldsymbol{A}_i} \mathbb{X}_j$, suma vektora kovarijata svih subjekata kod kojih se dogodio događaj u vremenu t_i. Kao i ranije, uzmimo da je R_i skup svih rizičnih subjekata u vremenu t_i.
Postoji puno različitih prijedloga kako računati parcijalnu vjerodostojnost u ovom slučaju. Najpopularniji su Breslowov, Efronov te Coxov.
Breslow (1974) predlaže da se svaki od d_i događaja u vremenu t_i smatra različitim i pridonosi vjerodostojnosti množeći sa svim događajima u vremenu t_i.

$$L_B(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{s}_i}}{\left(\sum_{j \in R_i} e^{\boldsymbol{\beta}^T \mathbb{X}_j} \right)^{d_i}}$$

Ova aproksimacija radi dobro kada je malo podudaranja u vremenima događaja i ona se koristi u većini statističkih paketa.
Efron (1977) predlaže parcijalnu vjerodostojnost koja je bliža vjerodostojnosti bez podudaranja.

$$L_E(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{s}_i}}{\prod_{j=1}^{d_i} \left(\sum_{k \in R_i} e^{\boldsymbol{\beta}^T \mathbb{X}_k} - \frac{j-1}{d_i} \sum_{k \in \boldsymbol{A}_i} e^{\boldsymbol{\beta}^T \mathbb{X}_k} \right)}$$

Kada je broj podudaranja mali, Breslowove i Efronove parcijalne vjerodostojnosti su slične.
Cox uzima da $h(t|\mathbb{X})$ označava uvjetnu vjerojatnost događaja u intervalu $(t, t+1)$ uz dano doživljenje do početka intervala i pretpostavlja da je

$$\frac{h(t|\mathbb{X})}{1 - h(t|\mathbb{X})} = \frac{h_0(t)}{1 - h_0(t)} e^{\boldsymbol{\beta}^T \mathbb{X}}.$$

Označimo s Q_i skup svih podskupova od d_i subjekata iz skupa rizičnih R_i. Neka je $q = (q_1, ..., q_{d_i})$ jedan od elemenata iz Q_i i neka je $\boldsymbol{s}_q^* = \sum_{j=1}^{d_j} \mathbb{X}_{qj}$. Tada je

$$L_C(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{e^{\boldsymbol{\beta}^T \boldsymbol{s}_i}}{\sum_{q \in Q_i} e^{\boldsymbol{\beta}^T \boldsymbol{s}_q^*}}.$$

2.5 Procjena funkcije doživljenja

Jednom kada smo procijenili parametar $\boldsymbol{\beta}$ iz Coxovog regresijskog modela, od interesa nam je procjeniti vjerojatnost doživljenja nekog subjekta sa zadanim vektorom kovarijata $\mathbb{X}$, tj. hoćemo procijeniti funkciju doživljenja. Procjenitelj se temelji na Breslowovom procjenitelju za kumulativnu funkciju hazarda.
Neka je $\hat{\boldsymbol{\beta}}$ procjena parametra Coxovog regresijskog modela metodom parcijalne maksimalne vjerodostojnosti i neka je $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})$ procjenjena kovarijacijska matrica dobivena iz inverza matrice informacija. Neka su $t_1 < t_2 < ... < t_D$ različita vremena događaja i neka je d_i broj događaja u vremenu t_i, za $i = 1, ..., D$. Definiramo

$$W(t_i; \hat{\boldsymbol{\beta}}) = \sum_{j \in R(t_i)} e^{\sum_{l=1}^{p} \hat{\beta}_l X_{jl}}.$$

Procjenitelj kumulativne bazne funkcije hazarda $H_0(t) = \int_0^t h_0(x)\, dx$ je dan s

$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i; \hat{\boldsymbol{\beta}})}$, što je *step* funkcija sa skokovima u opaženim vremenima događaja.

Procjenitelj bazne funkcije doživljenja $S_0(t) = e^{-H_0(t)}$, što je zapravo funkcija doživljenja za subjekta s vektorom kovarijata $\mathbb{X} = \mathbf{0}$, je

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)}.$$

Da bi procijenili funkciju doživljenja za subjekta s vektorom kovarijata $\mathbb{X}_0$ koristimo procjenitelj

$$\hat{S}(t|\mathbb{X} = \mathbb{X}_0) = \hat{S}_0(t)^{exp\left(\hat{\boldsymbol{\beta}}^T \mathbb{X}_0\right)}.$$

Pod nekim uvjetima ovaj procjenitelj, za fiksni t, ima asimptotsku normalnu distribuciju s očekivanjem $S(t|\mathbb{X} = \mathbb{X}_0)$ i varijancom

$$\hat{V}\left(\hat{S}(t|\mathbb{X} = \mathbb{X}_0)\right) = \left(\hat{S}(t|\mathbb{X} = \mathbb{X}_0)\right)^2 (Q_1(t) + Q_2(t; \mathbb{X}_0)).$$

Ovdje je

$Q_1(t) = \sum_{t_i \le t} \frac{d_i}{W(t_i; \hat{\boldsymbol{\beta}})^2}$, procjenitelj varijance od $\hat{H}_0(t)$ u slučaju kad je $\hat{\boldsymbol{\beta}}$ prava vrijednosti od $\boldsymbol{\beta}$.

$$Q_2(t; \mathbb{X}_0) = Q_3(t; \mathbb{X}_0)^T \hat{V}(\hat{\boldsymbol{\beta}}) Q_3(t; \mathbb{X}_0)$$

Q_2 odražava nepouzdanost u procjeni od $\boldsymbol{\beta}$ s $\hat{\boldsymbol{\beta}}$. Q_3 je p-dimenzionalni vektor čiji je k-ti element dan s

$$Q_3(t; \mathbb{X}_0)_k = \sum_{t_i \le t} \left(\frac{W^{(k)}(t_i; \hat{\boldsymbol{\beta}})}{W(t_i; \hat{\boldsymbol{\beta}})} - X_{0k} \right) \left(\frac{d_i}{W(t_i; \hat{\boldsymbol{\beta}})} \right), \; k = 1, ..., p,$$

gdje je

$$W^{(k)}(t_i; \hat{\boldsymbol{\beta}}) = \sum_{j \in R(t_i)} X_{jk} e^{\hat{\boldsymbol{\beta}}^T \mathbb{X}_j}.$$

Koristeći ovu procjenu varijance možemo konstruirati pouzdane intervale za funkciju doživljenja subjekta s vektorom kovarijata $\mathbb{X}_0$ označenu s $S(t|\mathbb{X} = \mathbb{X}_0)$.

2.6 Interakcija

Kada želimo predvidjeti vrijednosti zavisne varijable za neke zadane kovarijate ključno je znati kako pojedina nezavisna varijabla utječe na zavisnu varijablu. Tada veliku ulogu imaju interakcije. Kažemo da su dvije nezavisne varijable u interakciji ako utjecaj jedne nezavisne varijable na zavisnu varijablu ovisi o vrijednosti druge nezavisne varijable. Jednostavnije rečeno, jedna kovarijata mijenja efekt druge na zavisnu varijablu. U tom slučaju,

da bi predvidjeli vrijednosti zavisne varijable za neke zadane kovarijate, nije dovoljno znati samo koeficijente u modelu uz kovarijate.
Zbog jednostavnosti zapisa, pretpostavimo da imamo model s dvije nezavisne varijable X_1 i X_2 te interakcijom između njih. Tada je

$$h(t|\boldsymbol{X}) = h_0(t)e^{\beta_1 X_1+\beta_2 X_2+\beta_3(X_1\times X_2)}, \text{ iz čega slijedi}$$

$$h(t|\boldsymbol{X}) = h_0(t)e^{(\beta_1+\beta_3 X_2)X_1+\beta_2 X_2}.$$

Iz modela lako uočimo da ako je interakcija pristuna, efekt jedne kovarijate mijenja se promjenom druge i obratno. Kada želimo vidjeti ima li u nekom modelu interakcije, testiramo hipotezu $H_0 : \beta_3 = 0$. U kompliciranijim modelima možemo imati više interakcija i veći broj (tri ili više) kovarijata u interakciji.

2.7 Karakteristike Coxovog regresijskog modela

Najvažnija karakteristika zaslužna za popularnost Coxovog regresijskog modela je njegova robusnost. Čak i kada ne znamo baznu funkciju hazarda te samim time ni funkciju hazarda, Coxov regresijski model daje relativno dobre procjenitelje parametara i predikcije zavisne varijable. U idealnoj situaciji znamo točan parametarski model (npr. Weibullov ili eksponencijalni), ali to većinom nije slučaj. Coxov regresijski model tada predstavlja "siguran" izbor u smislu da znamo da imamo relativno dobar model.
Također, Coxov regresijski model je bolji izbor od logistički regresije u slučaju kad imamo dostupne podatke o vremenima doživljenja i kad imamo cenzuriranje.
S druge strane, mana ovog modela je ovisnost o pretpostavci proporcionalnog hazarda. Da bi provjerili pretpostavku proporcionalnosti hazarda za fiksnu kovarijatu X_1, definiramo vremenski ovisnu kovarijatu X_2

$$X_2(t) = X_1 \times g(t), \text{ gdje je } g(t) \text{ poznata funkcija vremena } t \text{ (npr. } \ln t).$$

Uzimanjem Coxovog regresijskog modela s varijablama X_1 i X_2, dobijemo parametre β_1 i β_2 te gledamo omjer hazarda za dva subjekta s različitim vrijednostima kovarijata X_1 i X_1^*

$$\frac{h(t|\mathbb{X})}{h(t|\mathbb{X}^*)} = e^{\beta_1\left(X_1-X_1^*\right)+\beta_2 g(t)\left(X_1-X_1^*\right)}.$$

Ovaj omjer ovisi o vremenu ako $\beta_2 \neq 0$. Dakle, proporcionalnost hazarda testiramo hipotezom $H_0 : \beta_2 = 0$.
Kada pretpostavka nije zadovoljena i X_1 je binarna varijabla (poprima vrijednosti 0 ili 1) možemo definirati vremenski ovisnu varijablu X_2

$$X_2(t) = \begin{cases} 0, & t \leq \tau \\ X_1, & t > \tau. \end{cases}$$

Tada imamo

$$h(t|\mathbb{X}(t)) = \begin{cases} h_0(t)e^{\beta_1 X_1}, & t \leq \tau \\ h_0(t)e^{\beta_1 + (\beta_1 + \beta_2)X_1}, & t > \tau, \end{cases}$$

gdje je $h_0(t)$ bazna funkcija hazarda i τ fiksirano vrijeme. Uočimo da sada imamo model proporcionalnog hazarda.

Poglavlje 3

Analiza ponovnog pojavljivanja raka dojke analizom doživljenja

3.1 Podaci

Rak dojke je zloćudna bolest koja nastaje kad normalne žljezdane stanice dojke promijene svoja svojstva te počnu nekontrolirano rasti, umnožavati se i uništavati okolno zdravo tkivo. Takve promijenjene stanice mogu potom otići u limfne i/ili krvne žile te tako proširiti bolest u druge dijelove tijela. Od raka dojke najčešće obolijevaju žene iznad pedesete godine života, ali u najnovije vrijeme sve češće obolijevaju i mlađe žene. Muškarci također mogu oboljeti od raka dojke, ali znatno rjeđe nego žene. Jedan posto svih zabilježenih slučajeva raka dojke zabilježen je kod muškog spola. Kada stanice raka otiđu u limfne čvorove radi se o raku dojke s pozitivnim čvorovima. Postojanje stanica raka u limfnim čvorovima obično znači da postoji veća šansa ponovnog pojavljivanja i širenja raka nakon operacije. Zbog toga je kod analize doživljenja bitno razlikovati radi li se o pacijentima s pozitivnim čvorovima.
Podaci koji se analiziraju u ovom radu prikupljeni su od strane *German Breast Study Group* (GBSG) od srpnja 1984. do prosinca 1989. (podacima se može pristupiti na linku `https://stat.ethz.ch/R-manual/R-devel/library/survival/html/gbsg.html`). U istraživanju je sudjelovalo 720 pacijenata s rakom dojke s pozitivnim čvorovima, od čega su za njih 686 prikupljeni potpuni podaci. Svi pacijenti u istraživanju bile su žene. Svrha istraživanja bila je da se ispita učinkovitost tri naspram šest ciklusa kemoterapije i dodatnog hormonalnog liječenja s tamoksifenom. Prikupljali su se podaci za vrijeme do ponovnog pojavljivanja raka i još deset drugih varijabli. Vrijeme do ponovnog pojavljivanja raka definira se kao vrijeme (u danima) od primarne operacije do prvog ponovnog pojavljivanja raka dojke ili smrti bilo kojeg uzroka. Maksimalno vrijeme praćenja bilo je sedam godina. Promatrane su sljedeće varijable:

- *status* - događaj koji pratimo, ponovno pojavljivanje raka dojke ili smrt nakon operacije (0 za pacijenta koji je živ bez ponovnog pojavljivanja raka, 1 za pacijenta koji je umro ili je došlo do ponovnog pojavljivanja raka)
- *rfstime* - (engl. *recurrence free survival time*) dani od operacije do ponovnog pojavljivanja raka dojke ili smrti
- *pid* - (engl. *patient identifier*) identifikacijski broj pacijenta
- *age* - starost (u godinama)
- *meno* - menopauzalni status (0 prije menopauze, 1 postmenopauza)
- *size* - veličina tumora (u milimetrima)
- *grade* - stadij tumora, moguće vrijednosti su 1, 2 i 3
- *nodes* - broj pozitivnih limfnih čvorova
- *pgr* - progesteronski receptori (u fmol/l)
- *er* - estrogen receptori (u fmol/l)
- *hormon* - hormonska terapija (0 ne, 1 da)

Opisnom statistikom opisat ćemo varijable te proučiti veze između njih da bi imali bolju predodžbu koje kovarijate imaju utjecaj na zavisnu varijablu. Nadalje, provodit ćemo analizu doživljenja kojom ćemo pokušati procijeniti i interpretirati funkcije doživljenja te objasniti utjecaj kovarijata na vjerojatnost doživljenja. Coxovom regresijom pokušat ćemo odrediti statistički značajne prediktore za događaj smrti ili ponovnog pojavljivanja raka (recidiva). Modelom koji ćemo dobiti mogli bi predviđati vrijeme do događaja za pacijente s određenim vrijednostima kovarijata.

3.2 Opisna statistika i veze između varijabla

U tablici 3.1 nalaze se osnovne informacije o dvije najvažnije varijable, *status* i *rfstime*. Primijetimo da imamo malo više pacijenata za koje se događaj nije dogodio (56.41%) od onih koji su doživjeli događaj. Raspon vremena do događaja je od 8 dana do 2659 dana, a artimetička sredina je 1124.49 dana.

status				
status	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	387	56.41	387	56.41
1	299	43.59	686	100.00

Analysis Variable : rfstime						
Minimum	Lower Quartile	Mean	Median	Variance	Upper Quartile	Maximum
8.000	567.000	1124.490	1084.000	413181.488	1685.000	2659.000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
8	460	2551	316
15	215	2556	569
16	329	2563	283
17	632	2612	253
17	95	2659	609

Tablica 3.1: Opisna statistika za varijable *status* i *rfstime* (ispis iz SAS-a)

U tablici 3.2 možemo iščitati osnovne statistike za varijable *age, size, nodes, pgr* i *er*. Kontingencijske tablice 3.3, 3.4 i 3.5 redom prikazuju broj, postotak, redčani postotak i stupčani postotak pacijenata po grupama određenim s varijablom *status* te *meno, hormon* i *grade*.

Variable	Minimum	Lower Quartile	Mean	Median	Skewness	Kurtosis	Variance	Upper Quartile	Maximum
age	21.000	46.000	53.052	53.000	-0.146	-0.363	102.429	61.000	80.000
size	3.000	20.000	29.329	25.000	1.776	5.325	204.382	35.000	120.000
nodes	1.000	1.000	5.010	3.000	2.885	13.313	29.981	7.000	51.000
pgr	0.000	7.000	109.996	32.500	4.786	35.073	40938.057	132.000	2380.000
er	0.000	8.000	96.252	36.000	3.088	12.505	23434.700	115.000	1144.000

Tablica 3.2: Opisna statistika za varijable *age, size, nodes, pgr* i *er* (ispis iz SAS-a)

Table of meno by status			
meno	status		
Frequency Percent Row Pct Col Pct	0	1	Total
0	171 24.93 58.97 44.19	119 17.35 41.03 39.80	290 42.27
1	216 31.49 54.55 55.81	180 26.24 45.45 60.20	396 57.73
Total	387 56.41	299 43.59	686 100.00

Statistic	DF	Value	Prob
Chi-Square	1	1.3301	0.2488
Likelihood Ratio Chi-Square	1	1.3322	0.2484
Continuity Adj. Chi-Square	1	1.1564	0.2822
Mantel-Haenszel Chi-Square	1	1.3282	0.2491
Phi Coefficient		0.0440	
Contingency Coefficient		0.0440	
Cramer's V		0.0440	

Tablica 3.3: Kontingencijska tablica i χ^2 (*Chi-Square*) test za varijablu *meno* (ispis iz SAS-a)

Table of hormon by status			
hormon	status		
Frequency Percent Row Pct Col Pct	0	1	Total
0	235 34.26 53.41 60.72	205 29.88 46.59 68.56	440 64.14
1	152 22.16 61.79 39.28	94 13.70 38.21 31.44	246 35.86
Total	387 56.41	299 43.59	686 100.00

Statistic	DF	Value	Prob
Chi-Square	1	4.5058	0.0338
Likelihood Ratio Chi-Square	1	4.5316	0.0333
Continuity Adj. Chi-Square	1	4.1714	0.0411
Mantel-Haenszel Chi-Square	1	4.4992	0.0339
Phi Coefficient		-0.0810	
Contingency Coefficient		0.0808	
Cramer's V		-0.0810	

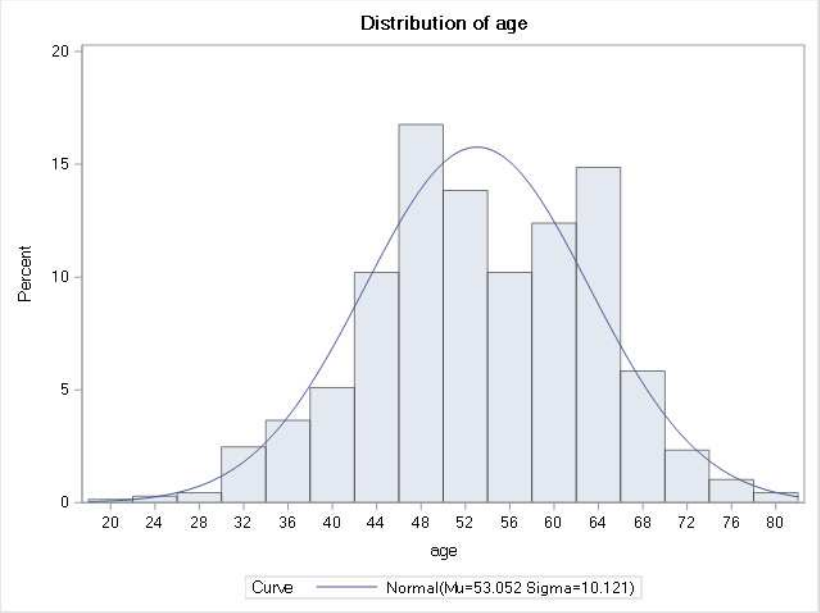
Tablica 3.4: Kontingencijska tablica i χ^2 (*Chi-Square*) test za varijablu *hormon* (ispis iz SAS-a)

Table of grade by status			
grade	status		
Frequency Percent Row Pct Col Pct	0	1	Total
1	63 9.18 77.78 16.28	18 2.62 22.22 6.02	81 11.81
2	242 35.28 54.50 62.53	202 29.45 45.50 67.56	444 64.72
3	82 11.95 50.93 21.19	79 11.52 49.07 26.42	161 23.47
Total	387 56.41	299 43.59	686 100.00

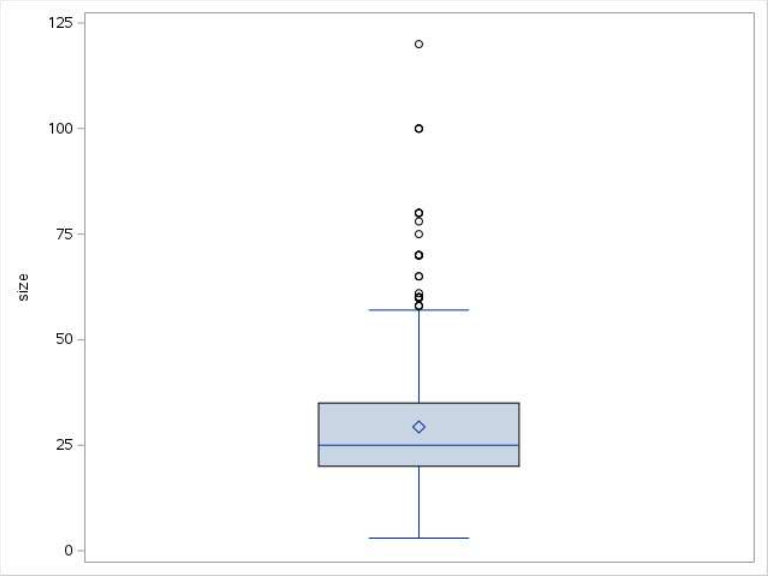
Statistic	DF	Value	Prob
Chi-Square	2	17.6615	0.0001
Likelihood Ratio Chi-Square	2	18.8220	<.0001
Mantel-Haenszel Chi-Square	1	11.9182	0.0006
Phi Coefficient		0.1605	
Contingency Coefficient		0.1584	
Cramer's V		0.1605	

Tablica 3.5: Kontingencijska tablica i χ^2 (*Chi-Square*) test za varijablu *grade* (ispis iz SAS-a)

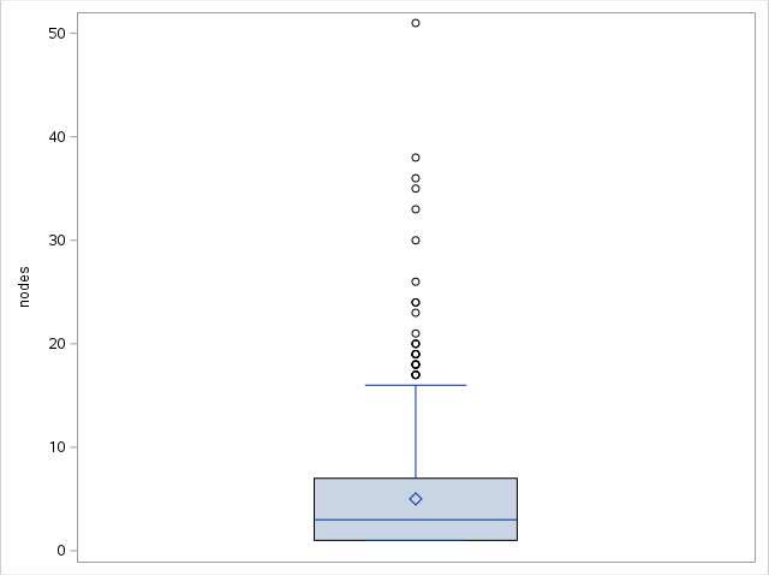
U tablici 3.2 vidimo da su najmlađe pacijentice imale tek dvadesetak godina. Promatranjem distribucije pacijentica po starosti na slici 3.1 uočavamo da je najviše pacijentica u istraživanju imalo između 42 i 70 godina.
Raspon veličina tumora je od 3 mm pa čak do 120 mm. Iz dijagrama pravokutnika (engl. *box and whisker plota*) na slici 3.2 možemo primjetiti da je srednja veličina 29.33 mm, ali imamo dosta izuzetaka (engl. *outliera*) s mnogo većim veličinama.
Za varijable *nodes, pgr* i *er* također imamo puno *outliera* s velikim vrijednostima, pogotovo za *pgr* i *er*. Dijagrame pravokutnika za navedene varijable možemo vidjeti na slikama 3.3, 3.4 i 3.5.



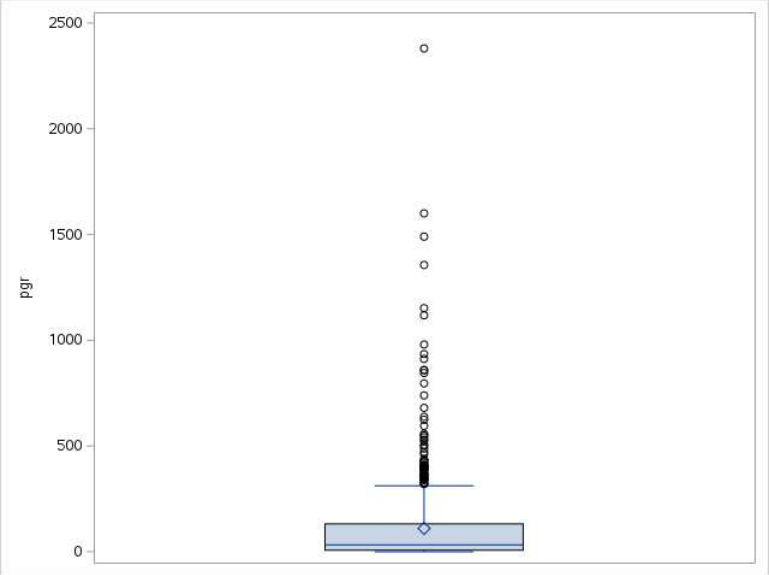
Slika 3.1: Histogram za varijablu *age*



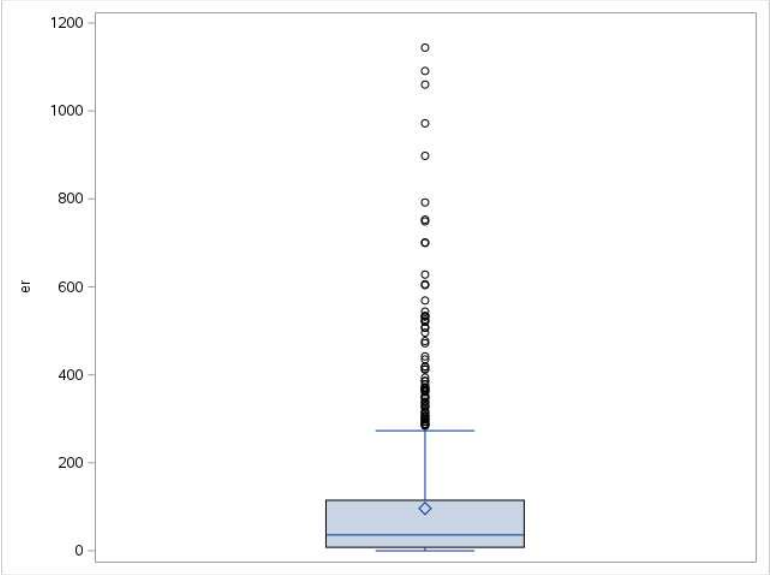
Slika 3.2: Dijagram pravokutnika za varijablu *size*



Slika 3.3: Dijagram pravokutnika za varijablu *nodes*



Slika 3.4: Dijagram pravokutnika za varijablu *pgr*



Slika 3.5: Dijagram pravokutnika za varijablu *er*

Za varijable *age, size, nodes, pgr* i *er* može se lako provjeriti da ne zadovoljavaju pretpostavku normalnosti distribucije. Zbog toga, ako bi željeli istražiti ponaša li se npr. starost pacijentica jednako s obzirom na varijablu *status* trebali bismo, umjesto *t testa* koji pretpostavlja normalnost distribucije, provoditi Wilcoxonov test za dva uzorka. Rezultate testova možemo vidjeti u tablici 3.6. Zaključak donosimo na temelju p-vrijednosti testa (stupac Pr > Z). Na razini značajnosti od 5%, za varijablu *age* ne odbacujemo nultu hipotezu testa te zaključujemo da nema statistički značajne razlike u starosti pacijentica po grupama s obzirom na događaj. Za varijable *size, nodes, pgr* i *er* odbacujemo nultu hipotezu na istoj razini značajnosti i zaključujemo da distribucija tih varijabla nije ista za pacijentice koje su doživjele događaj i one koje nisu doživjele događaj.
Za varijable *meno, hormon* i *grade*, pomoću χ^2 (*chi-square*) testa možemo testirati postoji li značajna veza između varijable *status* i konkretne diskretne varijable. Tablice 3.3, 3.4 i 3.5 pokazuju da na razini značajnosti od 5% za varijablu *meno* ne odbacujemo nezavisnost s varijablom *status*. S druge strane, zaključujemo da postoji veza između varijable *status* i varijabla *hormon* te *grade*. Te veze ćemo detaljnije promatrati u nastavku.

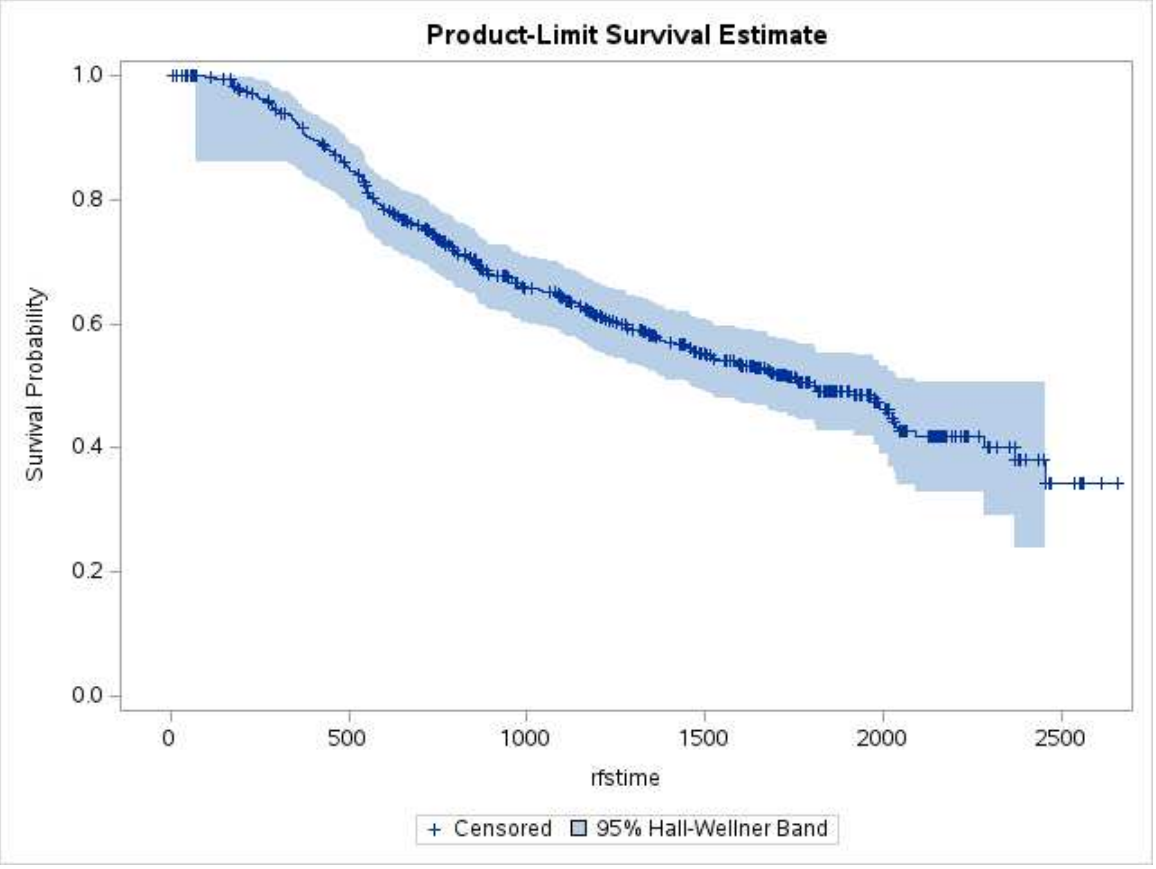
Wilcoxonov test za dva uzorka			
Varijabla	Statistika	Z	Pr > Z
age	103748.0	0.4047	0.3429
size	110853.5	3.1699	0.0008
nodes	120383.0	6.9624	<0.0001
pgr	87955.00	-5.7376	<0.0001
er	94299.50	-3.2691	0.0006

Tablica 3.6: Rezultati Wilcoxonovog testa za dva uzorka za varijable *age, size, nodes, pgr* i *er*

Ipak, u ovom radu prvenstveno se bavimo analizom doživljenja. Cilj nam je promatrati kako vrijeme do događaja (varijabla *rfstime*) utječe na vjerojatnost doživljenja te kako ostale varijable (kovarijate) utječu na vjerojatnost doživljenja nekog trenutka.

3.3 Kaplan-Meierove procjene funkcije doživljenja

Kaplan-Meierov procjenitelj najpoznatiji je procjenitelj funkcije doživljenja. Osim njega, funkciju doživljenja možemo procjenjivati Nelson-Aalenovim procjeniteljem te aktuarskom procjenom. U početku, Kaplan-Meierova procjena iznosi 1 i ona se smanjuje kako raste broj događaja. Na slici 3.6 vidimo da vjerojatnost doživljenja ne pada sve do 0 budući da je nekoliko pacijentica s najvećim vrijednostima varijable *rfstime* cenzurirano. Zbog toga je vjerojatnost doživljenja u krajnjem desnom rubu konstantna i iznosi 0.3428. Procijenjena vjerojatnost doživljenja za vrijednosti varijable *rfstime* veće od otprilike 2000 pada manjom brzinom zbog malog broja događaja s velikim vremenima doživljenja. U tablici 3.7 možemo vidjeti do kojeg vremena određeni postotak pacijentica doživi događaj te ukupan broj i postotak cenzuriranih događaja.



Slika 3.6: Kaplan-Meierova procjena funkcije doživljenja

Summary Statistics for Time Variable rfstime

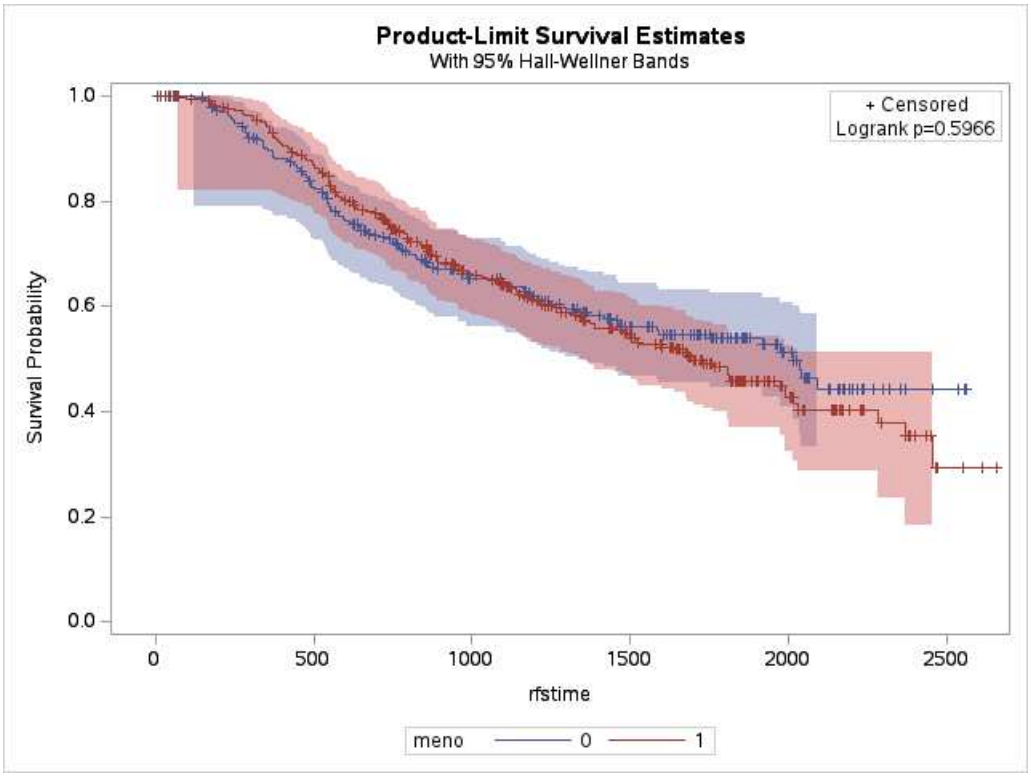
Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2456.00	.
50	1807.00	LOGLOG	1528.00	2018.00
25	727.00	LOGLOG	600.00	799.00

Mean	Standard Error
1591.69	36.23

Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
686	299	387	56.41

Tablica 3.7: Položajne vrijednosti i deskriptivna statistika za varijablu *rfstime* (ispis iz SAS-a)

Pomoću Kaplan-Meierovih procjena funkcija doživljenja i *log-rank* testa ćemo ispitivati imaju li varijable *meno*, *grade* i *hormon* utjecaja na vremena doživljenja. Podijelit ćemo pacijentice po grupama s obzirom na vrijednosti navedenih varijabla pa onda uspoređivati procjene funkcija doživljenja.
Nakon podjele u grupe po varijabli *meno*, sa slike 3.7 uočavamo da se procijenjene funkcije doživljenja i njihovi intervali pouzdanosti isprepleću i ne vidimo jasno odstupanje jedne od druge. Zaključujemo da ne možemo tvrditi da postoji statistički značajna razlika u funkcijama doživljenja dviju grupa. Zaključak potvrđuje *log-rang* test iz tablice 3.8 u kojem ne odbacujemo nultu hipotezu koja kaže da nema razlike u funkcijama doživljenja na razini značajnosti od 5%. Odluku o odbacivanju ili neodbacivanju donosimo iščitavanjem p vrijednosti (*Pr > Chi-Square*) za *log rank* test.



Slika 3.7: Kaplan-Meierove procjene funkcija doživljenja po grupama za varijablu *meno*

Summary Statistics for Time Variable rfstime (meno=0)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	2015.00	LOGLOG	1463.00	.
25	648.00	LOGLOG	542.00	801.00

Mean	Standard Error
1452.67	46.25

Summary Statistics for Time Variable rfstime (meno=1)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2456.00	.
50	1701.00	LOGLOG	1449.00	1990.00
25	745.00	LOGLOG	624.00	859.00

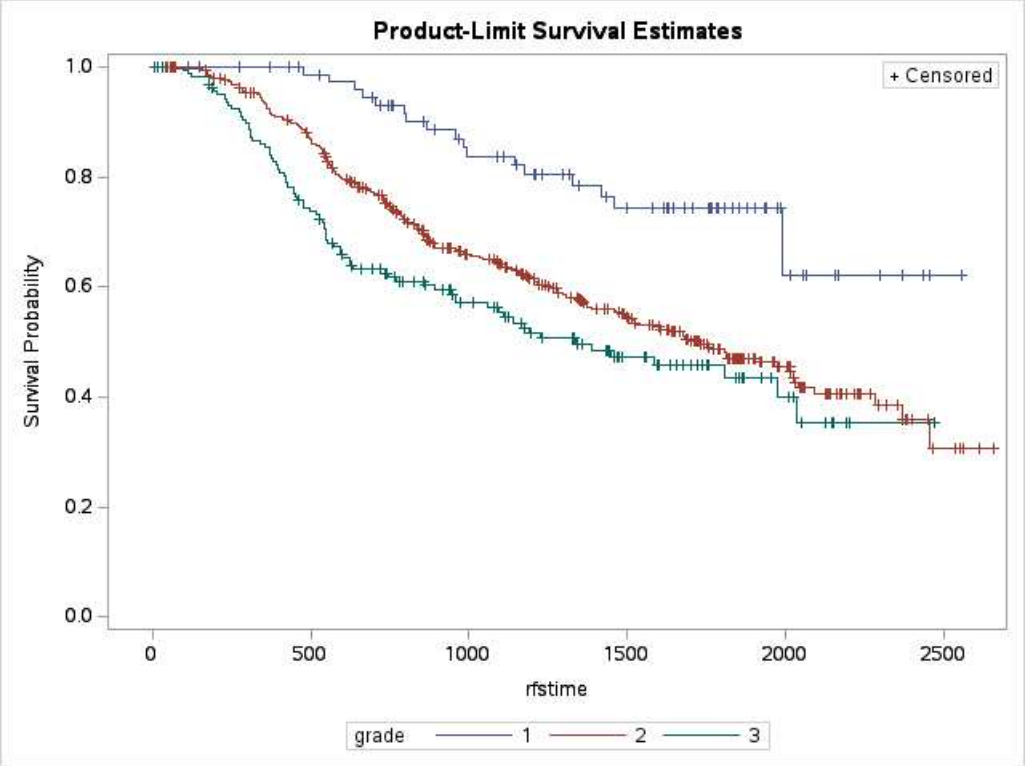
Mean	Standard Error
1578.36	46.60

Summary of the Number of Censored and Uncensored Values					
Stratum	meno	Total	Failed	Censored	Percent Censored
1	0	290	119	171	58.97
2	1	396	180	216	54.55
Total		686	299	387	56.41

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.2802	1	0.5966
Wilcoxon	0.1176	1	0.7317
-2Log(LR)	0.3906	1	0.5320

Tablica 3.8: Položajne vrijednosti, deskriptivna statistika i *Log-rang* test za *meno* (ispis iz SAS-a)

Kod podjele u grupu po stadijima tumora uočavamo jasnu razliku između najnižeg stadija (1) i viših stadija kod procijenjenih funkcija doživljenja (slika 3.8). Za stadije tumora 2 i 3 razlika nije toliko uočljiva. Za svaki trenutak, pacijentica sa stadijem tumora 1 ima veću vjerojatnost doživljenja od pacijentica sa stadijem tumora 2 ili 3. U tablici 3.9 uočavamo da postoji statistički značajna razlika u funkcijama doživljenja za grupe po stadijima tumora. Također, tablica sadrži rezultate analize koja se vrši kako bi se vidjelo koje grupe prave razliku (*post hoc* analiza). Vidimo da najznačajniju razliku imamo između grupa s najvećim i najmanjim stadijem tumora.



Slika 3.8: Kaplan-Meierove procjene funkcija doživljenja po grupama za varijablu *grade*

Summary Statistics for Time Variable rfstime (grade=1)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	.	LOGLOG	1990.00	.
25	1459.00	LOGLOG	982.00	.

Mean	Standard Error
1729.76	59.71

Summary Statistics for Time Variable rfstime (grade=2)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2456.00	.
50	1730.00	LOGLOG	1481.00	2018.00
25	745.00	LOGLOG	629.00	838.00

Mean	Standard Error
1581.85	43.96

Summary Statistics for Time Variable rfstime (grade=3)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2034.00	.
50	1337.00	LOGLOG	956.00	2034.00
25	476.00	LOGLOG	394.00	552.00

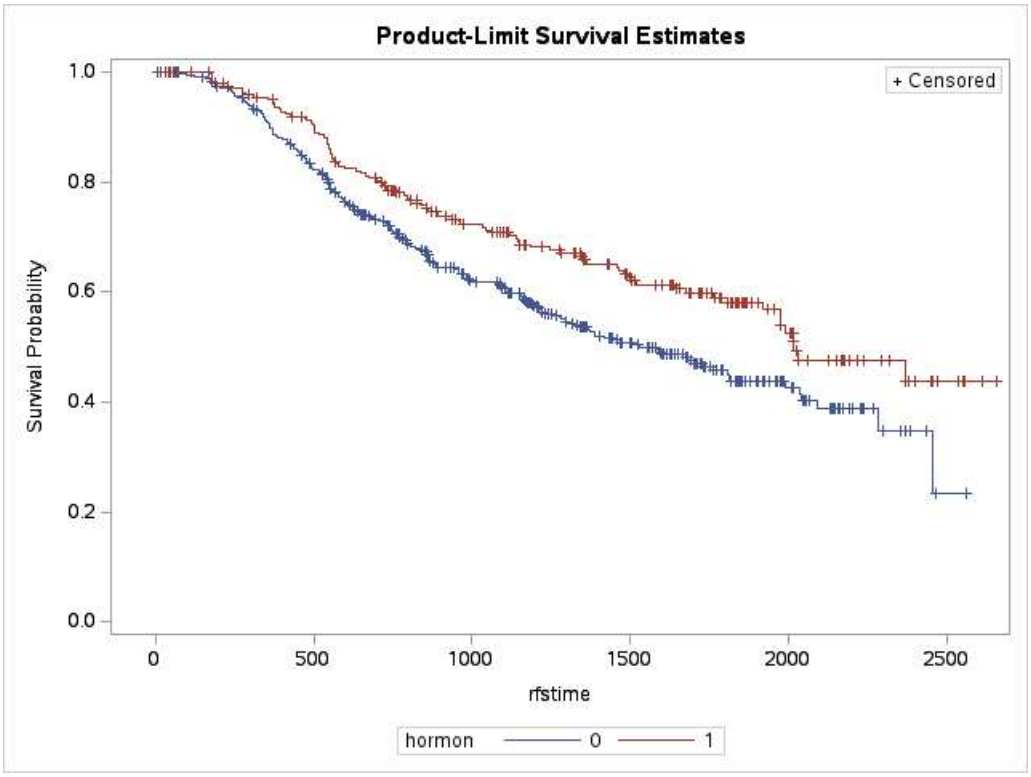
Mean	Standard Error
1265.29	62.86

Summary of the Number of Censored and Uncensored Values					
Stratum	grade	Total	Failed	Censored	Percent Censored
1	1	81	18	63	77.78
2	2	444	202	242	54.50
3	3	161	79	82	50.93
Total		686	299	387	56.41

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	21.0944	2	<.0001
Wilcoxon	27.2049	2	<.0001
-2Log(LR)	23.0912	2	<.0001

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
grade	grade	Chi-Square	Raw	Bonferroni
1	2	4.9210	0.0265	0.0796
1	3	19.9609	<.0001	<.0001
2	3	1.4387	0.2303	0.6910

Tablica 3.9: Položajne vrijednosti, deskriptivna statistika, *Log-rang* test i *post hoc* analiza za *grade* (ispis iz SAS-a)



Slika 3.9: Kaplan-Meierove procjene funkcija doživljenja po grupama za varijablu *hormon*

Za varijablu *hormon* imamo sličan zaključak. Na slici 3.9 uočavamo jasnu razliku između procjena funkcija doživljenja. Za *hormon* = 0 (pacijentice koje nisu primale hormonsku terapiju) imamo manju vjerojatnost doživljenja kroz cijelo vrijeme. Drugim riječima, postoji statistički značajna razlika u funkcijama doživljenja dviju grupa. Zaključak potvrđujemo *log-rang* testom (tablica 3.10) za koji p-vrijednost iznosi 0.0034 pa odbacujemo nultu hipotezu na svakoj razumnoj razini značajnosti.

Summary Statistics for Time Variable rfstime (hormon=0)

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	2456.00	LOGLOG	2456.00	.
50	1528.00	LOGLOG	1280.00	1814.00
25	629.00	LOGLOG	550.00	754.00

Mean	Standard Error
1512.62	46.12

Summary Statistics for Time Variable rfstime (hormon=1)

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	2018.00	LOGLOG	1918.00	.
25	859.00	LOGLOG	675.00	1146.00

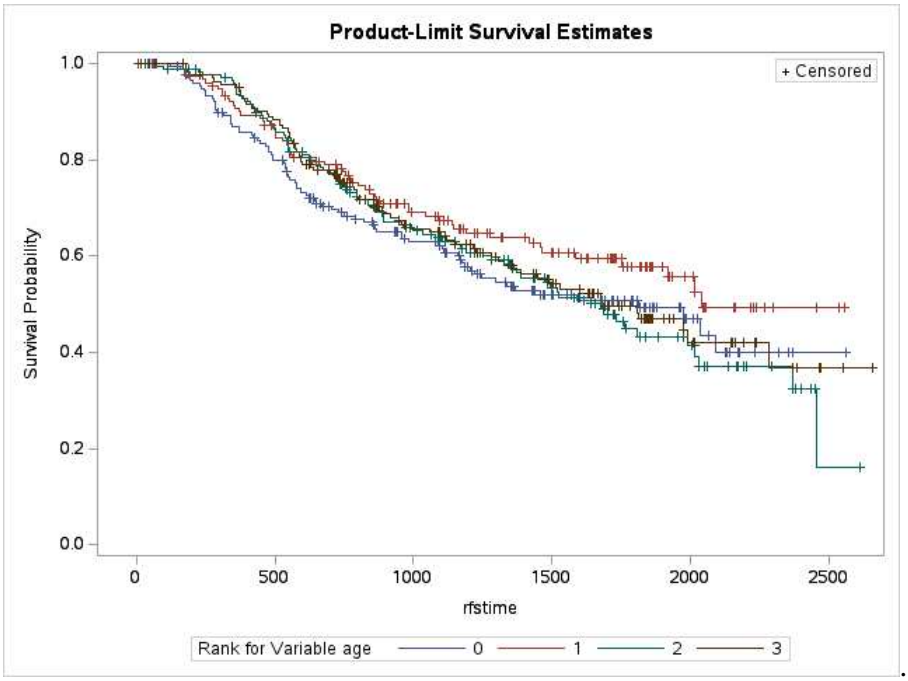
Mean	Standard Error
1690.81	55.07

Summary of the Number of Censored and Uncensored Values					
Stratum	hormon	Total	Failed	Censored	Percent Censored
1	0	440	205	235	53.41
2	1	246	94	152	61.79
Total		686	299	387	56.41

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	8.5648	1	0.0034
Wilcoxon	8.3614	1	0.0038
-2Log(LR)	8.4808	1	0.0036

Tablica 3.10: Položajne vrijednosti, deskriptivna statistika i *Log-rang* test za *hormon* (ispis iz SAS-a)

Da bi mogli proučavati kako varijable *age, size* i *nodes* utječu na vjerojatnost doživljenja, podijelit ćemo vrijednosti varijable u četiri kategorije. U nultoj kategoriji bit će vrijednosti od najmanje do donjeg kvartila, u prvoj od donjeg kvartila do medijana, u drugoj od medijana do gornjeg kvartila te u trećoj od gornjeg kvartila do najveće vrijednosti.
Na slici 3.10 ne vidimo nikakvu jasnu razliku u procijenjenim funkcijama doživljenja po grupama za varijablu *age*. Zaključujemo da ne možemo tvrditi da postoji statistički značajna razlika u funkcijama doživljenja tih grupa. Zaključak potvrđuje *log-rang* test iz tablice 3.12 u kojem ne odbacujemo nultu hipotezu koja kaže da nema razlike u funkcijama doživljenja.



Slika 3.10: Kaplan-Meierove procjene funkcija doživljenja po grupama za varijablu *age*. Grupe: 0 (21-46 godina), 1 (47-53 godina), 2 (54-61 godina), 3 (62-80 godina).

Summary Statistics for Time Variable rfstime (age=0)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	1814.00	LOGLOG	1183.00	.
25	578.00	LOGLOG	476.00	748.00

Mean	Standard Error
1390.18	59.54

Summary Statistics for Time Variable rfstime (age=1)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	2039.00	LOGLOG	1753.00	.
25	801.00	LOGLOG	554.00	1090.00

Mean	Standard Error
1496.69	59.13

Summary Statistics for Time Variable rfstime (age=3)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	2456.00	LOGLOG	2372.00	.
50	1684.00	LOGLOG	1352.00	2018.00
25	729.00	LOGLOG	548.00	889.00

Mean	Standard Error
1556.74	68.47

Summary Statistics for Time Variable rfstime (age=3)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2286.00	.
50	1679.00	LOGLOG	1329.00	2286.00
25	772.00	LOGLOG	577.00	918.00

Mean	Standard Error
1529.22	64.72

Summary of the Number of Censored and Uncensored Values

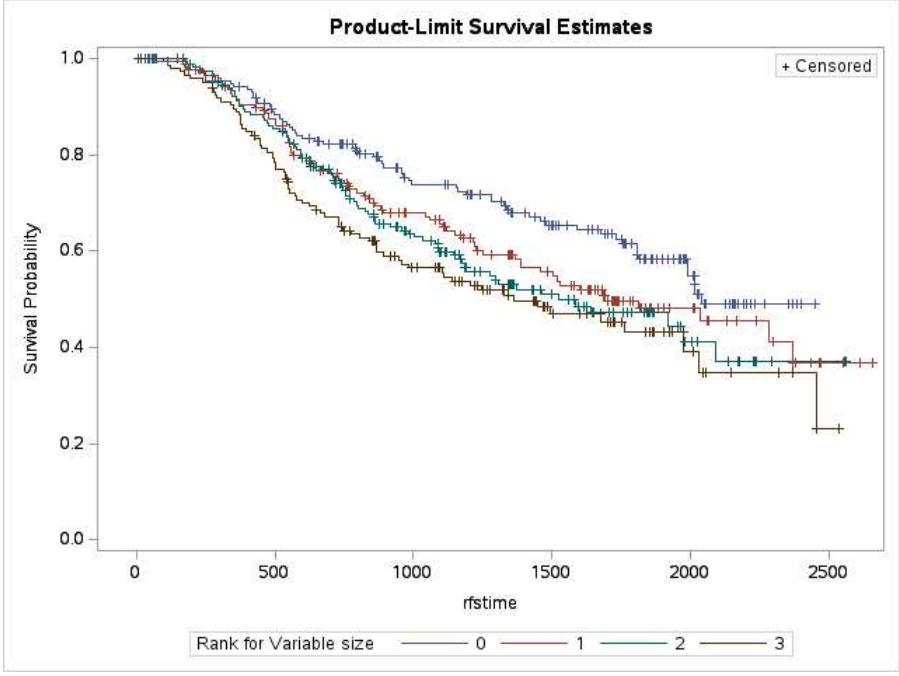
Stratum	age_rank	Total	Failed	Censored	Percent Censored
1	0	181	81	100	55.25
2	1	157	58	99	63.06
3	2	180	85	95	52.78
4	3	168	75	93	55.36
Total		686	299	387	56.41

Tablica 3.11: Položajne vrijednosti i deskriptivna statistika za *age*.
Grupe: 0 (21-46 godina), 1 (47-53 godina), 2 (54-61 godina), 3 (62-80 godina). (ispis iz SAS-a)

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	3.0676	3	0.3813
Wilcoxon	3.1404	3	0.3705
-2Log(LR)	3.0834	3	0.3789

Tablica 3.12: *Log-rang* test za *age*.
Grupe: 0 (21-46 godina), 1 (47-53 godina), 2 (54-61 godina), 3 (62-80 godina). (ispis iz SAS-a)

Za varijablu *size* primjećujemo odstupanje funkcije doživljenja za pacijentice s najmanjim veličinama tumora (od 3 mm do 20 mm) od ostalih (slika 3.11). Te pacijentice imaju veće vjerojatnosti doživljenja za sva vremena veća od otprilike 500 dana od ostalih pacijentica. Za ostale grupe ne primjećujemo jasne razlike. *Log-rang* test (tablica 3.14) potvrđuje da postoje razlike u funkcijama doživljenja. *Post hoc* analizom dobivamo da je najveća razlika u grupama s najmanjim i najvećim veličinama, a značajna razlika je još u grupama s najmanjim i veličinama od medijana do gornjeg kvartila.



Slika 3.11: Kaplan-Meierove procjene funkcija doživljenja po grupama za varijablu *size*.
Grupe: 0 (3-20 mm), 1 (21-25 mm), 2 (26-35 mm), 3 (36-120 mm).

Summary Statistics for Time Variable rfstime (size=0)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	2039.00	LOGLOG	1814.00	.
25	983.00	LOGLOG	790.00	1352.00

Mean	Standard Error
1570.57	51.54

Summary Statistics for Time Variable rfstime (size=1)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2372.00	.
50	1701.00	LOGLOG	1366.00	2372.00
25	730.00	LOGLOG	554.00	891.00

Mean	Standard Error
1574.09	68.27

Summary Statistics for Time Variable rfstime (size=2)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2093.00	.
50	1525.00	LOGLOG	1174.00	2093.00
25	714.00	LOGLOG	578.00	827.00

Mean	Standard Error
1397.82	56.45

Summary Statistics for Time Variable rfstime (size=3)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	2456.00	LOGLOG	2030.00	.
50	1363.00	LOGLOG	945.00	2030.00
25	537.00	LOGLOG	448.00	662.00

Mean	Standard Error
1436.78	81.40

Summary of the Number of Censored and Uncensored Values					
Stratum	size_rank	Total	Failed	Censored	Percent Censored
1	0	180	65	115	63.89
2	1	173	76	97	56.07
3	2	185	83	102	55.14
4	3	148	75	73	49.32
Total		686	299	387	56.41

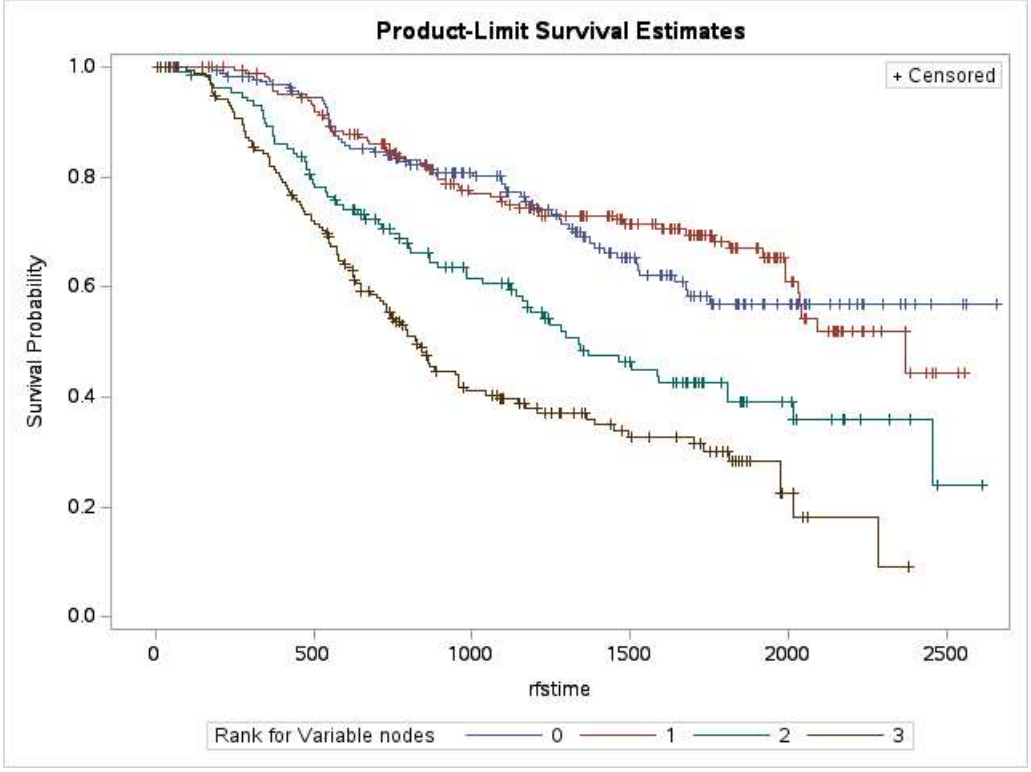
Tablica 3.13: Položajne vrijednosti i deskriptivna statistika za *size*.
Grupe: 0 (3-20 mm), 1 (21-25 mm), 2 (26-35 mm), 3 (36-120 mm). (ispis iz SAS-a)

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	11.6312	3	0.0088
Wilcoxon	13.1548	3	0.0043
-2Log(LR)	11.3880	3	0.0098

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
size_rank	size_rank	Chi-Square	Raw	Tukey-Kramer
0	1	2.9882	0.0839	0.3087
0	2	5.7050	0.0169	0.0793
0	3	11.0905	0.0009	0.0048
1	2	0.4472	0.5037	0.9089
1	3	2.3102	0.1285	0.4254
2	3	0.6566	0.4178	0.8496

Tablica 3.14: Log-rang test i *post hoc* analiza za *size*.
Grupe: 0 (3-20 mm), 1 (21-25 mm), 2 (26-35 mm), 3 (36-120 mm). (ispis iz SAS-a)

Jasnu razliku u procijenjenim funkcijama doživljenja primjećujemo i na slici 3.12. Kod pacijentica, podijeljenih u grupe po broju pozitivnih limfnih čvorova (*nodes*), u prve dvije grupe (s najmanje pozitivnih limfnih čvorova) ne vidimo nikakvu statistički značajnu razliku u vjerojatnostima doživljenja. Prelaskom u drugu grupu (od medijana do gornjeg kvartila) vidljiva je očita razlika. Vjerojatnost doživljenja se znatno smanjuje. Također, vjerojatnost doživljenja je značajno manja za pacijentice u drugoj grupi od onih u trećoj grupi. Razlike između grupa vidimo i u 50-tim percentilima varijable *rfstime* čija procjena pada prelaskom u grupe s većim brojem čvorova. Za nultu grupu nema procjene jer nema cenzuriranja s vjerojatnosti doživljenja manjom od 0.5, za prvu kategoriju iznosi 2372, za drugu 1337 te za treću 827. Zaključak potvrđuje *log-rank* test iz tablice 3.16, ali i postotak cenzuriranih pacijentica u tablici 3.15, gdje povećanje broja pozitivnih čvorova dovodi do očitog pada postotaka cenzuriranih. *Post hoc* analizom zaključujemo da između svih grupa osim dviju s najmanje čvorova postoji statistički značajna razlika.



Slika 3.12: Kaplan-Meierove procjene funkcija doživljenja po grupama za varijablu *nodes*. Grupe: 0 (1 čvor), 1 (2-3 čvorova), 2 (4-7 čvorova), 3 (8-51 čvorova).

Summary Statistics for Time Variable rfstime (nodes=0)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	.	LOGLOG	1684.00	.
25	1192.00	LOGLOG	855.00	1387.00

Mean	Standard Error
1432.45	37.58

Summary Statistics for Time Variable rfstime (nodes=1)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	2372.00	LOGLOG	2030.00	.
25	1105.00	LOGLOG	861.00	1814.00

Mean	Standard Error
1823.98	58.59

Summary Statistics for Time Variable rfstime (nodes=2)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	2456.00	LOGLOG	2015.00	.
50	1337.00	LOGLOG	1140.00	1807.00
25	573.00	LOGLOG	475.00	799.00

Mean	Standard Error
1455.70	82.77

Summary Statistics for Time Variable rfstime (nodes=3)

Quartile Estimates				
	Point	95% Confidence Interval		
Percent	Estimate	Transform	[Lower	Upper)
75	1977.00	LOGLOG	1493.00	.
50	827.00	LOGLOG	712.00	960.00
25	455.00	LOGLOG	377.00	547.00

Mean	Standard Error
1113.70	63.39

Summary of the Number of Censored and Uncensored Values					
Stratum	nodes_rank	Total	Failed	Censored	Percent Censored
1	0	187	59	128	68.45
2	1	189	60	129	68.25
3	2	131	68	63	48.09
4	3	179	112	67	37.43
Total		686	299	387	56.41

Tablica 3.15: Položajne vrijednosti i deskriptivna statistika za *nodes*.
Grupe: 0 (1 čvor), 1 (2-3 čvorova), 2 (4-7 čvorova), 3 (8-51 čvorova). (ispis iz SAS-a)

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	82.7291	3	<.0001
Wilcoxon	82.1307	3	<.0001
-2Log(LR)	69.1535	3	<.0001

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
nodes_rank	nodes_rank	Chi-Square	Raw	Tukey-Kramer
0	1	0.2696	0.6036	0.9545
0	2	14.0229	0.0002	0.0010
0	3	51.6624	<.0001	<.0001
1	2	18.1582	<.0001	0.0001
1	3	58.4804	<.0001	<.0001
2	3	15.4556	<.0001	0.0005

Tablica 3.16: *Log-rang* test i *post hoc* analiza za *nodes*.
Grupe: 0 (1 čvor), 1 (2-3 čvorova), 2 (4-7 čvorova), 3 (8-51 čvorova). (ispis iz SAS-a)

3.4 Coxov regresijski model

Želimo ispitati koje varijable najviše utječu na funkciju doživljenja te kakav je njihov utjecaj na relativan rizik ponovnog pojavljivanja raka dojke ili smrt. Za to koristimo Coxov regresijski model. Testiramo pretpostavku proporcionalnosti hazarda za svaku varijablu. Rezultate testova vidimo u tablici 3.17, gdje *PH-pvalue* označava p-vrijednost testa za testiranje proporcionalnosti hazarda. Na razini značajnosti od 5% za varijable *age, grade, meno, pgr* i *er* odbacujemo nultu hipotezu koja kaže da vrijedi pretpostavka proporcionalnosti hazarda, dok za ostale varijable ne odbacujemo nultu hipotezu.
Iako za pojedine varijable pretpostavka proporcionalnosti nije ispunjena, uključit ćemo ih u daljnje modele.

p-vrijednost testa za testiranje proporcionalnosti hazarda								
	age	*size*	*grade*	*nodes*	*meno*	*pgr*	*er*	*hormon*
PH-pvalue	0.0100	0.5330	<0.0001	0.3430	0.0370	0.0270	0.0160	0.6060

Tablica 3.17: Testiranje pretpostavke proporcionalnosti hazarda

U punom modelu (promatranom u tablici 3.18) vidimo da je model statistički značajan te da su značajne varijable *grade, nodes, pgr* i *hormon*. Značajne varijable su one kod kojih odbacujemo nultu hipotezu koja kaže da je parametar jednak 0. Također, možemo iščitati utjecaj svake od varijabla na relativan rizik događaja (smrt ili recidiv). Povećanjem broja limfnih čvorova za 1, relativan rizik za ponovno pojavljivanje raka ili smrt se povećava

za 5.1%. Povećanjem stadija tumora za 1, relativan rizik događaja povećava se za 32.3%. S druge strane, pacijentice koje primaju hormonsku terapiju imaju 28.6% manji relativan rizik događaja od pacijentica koje ne primaju hormonsku terapiju. Povećanje progesteronskih receptora za 10 fmol/l dovodi do smanjenja relativnog rizika za 2.2%.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	3576.346	3474.522
AIC	3576.346	3490.522
SBC	3576.346	3520.126

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	101.8241	8	<.0001
Score	120.0240	8	<.0001
Wald	115.2078	8	<.0001

Analysis of Maximum Likelihood Estimates									
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
age	1	-0.00939	0.00927	1.0247	0.3114	0.991	0.973	1.009	age
meno	1	0.26699	0.18334	2.1207	0.1453	1.306	0.912	1.871	meno
size	1	0.00772	0.00395	3.8191	0.0507	1.008	1.000	1.016	size
grade	1	0.28013	0.10606	6.9768	0.0083	1.323	1.075	1.629	grade
nodes	1	0.04989	0.00741	45.3226	<.0001	1.051	1.036	1.067	nodes
pgr	1	-0.00224	0.0005757	15.1117	0.0001	0.998	0.997	0.999	pgr
er	1	0.0001680	0.0004477	0.1407	0.7075	1.000	0.999	1.001	er
hormon	1	-0.33718	0.12896	6.8360	0.0089	0.714	0.554	0.919	hormon

Tablica 3.18: Puni Coxov regresijski model (sa svim varijablama) (ispis iz SAS-a)

Osim punog modela, možemo promatrati univarijatne (jednostruke) Coxove regresijske modele za svaku varijablu. Iz tablice 3.19 uočavamo da, na razini značajnosti od 5%, koeficijenti u modelima uz varijable *age* i *meno* nisu statistički značajno različiti od 0. Za ostale varijable, na razini značajnosti od 5%, zaključujemo da su koeficijenti uz njih različiti od nule. U modelu u kojem je *size* jedina varijabla koeficijent iznosi 0.01484 pa se uz povećanje veličine tumora za 1 mm relativan rizik događaja povećava za 1.5%. Budući

da 95% pouzdani interval za *hazard ratio* ne sadrži 1 možemo reći da je ta razlika statistički značajna. Kod univarijatnog Coxovog modela za varijablu *grade* utjecaj je još veći nego u punom modelu. Povećanje stadija tumora za 1 dovodi do povećanja relativnog rizika događaja za 56.7%. Povećanje broja pozitivnih limfnih čvorova za 1 dovodi do povećanja relativnog rizika za 6%. S druge strane, povećanje progesteronskih ili estrogen receptora za 10 fmol/l dovodi do smanjenja relativnog rizika događaja za redom 2.8% i 0.9%. Pacijentice koje se liječe hormonskom terapijom imaju relativan rizik događaja za 30.5% manji od pacijentica na nehormonskoj terapiji. Pomoću tablice 3.20 možemo uspoređivati modele te odrediti koji model bolje opisuje dobivene podatke. To radimo pomoću -2log vjerodostojnost s kovarijatama (engl. $-2LOG\ L\ With\ Covariates$). Ako želimo da model bude značajan, cilj nam je dobiti čim manje vrijednosti. Pomoću p-vrijednosti omjera vjerodostojnosti (engl. *Likelihood Ratio p-value*) testiramo je li koeficijent uz varijablu u modelu jednak nuli.

Analiza procjene metodom maksimalne vjerodostojnosti

Parametar	DF	Procjena parametra	Standardna greška	p vrijednost	Hazard Ratio	95% HR ci
age	1	-0.00449	0.00589	0.4460	0.996	[0.984,1.007]
size	1	0.01484	0.00351	<0.0001	1.015	[1.008,1.022]
grade	1	0.44884	0.10057	<0.0001	1.567	[1.286,1.908]
nodes	1	0.05860	0.00674	<0.0001	1.060	[1.046,1.074]
meno	1	0.06221	0.11824	0.5988	1.064	[0.844,1.342]
pgr	1	-0.00277	0.0005759	<0.0001	0.997	[0.996,0.998]
er	1	-0.000946	0.0004632	0.0411	0.999	[0.998,1.000]
hormon	1	-0.36385	0.12504	0.0036	0.695	[0.544,0.888]

Tablica 3.19: Univarijatni Coxovi modeli za sve kovarijate

	−2LOG L With Covariates	Likelihood Ratio p-value
age	3575.767	0.4464
size	3560.665	< 0.0001
grade	3556.374	< 0.0001
nodes	3526.331	< 0.0001
meno	3576.065	0.5960
pgr	3542.307	< 0.0001
er	3571.648	0.0302
hormon	3567.530	0.0030

Tablica 3.20: Testiranje značajnosti univarijatnih Coxovih modela

Jedan od načina za odabir varijabla u multivarijatnom (višestrukom) modelu je pomoću *stepwise* procedure. Krećemo od modela bez varijabli. U svakom koraku se za svaku varijablu razmatra dodavanje ili uklanjanje iz modela. Na kraju koraka odabere se varijabla čijim dodavanjem ili uklanjanjem najviše poboljšavamo model. Procedura završava kada se model više ne može poboljšati dodavanjem ili uklanjanjem varijabli. Model koji dobivamo nalazi se u tablici 3.21.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	3576.346	3480.162
AIC	3576.346	3488.162
SBC	3576.346	3502.964

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
hormon	1	-0.31036	0.12559	6.1069	0.0135	0.733	0.573	0.938	hormon
grade	1	0.29339	0.10550	7.7342	0.0054	1.341	1.090	1.649	grade
nodes	1	0.05524	0.00677	66.6106	<.0001	1.057	1.043	1.071	nodes
pgr	1	-0.00224	0.0005600	15.9320	<.0001	0.998	0.997	0.999	pgr

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Effect Label
1	nodes		1	1	78.4115		<.0001	nodes
2	pgr		1	2	19.6329		<.0001	pgr
3	grade		1	3	9.0320		0.0027	grade
4	hormon		1	4	6.1538		0.0131	hormon

Tablica 3.21: Coxov regresijski model - *stepwise* procedura

Funkcija hazarda za ovaj model glasi

$$h(t|\mathbb{X}) = h_0(t)e^{0.05524\times nodes - 0.00224\times pgr + 0.29339\times grade - 0.31036\times hormon},$$

gdje je h_0 bazna funkcija hazarda i $\mathbb{X}$ je vektor kovarijata *nodes, pgr, grade, hormon*. Odmah možemo iščitati da se povećanjem broja čvorova (*nodes*) i stadija tumora (*grade*) za 1 statistički značajno povećava relativan rizik ponovnog pojavljivanja bolesti ili smrti (za broj čvorova za 5.7%, a za stadij tumora za 34.1%). S druge strane, povećanjem progesteronskih receptora (*pgr*) za 10 fmol/l smanjuje se relativan rizik događaja (za 2.2%). Rizik je manji i za pacijentice koje primaju hormonsku terapiju u odnosu na one koje ju ne primaju (za 26.7%).

Pogledajmo jedan model s interakcijom. U model dobiven *stepwise* metodom dodajmo interakciju varijabla *hormon* i *nodes*. Sada bolje možemo opisati kako se ponaša omjer hazarda varijable *nodes* s obzirom na terapiju (hormonska ili nehormonska).

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	3576.346	3475.566
AIC	3576.346	3485.566
SBC	3576.346	3504.069

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	100.7798	5	<.0001
Score	116.9585	5	<.0001
Wald	111.9212	5	<.0001

Analysis of Maximum Likelihood Estimates									
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		Label
pgr	1	-0.00220	0.0005559	15.6367	<.0001	0.998	0.997	0.999	pgr
grade	1	0.28519	0.10556	7.2988	0.0069	1.330	1.081	1.636	grade
nodes	1	0.04744	0.00829	32.7409	<.0001	1.049	1.032	1.066	nodes
hormon	1	-0.53149	0.16416	10.4819	0.0012	0.588	0.426	0.811	hormon
hormonnodes	1	0.03342	0.01501	4.9584	0.0260	1.034	1.004	1.065	

Tablica 3.22: Coxov regresijski model s interakcijom varijabla *hormon* i *nodes*

U tablici 3.22 vidimo da je koeficijent uz interakciju statistički značajan. Iz tablice čitamo da je ukupni efekt broja čvorova $0.04744 + \textit{hormon} \times 0.03342$. Ako pacijentice primaju hormonsku terapiju te imaju veći broj pozitivnih limfnih čvorova, onda je relativan rizik događaja veći. Hormonska terapija kod pacijentica s većim brojem pozitivnih limfnih čvorova pokazuje se manje uspješnom nego u slučaju manjeg broja čvorova. Također, kada bi uspoređivali modele po kriteriju -2log vjerodostojnosti (engl. *-2 log likelihood*) model s interakcijom ispada bolji od modela bez interakcije.

3.5 Zaključak

Kod pacijenata koji boluju od raka dojke, nakon operacije bitno je pratiti i pokušati spriječiti ponovno pojavljivanje raka. Zbog toga se, da bi se bolje razumio i efektivnije mogao sprječavati, u medicinskim istraživanjima prati događaj ponovnog pojavljivanja raka i vrijeme

koje prođe do događaja te još neki faktori koji mogu utjecati na ponovno pojavljivanje. Takve podatke pogodno je analizirati metodama analize doživljenja.
Pomoću Kaplan-Meierove procjene vidjeli smo važnost hormonske terapije, stadija tumora, veličine tumora i broja pozitivnih limfnih čvorova. Sve te varijable imale su statistički značajan utjecaj na vjerojatnost doživljenja. Npr., pacijentice koje su primale hormonsku terapiju imale su značajno veću vjerojatnost doživljenja od pacijentica koje nisu primale hormonsku terapiju. S druge strane, menopauzalni status i starost nisu imali značajan utjecaj na vjerojatnost doživljenja. Coxovom regresijom ispitali smo koje vrste liječenja te koje karakteristike bolesti imaju najveći utjecaj na vrijeme doživljenja. Zaključili smo da su to broj pozitivnih limfnih čvorova, stadij tumora, progesteronski receptori te vrsta terapije (hormonska ili nehormonska). Također, uočili smo da se hormonska terapija u slučajevima s većim brojem pozitivnih limfnih čvorova pokazuje manje uspješnom nego u slučaju manjeg broja čvorova.

Bibliografija

[1] Paul D Allison, *Survival analysis using SAS: a practical guide*, Sas Institute, 2010.

[2] John P Klein, Melvin L Moeschberger et al., *Survival analysis: techniques for censored and truncated data*, sv. 1230, Springer, 2003.

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

[4] Patrick Royston i Douglas G Altman, *External validation of a Cox prognostic model: principles and methods*, BMC medical research methodology **13** (2013).

Sažetak

U ovom diplomskom radu proučavali smo analizu doživljenja, skup statističkih metoda za analizu podataka kod kojih je varijabla od interesa vrijeme do nekog promatranog događaja. Osim toga, Coxovim regresijskim modelom objasnili smo kako razne kovarijate utječu na vrijeme do događaja. Tim metodama analizirali smo podatke o pacijentima za koje je promatrani događaj bio ponovno pojavljivanje raka dojke ili smrt bilo kojeg uzroka. Ispitali smo imaju li pojedine karakteristike pacijenata i vrsta terapije utjecaj na vrijeme do događaja. Također, opisali smo kako se promjena pojedinih karakteristika odražava na rizičnost događaja.

Summary

In this master's thesis we studied survival analysis, collection of statistical methods for data analysis in which variable of interest is time to the observed event. Apart from that, using Cox regression model we explained how different covariates affect time to the event. With those methods, we analyzed data about patients for which observed event was the earlier of breast cancer recurrence or death from any cause. We examined do some patients' characteristics and type of therapy have affect on time to the event. Also, we described how does change of some characteristics reflect on riskiness of the event.

Životopis

Rođen sam 30. rujna 1998. godine u Zaboku. Nakon osnovne škole, upisujem Gimnaziju Antuna Gustava Matoša u Zaboku koju završavam 2017. Te godine upisujem preddiplomski studij Matematike na matematičkom odjelu PMF-a Sveučilišta u Zagrebu. Preddiplomski studij završavam 2021. godine te iste upisujem diplomski studij Matematička statistika na istom fakultetu.