Coxov regresijski model

Paradžik, Filip

Master's thesis / Diplomski rad

2017

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Mathematics / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:126:052172

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**



Repository / Repozitorij:

Repository of School of Applied Mathematics and Computer Science







Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij financijske matematike i statistike

Filip Paradžik

Coxov regresijski model

Diplomski rad

Osijek, 2017.

Sveučilište J. J. Strossmayera u Osijeku
Odjel za matematiku
Sveučilišni diplomski studij financijske matematike i statistike

Filip Paradžik

Coxov regresijski model

Diplomski rad

Mentor: prof.dr.sc. Mirta Benšić

Osijek, 2017.

Sadržaj

1 Uvod **1**

2 Osnovni pojmovi i analiza preživljenja **2**
2.1 Posebne značajke podataka 2
2.2 Važne funkcije u analizi preživljenja 2
2.3 Procjena funkcije preživljenja 4
2.4 Testiranje hipoteza o jednakosti funkcija preživljenja 5
2.5 Log-rang test 6
2.6 Wilcoxonov test 8

3 Coxov regresijski model **10**
3.1 Definicija Coxova regresijskog modela 10
3.2 Procjena parametara 10
3.3 Opravdanje funkcije vjerodostojnosti 13
3.4 Intervali pouzdanosti i testiranje hipoteza 14
3.5 Standardne greške i intervali pouzdanosti za omjere rizika 14
3.6 Poveznica s log-rang testom 15

4 Odabir varijabli **18**
4.1 Statistika logaritma vjerodostojnosti 18
4.2 Usporedba s ugnježđenim modelom 18
4.3 Odabir varijabli 19

5 Provjera pretpostavki modela proporcionalnog rizika **21**
5.1 Cox-Snell reziduali 21
5.2 Martingalni reziduali 22
5.3 Reziduali temeljeni na devijanci 23
5.4 Metode provjere pretpostavke proporcionalnog rizika 24
5.5 Schoenfeld reziduali 25
5.6 Skor reziduali 26

6 Istraživanje **28**
6.1 Opis podataka 28
6.2 Cilj istraživanja i statističke metode 28
6.3 Selekcija varijabli 28
6.4 Opis varijabli 29
6.5 Dijagnostika modela 32
6.6 Stršeća i utjecajna mjerenja 35
6.7 Interpretacija 39

Literatura **40**

Sažetak **41**

Životopis **42**

1 Uvod

Analiza preživljenja podrazumijeva statističku analizu vremena koje prođe prije nego pojedina opservacija dođe u stanje definiranog događaja. Taj događaj se često odnosi na smrt, kvar uređaja ili slično. Analiza preživljenja ima primjenu u mnogim granama znanosti kao što su biomedicina, inženjerstvo i društvene znanosti. U inženjerstvu ona svoju primjenu pronalazi u analizi trajnosti pojedinih strojeva ili njihovih komponenti. Takva je analiza smislena budući da proizvođači strojeva obično na svoje proizvode daju određena jamstva te je stoga poželjno znati očekivani vijek trajanja takvih proizvoda. U području društvenih znanosti moguće njome možemo proučavati duljinu trajanja brakova, od njihova sklapanja pa sve do njihova poništenja, razvoda ili smrti supružnika. Što se tiče primjene u biomedicini, tu analiza preživljenja pokazuje svoju možda i najveću upotrebu. Kao primjer njena korištenja navodimo studiju o vremenu doživljenja pacijenata s dijagnosticiranim rakom debelog crijeva.
Rad je podijeljen na dva dijela, teorijski i praktični. U teorijskom dijelu se u početku bavimo osnovnim pojmovima vezanima uz analizu preživljenja, problematikom cenzuriranja, procjenom te usporedbom funkcija preživljenja. Također, obrađujemo teorijsku pozadinu Coxova regresijskog modela proporcionalnog rizika te procedure koje koristimo pri zaključivanju o valjanosti modela. Praktični dio diplomskog rada sadržava primjenu Coxova regresijskog modela za opisivanje funkcije hazarda pacijenata kojima je dijagnosticiran i operiran rak debelog crijeva.

2 Osnovni pojmovi i analiza preživljenja

2.1 Posebne značajke podataka

Odmah na početku postavlja se pitanje zašto modeliranju preživljenja ne pristupiti klasičnim regresijskim metodama. Na prvi pogled njihovo korištenje izgleda itekako smisleno, no sama priroda podataka o preživljenju je takva da se njihova primjena pokazuje ili neprikladna ili inferiorna drugim procedurama. Kratko ćemo objasniti ključne karakteristike podataka o preživljenju koji dolaze iz, nama za ovaj rad interesantnih, medicinskih istraživanja.
Jedna takva karakteristika podataka jest cenzuriranje i usko je promatrana uz samo vrijeme doživljenja pacijenta. Premda postoji više vrsta cenzuriranja, nama će od značaja biti samo desno cenzuriranje. Ono je najčešće pojavljivana vrsta cenzuriranja u medicinskim istraživanjima.
Vrijeme doživljenja pacijenta je desno cenzurirano ukoliko događaj od značaja (primjerice smrt) nije zabilježen do završetka studije ili ukoliko je iz pacijent izgubljen iz nje prije no što se realizirao događaj od značaja. U nastavku rada, pod cenzuriranjem smatramo isključivo desno cenzuriranje.
Još jedna karakteristika podataka o preživljenju jest to da distribucija vremena doživljenja generalno nije simetrična. Ukoliko bismo konstruirali njen histogram, uočili bismo pozitivnu zakrivljenost distribucije, odnosno dugi desni 'rep'. Zbog toga nije razumno takvim podacima pretpostavljati normalnu distribuiranost.
Kako je slučajna varijabla koja modelira vrijeme doživljenja nenegativna, problem s dugim desnim repom distribucije često može biti popravljen koristeći neku transformaciju vremena doživljenja. Jedna od popularnih transformacija jest ona logaritamska. Problem se javlja kada je prisutno cenzuriranje koje bitno narušava pretpostavke klasičnih regresijskih modela. Kao odgovor na problem, razvile su se metode iz analize preživljenja.

2.2 Važne funkcije u analizi preživljenja

Započet ćemo uzevši u obzir najjednostavniju situaciju, tj. onu bez cenzuriranja. Neka je T neprekidna, nenegativna slučajna varijabla koja opisuje vrijeme proteklo do nekog specifičnog događaja. Od sada pa nadalje ćemo specifični događaj nazivati "smrt". Kako trenutno opisujemo situaciju bez cenzuriranja, za svakog ispitanika poznato je vrijeme do smrti. Označimo s f funkciju gustoće neprekidne slučajne varijable T. Tada je funkcija distribucije slučajne varijable T, u oznaci F, definirana kao

$$F(t) = P(T \leq t) = \int_0^t f(x)dx.$$

U sadašnjim uvjetima, $F(t)$ nam predstavlja vjerojatnost smrti pojedinog ispitanika zaključno s vremenom t. Više informacija o funkciji distribucije kao i drugim osnovnim pojmovima iz teorije vjerojatnosti mogu se pronaći u [2]. Istu će informaciju, no s drugog gledišta, ponuditi funkcija preživljenja. Funkciju preživljenja definiramo kao

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(x)dx. \tag{2.1}$$

Iz 2.1 vidimo kako je ona definirana preko F. Koristeći generalna svojstva funkcije distribucije neprekidne slučajne varijable T, slijedi da je funkcija preživljenja nenegativna, monotono padajuća i neprekidna funkcija za koju vrijedi $S(0) = 1$ i $\lim_{t\to\infty} S(t) = 0$. Ona predstavlja

vjerojatnost preživljenja vremena t pojedinog ispitanika.

Sljedeća važna funkcija je funkcija hazarda ili rizika, odnosno intenzitet smrtnosti. Funkcija hazarda definira se kao vjerojatnost smrti ispitanika u trenutku t, uvjetno na to da je preživio do tog trenutka. Moglo bi se reći kako ona predstavlja trenutnu stopu smrtnosti u trenutku t. Označavamo ju s $h(t)$ i formalnije definiramo kao

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.2}$$

Iz teorije vjerojatnosti znamo da za događaje A, B vrijedi formula uvjetne vjerojatnosti

$$P(A|B) = \frac{P(A,B)}{P(B)}. \tag{2.3}$$

Primjenom formule 2.3 za neprekidnu slučajnu varijablu T vrijedi

$$\begin{aligned} P(t \leq T < t + \Delta t | T \geq t) &= \frac{P(t \leq T < t + \Delta t, T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} \\ &= \frac{F(t + \Delta t) - F(t)}{S(t)}. \end{aligned} \tag{2.4}$$

Tada je

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

Kako je

$$\lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t},$$

definicija derivacije od $F(t)$ s obzirom na t, što je zapravo $f(t)$, sada imamo

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.5}$$

Slijedi

$$h(t) = -\frac{d}{dt}\big(lnS(t)\big), \tag{2.6}$$

nakon čega integracijom i sređivanjem dolazimo do

$$S(t) = exp\big(-H(t)\big), \tag{2.7}$$

gdje je

$$H(t) = \int_0^t h(u)du.$$

Funkcija $H(t)$ također se koristi u analizi preživljenja te ju zovemo funkcija kumulativnog hazarda. Funkcija kumulativnog hazadra se prema formuli 2.7 lako dobiva iz funkcije preživljenja kao

$$H(t) = -lnS(t). \tag{2.8}$$

2.3 Procjena funkcije preživljenja

Prilikom procjene funkcije preživljenja postoje razlike u tome radi li se o podatcima u kojima je prisutno cenzuriranje ili ne. Započnimo s najjednostavnijem situacijom, tj. onoj kada nema cenzuriranja. Kao procjenitelja funkcije preživljenja koristimo empirijsku funkciju preživljenja koja se u slučaju necenzuriranog uzorka definira kao

$$\hat{S}(t) = \frac{Broj\ ispitanika\ s\ vremenom\ doživljenja \geq t}{Ukupan\ broj\ ispitanika}. \tag{2.9}$$

Iz same formule 2.9 vidimo kako je ona padajuća stepenasta funkcija koja poprima vrijednosti iz $[0, 1]$. U trenutku $t = 0$ njena je vrijednost jednaka 1 te bilježi padove u onim trenutcima u kojima dođe do događaja, odnosno u našem slučaju smrti ispitanika. Ako pretpostavimo da je ukupan broj pacijenata u studiji jednak n, tada ukoliko u trenutku t dođe do smrti točno jedne osobe, vrijednost funkcije $\hat{S}$ će se smanjiti za $1/n$. Ukoliko bi u trenutku t umrlo d osoba, tada bi pad vrijednosti funkcije $\hat{S}$ bio d/n.

Kada radimo sa cenzuriranim podatcima formula 2.9 više nije prikladna jer nakon što je određeni ispitanik cenzuriran u trenutku s, tada za svaki trenutak $t > s$ ne znamo je li on još uvijek živ ili ne. Stoga će pri definiranju empirijske funkcije preživljenja biti potrebno ukomponirati informacije o cenzuriranju za svakog ispitanika. Ovdje predložena modifikacija naziva se Kaplan-Meier procjenitelj, prema njihovim autorima (Kaplan i Meier, 1958).

Važna pretpostavka koju moramo uvesti prije definicije Kaplan-Meier procjenitelja jest međusobna nezavisnost smrti u uzorku. Pretpostavimo da $(t_i, \delta_i), i = 1, \ldots, n$, predstavlja cenzurirani uzorak vremena doživljenja, gdje su t_i zabilježena vremena doživljenja, a δ_i je indikator smrtnosti definiran kao $\delta_i = I(t_i\ je\ vrijeme\ smrti)$. Neka su $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, $r \leq n$, različita vremena u kojima je zabilježena smrt. Dopuštena je mogućnost pojavljivanja više smrti u istom trenutku. S $d_j = \sum_{i=1}^{n} I(t_i = t_{(j)}, \delta_i = 1)$ označen je broj smrti u trenutku $t_{(j)}$. Kako imamo cenzurirani uzorak, osim vremena $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, u kojima su se realizirale smrti, pojavljuju se i vremena u kojima su pojedina mjerenja cenzurirana. Sada, Kaplan-Meier procjenitelj funkcije preživljenja definiramo kao

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j}, \tag{2.10}$$

gdje je $n_j = \sum_{i=1}^{n} I(t_i \geq t_{(j)})$ broj pacijenata koji su do trenutka $t_{(j)}$ živi i necenzurirani, tj. kažemo da su oni u trenutku $t_{(j)}$ pod rizikom. Ukoliko je za ispitanika vrijeme cenzuriranja ili vrijeme smrti jednako upravo $t_{(j)}$, njega također uključujemo u onih n_j ispitanika koji su pod rizikom u trenutku $t_{(j)}$. Iz jednadžbe 2.10 uočavamo kako je $\hat{S}(t) = 1$ za $t < t_{(1)}$, a ukoliko je najveća zabilježena vrijednost t^* necenzurirana, tada je $\hat{S}(t) = 0$ za sve $t \geq t^*$. Ukoliko je pak najveća vrijednost t^* cenzurirano vrijeme, tada kažemo da je $\hat{S}(t)$ nedefinirana za $t > t^*$. Kaplan-Meier procjenitelj funkcije preživljenja je monotono padajuća stepenasta funkcija s padovima u vremenima kada je zabilježena smrt, a između dvaju susjednih vremena smrti ima konstantne vrijednosti.

Premda ih nećemo ovdje izvoditi, od interesa je za dobiveni procjenitelj znati standardne greške te pouzdane intervale. Aproksimacija koja se koristi za standardne greške Kaplan-Meier procjenitelja funkcije preživljenja dana je s

$$s.e.\big(\hat{S}(t)\big) \approx \hat{S}(t)\sqrt{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}}, \tag{2.11}$$

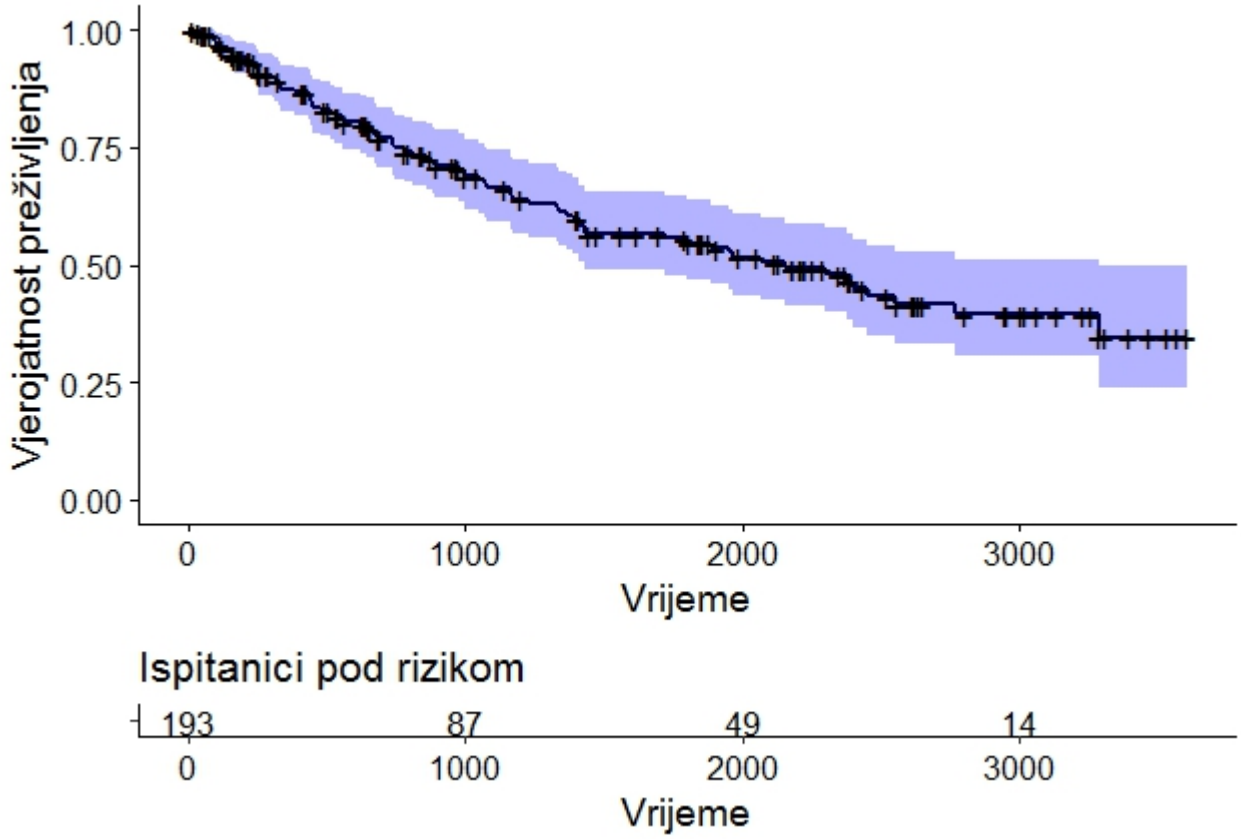
gdje je $t_{(k)} \leq t < t_{(k+1)}$. Ovaj razultat poznat je kao Greenwoodova formula.

Nakon što je izračunata standardna greška za $\hat{S}(t)$ mogu se pronaći i pouzdani intervali. Pod pretpostavkom da je $\hat{S}(t)$ normalno distribuiran s očekivanjem $S(t)$ i varijancom jednakom kvadratu standardne greške, tada je $100(1-\alpha)\%$ pouzdani interval za $S(t)$ dan s

$$\left[\hat{S}(t) - z_{\alpha/2} s.e.\big(\hat{S}(t)\big), \hat{S}(t) + z_{\alpha/2} s.e.\big(\hat{S}(t)\big)\right],$$

gdje $z_{\alpha/2}$ predstavlja $(1 - z_{\alpha/2})$ kvantil standardne normalne distribucije. Izvod Greenwoodove formule te alternativne aproksimacije standardnih grešaka i pouzdanih intervala mogu se pronaći u [4].

Primjer 2.1. *Dani su podaci o pacijentima s dijagnosticiranim i operiranim karcinomom debelog crijeva. Slika 1 predstavlja grafički prikaz Kaplan-Meier procjenitelja funkcije preživljenja s označenim 95%-tnim pouzdanim intervalima. Vertikalne linije koje sijeku krivulju preživljenja predstavljaju zabilježena cenzurirana vremena.*



Slika 1: Kaplan-Meier procjenitelj funkcije preživljenja

2.4 Testiranje hipoteza o jednakosti funkcija preživljenja

U mnogim je situacijama, posebice u medicinskim istraživanjima potrebno usporediti dvije ili više funkcija preživljenja. Jedna takva situacija bi se dogodila ukoliko bismo za određenu bolest htjeli provjeriti rezultate nove terapije u odnosu na staru ili usporediti vremena doživljenja osoba koji primaju određenu terapiju u odnosu na one koji su na placebu. Tada više

nemamo situaciju kao na slici 1, nego imamo onoliko različitih krivulja preživljenja koliko je grupa pacijenata. Od interesa je usporediti funkcije preživljenja i donijeti neke zaključke.

Pretpostavimo situaciju kad imamo dvije grupe pacijenata. Iz grafičkog prikaza Kaplan-Meier procjenitelja dvaju funkcija preživljenja možemo dobiti osjećaj o mogućem postojanju razlika između promatranih, no na temelju njega ne možemo donositi mjerljive zaključke. Do njih dolazimo koristeći se valjanim statističkim procedurama. U svrhu proučavanja jednakosti, odnosno nejednakosti funkcija preživljenja, potrebno je formirati statistički test. Za uspoređivanje podataka o preživljenju dvaju grupa postoje mnoge metode kojim možemo kvantificirati razliku između grupa. Ovdje predstavljamo dvije popularne neparametarske procedure, log-rang test i Wilcoxonov test.

2.5 Log-rang test

Najraširenija metoda korištena u svrhu uspoređivanja funkcija preživljenja je upravo log-rang test, ponekad zvan i Mantel-Coxov test ili Peto-Mantel-Haenszelov test. Premda ćemo mi njegovu primjenu ilustrirati samo na usporedbi dvaju funkcija preživljenja, on se jednostavno generalizira i za upotrebu nad više njih.

Kao i dosad, ako se pojavi situacija da u istom trenutku imamo "smrt" i "cenzuriranje", cenzuriranom vremenu dajemo veći rang. Pretpostavimo da unutar dvije grupe imamo ukupno r različitih vremena smrti $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. S d_{1j} označavamo broj ispitanika u Grupi 1 koji su umrli u trenutku $t_{(j)}$. Analogno, s d_{2j} označavamo broj ispitanika u Grupi 2 koji su umrli u trenutku $t_{(j)}$. Jasno je, ukoliko u grupi u istom trenutku nisu umrle dvije ili više osoba, da će d_{1j} i d_{2j} postizati vrijednosti nula ili jedan. S n_{1j} označavamo osobe iz Grupe 1 koje su pod rizikom smrti u trenutku $t_{(j)}$ te analogno s n_{2j} označimo broj osoba pod rizikom smrti iz Grupe 2 u istom trenutku. Dakle, u trenutku $t_{(j)}$ ukupno imamo $d_j = d_{1j} + d_{2j}$ smrti od ukupno $n_j = n_{1j} + n_{2j}$ osoba pod rizikom.

Grupa	Broj smrti u trenutku $t_{(j)}$	Broj osoba preživjelih nakon $t_{(j)}$	Broj osoba pod rizikom u trenutku neposredno prije $t_{(j)}$
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Ukupno	d_j	$n_j - d_j$	n_j

Tablica 1: Broj smrti u trenutku $t_{(j)}$ u pojedinoj grupi pacijenata

Za nul-hipotezu log-rang testa uzima se nepostojanje razlika u distribucijama vremena doživljenja dvaju grupa, odnosno nul-hipoteza smatra da je tempo umiranja pacijenata u obje grupe jednak. Drugim riječima, za nul-hipotezu zapravo uzimamo nezavisnost vjerojatnosti preživljenja o grupi kojoj pacijenti pripadaju. Validnost nul-hipoteze provjeravamo promatrajući razlike između stvarnog broja umrlih po grupama i broja umrlih koji bismo očekivali unutar nul-hipoteze i to u svim trenutcima kada je zabilježena smrt.

Ukoliko marginalne vrijednosti tablice 1 uzmemo kao fiksne i ako je nul-hipoteza istinita, tada očito sve vrijednosti ovise samo o vrijednosti d_{1j} pa d_{1j} možemo smatrati slučajnom varijablom. Ta slučajna varijabla modelira broj umrlih pacijenata u Grupi 1 u trenutku $t_{(j)}$, a može poprimiti vrijednosti između 0 i $min(d_j, n_{1j})$. Zapravo, d_{1j} ima *hipergeometrijsku distribuciju*. Stoga vjerojatnost realiziranja d_{1j} smrti unutar Grupe

1 u trenutku $t_{(j)}$ iznosi

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j-d_j}{n_{1j}-d_{1j}}}{\binom{n_j}{n_{1j}}}. \tag{2.12}$$

Ovako definirana slučajna varijabla s hipergeometrijskom distribucijom ima očekivanje dano formulom

$$e_{1j} = n_{1j}dj/n_j. \tag{2.13}$$

Dakle, e_{1j} predstavlja očekivani broj smrti unutar Grupe 1 u trenutku $t_{(j)}$. Uočimo da je u uvjetima nul-hipoteze vjerojatnost umiranja pacijenta u trenutku $t_{(j)}$ dana s d_j/n_j te posljedično u trenutku $t_{(j)}$ očekivani broj smrti u Grupi 1 iznosi $n_{1j}d_j/n_j$, a u Grupi 2 $n_{2j}d_j/n_j$. Varijanca slučajne varijable d_{1j} dana je s

$$Var(d_{1j}) = \frac{n_{1j}n_{2j}d_j(n_j-d_j)}{n_j^2(n_j-1)}. \tag{2.14}$$

Sada želimo pronaći skupnu mjeru devijacije izmjerenih vrijednosti d_{1j} od onih očekivanih za sve trenutke u kojima je nastupila barem jedna smrt. Najdirektniji pristup tome je sumirati razlike $d_{1j} - e_{1j}$ po svim različitim trenutcima smrti. Rezultirajuća statistika dana je s

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}). \tag{2.15}$$

Statistika U_L ima očekivanje 0 jer je $E(d_{1j}) = e_{1j}$. Štoviše, kako su vremena smrti nezavisna jedna od drugih, tada varijancu statistike U_L, u oznaci V_L, možemo jednostavno računati kao sumu varijanci od d_{1j}

$$V_L = Var(U_L) = \sum_{j=1}^{r} Var(d_{1j}). \tag{2.16}$$

Ukoliko broj smrti r nije premali, može se pokazati da je U_L dobro aproksimirana normalnom distribucijom. Kako smo već pokazali da U_L ima očekivanje 0, tada $U_L/\sqrt{V_L}$ ima standardnu normalnu distribuciju što ćemo zapisivati kao

$$\frac{U_L}{\sqrt{V_L}} \sim N(0,1).$$

Iz teorije vjerojatnosti poznato je da kvadrat standardne normalne slučajne varijable ima Hi-kvadrat distribuciju s jednim stupnjem slobode, u oznaci χ_1^2 i pišemo

$$\frac{U_L^2}{V_L} \sim \chi_1^2. \tag{2.17}$$

Dobivenu statistiku označavamo s W_L. Ona sumira razinu odstupanja izmjerenih od očekivanih vrijednosti vremena doživljenja dvaju grupa, u uvjetima nul-hipoteze o nepostojanju razlika između grupa. Što je veća vrijednost statistike W_L, to je dokaz protiv nul-hipoteze snažniji. Kako u slučaju nul-hipoteze W_L ima približno χ_1^2 distribuciju, p-vrijednosti povezane sa test statistikom dobivamo iz funkcije distribucije tako distribuirane slučajne varijable.

2.6 Wilcoxonov test

Wilcoxonov test je veoma sličan dosad predstavljenom log-rang testu te slijedi sličnu logiku. Nul-hipoteza o nepostojanju razlika u funkcijama preživljenja dvaju grupa je prisutna i ovdje. Razlika se pojavljuje u definiciji test statistike. Wilcoxonov test baziran je na statistici

$$U_W = \sum_{j=1}^{r} n_j(d_{1j} - e_{1j}),$$

pri čemu su vrijednosti n_j, d_{1j}, e_{1j} definirane kao u dijelu koji se tiče log-rang testa. Statistika U_W sumira razlike između stvarnog i očekivanog broja smrti za trenutke u kojima se smrt dogodila. Razlika između nje i statistike U_L iz log-rang testa jest u tome što U_W pridodaje različite težine različitim vremenima smrti. Kako se n_j smanjuje sa svakim novim pojavljivanjem smrti, statistika U_W će manju važnost dati onim devijacijama stvarnih od očekivanih vrijednosti koje se događaju kasnije. Ovako definirana statistika će biti manje osjetljiva na devijacije d_{1j} od e_{1j} u repnim dijelovima distribucije vremena doživljenja. Prema svojstvu linearnosti očekivanja, ona također ima očekivanje jednako nula, a jednostavnim računom pokazuje se da je varijanca statistike dana kao

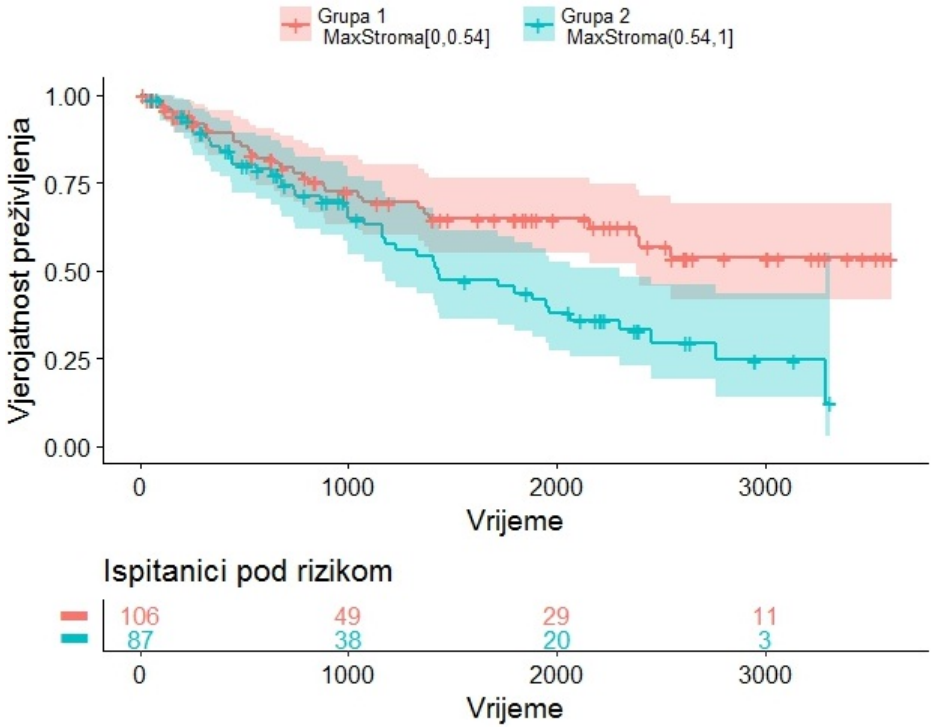
$$V_W = \sum_{j=1}^{r} n_j^2 Var(d_{1j}),$$

gdje je $Var(d_{1j})$ dana jednadžbom 2.14. Definiramo Wilcoxonovu test statistiku kao

$$W_W = \frac{U_W^2}{V_W}. \tag{2.18}$$

U slučaju istinitosti nul-hipoteze W_W ima χ_1^2 distribuciju. Zaključke o jednakostima funkcija preživljenja dvaju grupa donosimo analogno kao i u log-rang testu.

Primjer 2.2. *U bazi podataka o pacijentima s dijagnosticiranim i operiranim karcinomom debelog crijeva zanima nas postoji li veza između maksimalnog udjela tumorske strome i preživljenja pacijenata. Ispitanike dijelimo u dvije skupine prema maksimalnom udjelu tumorske strome. Ispitanici kojima je taj udio manji ili jednak 0.54 spadaju u prvu skupinu, a ostali ispitanici u drugu. Slika 2 prikazuje krivulje preživljenja dvaju skupina.*



Slika 2: Kaplan-Meier procjenitelji funkcija preživljenja dvaju grupa pacijenata. p-vrijednost log-rang testa iznosi 0.0068, a Wilcoxonovog testa 0.087.

3 Coxov regresijski model

Korištenje Kaplan-Meier procjenitelja funkcije doživljenja te log-rang i Wilcoxonovog testa može se pokazati korisnima pri analizi preživljenja jednog uzorka ili prilikom usporedbi distribucija doživljenja dvaju ili više grupa. U medicinskim istraživanjima se u praksi o svakom pojedinom sudioniku bilježe dodatne informacije o drugim veličinama, primjerice spolu, dobi, životnom stilu, prehrani, razini hormona, tretmanu kojem je on podvrgnut i mnoge druge. S ciljem pronalaska veze između vremena doživljenja pacijenta i opisnih varijabli koristi se statističko modeliranje. Ovakvim pristupom prema analizi vremena doživljenja možemo otkriti kako za pojedine grupe pacijenata vrijeme doživljenja ovisi o opisnim varijablama. Primjerice, u istraživanju o karcinomu debelog crijeva, od interesa je saznati kako je maksimalna vrijednost tumorske strome povezana s vremenom doživljenja i uz to otkriti postoji li i neka dodatna veza sa starošću pacijenta ili njegovim spolom.

Pristup modeliranju preživljenja regresijom, kojim se bavimo dalje u radu, koristi modeliranje funkcije hazarda na temelju danih regresora. Najuobičajeniji od modela s ovakvim pristupom su modeli proporcionalnog rizika. Kao što ime govori, temeljna pretpostavka na kojoj oni počivaju je upravo ona o proporcionalnom riziku. Takva pretpostavka podrazumijeva da su funkcije hazarda jednake do na množenje konstantom. U ovakvim modelima funkcija hazarda za vrijeme doživljenja t, uz dane vrijednosti regresora $\boldsymbol{x}$, dana je formulom

$$h(t|\boldsymbol{x}) = h_0(t)r(\boldsymbol{x}), \tag{3.1}$$

pri čemu $r(\boldsymbol{x})$ i $h_0(t)$ poprimaju isključivo pozitivne vrijednosti. Uobičajen naziv za funkciju $h_0(t)$ jest bazna funkcija hazarda. Ona predstavlja funkciju hazarda za onog ispitanika čije vrijednosti regresora su takve da je vrijednost funkcije $r(\boldsymbol{x}) = 1$. Izbor funkcije $r(\boldsymbol{x})$, naravno, nije jedinstven. U nastavku ćemo $h(t|\boldsymbol{x})$ kraće označavati $h(t)$.

Primijetimo, klasa modela dana jednadžbom 3.1 ima različitu formu u odnosu na klasične regresijske modele u kojima smo očekivanje ili neke funkcije očekivanja dovodili u vezu s opisnim varijablama. Unatoč tome, mnogi od principa korištenih u linearnom modeliranju prenose se i na modeliranje ovakvih podataka.

3.1 Definicija Coxova regresijskog modela

Jednadžbom 3.1 definirana je široka klasa modela koji ovise isključivo o izboru funkcije $r(\boldsymbol{x})$. Nama će od interesa biti jedan od tih modela koji se dobiva za izbor funkcije $r(\boldsymbol{x}) = exp(\boldsymbol{\beta}'\boldsymbol{x})$, gdje je $\boldsymbol{x}$ vektor danih vrijednosti regresora, a $\boldsymbol{\beta}$ vektor nepoznatih parametara. Uvrstivši tako definiranu funkciju $r(\boldsymbol{x})$ u jednadžbu 3.1 dobivamo model

$$h(t) = h_0(t)exp(\boldsymbol{\beta}'\boldsymbol{x}), \tag{3.2}$$

a koji je poznat pod imenom Coxov regresijski model proporcionalnog rizika. Uočimo da je za vrijednost regresora $\boldsymbol{x} = 0$ vrijednost tako odabrane funkcije r jednaka 1. Drugim riječima, bazna vrijednost funkcije hazarda postiže se kod onog individualca čija je vrijednost regresora $\boldsymbol{x} = 0$.

3.2 Procjena parametara

Pretpostavimo da je dan desno cenzurirani slučajan uzorak $(t_i, \delta_i), i = 1, \ldots, n$, pri čemu je $\delta_i = I(t_i \; je \; vrijeme \; smrti)$ indikator smrtnosti. Neka su s $t_{(1)} < \cdots < t_{(r)}$ označena uređena

vremena smrti ispitanika, dok su preostalih $n-r$ cenzurirana vremena. Za sada pretpostavljamo kako su vremena smrti ispitanika međusobno različita, što je valjan zahtjev u slučaju neprekidnog vremena. Neka $R(t_{(j)}), j = 1, 2, \ldots, r$, označava skup svih onih pacijenata koji su neposredno do trenutka $t_{(j)}$ živi i necenzurirani. Skup $R(t_{(j)})$ nazivamo skup pod rizikom u trenutku $t_{(j)}$ jer su u njega uključeni svi ispitanici koji su u tom trenutku pod rizikom smrti. Primjerice, $i-$ti ispitanik će upasti u skup $R(t_{(j)})$ ako za pripadno vrijeme doživljenja t_i vrijedi $t_i \geq t_{(j)}$. Pretpostavimo dodatno da su $X_1, X_2, \ldots, X_p$ slučajne varijable kojima opisujemo funkciju hazarda te neka su $\beta_1, \beta_2, \ldots, \beta_p$ nepoznati parametri Coxova regresijskog modela. Uz dane vrijednosti regresora $x_1, x_2, \ldots, x_p$, Coxov regresijski model poprima oblik

$$h(t) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p).$$

U ovakvom se modeliranju ne pojavljuje slobodni član (β_0) jer će njegovo djelovanje na funkciju rizika biti ugrađeno unutar bazne funkcije rizika. Jasno se vidi da će za procjenu funkcije rizika ispitanika u određenom trenutku biti potrebno procijeniti vrijednosti $\beta_1, \ldots, \beta_p$, ali isto tako i vrijednosti bazne funkcije rizika kroz vrijeme. Nama je od interesa isključivo doći do procjene za $\boldsymbol{\beta}$ dok nas procjenjivanje $h_0(t)$ ne zanima. Cox pokazuje, vidi [5], da ukoliko je pretpostavka o proporcionalnom riziku zadovoljena, tada $h_0(t)$ nije korisna za procjenu parametara od interesa, $\boldsymbol{\beta}$. Ujedno je pokazao da se procjenjivanje koeficijenata $\boldsymbol{\beta}$ može vršiti odvojeno od procjenjivanja $h_0(t)$. Ovakav rezultat je koristan jer ukoliko možemo procijeniti parametre $\boldsymbol{\beta}$ neovisno o $h_0(t)$, tada već iz njih možemo zaključivati o $h_i(t)/h_0(t)$. Odnosno, bit će moguće reći nešto o omjerima rizika smrti i-tog ispitanika u odnosu na baznu vrijednost funkcije rizika ili omjerima rizika ispitanika međusobno.

Vektor koeficijenata $\boldsymbol{\beta}$ može se u modelima proporcionalnog rizika procijeniti metodom maksimalne vjerodostojnosti. Premda se ne radi o pravoj funkciji vjerodostojnosti, relevantna funkcija vjerodostojnosti za model proporcionalnog rizika 3.2 dana je izrazom

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{exp(\boldsymbol{\beta}' \boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}, \tag{3.3}$$

gdje je $\boldsymbol{x}_{(j)}$ vektor regresora onog ispitanika čija je smrt zabilježena u trenutku $t_{(j)}$ (vidi [5]). Opravdanje za korištenjem $L(\boldsymbol{\beta})$ kao funkcije vjerodostojnosti navodimo u poglavlju 3.3. Suma u nazivniku ovako definirane funkcije vjerodostojnosti ide po svim ispitanicima koji su u trenutku $t_{(j)}$ živi i necenzurirani, tj. pod rizikom, dok se produkt vrši samo po onim mjerenjima za koje je zabilježena smrt. Dakle, cenzurirana mjerenja ne pridonose brojniku, ali ulaze u sumu unutar nazivnika. Štoviše, može se uočiti da funkcija vjerodostojnosti ovisi samo o poretku vremena smrti budući da ono određuje skupove pod rizikom za svako vrijeme smrti. Kao posljedica toga će i zaključci o efektima opisnih varijabli na funkciju rizika ovisiti samo o poretku vremena doživljenja.

Neka $R(t_i), i = 1, 2, \ldots, n$, označava skup pod rizikom u trenutku doživljenja t_i. Ukoliko se poslužimo indikatorom smrtnosti δ_i, izraz 3.3 može se zapisati kao

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{exp(\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \right)^{\delta_i}. \tag{3.4}$$

Uočimo, ukoliko t_i nije vrijeme smrti, tada je $\delta_i = 0$ pa odgovarajući faktor iznosi jedan i ne doprinosi produktu.
Uz ovakav zapis, logaritmirana funkcija vjerodostojnosti poprima oblik

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \Big(\boldsymbol{\beta}' \boldsymbol{x}_i - ln \sum_{l \in R(t_i)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l) \Big). \tag{3.5}$$

Skor vektor $\mathbf{U}(\boldsymbol{\beta}) = (\partial l/\partial \beta_1, \ldots, \partial l/\partial \beta_p)'$ i informacijska matrica poprimaju jednostavne forme. Derivirajući izraz 3.5 po $\boldsymbol{\beta}$ dobivamo vrijednost skor vektora

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \Big(\boldsymbol{x}_i - \frac{\sum_{l \in R(t_i)} \boldsymbol{x}_l exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \Big). \tag{3.6}$$

Sada, ukoliko za svaki $t > 0$ definiramo $p \times 1$ vektor

$$\overline{\boldsymbol{x}}(t, \boldsymbol{\beta}) = \frac{\sum_{l \in R(t)} \boldsymbol{x}_l exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}{\sum_{l \in R(t)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)}, \tag{3.7}$$

izraz 3.6 postaje

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \big(\boldsymbol{x}_i - \overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta}) \big). \tag{3.8}$$

Jednostavnim računom može se dobiti i vrijednost $p \times p$ dimenzionalne informacijske matrice

$$I(\boldsymbol{\beta}) = \frac{-\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = \sum_{i=1}^{n} \delta_i \Big(\frac{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l) \big(\boldsymbol{x}_l - \overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta}) \big) \big(\boldsymbol{x}_l - \overline{\boldsymbol{x}}(t_i, \boldsymbol{\beta}) \big)'}{\sum_{l \in R(t_i)} exp(\boldsymbol{\beta}' \boldsymbol{x}_l)} \Big). \tag{3.9}$$

U Coxovom regresijskom modelu proporcionalnog rizika se rješava problem procjene parametara $\boldsymbol{\beta}$ maksimizacijom funkcije 3.5 koristeći numeričke metode. Najčešće korištena metoda je Newton-Raphsonova (vidi [4]). Maksimizacija rezultira procjeniteljem $\hat{\boldsymbol{\beta}}$ koji je konzistentan i asimptotski normalan uz prikladne uvjete, a statistike bazirane na $L(\boldsymbol{\beta})$ kao što su skor statistika, informacija Fishera te omjer vjerodostojnosti, ponašaju se kao kod obične funkcije vjerodostojnosti (vidi [10]). Statistički paketi uz procijenjene vrijednosti parametara $\boldsymbol{\beta}$ obično daju i vrijednosti standardnih grešaka, test-statistika i intervala pouzdanosti koji se oslanjaju na asimptotsku aproksimaciju standardnom normalnom distribucijom,

$$\hat{\boldsymbol{\beta}} \overset{as}{\sim} N\big(\boldsymbol{\beta}, I(\hat{\boldsymbol{\beta}})^{-1}\big).$$

Valja napomenuti da je zaključke o $\boldsymbol{\beta}$ moguće donositi i na temelju drugih statistika. Neke od takvih statistika su statistika omjera vjerodostojnosti $\Lambda(\boldsymbol{\beta}) = 2l(\hat{\boldsymbol{\beta}}) - 2l(\boldsymbol{\beta})$ ili već definirana skor statistika $U(\boldsymbol{\beta})$. Načini korištenja navedenih statistika bit će prikazani kasnije.

Premda modeli proporcionalnog hazarda pretpostavljaju da je funkcija hazarda neprekidna, u primjeni se vremena doživljenja obično zaokružuju na najbliži dan, mjesec ili godinu i pritom je moguće da više pacijenata ima ista vremena smrti. Kao i dosad, ukoliko istu vrijednost postignu cenzurirano i necenzurirano vrijeme, za cenzurirano mjerenje smatramo da se dogodilo poslije. U situaciji kada omogućavamo istovremena mjerenja, za pronalazak procjene koeficijenata $\boldsymbol{\beta}$ više se ne možemo koristiti istom funkcijom vjerodostojnosti. Prisjetimo se, u formuli 3.3 $\boldsymbol{x}_{(j)}$ označava vrijednost vektora regresora za onog pojedinca koji umire u trenutku $t_{(j)}$. Ukoliko u tome trenutku umire više ispitanika, tada više nije jasno što je $\boldsymbol{x}_{(j)}$. Kalbfleish i Prentice (1980.) daju prikladnu funkciju vjerodostojnosti, no zbog kompliciranе forme njena upotreba nije pogodna. Postoje različite varijante aproksimacija

funkcije vjerodostojnosti koje nisu toliko numerički zahtjevne te se one koriste u primjenama. Navodimo jednu takvu aproksimaciju.

Neka je $\boldsymbol{s}_j$ $p-$dimenzionalni vektor dobiven sumiranjem regresora pacijenata koji umiru u trenutku $t_{(j)}, j = 1, 2, \ldots, r$. Ukoliko u trenutku $t_{(j)}$ imamo zabilježeno d_j smrti, tada je vrijednost h-tog elementa vektora $\boldsymbol{s}_j$ jednaka $s_{hj} = \sum_{k=1}^{d_j} x_{hjk}$, gdje je x_{hjk} vrijednost h-tog regresora $(h = 1, 2, \ldots, p)$ za k-tog od d_j pacijenata $(k = 1, 2, \ldots, d_j)$ koji su umrli u trenutku $t_{(j)}$ $(j = 1, 2, \ldots, r)$. Breslow (1974) daje najjednostavniju aproksimaciju funkcije vjerodostojnosti definiranu sljedećim izrazom

$$\prod_{j=1}^{r} \frac{exp(\boldsymbol{\beta}'\boldsymbol{s}_j)}{\left(\sum_{l \in R(t_{(j)})} exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right)^{d_j}}. \tag{3.10}$$

Ovakva aproksimacija nije numerički zahtjevna, a njena primjena je adekvatna ako ni jedan d_j nije prevelik. Još neke od aproksimacija daju Efron (1977.) i Cox (1972.) (vidi [4]). U praksi je najčešće korištena upravo 3.10, a rezultati dobiveni navedenim aproksimacijama pokazuju se sličnima. U slučaju kada je $d_j = 1$ za svako od vremena smrti, sve navedene aproksimacije podudaraju se s funkcijom vjerodostojnosti 3.3.

3.3 Opravdanje funkcije vjerodostojnosti

Konstrukciju funkcije vjerodostojnosti 3.3 za modele proporcionalnog rizika temeljimo na argumentu da intervali između vremena u kojima se dogodila smrt ne daju informacije o efektima regresora na funkciju hazarda. Opravdanje toga je što bazna funkcija rizika ima proizvoljan oblik te stoga ima smisla za vrijednost $h_0(t)$, pa tako i $h(t)$, uzeti nula u intervalima u kojima nije zabilježena smrt. Ovakav izbor sugerira da ti intervali neće nositi nikakve informacije o nepoznatom vektoru parametara $\boldsymbol{\beta}$. Iz podataka ćemo znati koje vrijednosti regresora postiže ispitanik koji umire u trenutku $t_{(j)}$. Označimo te vrijednosti regresora s $x_{(j)}$. Pretpostavimo kako taj vektor regresora pripada $i-$tom ispitaniku, tj. $\boldsymbol{x}_i = \boldsymbol{x}_{(j)}$.

Promatramo vjerojatnost da i-ti ispitanik umre u trenutku $t_{(j)}$, uvjetno na to da je trenutak $t_{(j)}$ jedan od onih r trenutaka u kojima se dogodila (točno jedna) smrt,

$$P(i - ti\ ispitanik\ umire\ u\ t_{(j)} | jedna\ smrt\ u\ t_{(j)}). \tag{3.11}$$

Događaj "$i-$ti ispitanik umire u $t_{(j)}$" je podskup događaja "jedna smrt u $t_{(j)}$". Koristeći tu činjenicu i formulu 2.3, izraz 3.11 postaje

$$\frac{P(i - ti\ ispitanik\ umire\ u\ t_{(j)})}{P(jedna\ smrt\ u\ t_{(j)})}.$$

Brojnik u izrazu iznad zapravo predstavlja funkciju hazarda $i-$tog ispitanika u trenutku $t_{(j)}$. Ova funkcija hazarda može se zapisati kao $h_i(t_{(j)})$. Nazivnik je jednak sumi hazarda smrti u $t_{(j)}$, za sve ispitanike koji su u tom trenutku pod rizikom. Izraz 3.11 sada postaje

$$\frac{h_i\big(t_{(j)}\big)}{\sum_{l \in R(t_{(j)})} h_l\big(t_{(j)}\big)}.$$

Kraćenjem baznih funkcija rizika, imamo

$$\frac{exp\big(\boldsymbol{\beta}'\boldsymbol{x}_{(j)}\big)}{\sum_{l \in R(t_{(j)})} exp\big(\boldsymbol{\beta}'\boldsymbol{x}_l\big)}.$$

U konačnici, produktom ovako dobivenih uvjetnih vjerojatnosti po svim vremenima smrti $t_{(j)}, j = 1, 2, \ldots, r$, dobivamo funkciju vjerodostojnosti danu formulom 3.3. Tako dobivena funkcija nije prava vjerodostojnost jer ona ne koristi stvarne vrijednosti cenzuriranih i necenzuriranih vremena, nego ovisi isključivo o poretku vremena.

3.4 Intervali pouzdanosti i testiranje hipoteza

Prilikom procjene koeficijenata $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ statistički programi daju i standardne greške za svaki od procjenitelja. Na temelju standardnih grešaka pronalazimo pouzdane intervale za nepoznate parametre $\boldsymbol{\beta}$. Primjerice, $100(1-\alpha)\%$ interval pouzdanosti za parametar β_j je oblika

$$[\hat{\beta}_j - z_{\alpha/2} s.e.(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2} s.e.(\hat{\beta}_j)],$$

gdje je $\hat{\beta}_j$ procijenjena vrijednost nepoznatog parametra β_j, a $z_{\alpha/2}$ predstavlja gornji $\alpha/2$ kvantil standardne normalne distribucije. Tada, ukoliko $100(1-\alpha)\%$ pouzdani interval ne sadrži 0, s tom sigurnošću možemo reći da je stvarna vrijednost parametra β_j različita od 0. Točnije, za svaki od parametara $\beta_j, j = 1, 2, \ldots, p$, definiramo statistički test s nultom i alternativnom hipotezom definiranom kao

$$H_0 : \beta_j = 0,$$

$$H_A : \beta_j \neq 0.$$

Za testiranje nul-hipoteze koristimo se test statistikom $\hat{\beta}_j / s.e.(\hat{\beta}_j)$. U uvjetima nul-hipoteze takva statistika ima standardnu normalnu distribuciju. Vrijednost statistike dobivene na temelju danih podataka uspoređuje se s percetilima standardne normalne distribucije u svrhu pronalaska pripadne p-vrijednosti. Ekvivalentno, možemo promatrati kvadriranu vrijednost ove statistike te ju uspoređivati s percentilima χ^2_1 distribucije. Taj postupak je poznat kao Waldov test.

Za interpretaciju p-vrijednosti za dani parametar β_j, potrebno je najprije prepoznati činjenicu da je hipoteza H_0 testirana uz prisutnost preostalih varijabli koje se pojavljuju u modelu. Primjerice, pretpostavimo da model sadrži tri opisne varijable X_1, X_2, X_3 uz koje stoje koeficijenti $\beta_1, \beta_2, \beta_3$. Recimo da promatramo test statistku $\hat{\beta}_2 / s.e.(\hat{\beta}_2)$ kojom želimo testirati hipotezu $\beta_2 = 0$ u prisutnosti β_1, β_3. Ako nemamo dovoljno dokaza za odbacivanje nul-hipoteze, zaključujemo da varijabla X_2 nije potrebna u modelu u kojem su prisutne X_1 i X_3. Pretpostavimo slučaj kad opet imamo iste tri varijable X_1, X_2, X_3. Ako se prilikom procjene koeficijenti β_1 i β_2 ne pokažu statistički značajnima, tada nije moguće donijeti zaključak da samo X_3 treba biti uključena u model. Razlog toga je što izbacivanje varijable X_1 iz modela može promijeniti vrijednost koeficijenta β_2 te ga učiniti statistički značajnim. Analogno, izbacivanje X_2 može učiniti β_1 značajnim. Ovakav scenarij će se događati u slučaju korelacije varijabli X_1 i X_2.

Zbog poteškoća pri interpretiranju rezultata testova koji se tiču koeficijenata uz opisne varijable u modelu, za uspoređivanje različitih modela proporcionalnog rizika zgodnije je koristiti neke alternativne metode.

3.5 Standardne greške i intervali pouzdanosti za omjere rizika

Standardne greške i intervale pouzdanosti za omjere rizika konstruiramo za slučaj u kojem uspoređujemo vremena doživljenja dvaju grupa pacijenata. Odnosno, promatramo situaciju

kada imamo samo jedan regresor koji je kategorijalna varijabla s mogućim realizacijama 0,1. Prema vrijednostima tog regresora određene su pripadnosti grupama. Jednadžba modela dana je formulom

$$h(t) = h_0(t)exp(\beta x).$$

Uočimo, funkcija hazarda Grupe 0 je zapravo bazna funkcija hazarda. Vrijednost parametra β jednaka je logaritmu omjera rizika u trenutku t ispitanika iz Grupe 1 u odnosu na one iz Grupe 0. Ukoliko promatramo baš omjer rizika ψ, on je s parametrom β povezan izrazom $\psi = exp(\beta)$. Iz ovog odnosa možemo lako doći do procjenitelja za omjer rizika. Ukoliko procjenitelj parametra β označimo $\hat{\beta}$, procjenitelj omjera rizika je dan s $\hat{\psi} = exp(\hat{\beta})$. Kako bismo i vrijednosti standardne greške od $\hat{\psi}$ dobili preko standardne greške od $\hat{\beta}$, koristimo se Taylorovom aproksimacijom varijance funkcije slučajne varijable (vidi[1]). Ona nam kaže, ukoliko imamo slučajnu varijalu X i funkciju g, tada je aproksimacija varijance slučajne varijable $g(X)$ dana sljedećim

$$Var\big(g(X)\big) \approx \left(\frac{dg(X)}{dX}\right)^2 Var(X). \tag{3.12}$$

Koristeći formulu 3.12, za varijancu procjenitelja $\hat{\psi}$ vrijedi

$$Var(\hat{\psi}) \approx \big(exp(\hat{\beta})\big)^2 Var(\hat{\beta}),$$

iz čega dobivamo aproksimaciju standardne greške procjenitelja $\hat{\psi}$,

$$s.e.(\hat{\psi}) \approx \hat{\psi} s.e(\hat{\beta}).$$

Interval pouzdanosti za stvarni omjer rizika generalno daje bolje informacije od standardne greške procjenitelja omjera rizika. $100(1-\alpha)\%$ pouzdani interval za omjer rizika ψ možemo jednostavno dobiti eksponencirajući pripadni pouzdani interval za β. Dobiveni pouzdani interval je oblika $[exp(\hat{\beta} - z_{\alpha/2}s.e.(\hat{\beta})), exp(\hat{\beta} + z_{\alpha/2}s.e.(\hat{\beta}))]$. Ovakav odabir pouzdanog intervala preferira se u odnosu na također korišten interval $[\hat{\psi} - z_{\alpha/2}s.e.(\hat{\psi}), \hat{\psi} + z_{\alpha/2}s.e.(\hat{\psi})]$. Razlog tome je što se distribucija logaritmiranog procjenitelja omjera rizika bolje aproksimira normalnom distribucijom nego sam procjenitelj omjera rizika.

3.6 Poveznica s log-rang testom

Pretpostavimo da imamo indikator varijablu X koja poprima vrijednost jedan ukoliko je pacijent iz Grupe 1, a vrijednost nula ako je pacijent iz Grupe 2. U poglavlju 2.5 pokazali smo kako usporediti distribucije vremena doživljenja dvaju grupa koristeći log-rang test. Nul-hipoteza log-rang testa jest jednakost distribucija doživljenja za te dvije grupe. Jasno je kako Coxov regresijski model također može služiti svrsi uspoređivanja distribucija doživljenja. Pristup koji Coxov model koristi drukčiji je nego u log-rang testu jer Coxov model opisuje funkciju hazarda. U nastavku pokazujemo postojanje uske veze između dvaju navedenih procedura. Oznake u nastavku preuzete su iz poglavlja 2.5.

Coxov model proporcionalnog hazarda za i-tog ispitanika možemo zapisati kao

$$h_i(t) = e^{\beta x_i} h_0(t),$$

gdje je x_i vrijednost indikator varijable X za i-tog ispitanika, $i = 1, 2, \ldots, n$. Nul-hipoteza koju ćemo testirati je $\beta = 0$. Kada bi nul-hipoteza vrijedila, tada bi funkcija hazarda

svakog ispitanika bila dana s h_0, bez obzira kojoj grupi on pripadao. Ako između grupa nema razlike u funkcijama hazarda, tada nema razlike ni u distribucijama doživljenja. U slučaju da su vremena smrti ispitanika međusobno različita, tj. kada vrijedi $d_{1j} + d_{2j} = 1$, do vrijednosti $\hat{\beta}$ dolazimo maksimizirajući funkciju vjerodostojnosti iz jednadžbe 3.3. Ako s $x_{(j)}$ označimo vrijednost varijable X za ispitanika koji umire u trenutku $t_{(j)}$, tada je funkcija vjerodostojnosti dana s

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta x_{(j)})}{\sum_{l=1}^{n_j} exp(\beta x_l)}, \tag{3.13}$$

kako je $n_j = n_{j1}+n_{j2}$ broj ispitanika pod rizikom u trenutku $t_{(j)}$. Odgovarajuća logaritmirana funkcija vjerodostojnosti poprima oblik

$$l(\beta) = \sum_{i=1}^{r} \beta x_{(j)} - \sum_{l=1}^{r} ln\Big(\sum_{l=1}^{n_j} exp(\beta x_l)\Big).$$

Kako je za ispitanike u Grupi 2 vrijednost $x_{(j)}$ jednaka nuli, prva sumacija ide po vremenima smrti u Grupi 1. Ako d_1 označava ukupan broj smrti u Grupi 1, tada vrijednost prvog sumanda iznosi $d_1\beta$. Također,

$$\sum_{l=1}^{n_j} exp(\beta x_l) = n_{1j}e^{\beta} + n_{2j}e^{0} = n_{1j}e^{\beta} + n_{2j},$$

te imamo

$$l(\beta) = d_1\beta - \sum_{j=1}^{r} ln\big(n_{1j}e^{\beta} + n_{2j}\big). \tag{3.14}$$

Maksimizirajući izraz 3.14 po parametru β, dolazimo do procjene za istog. Za maksimizaciju je potrebno koristiti nelinearne optimizacijske rutine. Nakon što dobijemo procjenu za β, nul-hipotezu $\beta = 0$ testiramo uspoređujući vrijednosti statistika logaritma vjerodostojnosti $-2\hat{l}(\hat{\beta})$ i $-2\hat{l}(0)$. Vrijednost statistike $-2\hat{l}(0)$ je $2\sum_{j=1}^{r} ln(n_j)$.

Izračun za $\hat{\beta}$ može se izbjeći koristeći skor test čija je nul-hipoteza također $\beta = 0$. Test statistika na kojoj je on baziran dana je s

$$\frac{U^2(0)}{I(0)},$$

gdje je

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta}$$

skor vektor, a

$$I(\beta) = -\frac{\partial^2 l(\beta)}{\partial(\beta^2)}$$

Fisherova informacijska funkcija. U uvjetima nul-hipoteze $U^2(0)/I(0)$ ima $\chi^2(1)$ distribuciju. Sada, deriviranjem jednadžbe 3.14 dobivamo

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{j=1}^{r} \left(d_{1j} - \frac{n_{1j}e^{\beta}}{n_{1j}e^{\beta} + n_{2j}}\right),$$

$$\frac{\partial^2 l(\beta)}{\partial \beta^2} = -\sum_{j=1}^{r} \frac{(n_{1j}e^{\beta} + n_{2j})n_{1j}e^{\beta} - (n_{1j}e^{\beta})^2}{(n_{1j}e^{\beta} + n_{2j})^2} = -\sum_{j=1}^{r} \frac{n_{1j}n_{2j}e^{\beta}}{(n_{1j}e^{\beta} + n_{2j})^2}.$$

Ukoliko skor vektor i informacijsku funkciju promatramo za vrijednost $\beta = 0$, dobivamo

$$U(0) = \sum_{j=1}^{r} \left(d_{1j} - \frac{n_{1j}}{n_{1j} + n_{2j}} \right),$$

i

$$I(0) = \sum_{j=1}^{r} \frac{n_{1j}n_{2j}}{(n_{1j} + n_{2j})^2}.$$

Uočimo, za specijalan slučaj kada su vremena smrti različita, tj. kada je $d_j = 1, j = 1, 2 \ldots, r$, vrijednosti $u(0)$ i $i(0)$ se podudaraju s vrijednostima statistika U_L i V_L definiranih jednadžbama 2.15 i 2.16. Odnosno, zaključujemo da se u ovako zadanoj situaciji log-rang test podudara sa skor testom koji proizlazi iz Coxovog modela. Kada imamo situaciju u kojoj dozvoljavamo da u istom trenutku umre više osoba, tada je funkciju vjerodostojnosti 3.13 potrebno zamijeniti nekom od aproksimacija koje to dozvoljavaju.
Prilikom analize ovakvih podataka prednost se daje Coxovom regresijskom modelu jer on, za razliku od log-rang testa, ujedno daje i procjene za omjere hazarda.

4 Odabir varijabli

Već smo napomenuli da u pristupu modeliranja podataka o doživljenju, model izgrađujemo povezujući funkciju hazarda s različitim regresorima. U potrazi za konačnim modelom izgrađeni su mnogi modeli proporcionalnog hazarda, a čije linearne komponente sadrže različite varijable. Odgovor na to koji od mnoštva modela izabrati kao najbolji nije nimalo jednostavan. Kroz ovo poglavlje najprije prikazujemo načine usporedbe modela s njemu ugnježđenim modelom, zatim usporedbe modela općenito i u konačnici navodimo proceduru za selekciju varijabli u model.

4.1 Statistika logaritma vjerodostojnosti

Kako bismo usporedili alternativne modele potrebno je pronaći statistiku koja će na neki način mjeriti koliko kvalitetno model prati dane podatke. Funkcija vjerodostojnosti je ta koja sumira informacije koje podaci sadrže o nepoznatim parametrima danog modela. Stoga, kao prirodna statistika za ovakav problem nameće se statistika dobivena metodom maksimalne vjerodostojnosti. Traženu statistiku dobivamo tako da u funkciji vjerodostojnosti nepoznate parametre zamijenimo s vrijednostima procjena koji takvu funkciju maksimiziraju. U našem konkretnom slučaju funkcija vjerodostojnosti dana je izazom 3.3 pa do vrijednosti statistike $\hat{L} = L(\hat{\boldsymbol{\beta}})$ dolazimo uvrštavajući u navedeni izraz one vrijednosti $\hat{\boldsymbol{\beta}}$ koji maksimiziraju $L(\boldsymbol{\beta})$. Za dani skup podataka vrijedi što je veća vrijednost maksimizirane vjerodostojnosti $\hat{L}$, to model bolje prati dane podatke.

Umjesto statistike $\hat{L}$ zgodnije je za usporedbu modela koristiti statistiku $-2\hat{l} = -2\ln\hat{L}$. $\hat{L}$ je dobivena kao produkt uvjetnih vjerojatnosti te je stoga njena vrijednost uvijek manja od 1. Zbog toga je vrijednost statistike $-2\hat{l}$ uvijek pozitivna, no sada je za dane podatke manja vrijednost statistike $-2\hat{l}$ vezana uz bolji model.

Uočimo da ona sama po sebi ne može reći koliko je model prikladan, nego će služiti tome da od više alternativa odaberemo onu najbolju. Jasno je iz definicije da $\hat{L}$ pa tako i $-2\hat{l}$, uveliko ovise o broju podataka koje imamo na raspolaganju te će se vrijednost navedenih statistika mijenjati kako novi podaci ulaze u model. Prilikom usporedbe dvaju modela korištenjem statistike $-2\hat{l}$ treba paziti da su uspoređivani modeli izgrađeni nad istim skupom mjerenja. U idućim poglavljima ilustriramo primjenu statistike $-2\hat{l}$ za testiranje hipoteza o jednakosti kvalitete dvaju modela.

4.2 Usporedba s ugnježđenim modelom

Pretpostavimo da smo nad istim skupom mjerenja izgradili dva modela koje označavamo Model(1) i Model(2). Označimo $\hat{L}(1)$ i $\hat{L}(2)$ vrijednosti pripadnih statistika dobivenih metodom maksimalne vjerodostojnosti. U prošlom poglavlju smo naznačili da ćemo se umjesto njih ipak koristiti statistikama $-2\hat{l}(1)$, $-2\hat{l}(2)$.

Pretpostavimo da je Model(1) izgrađen na temelju varijabli $X_1, X_2, \ldots, X_p$, a Model(2) na temelju $X_1, X_2, \ldots, X_p, X_{p+1}, \ldots, X_{p+q}$. Vidimo da Model(2) sadrži sve varijable kao i Model(1) te još q dodatnih. Za Model(1) kažemo da je ugnježđen u Model(2). Postavlja se pitanje koji od ponuđenih modela je bolji za upotrebu. Jasno je da će Model(2) bolje pratiti dane podatke, no je li uvođenje dodatnih varijabli $X_{p+1}, \ldots, X_{p+q}$, zaista dalo značajan doprinos kvaliteti modela. Taj doprinos ćemo mjeriti statistikom $-2\hat{l}(1) + 2\hat{l}(2)$. Ovako definiranu statistiku možemo zapisati i kao

$$-2ln\{\hat{L}(1)/\hat{L}(2)\},$$

te ju zovemo statistika log-omjera vjerodostojnosti. Njom se služimo u statističkom testu koji za nultu i alternativnu hipotezu ima sljedeće:

$$H_0 : \beta_{p+1} = 0, \beta_{p+2} = 0, \ldots, \beta_{p+q} = 0,$$

$$H_A : \beta_j \neq 0, \text{ za neki } j = p+1, \ldots, p+q.$$

U uvjetima nul-hipoteze ona asimptotski ima χ^2 distribuciju s onoliko stupnjeva slobode kolika je razlika u broju parametara dvaju modela, što je u ovom slučaju q. Vrijednost statistike log-omjera vjerodostojnosti promatramo u odnosu na pripadnu χ^2_q distribuciju. Ako dobivena vrijednost nije dovoljno velika, tada za modele smatramo da su jednako prikladni, no odlučujemo se za korištenje manjeg modela. Ukoliko vrijednost statistike pokaže da su modeli značajno različiti, odnosno da je Model(2) statistički značajno bolji od Model(1), tada je potrebno tražiti složeniji model dodavajući neke od q preostalih varijabli u Model(1).

4.3 Odabir varijabli

Prilikom analiziranja vremena doživljenja potrebno je najprije pronaći one varijable koje imaju potencijal da ih se uključi u linearnu komponentu modela proporcionalnog hazarda. Nakon toga potrebno je između svih tih varijabli odabrati one koje ćemo koristiti pri modeliranju funkcije hazarda. U praksi je često situacija takva da funkcija hazarda ne ovisi o jedinstvenoj kombinaciji regresora, nego postoji više jednako dobrih modela. Iz tog razloga, poželjno je razmatrati širok spektar modela. Pretpostavimo trenutno da smo nekim procedurama pronašli više kombinacija regresora koji dobro opisuju funkciju hazarda. Takve kombinacije vode različitim modelima, no kako sada usporediti sve te modele.

U prethodnom poglavlju smo pokazali kako usporediti model s njemu ugnježđenim modelom koristeći statistiku $-2\hat{l}(1) + 2\hat{l}(2)$. Za usporedbu modela kod kojih nije nužno da je jedan ugnježđen u drugi možemo se poslužiti statistikom

$$AIC = -2\hat{l} + \alpha q, \tag{4.1}$$

gdje je q broj nepoznatih parametara $\boldsymbol{\beta}$ unutar modela, a α unaprijed određena konstanta. Za vrijednost α najčešće se uzima vrijednost između 2 i 6, s tim da je generalna preporuka uzeti $\alpha = 3$. Statistika dana formulom 4.1 poznata je pod nazivom Akaike informacijski kriterij (AIC). Što manju vrijednost AIC poprima to je model bolji. Iz njegove definicije vidimo da AIC kažnjava ubacivanje nepotrebnih varijabli u model.

Česta je situacija da uključivanje različitih skupova regresora daje slične vrijednosti AIC sugerirajući njihovu jednaku prikladnost, tako da nije moguće donositi zaključke samo na temelju AIC-a. Svakako je u takvim situacijama potrebno analizirati smislenost uključenih varijabli. U svakom slučaju potrebno je pri odabiru provjeriti i za koji su skup regresora najbolje zadovoljene teorijske pretpostavke modela.

Uočimo, do sada ni u jednom trenutku nismo pojasnili način kako doći do neke od kombinacija regresora koje bismo uključili u model. Postoje razne automatske procedure za selekciju varijabli kao što su selekcija unaprijed, selekcija unazad te njihova kombinacija. Sve one ubacuju ili izbacuju varijable na temelju $-2\hat{l}$. Premda se one često pokazuju uspješnima, automatske procedure imaju svoje nedostatke. Neki od nedostataka su što one kao rezultat daju točno jedan model dok je stvarna situacija često takva da postoji više jednako dobrih modela. Također, općenito ne poštuju pravilo hijerarhije i izbor varijabli uvelike ovisi o tome kako se definira pravilo zaustavljanja koje odlučuje hoće li varijabla ući u model ili ne. Uz sve to, moguće je da postoje neke od varijabli čije je uključivanje u model od primarnog značaja, a koje takve procedure možda odluče ne ubaciti u model.

Uzevši u obzir nedostatke automatskih selekcijskih procedura, Collett predlaže korištenje sljedeće strategije selekcije (vidi [4]).

1. Najprije se za svakog od potencijalnih regresora napravi poseban model. Vrijednosti statistike $-2\hat{l}$ svakog od tih modela uspoređuju se s vrijednošću iste iz nul-modela te se tako utvrđuje koje varijable same po sebi značajno doprinose smanjenju statistike.

2. Varijable koje se u koraku 1 pokažu značajnima se zajedno uključuju u novi model. U prisutnosti drugih, neke od varijabli mogu prestati biti značajne. Ukoliko trenutno izbacivanje varijable iz modela ne dovede do značajnog povećanja vrijednosti statistike $-2\hat{l}$, tada ju možemo odbaciti. Značajnost svake sljedeće varijable promatra se u odnosu na model u kojem su izbačene dotad beznačajne varijable.

3. Varijable koje se u koraku 1 nisu pokazale značajne samostalno, i kao takve nisu sudjelovale u koraku 2, mogu postati značajne u prisutnosti ostalih varijabli. Ove varijable dodajemo u model iz koraka 2, jednu po jednu. Ako neke od njih značajno smanje $-2\hat{l}$, onda ih ubacujemo u model. Moguće je da nakon njihova ubacivanja neki od koeficijenata uz varijable iz koraka 2 prestanu biti značajni.

4. Zadnja provjera vrši se kako bi se osiguralo da niti jedna od varijabli iz modela ne može biti izbačena bez značajnog povećanja $-2\hat{l}$ te da dodavanje preostalih varijabli ne vodi značajnom smanjenju $-2\hat{l}$.

Treba napomenuti da prilikom ovakvog selektiranja nije preporučljivo držati se neke fiksne razine značajnosti. Preporuča se korištenje značajnosti veličine otprilike 10%. Ponekad je potrebno razmotriti i uključivanje interakcija ili primjerice potencija varijabli u model. Njih bismo u strategiji selekcije ubacili u koraku 3 nakon što smo se uvjerili da su izrazi koji su potrebni za poštivanje hijerarhije već ubačeni u model. Ako oni vode značajnom smanjenju $-2\hat{l}$, tada ih ubacujemo u model.

5 Provjera pretpostavki modela proporcionalnog rizika

Pretpostavimo da je dosad opisanim metodama pronađen model koji služi kao kandidat za opisivanje distribucije podataka. Prije no što model stavimo u upotrebu, potrebno je provjeriti njegovu adekvatnost. Proučit ćemo četiri aspekta provjere adekvatnosti predloženog modela proporcionalnog rizika. Prvi aspekt tiče se najbolje funkcionalne forme regresora, tj. provjere treba li neke od uključenih varijabli transformirati prije uključivanja u model. Drugi aspekt podrazumijeva provjeru pretpostavke o proporcionalnom riziku. Ukoliko ona nije zadovoljena, tada rezultati modela mogu dovesti do krivih zaključaka. U trećem aspektu provjeravamo točnost predikcija doživljenja za pacijente. Tu nam je od interesa pronaći one pacijente čija se vremena doživljenja razlikuju od očekivanog vremena dobivenog modelom. Posljednji aspekt provjerava utjecaje pacijenata na model.

Mnoge od procedura koje služe za provjeru valjanosti modela koriste se rezidualima. Reziduali su veličine koje je moguće izračunati za svakog pacijenta te je njihovo ponašanje poznato ili aproksimativno poznato u slučaju adekvatnosti promatranog modela. Za Coxov regresijski model do danas je predloženo mnogo različitih definicija reziduala koji svoju primjenu nalaze u prethodno navedenim aspektima. U nastavku su definirani i objašnjeni neki od standardno korištenih reziduala.

Kao i dosad, promatramo slučaj kada imamo ukupno n dostupnih mjerenja od kojih je ukupno r puta zabilježena smrt ispitanika, a $n-r$ podataka je desno cenzurirano. Linearnu komponentu modela opisujemo s ukupno p regresora $X_1, X_2, \ldots, X_p$, koji su vremenski nepromjenjivi. Tada za i-tog pacijenta procijenjenju funkciju rizika u trenutku t definiramo kao

$$\hat{h}_i(t) = exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)\hat{h}_0(t),$$

gdje je $\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i = \hat{\boldsymbol{\beta}}_1' x_{1i} + \hat{\boldsymbol{\beta}}_2' x_{2i} + \cdots + \hat{\boldsymbol{\beta}}_p' x_{pi}$ vrijednost linearne komponenta modela (još zvana skor rizika) za i-tog pacijenta, a $\hat{h}_0(t)$ procijenjena bazna funkcija rizika.

5.1 Cox-Snell reziduali

Cox-Snell reziduali često su korišteni prilikom analize preživljenja te svoju primjenu imaju prvenstveno pri provjeri ukupne kvalitete Coxova modela proporcionalnog rizika.
Za $i = 1, 2, \ldots, n$, je i-ti Cox-Snell rezidual definiran s

$$r_{C_i} = exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)\hat{H}_0(t_i), \tag{5.1}$$

pri čemu je $\hat{H}_0(t_i) = -ln(\hat{S}_0(t_i))$ procijenjena kumulativna bazna funkcija rizika u trenutku t_i. Uočimo da je $r_{C_i} = \hat{H}_i(t_i) = -ln(\hat{S}_i(t_i))$. Iz same definicije vidimo da ovako definirani reziduali poprimaju isključivo nenegativne vrijednosti.

Navedimo teorem kojim se koristimo pri zaključivanju o korektnosti modela na temelju Cox-Snell reziduala.

Teorem 5.1. *Neka je T neprekidna nenegativna slučajna varijabla s kumulativnom funkcijom hazarda H. Tada slučajna varijabla $Y = H(T)$ ima eksponencijalnu distribuciju s parametrom $\lambda = 1$.*

Dokaz. Prisjetimo se, prema formuli 2.8 vrijedi $Y = H(T) = -ln(S(T))$. Sada koristimo rezultat iz teorije vjerojatnosti koji kaže da, ako je $f_X(x)$ funkcija gustoće slučajne varijable X, a slučajna varijabla Y zadana s $Y = g(X)$, tada je

$$f_Y(y) = f_X\big(g^{-1}(y)\big) \Big/ \left|\frac{dy}{dx}\right|.$$

Koristeći gornji rezultat imamo

$$f_Y(y) = f_T\big(S^{-1}(e^{-y})\big) / \left|\frac{dy}{dt}\right|, \tag{5.2}$$

gdje je f_T funkcija gustoće slučajne varijable T. Sada je

$$\frac{dy}{dt} = \frac{d(-ln(S(t)))}{dt} = \frac{f_T(t)}{S(t)},$$

a kada apsolutnu vrijednost gornjeg izraza izrazimo u terminima y, derivacija postaje

$$\frac{f_T(S^{-1}(e^{-y}))}{S(S^{-1}(e^{-y}))} = \frac{f_T(S^{-1}(e^{-y}))}{e^{-y}}.$$

Dobiveno uvrstimo u jednadžbu 5.2 te imamo

$$f_Y(y) = e^{-y},$$

što odgovara funkciji gustoće slučajne varijable koja ima eksponencijalnu distribuciju s parametrom 1, tj. $Y \sim \varepsilon(1)$. □

Iz teorema 5.1 i jednakosti 2.8 slijedi da $-ln(S(T))$ ima eksponencijalnu distribuciju s parametrom $\lambda = 1$. Ako su vrijednosti $\hat{\beta}_1, \ldots, \hat{\beta}_p$ bliske stvarnim vrijednostima $\beta_1, \ldots, \beta_p$, tada bi prema prethodnom teoremu Cox-Snell reziduali trebali izgledati kao cenzurirani uzorak iz jedinične eksponencijalne distribucije. Razlog toga je što je u slučaju zadovoljavajućeg modela, procjenitelj funkcije preživljenja $\hat{S}_i(t_i)$ blizak stvarnoj vrijednosti $S_i(t_i)$. To nam sugerira da će $\hat{S}_i(t_i)$ imati slična svojstva kao i $S_i(t_i)$. Posljedično će $\hat{H}_i(t_i)$ imati svojstva slična $H_i(t_i)$ pa prema teoremu $\hat{H}_i(t_i), i = 1, 2, \ldots, n$, čine uzorak iz eksponencijalne distribucije.

Kumulativnu funkcija hazarda slučajne varijable s jediničnom eksponencijalnom distribucijom zadovoljava $H_E(t) = t$. Tada, ukoliko r_{C_i} dolaze iz takve distribucije, treba vrijediti $H_r(r_{C_i}) = r_{C_i}$. O prikladnosti upotrebe modela možemo zaključivati iz grafičkog prikaza kojemu je na apscisi r_{C_i}, a na ordinati vrijednost procijenjenog kumulativnog hazarda za $r_{C_i}-$ove, odnosno $\hat{H}_r(r_{C_i})$. U slučaju da r_{C_i} zaista dolaze iz eksponencijalne distribucije tada će uređeni parovi$(r_{C_i}, \hat{H}_r(r_{C_i}))$ ležati blizu pravca koji prolazi kroz ishodište i ima nagib 1. Ukoliko oni ne leže oko tog pravca, to nam sugerira da model nije dobar, no ne daje informacije o kakvom tipu odstupanja je riječ.

5.2 Martingalni reziduali

U ovom poglavlju ćemo proučiti problematiku odabira ispravne forme regresora prilikom uključivanja u model, odnosno pronalaska transformacije regresora koja najbolje objašnjava njen efekt na preživljenje unutar Coxova regresijskog modela. Neke od transformacija koje se najčešće upotrebljavaju su logaritamska, kvadratna ili pak diskretizacija. Diskretizacija je proces koji se često koristi u praksi, a martingalni reziduali koje ovdje predstavljamo mogu se koristiti prilikom odabira točaka prekida regresora. Martingalne reziduale ovdje definiramo specijalno za slučaj kada radimo s desno cenzuriranim podacima i regresorima koji su vremenski nepromjenjivi (vidi [9]). Oni su zapravo modifikacija Cox-Snell reziduala definiranih u prethodnom poglavlju, a definiramo ih kao

$$r_{M_i} = \delta_i - r_{C_i}, \tag{5.3}$$

pri čemu je r_{C_i} i-ti Cox-Snell rezidual, a $\delta_i = 1$ ako je i-to mjerenje necenzurirano, a inače jednako 0. Martingalni reziduali općenito poprimaju vrijednosti iz intervala $(-\infty, 1)$, a oni za cenzurirana vremena isključivo negativne vrijednosti. Reziduali imaju svojstvo $\sum_{j=1}^{n} r_{M_j}=0$, a za velike uzorke $r_{M_j}, j = 1, 2, \ldots, n$, su nekorelirani iz populacije s očekivanjem nula. Motivacija za uvođenje martingalnih reziduala leži u činjenici da, ukoliko u njihovoj definiciji umjesto procijenjenih vrijednosti iz uzorka stavimo stvarne vrijednosti $\boldsymbol{\beta}$ i H_0, tako dobiveni proces reziduala bio bi martingal (vidi [9]). Mogli bismo ih interpretirati kao razlike između zabilježenog broja smrti i očekivanog broja smrti pod pretpostavljenim Coxovim modelom, tijekom vremena.

Označimo sada s $\boldsymbol{X}$ vektor regresora odabranih u modelu te razdvojimo vektor $\boldsymbol{X}$ na X_1 i $\boldsymbol{X}^*$. Pretpostavimo da za $\boldsymbol{X}^*$ znamo ispravnu funkcionalnu formu unutar Coxova modela, a za X_1 to tek pokušavamo ustanoviti. Nadalje, pretpostavimo da su X_1 i $\boldsymbol{X}^*$ međusobno nezavisni. Označimo s f funkciju koja najbolje transformira X_1 za potrebe opisivanja preživljenja. Tada naš optimalni Coxov model poprima oblik

$$H(t) = H_0(t)exp(\boldsymbol{\beta}^*\boldsymbol{X}^*)exp(f(X_1)).$$

Kako bismo pronašli f, Coxov model dobiven na temelju $\boldsymbol{X^*}$ usklađujemo s podacima te računamo martingalne reziduale. Dobivene reziduale stavljamo u odnos s vrijednostima X_1 na dijagram raspršenja. Na dijagram raspršenja dodajemo zaglađenu krivulju dobivenu LOESS procedurom koja nam sugerira kakvog oblika je funkcija f. Više o LOESS proceduri može se naći u [3]. Primjerice, ukoliko je krivulja linearna, tada X_1 nije potrebno transformirati. Ako se na različitim intervalima krivulja različito ponaša, to sugerira diskretnu transformaciju varijable X_1.

Neki autori predlažu korištenje grafičkih prikaza martinglanih reziduala u odnosu na vremena doživljenja ili u odnosu na rangove vremena doživljenja za indikacije o adekvatnosti modela. Mi ih ne koristimo za ovu svrhu iz razloga što oni nisu simetrično distribuirani oko nule i što ta zakrivljenost čini grafičke prikaze bazirane na njima teškim za interpretaciju.

5.3 Reziduali temeljeni na devijanci

U ovom poglavlju uvodimo novu defincju reziduala koje upotrebljavamo pri utvrđivanju stršećih mjerenja u modelu, nakon što smo se odlučili za konačni model. Reziduale temeljene na devijanci definiraju Therneau, Grambsch i Fleming kao

$$r_{D_i} = sgn(r_{M_i})\sqrt{-2\big(r_{M_i} + \delta_i ln(\delta_i - r_{M_i})\big)}, \tag{5.4}$$

gdje je $sgn()$ funkcija predznaka kojom se osiguravamo da rezidual bude istog predznaka kao i pripadni martingalni rezidual (vidi [14]). U odnosu na martingalne, ovako definirani reziduali imaju distribuciju koja sliči normalnoj. Možemo ih smatrati martingalnim rezidualima koji su transformirani tako da u slučaju prikladnosti modela budu simetrično distribuirani oko nule. Iako će tada njihova distribucija biti simetrična oko nule, to i dalje ne znači da im je suma jednaka nula. Kao i kod linearne regresije, u interesu nam je pronaći odgovarajuće grafičke prikaze iz kojih možemo vidjeti postoje li mjerenja za koja model ne daje dobru predikciju. Kako bismo ustanovili efekt pojedinog mjerenja na model, preporuča se korištenje grafičkog prikaza u kojem reziduale r_{D_i} stavljamo u odnos sa skorom rizika $\sum_{k=1}^{p} \hat{\boldsymbol{\beta}}_k x_{jk}$. Ako uzorak ne sadrži velik broj cenzuriranih mjerenja, tada on izgleda kao uzorak iz normalne distribucije. Ukoliko imamo dosta cenzuriranih mjerenja, tada će velik broj reziduala imati vrijednosti bliske nuli što će pokvariti normalnost distribucije. U oba će slučaja, neovisno

o razini cenzuriranja, potencijalna stršeća mjerenja imati reziduale s visokim apsolutnim vrijednostima. Mjerenja s visokom pozitivnom vrijednošću reziduala predstavljaju ispitanike koji umiru znatno ranije u odnosu na modelom prediktirano vrijeme. Negativni reziduali s visokom apsolutnom vrijednosti predstavljaju one ispitanike koji žive duže nego to model sugerira.

5.4 Metode provjere pretpostavke proporcionalnog rizika

U ovom poglavlju dajemo neke od metoda kojima možemo ispitati remeti li kategorijalna varijabla pretpostavku proporcionalnosti rizika.

Jedan od načina na koji je moguće provjeriti pretpostavku proporcionalnog rizika jest ubacivanje vremenski zavisne varijable u Coxov regresijski model. Primjenu ove metode objašnjavamo u specifičnoj situaciji kada su pacijenti podijeljeni u dvije grupe, primjerice prilikom testiranja nove terapije, kada jedna grupa pacijenata prima standardnu, a druga grupa novu. Pretpostavimo da je X_1 indikator varijabla koja poprima vrijednost nula ukoliko pacijent prima standardnu terapiju, a vrijednost jedan ako prima novu terapiju. Od interesa je saznati zavisi li omjer rizika smrti u trenutku t dvaju grupa o vremenu doživljenja. Za i-tog pacijenta vrijednost funkcije rizika u trenutku t dana je s

$$h_i(t) = exp(\beta_1 x_{1i}) h_0(t).$$

U slučaju istinitosti pretpostavke o proporcionalnosti hazarda je omjer rizika smrti pacijenata na novoj terapiji u odnosu na one na standardnoj u svakom trenutku t jednak e^{β_1}.

Uvedimo sada novu opisnu varijablu $X_2 = X_1 t$. Nakon uvođenja varijable X_2 naš model za i-tog pacijenta postaje

$$h_i(t) = exp(\beta_1 x_{1i} + \beta_2 x_{2i}) h_0(t).$$

U trenutku t omjer rizika sada iznosi $exp(\beta_1 + \beta_2 t)$, što očito ovisi o vremenu t, stoga gornji model više nije model proporcionalnog rizika. Ukoliko procjenom dobijemo vrijednost $\beta_2 < 0$, tada smo u situaciji kada se omjer rizika smanjuje tijekom vremena, a interpretiracija bi glasila da djelotvornost nove terapije tijekom vremena postaje izraženija. Ukoliko je $\beta_2 > 0$, tada se omjer rizika povećava s vremenom, što bi značilo da rizik smrti pacijenata na novoj terapiji tijekom vremena postaje veći u odnosu na onaj standardne terapije. U specijalnom slučaju ovog modela, kada je $\beta_2 = 0$, omjer rizika je konstantan i iznosi e^{β_1}. Tada testiranje hipoteze $\beta_2 = 0$ ujedno predstavlja testiranje pretpostavke proporcionalnosti rizika. Kako bismo izbjegli probleme s interpretacijom parametara modela i onima numeričke prirode, moguće je varijablu X_2 definirati u terminima otklona od nekog fiksiranog vremena t_0, $X_2 = X_1(t - t_0)$. Za t_0 obično se uzima prosječno ili medijalno vrijeme doživljenja. Ovakvi odabiri za t_0 pružaju mogućnost jednostavne interpretacije parametara β_1, β_2. Omjer rizika sada iznosi

$$exp(\beta_1 + \beta_2(t - t_0)).$$

U ovako definiranom modelu e^{β_1} predstavlja omjer rizika smrti pacijenata na novoj terapiji u odnosu na one na standardnoj u trenutku t_0. U mnogim primjenama, vremenski zavisna varijabla X_2 definira se kao $X_2 = X_1 ln(t)$. Razlog tome jest što vrijednosti vremena doživljenja znaju biti velike i stoga se pojavljuju problemi pri računanju $exp(\beta_2 x_{2i})$. No, testiranje hipoteze $\beta_2 = 0$ je i dalje test za provjeru pretpostavke proporcionalnog rizika.

Još jedna od metoda provjere proporcionalnosti rizika bila bi dati grafički prikaz logaritmiranih funkcija kumulativnog hazarda po grupama u odnosu na vrijeme. Razloge za upotrebom ovakve metode objašnjavamo za situaciju kada imamo samo jednu varijablu uključenu

u model, i to indikator varijablu X_1. Coxov regresijski model zapisujemo u sljedećem obliku

$$H(t) = H_0(t)exp(\beta x_1).$$

Tada logaritmirajući imamo

$$ln(H(t)) = ln(H_0(t)) + \beta x_1. \tag{5.5}$$

Neka $\hat{H}_0(t)$ označava procijenjenu vrijednost funkcije kumulativnog hazarda u trenutku t za one ispitanike kojima je $x_1 = 0$, a $\hat{H}_1(t)$ procijenjenu vrijednost funkcije kumulativnog hazarda u t za one ispitanike kojima je $x_1 = 1$. Jednadžba 5.5 nam sugerira, ukoliko varijabla X_1 podržava pretpostavku proporcionalnosti, tada će procijenjene krivulje kumulativnih hazarda dvaju grupa biti aproksimativno paralelne, a vertikalna udaljenost između njih daje grubu procjenu parametra β. Alternativno, mogli bismo promatrati grafički prikaz na kojem apscisa i dalje predstavlja vrijeme, a ordinata $ln(\hat{H}_1(t)) - ln(\hat{H}_0(t))$. Ako je pretpostavka proporcionalnosti hazarda zadovoljena, tada će krivulja razlike biti otprilike jednaka konstanti, odnosno fluktuirat će oko konstantne vrijednosti koja je jednaka procjeni parametra β. Ovakvo testiranje se analogno generalizira za slučaj kad varijabla X_1 ima više od dvije kategorije, kao i za slučaj kada postoje i druge varijable uključene u model gdje se dodatno zahtijeva da ne postoje interakcije između varijable X_1 i preostalih varijabli.

Još neke od standardno korištenih grafičkih metoda za ispitivanje proporcionalnosti rizika su Andersenovi te Arjas prikazi. I oni funkcioniraju isključivo za provjeru proporcionalnosti hazarda za kategorijalne varijable. Više o njima može se pronaći u [9].

5.5 Schoenfeld reziduali

Glavna dva nedostatka dosad definiranih reziduala su zavisnost o opserviranim vremenima doživljenja te nužnost procjenjivanja funkcije kumulativnog hazarda. Oba nedostatka riješena su korištenjem reziduala koje predlaže Schoenfeld (vidi [12]). Još jedna od razlika s prethodno definiranim rezidualima je to što za svakog ispitanika ovdje imamo onoliko reziduala koliko je varijabli uključeno u Coxov regresijski model. $i-$ti Schoenfeldov rezidual za varijablu X_j, dan je kao

$$r_{SC_{ji}} = \delta_i(x_{ji} - \hat{a}_{ji}), \tag{5.6}$$

gdje je x_{ji} vrijednost $j-$te varijable, $j = 1, 2, \ldots, p$, za $i-$tog ispitanika, $i = 1, 2, \ldots, n$,

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_i)} x_{jl} exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}{\sum_{l \in R(t_i)} exp(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_l)}, \tag{5.7}$$

a $R(t_i)$ skup ispitanika pod rizikom u trenutku t_i. Primijetimo kako Schoenfeldovi reziduali poprimaju ne-nul vrijednosti isključivo za one ispitanike čija su vremena doživljenja necenzurirana. Također, uočimo kako je $i-$ti Schoenfeldov rezidual za varijablu X_j zapravo procjena $i-$te komponente prve derivacije logaritma funkcije vjerodostojnosti po parametru β_j. Odnosno, on je procjena $j-$te komponente skor vektora $\boldsymbol{U}(\boldsymbol{\beta})$ danog s 3.6, za $i-$tog ispitanika.

Metode iz poglavlja 5.4 nisu mogle reći zadovoljava li promatrani regresor pretpostavke proporcionalnog rizika bez prethodne diskretizacije. Sljedeća metoda omogućuje proučavanje otklona od proporcionalnosti za dani regresor, bez potrebe za diskretizacijom. Metoda se temelji na skaliranim Schoenfeld rezidualima koje predlažu Grambsch i Therneau (vidi [8]).

Neka je s $r_{SC_i} = (r_{SC_{1i}}, r_{SC_{2i}}, \ldots, r_{SC_{pi}})'$ dan vektor Schoenfeldovih reziduala za $i-$tog ispitanika. Skalirani Schoenfeldovi reziduali $r^*_{SC_{ij}}$ su komponente vektora

$$r^*_{SC_i} = rVar(\hat{\boldsymbol{\beta}})\boldsymbol{r}_{SC_i},$$

gdje je r broj smrti u uzorku od n ispitanika, a $Var(\hat{\boldsymbol{\beta}})$ matrica kovarijanci procjenitelja $\hat{\boldsymbol{\beta}}$.

Nadalje, pretpostavimo situaciju u kojoj na temelju varijabli $X_1, X_2, \ldots, X_p$, želimo izgraditi Coxov regresijski model proporcionalnog hazarda

$$h(t) = h_0(t)exp(\boldsymbol{\beta}'\boldsymbol{x}).$$

Postavlja se pitanje, mijenja li se efekt regresora $\boldsymbol{x}$ na funkciju hazarda tijekom vremena, tj. jesu li koeficijenti uz varijable vremenski promjenjivi. Za odgovor na to pitanje model zapisujemo u općenitijem obliku

$$h(t) = h_0(t)exp\big(\boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\gamma}'(\boldsymbol{g}(t) * \boldsymbol{x})\big), \tag{5.8}$$

gdje je $\boldsymbol{g}(t) * \boldsymbol{x} = (g_1(t)x_1, \ldots, g_p(t)x_p)$, funkcije $g_j(t)$ su poznate, a $\boldsymbol{\gamma}$ nepoznati $p \times 1$ dimenzionalni vektor koeficijenata. Uočimo, ako je $\gamma_j \neq 0$, tada je koeficijent β_j vremenski promjenjiv, odnosno $\beta_j(t) = \beta_j + \gamma_j g_j(t)$.

Za testiranje nul-hipoteze $\boldsymbol{\gamma} = \boldsymbol{0}$ u 5.8 postoji mnogo predloženih procedura, od kojih se mi odlučujemo za test koji predlažu Grambsch i Therneau. Oni pokazuju u [8], da ako je model

$$h(t) = h_0(t)exp(\boldsymbol{\beta}(t)'\boldsymbol{x})$$

ispravan, no model 3.2 s konstantnim $\boldsymbol{\beta}$ je korišten, tada za $j = 1, \ldots, p$, vrijedi

$$E[r^*_{SC_{ij}} + \hat{\beta}_j] = \beta_j(t_i).$$

Predložena procedura je za svaki od p regresora dati grafički prikaz točaka $(t_i, r^*_{SC_{ij}})$ ili $(g(t_i), r^*_{SC_{ij}})$ za neku funkciju g. Ako je pretpostavka o proporcionalnosti rizika zadovoljena za X_j, tada će dodana zaglađena krivulja na dijagramu raspršenja biti aproksimativno vodoravna na razini $\hat{\beta}_j$. Ako postoji određeni trend, to nam sugerira da je efekt $j-$tog regresora vremenski promjenjiv. Statistike bazirane na $(t_i, r^*_{SC_{ij}})$ i testovi za testiranje hipoteze $\boldsymbol{\gamma} = \boldsymbol{0}$ mogu se pronaći u [7].

5.6 Skor reziduali

U ovom poglavlju uvesti ćemo metode kojima možemo provjeriti utjecaj pojedinog mjerenja na procjenu vektora parametara $\boldsymbol{\beta}$. Najsmisleniji pristup je najprije procijeniti vrijednost parametara $\boldsymbol{\beta}$ na temelju svih podataka, u oznaci $\hat{\boldsymbol{\beta}}$, te potom izračunati procjenu dobivenu u modelu bez $i-$tog mjerenja, u oznaci $\hat{\boldsymbol{\beta}}_{(i)}$. Ukoliko je $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ blisko nul-vektoru, tada $i-$to mjerenje ima mali utjecaj na procjenu, dok visoka odstupanja sugeriraju velik utjecaj. Ovakvim pristupom potrebno je izraditi $n+1$ Coxovih modela što se može pokazati zahtjevnim ukoliko je broj podataka n velik. Stoga se koristimo aproksimacijom baziranom na Coxovom modelu izgrađenom nad svim podacima. Aproksimacija se bazira na skor rezidualima. Definiramo $i-$ti skor rezidual, $i = 1, 2, \ldots, n$, za varijablu X_j, $j = 1, 2, \ldots, p$, kao

$$r_{S_{ji}} = \delta_i(x_{ji} - \hat{a}_{ji}) + exp\big(\hat{\boldsymbol{\beta}}'\boldsymbol{x_i}\big)\sum_{t_k \leq t_i} \delta_k \frac{\hat{a}_{jk} - x_{ji}}{\sum_{l \in R_{(t_k)}} exp\big(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_l\big)}, \tag{5.9}$$

gdje je $R(t_k)$ skup ispitanika pod rizikom u trenutku t_k, a x_{ji} vrijednost $j-$tog regresora za $i-$tog ispitanika. Označimo s $\boldsymbol{r}_{S_i}$ vektor skor reziduala za $i-$tog ispitanika, $\boldsymbol{r}'_{S_i} = (r_{S_{1i}}, r_{S_{2i}}, \ldots, r_{S_{pi}})$. Pokazuje se da je $j-$ta komponenta vektora

$$\boldsymbol{r}'_{S_i} Var(\hat{\boldsymbol{\beta}}) \tag{5.10}$$

dobra aproksimacija za $\hat{\beta}_j - \hat{\beta}_{j(i)}$, gdje je $Var(\hat{\boldsymbol{\beta}})$ matrica kovarijanci procjenitelja $\hat{\boldsymbol{\beta}}$. Komponente vektora $\boldsymbol{r}_{S_i}$, za $i = 1, 2, \ldots, n$, zovemo delta-beta veličine. $(i, j)-$tu delta-beta veličinu označavamo s $\Delta_i\hat{\beta}_j$. Dakle, vrijedi $\Delta_i\hat{\beta}_j \approx \hat{\beta}_j - \hat{\beta}_{j(i)}$. Premda je izraz 5.10 kompliciran, velika prednost koju dobivamo njegovim korištenjem jest dovoljnost izrade samo jednog Coxovog modela i to onoga sa svim podacima.

Korisni grafički prikazi bili bi dijagrami raspršenja veličina $\Delta_i\hat{\beta}_j$ u odnosu na redni broj mjerenja za svaku od varijabli X_j. Ukoliko $i-$to mjerenje ima veliki utjecaj na procjenu $j-$tog parametra, tada će u odnosu na preostala mjerenja, $\Delta_i\hat{\beta}_j$ biti veliko po apsolutnoj vrijednosti. Također, korisno je promatrati prikaz $\Delta_i\hat{\beta}_j$ u odnosu na rangove vremena doživljenja kako bi se dobila informacija o vezi između utjecaja i vremena doživljenja.

6 Istraživanje

6.1 Opis podataka

Istraživanje obuhvaća ispitanike koji su 2006. i 2007. godine operirani na Klinici za kirurgiju KBC-a Osijek, a nakon toga liječeni ili praćeni na Zavodu za onkologiju KBC-a Osijek. Kriterij po kojemu su pacijenti zadržani u studiji mogu se vidjeti u [6]. U konačnici, u bazi podataka imamo 236 pacijenata s operiranim karcinomom debelog crijeva. Od toga, ukupno 206 pacijenata je liječeno i na Zavodu za onkologiju KBC-a Osijek, a za 205 ispitanika imamo zabilježeno vrijeme doživljenja te stoga samo oni dolaze u obzir pri daljnjem modeliranju. Od tih 205 ispitanika, za njih 84 imamo zabilježeno vrijeme smrti, dok su preostalih 121 desno cenzurirana mjerenja. Za svakog od ispitanika bilježene su ukupno 34 različite karakteristike, pri čemu je prisutan velik broj nedostajućih vrijednosti. Metode kojima su dobivene vrijednosti pojednih varijabli ovdje nisu od presudnog značaja te stoga nisu navedene, no također mogu se pogledati u [6].

6.2 Cilj istraživanja i statističke metode

Cilj istraživanja je pronalazak optimalnog skupa varijabli za modeliranje funkcije rizika Coxovim regresijskim modelom proporcionalnog rizika. Pri opisu varijabli koristimo standardne metode deskriptivne statistike, kao što su tablice frekvencija i relativnih frekvencija za kategorijalne varijable, dok za numeričke varijable navodimo numeričke karakteristike kao što su aritmetička sredina, medijan, standardna devijacija, maksimum i minimum. Analizu preživljenja provodimo koristeći Coxov regresijski model proporcinalnog rizika, a adekvatne zaključke donosimo prema metodama obrađenim u teorijskom dijelu rada. Za potrebe statističke analize podataka korišten je programski paket R, biblioteka "survival" za analizu podataka o preživljenju (https://cran.r-project.org/web/packages/survival/survival.pdf).

6.3 Selekcija varijabli

Kao što smo već naveli, cilj istraživanja je procijeniti rizik smrti pacijenata s obzirom na birani skup varijabli. Neki od podataka u bazi su nedostajući. Od sveukupno 236 podataka, njih 205 ima zabilježeno vrijeme doživljenja te zbog prirode cilja samo tih 205 mjerenja sudjeluje u nastavku analize. U procesu modeliranja potrebno je pronaći one karakteristike koje pridonose boljem opisivanju i točnijem prediktiranju rizika smrti. Kako početna baza podataka sadrži velik broj varijabli, potrebno je taj broj smanjiti na one koje zaista nose informacije o riziku smrti. Neke od varijabli imaju velik broj nedostajućih vrijednosti te bi uključivanje takvih varijabli rezultiralo dodatnim smanjenjem uzorka, što je nepoželjan scenarij. Kao prvi kriterij za izbor potencijalnih varijabli u model promatramo broj nedostajućih vrijednosti svake od njih. Također, promatramo i smislenost uključivanja varijable u model. Varijable koje sadrže više od 15 nedostajućih vrijednosti automatski eliminiramo iz daljnje analize. Varijable koje preostaju nakon ovakve selekcije i koje dolaze kao u obzir za uključivanje u model su sljedeće: maksimalni udio tumorske strome, spol, dob, datum operacije, veličina tumora, broj pregledanih limfnih čvorova, broj zauzetih limfnih čvorova, udaljenost tumora od ruba resekcije, postojanje invazije krvnih i limfnih žila, postojanje presadnice, stadij bolesti, udio tumorske strome, grupacija tumora na one na lijevom te desnom kolonu, lokalizacija tumora, udio strome grupiran u dvije kategorije (manji ili jednak 0.5 i oni s udjelom tumorske strome većim od 0.5), vrijeme doživljenja, indikator smrtnosti i dubina prodora kroz stijenku crijeva. Od navedenih 18 varijabli, sve osim indikatora smrtnosti i

vremena doživljenja potencijalni su prediktori funkcije rizika. Kako bismo dobili uzorak koji nema nedostajućih vrijednosti, ipak je bilo potrebno izbaciti određena mjerenja. Uzorak nakon reduciranja sadrži 193 mjerenja, od kojih je 115 cenzurirano, a kod preostalih 78 je registrirana smrt. Daljnje istraživanje rađeno je na tom uzorku.

Uključivanje nepotrebnih varijabli u model umanjuje efikasnost modela, što se očituje u većim standardnim greškama, širim intervalima pouzdanosti te većim p-vrijednostima. Od preostalih kandidata potrebno je pronaći podskup varijabli koji će najefikasnije opisati rizik smrti. Odnosno, želimo pronaći najmanji skup prediktora kojim su dobro obuhvaćene informacije o riziku smrti. Kao idući korak pri odabiru regresora koristimo se postupkom selekcije navedenim u poglavlju 4.3. Postupak sugerira korištenje maksimalne vrijednosti tumorske strome, spola, dobi, stadija i udaljenosti tumora od ruba resekcije za modeliranje funkcije rizika. U nastavku pod puni model smatramo onaj dobiven na temelju navedenih 16 varijabli, a reducirani model onaj dobiven s prethodno navedenih 5. Primijetimo, reducirani model je ugnježđen u puni model.

Tablica 2 prikazuje usporedbu punog i reduciranog modela. S χ^2 u tablici je označena statistika log-omjera vjerodostojnosti definirana u poglavlju 4.2. Ukoliko su modeli jednako dobri, tada statistika log-omjera vjerodostojnosti ima χ^2_7 distribuciju. Visoka p-vrijednost od 0.712 svjedoči tome da reducirani model nije statistički značajno lošiji od punog modela, odnosno koristeći reducirani model ne gubimo značajno na točnosti.

Model	$-2\hat{l}$	χ^2	Stupanj slobode	$P(>\|\chi\|)$
Puni model	-307.78			
Reducirani model	-310.07	4.569	7	0.712

Tablica 2: Usporedba modela

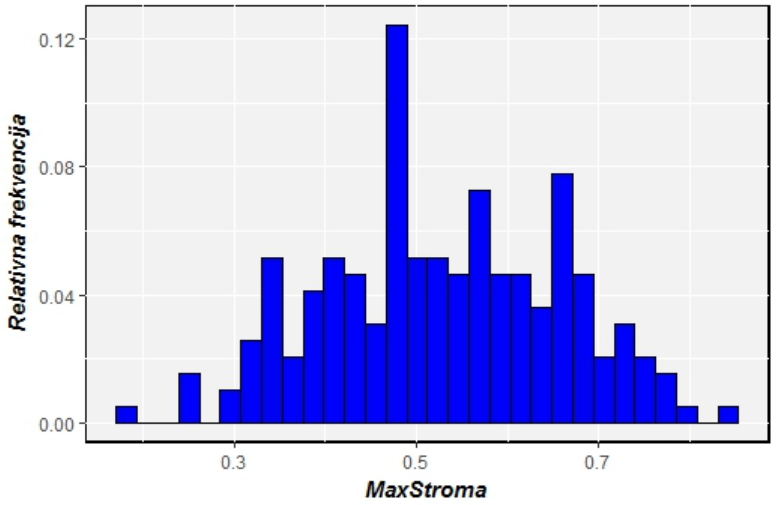
Model dobiven na temelju navedenih varijabli ima svojstvo da ga dodavanje bilo koje od preostalih varijabli ne poboljšava značajno, ali izbacivanje bilo koje od navedenih varijabli će rezultirati statistički značajno lošijim modelom. Ovakav odabir modela još nije konačan te će u nastavku biti predložene određene izmjene pri načinu na koji varijable ulaze u model. Ono što u ovom trenutku znamo, jest da će sve te varijable na neki način biti korištene pri modeliranju funkcije hazarda. U nastavku navedene varijable detaljnije opisujemo.

6.4 Opis varijabli

MaxStroma je numerička varijabla koja predstavlja maksimalan udio tumorske strome. Kako je ona definirana kao postotak, njene vrijednosti variraju između 0 i 1. Tablica 3 prikazuje vrijednosti njenih numeričkih karakteristika. Slika 3 prikazuje stupčasti dijagram relativnih frekvencija varijable MaxStroma. On nam daje dodatan uvid u distribuciju varijable.

Minimum	Medijan	Prosjek	Maksimum	Standardna devijacija
0.180	0.530	0.523	0.840	0.129

Tablica 3: Numeričke karakteristike varijable MaxStroma



Slika 3: Stupčasti dijagram relativnih frekvencija varijable MaxStroma

Varijablu Spol nije potrebno posebno pojašnjavati. Tablica 4 prikazuje frekvencije i relativne frekvencije spola ispitanika.

	Muškarci (1)	Žene (2)	Ukupno
Frekvencija	110	83	193
Postotak	56.995	43.005	100.000

Tablica 4: Tablica frekvencija varijable Spol

Varijabla Stadij je kategorijalna, a kategorije predstavljaju stadij napretka bolesti te poprimaju vrijednosti 1,2,3,4. Frekvencije i relativne frekvencije ispitanika prema stadiju bolesti dane su u tablici 5.

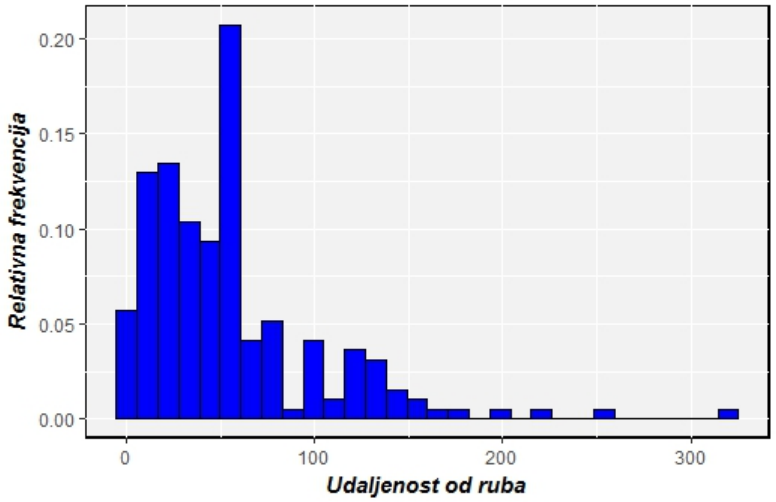
	Stadij 1	Stadij 2	Stadij 3	Stadij 4	Ukupno
Frekvencija	18	55	84	36	193
Postotak	9.326	28.497	43.523	18.653	100.000

Tablica 5: Tablica frekvencija varijable Stadij

Varijabla udOdRub je numeričkog tipa, a predstavlja udaljenost tumora od ruba resekcije. Uvid u distribuciju varijable dobivamo iz tablice 6 i slike 4.

Minimum	Medijan	Prosjek	Maksimum	Standardna devijacija
0.000	40.000	54.400	320.000	48.151

Tablica 6: Numeričke karakteristike varijable udOdRub

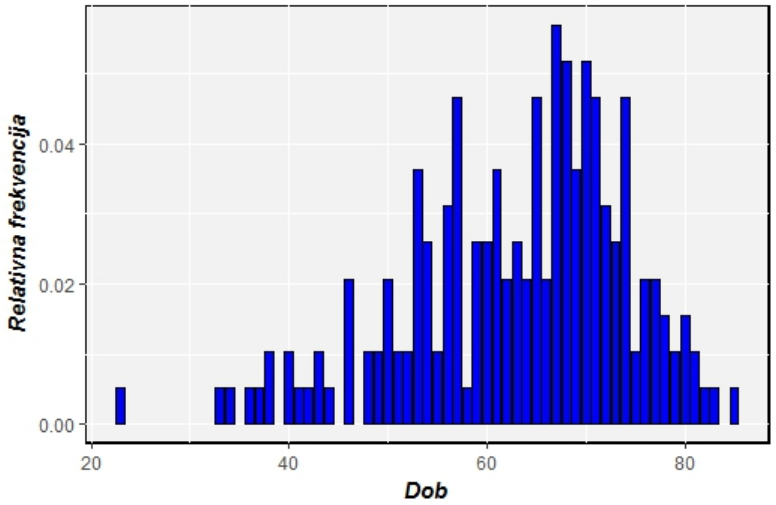


Slika 4: Stupčasti dijagram relativnih frekvencija varijable udOdRub

Varijabla Dob numeričkog je tipa i predstavlja starost ispitanika u godinama. Informacije o distribuciji dobi pacijenata vidimo iz tablice 7 te histograma relativnih frekvencija danim slikom 5.

Minimum	Medijan	Prosjek	Maksimum	Standardna devijacija
23.000	65.000	63.007	85.000	11.172

Tablica 7: Numeričke karakteristike varijable Dob

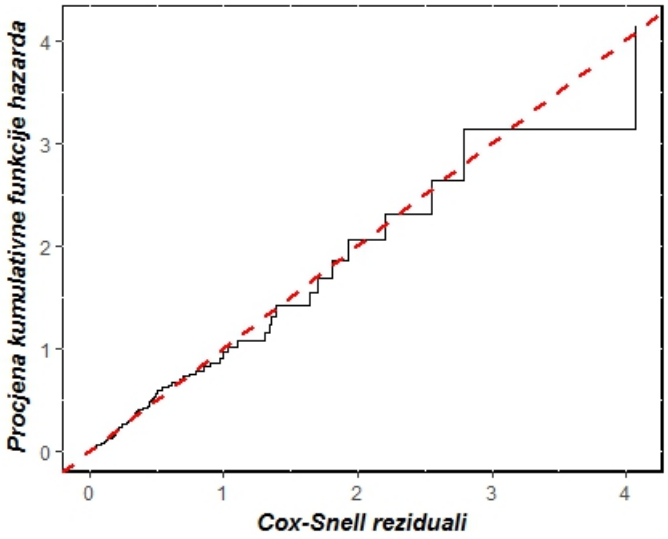


Slika 5: Stupčasti dijagram relativnih frekvencija varijable Dob

U nastavku nas zanima zadovoljava li navedeni model pretpostavku proporcionalnog rizika, jesu li varijable iskorištene na najoptimalniji način te postoje li stršeća ili utjecajna mjerenja unutar modela.

6.5 Dijagnostika modela

Nakon što smo došli do potencijalnog modela i opisali varijable koje u njega ulaze, na red dolazi ispitivanje pretpostavki na kojima počiva Coxov regresijski model proporcionalnog hazarda. Najprije nas zanima zadovoljava li naš model pretpostavku proporcionalnosti rizika. Za provjeru se služimo metodama iz poglavlja 5.1, 5.5. Za dani model računamo Cox-Snell reziduale i provjeravamo prate li oni zaista eksponencijalnu distribuciju s parametrom 1. Ukoliko se pokaže da to nije istina, tada postoji sumnja u pretpostavci proporcionalnosti hazarda.



Slika 6: Cox-Snell reziduali

Sumnja o odstupanjima od eksponencijalne distribucije postoji ukoliko bi grafički prikaz Cox-Snell reziduala u odnosu na vrijednost procijenjene funkcije kumulativnog hazarda odstupao od pravca koji $y = x$. Na osnovu slike 6 nema razloga sumnjati u proporcionalnost hazarda danog modela.

Kao dodatnu provjeru proporcionalnosti rizika koristimo se testom spomenutim u poglavlju 5.5, a baziranom na skaliranim Schoenfeld rezidualima.

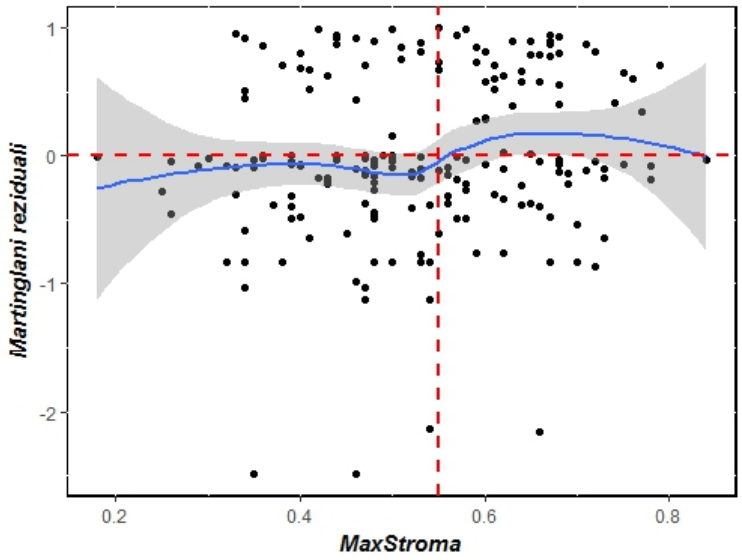
	ρ	χ^2	$P(>\|\chi\|)$
MaxStroma	0.2098	3.9270	0.0475*
Dob	0.0270	0.0702	0.7911
Spol2	-0.0919	0.6773	0.4105
udOdRub	0.0172	0.0263	0.8172
Stadij2	-0.1172	1.0382	0.3082
Stadij3	-0.1252	1.1832	0.2767
Stadij4	-0.1234	1.1398	0.2857
Ukupno	NA	6.2824	0.5072

Tablica 8: Rezultati testa o proporcionalnosti rizika modela

Iz tablice 8 vidimo da globalno možemo smatrati kako model zadovoljava pretpostavku proporcionalnosti. Za koeficijent uz varijablu MaxStroma na razini značajnosti 0.05 odba-

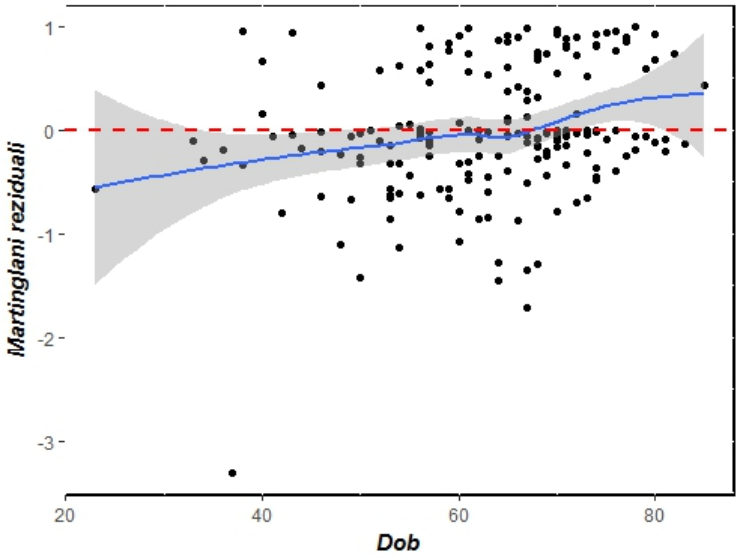
cujemo nul-hipotezu o nepostojanju linearne veze koeficijenta s vremenom. Odnosno, na razini značajnosti 0.05 smatramo kako je koeficijent uz MaxStroma u linearnoj vezi s vremenom. Taj rezultat donosi određenu nesigurnost u adekvatnost našeg modela. Situaciju ćemo pokušati popraviti u nastavku.

Dosad smo ustanovili koje varijable ćemo koristiti za modeliranje funkcije rizika, no nismo odlučili hoće li varijable u model ući u trenutnoj formi. To se isključivo odnosi na numeričke varijable MaxStroma, Dob i udOdRub. Za pronalazak optimalne forme numeričkih varijabli koristimo se martingalnim rezidualima, na način opisan u poglavlju 5.2. Najprije promatramo model u kome su uključene samo varijable Stadij i Spol te na temelju tog modela računamo martingalne reziduale, koje potom stavljamo u odnos s vrijednošću maksimalne strome ispitanika. Grafički prikaz tog odnosa vidimo na slici 7.



Slika 7: Martingalni reziduali u odnosu na MaxStroma

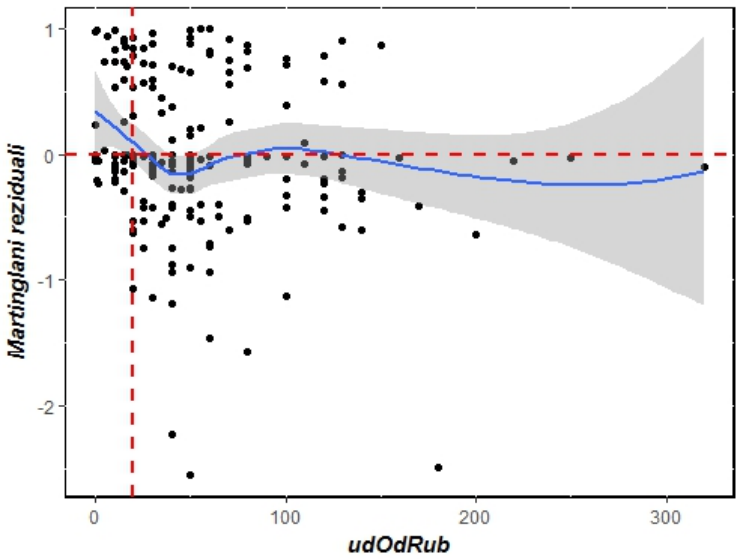
Loess krivulja nam sugerira kategorizaciju varijable MaxStroma u dvije skupine. Za graničnu vrijednost uzimamo 0.55. Definiramo novu varijablu MaxStroma_grup koja poprima vrijednost 0 ako je vrijednost MaxStrome manja od 0.55, inače poprima vrijednost 1. Sada gradimo novi model koji sadrži varijable Stadij, Spol i Maxstroma_grup. Zanima nas na koji način treba varijablu Dob uključiti u model. Prikaz martingalnih reziduala u odnosu na dob ispitanika dan je slikom 8.



Slika 8: Martingalni reziduali u odnosu na Dob

Krivulja dobivena loess procedurom linearnog je karaktera što nam sugerira da varijablu Dob prilikom uključenja nije potrebno transformirati.

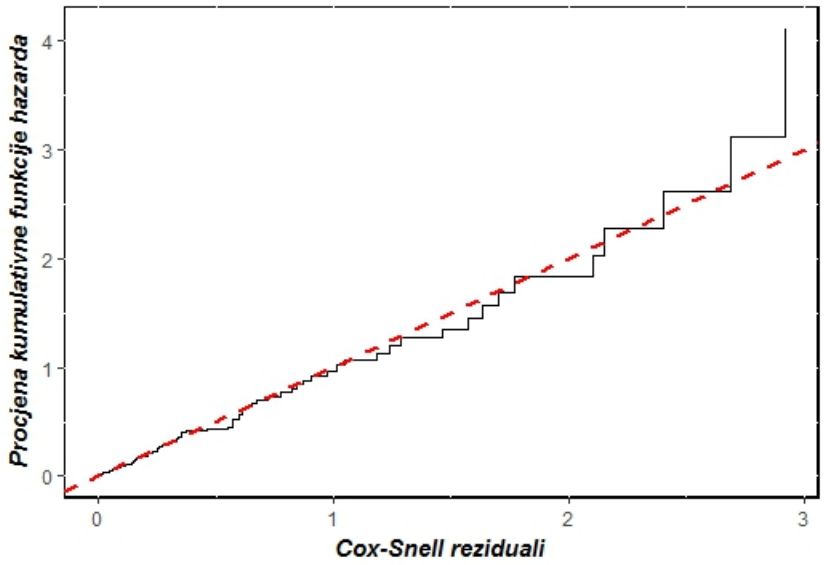
Na kraju, potrebno je još istražiti idealnu funkcionalnu formu varijable udOdRub. Opet gradimo novi model koji sada uključuje varijable Stadij, Spol, Maxstroma_grup i Dob. Slika 9 prikazuje martingalne reziduale u odnosu na udaljenost tumora od ruba resekcije.



Slika 9: Martingalni reziduali u odnosu na udOdRub

Iz slike nije najjasnije što učiniti po pitanju varijable udOdRub. Mogli bismo ju ostaviti u istom obliku ili pokušati diskretizirati. Mi se odlučujemo na diskretizaciju i to s dvije kategorije. Stvaramo varijablu udOdRub_grup koja za ispitanika poprima vrijednost 0 ukoliko je pripadna vrijednost udOdRub manja od 20, inače vrijednost 1. Na ovakav odabir smo se odlučili jer nam on smiruje martingalne reziduale, popravlja situaciju iz testa proporcionalnosti rizika, daje bolje vrijednosti AIC-a i statistike logaritma vjerodostojnosti.

U ovome trenutku smo došli do našeg krajnjeg modela. Model funkciju rizika opisuje na temelju varijabli Spol, Stadij, Dob, MaxStroma_grup, udOdRub_grup. Za provjeru proporcionalnosti rizika u modelu slikom 10 dan je prikaz Cox-Snell reziduala u odnosu na procjenu kumulativne funkcije hazarda. Također, tablicom 9 prikazujemo rezultate testa o proporcionalnosti rizika. Model prolazi kao valjan u obje provjere.



Slika 10: Cox-Snell reziduali

	ρ	χ^2	$P(>\|\chi\|)$
MaxStroma_grup1	0.1906	2.9165	0.0877.
Dob	0.0276	0.0752	0.7840
Spol2	-0.0765	0.5076	0.4762
udOdRub_grup1	-0.0317	0.0834	0.7727
Stadij2	-0.1136	0.9341	0.3338
Stadij3	-0.0939	0.6417	0.4231
Stadij4	-0.0948	0.6478	0.4209
Ukupno	NA	5.2511	0.6293

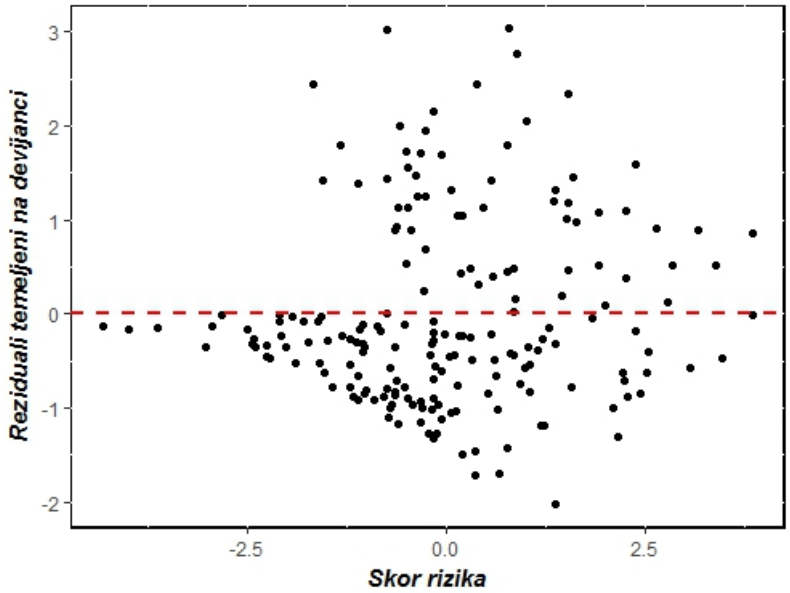
Tablica 9: Rezultati testa o proporcionalnosti rizika konačnog modela

Uočimo, naš konačni model ima 7 parametara. To je jednako broju parametara modela prije diskretizacije, što je dobro jer model nismo dodatno zakompicirali, a pokazuje se boljim od prvotnog modela prema mnogim kriterijima. Ima manji AIC, daje veću sigurnost u validnost pretpostavke proporcionalnog rizika te varijable uključene na ovaj način u model poprimaju niže $p-$vrijednosti nego u prvotnom modelu. Sve su to razlozi zbog kojih njega uzimamo kao konačni model.

6.6 Stršeća i utjecajna mjerenja

Prije nego krenemo s interpretacijom koeficijenata i donošenjem zaključaka, potrebno je istražiti postoje li stršeća te utjecajna mjerenja u modelu.

Metode kojima provjeravamo je li mjerenje stršeće koriste reziduale temeljene na devijanci. Mjerenje možemo smatrati stršećim ukoliko je pripadni rezidual po apsolutnoj vrijednosti veći od 2.5 [13]. Za takva mjerenja ne možemo očekivati dobru predikciju modelom.



Slika 11: Reziduali temeljeni na devijanci u odnosu na skor rizika

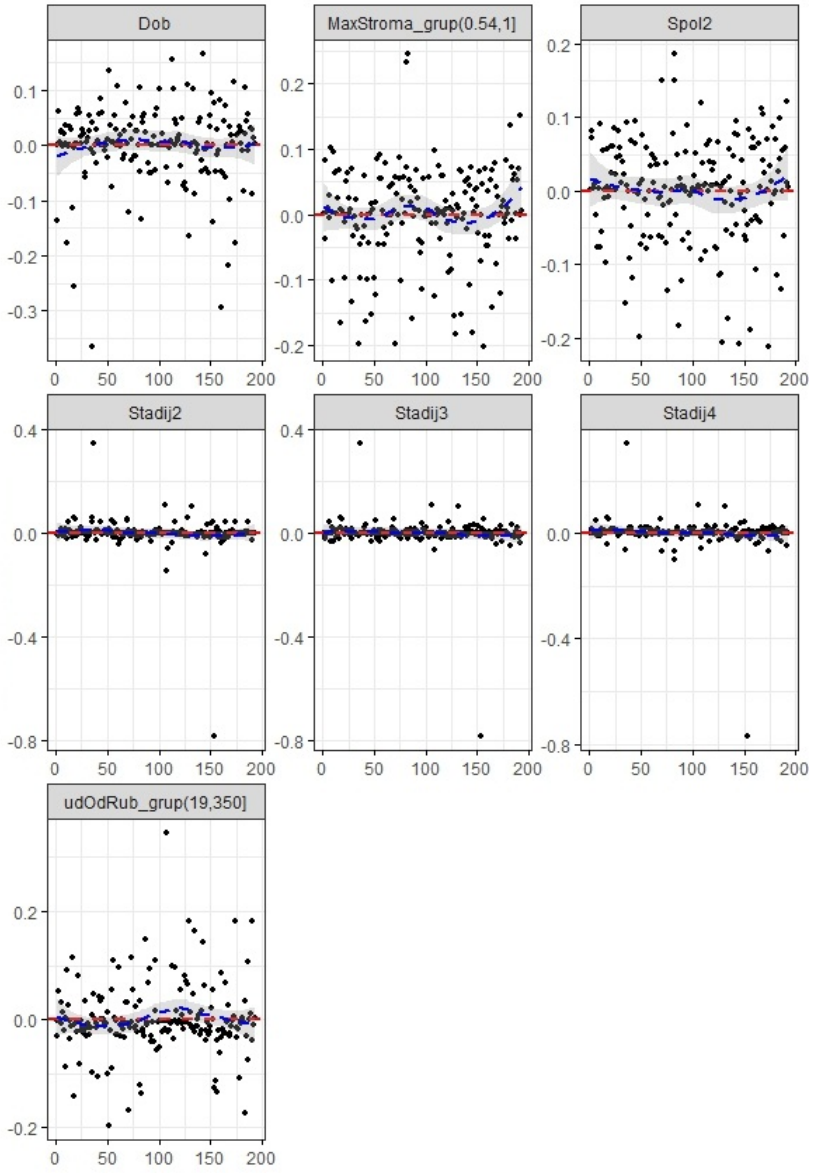
Na slici 11 vidimo nekoliko mjerenja koja možemo smatrati stršećim. U tablici 10 dane su karakteristike triju ispitanika s najvećim rezidualima po apsolutnoj vrijednosti.

Redni broj	MaxStroma_grup	Spol	Dob	Stadij	udOdRub_grup	VrDozivljenja	ICSmrt
122	1	2	78	2	1	46	1
180	1	1	61	2	1	100	1
135	0	1	56	4	1	23	1

Tablica 10: Potencijalna stršeća mjerenja

Mjerenja s rednim brojem 122 i 180 su ispitanici, koji u usporedbi s ispitanicima sličnih vrijednosti karakteristika, umiru znatno ranije. Ispitanik 135 je neobično mjerenje iz razloga što je on unutar skupa njemu sličnih ispitanika najmlađi, a umire prvi.

Preostaje ispitati postoje li unutar modela neka utjecajna mjerena, odnosno mjerenja čije prisustvo znatno mijenja koeficijente uz pojedine varijable. Metoda kojom pronalazimo utjecajna mjerenja oslanja se na delta-beta reziduale definirane u poglavlju 5.4. Slika 12 prikazuje aproksimirane utjecaje pojedinih mjerenja na procjene koeficijenata uz varijable. Tražimo ona mjerenja čije su apsolutne vrijednosti delta-beta veličina značajno veće od preostalih. Navedimo četiri mjerenja koja se najviše ističu, od toga dva mjerenja iz prikaza varijabli Stadij (160. i 37. mjerenje), jedno iz prikaza varijable udOdRub_grup (112. mjerenje) i jedno iz prikaza varijable Dob (35. mjerenje). Tablica 11 prikazuje vrijednosti karakteristika navedenih mjerenja.



Slika 12: Delta-beta reziduali za sve koeficijente u modelu

Redni broj	MaxStroma_grup	Spol	Dob	Stadij	udOdRub_grup	VrDozivljenja	ICSmrt
160	0	1	75	1	0	2380	1
37	1	1	70	1	0	2951	0
112	1	1	67	2	0	3287	1
35	0	1	38	2	0	532	1

Tablica 11: Utjecajna mjerenja

Ispitanik 160 pokazuje se kao izrazito utjecajno mjerenje za koeficijente uz varijable Stadij2, Stadij3, Stadij4. Opravdanje toga jest činjenica što je taj ispitnik jedini sa stadijom 1 kojemu je registrirana smrt. Stoga, izbacivanje ovog mjerenja iz baze rezultira eksplozijom

koeficijenata uz navedene varijable. Ispitanik 112 ima velik utjecaj na koeficijent uz udOdRub_grup jer je neobično da se smrt dogodi nakon toliko vremena. Nakon 2800 dana u studiji imamo zabilježenu samo jednu smrt od njih 17. Od tih 17, tek tri spadaju u grupu udOdRub[0,19] od kojih je jedno upravo mjerenje 112. Ispitanik 35 ima veliki utjecaj na procjenu koeficijenta uz varijablu Dob, iz razloga što je to izrazito mlada osoba koja umire relativno brzo nakon operacije, u odnosu na njemu slične ispitanike.

6.7 Interpretacija

Osnovne informacije o konačnom modelu vidimo u tablici 12.

	Koeficijent	Standardna greška	z	$Pr(>\|z\|)$
Spol2	0.72982	0.24390	2.992	0.002769**
MaxStroma_grup1	0.77990	0.23909	3.262	0.001106**
Dob	0.05300	0.01286	4.121	3.78e-05***
udOdRub_grup1	-1.06475	0.28134	-3.784	0.000154***
Stadij2	1.83795	1.03843	1.770	0.076739.
Stadij3	2.35210	1.03036	2.283	0.022442*
Stadij4	4.41852	1.04380	4.233	2.30e-05***
	exp(Koef)	exp(-Koef)	Donja granica	Gornja granica
Spol2	2.0747	0.48200	1.2863	3.3463
MaxStroma_grup1	2.1813	0.45845	1.3652	3.4851
Dob	1.0544	0.94838	1.0282	1.0814
udodrub_grup1	0.3448	2.90011	0.1987	0.5985
Stadij2	6.2836	0.15914	0.8209	48.0980
Stadij3	10.5076	0.09517	1.3946	79.1677
Stadij4	82.9732	0.01205	10.7264	641.8325

Tablica 12: Koeficijenti modela

Model je dan sljedećom formulom

$$\frac{h(t)}{h_0(t)} = exp\big(0.730Spol2 + 0.780MaxStroma_grup1 + 0.053Dob - 1.065udOdRub_grup1 + 1.838Stadij2 + 2.352Stadij3 + 4.419Stadij4\big). \tag{6.1}$$

Prilikom interpretacije modela oslanjamo se na eksponencirane vrijednosti koeficijenata. Interpretacija koeficijenata:

U prosjeku, uz iste vrijednosti preostalih varijabli, muškarci imaju 2.075 puta veći rizik smrti u odnosu na žene.

U prosjeku, uz iste vrijednosti preostalih varijabli, ispitanici s maksimalnim udjelom strome većim od 0.54 imaju 2.181 puta veći rizik smrti u odnosu na ispitanike čiji je udio manji ili jednak 0.54.

U prosjeku, uz iste vrijednosti preostalih varijabli, rizik smrti ispitanika u stadiju 2 je 6.284 puta veći u odnosu na ispitanike u stadiju 1.

U prosjeku, uz iste vrijednosti preostalih varijabli, rizik smrti ispitanika u stadiju 3 je 10.508 puta veći u odnosu na ispitanike u stadiju 1.

U prosjeku, uz iste vrijednosti preostalih varijabli, rizik smrti ispitanika u stadiju 4 je 82.973 puta veći u odnosu na ispitanike u stadiju 1.

U prosjeku je, uz iste vrijednosti preostalih varijabli, rizik smrti ispitanika čija je udaljenost tumora od ruba resekcije manja 20mm 2.900 puta veći u odnosu na ispitanike s udaljenošću većom ili jednakom 20mm.

U prosjeku, za ispitanike s ostalim varijablama na istim razinama, svaka dodatna godina starosti povezana je s 1.054 puta većim rizikom smrti.

Literatura

[1] A. H. S. Ang, W. H. Tang, *Probability concepts in engineering : emphasis on applications to civil and environmental engineering*, Hoboken, Wiley, 2007.

[2] M. Benšić, N. Šuvak, *Uvod u vjerojatnost i statistiku*, Sveučilište J.J. Strossmayera, Odjel za matematiku, Osijek, 2014. (javno dostupno na: `https://www.mathos.unios.hr/uvis/UVIS_knjiga_final/UVIS_knjiga_web.pdf`)

[3] W. S. Cleveland,*Robust Locally Weighted Regression and Smoothing Scatterplots*, Journal of the American Statistical Association Vol. 74, No. 368 (Dec., 1979.), pp. 829-836, (javno dostupno na: `http://www.stat.washington.edu/courses/stat527/s13/readings/Cleveland_JASA_1979.pdf`)

[4] D. Collett, *Modelling Survival Data in Medical Research*, Chapman and Hall/CRC, 2003.

[5] D. R. Cox, *Regression Models and Life-Tables*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, No. 2. (1972.), pp. 187-220, (javno dostupno na: `http://www.stat.cmu.edu/~ryantibs/journalclub/cox_1972.pdf`).

[6] J. Flam, *Udio strome kao prognostički čimbenik kod karcinoma debelog crijeva*, Doktorska disertacija, Osijek, 2016., (javno dostupno na: `https://tinyurl.com/yc3ov37v`).

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

[8] P. M. Grambsch, T. M. Therneau, *Proportional hazards tests and diagnostics based on weighted residuals*, Biometrika, 81:515- 526, 1994. (javno dostupno na: `http://escarela.com/archivo/anahuac/03o/coxdiag.pdf`)

[9] J. P. Klein, M. L. Moeschberger,*Survival analysis Techniques for censored and truncated data, Second Edition*, Springer-Verlag, New York, 2003.

[10] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Inc., Hoboken, New Yersey, 2003.

[11] D. Machin, Y. B. Cheung, M. Parmar,*Survival Analysis: A Practical Approach, Second Edition*, John Wiley & Sons, 2006.

[12] D. Schoenfeld, *Partial Residuals for The Proportional Hazards Regression Model*, Biometrika, Vol. 69, No. 1. (Apr., 1982), pp. 239-241 (javno dostupno na: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.544.6796&rep=rep1&type=pdf`)

[13] R. L. Ting, J. A. Fitrianto, *Several Types of Residuals in Cox Regression*, Int. Journal of Math. Analysis, Vol. 7(2013.), no. 53, 2645 - 2654, (javno dostupno na: `http://www.m-hikari.com/ijma/ijma-2013/ijma-53-56-2013/fitriantoIJMA53-56-2013.pdf`)

[14] T. M. Therneau, P. M. Grambsch, T. Fleming *Martingale based residuals for survival models*, Technical Report Series 840, (javno dostupno na: `http://www.mayo.edu/research/documents/40/doc-20309106`)

Sažetak. U radu uvodimo osnovne pojmove analize preživljenja kao što su cenzuriranje, funkcija preživljenja, funkcija rizika itd.
Upoznajemo se s Kaplan-Meier procjeniteljem funkcije preživljenja te s log-rang i Wilcoxonovim testom kao metodama usporedbe dvaju ili više funkcija preživljenja.
Detaljno obrađujemo teorijsku pozadinu Coxovog regresijskog modela, definiramo procedure za odabir varijabli u model te uvodimo različite definicije reziduala kojima se služimo u dijagnostici modela.
U konačnici, koristeći se principima definiranim u teorijskom dijelu, gradimo Coxov regresijski model preživljenja nad ispitanicima s operiranim karcinomom debelog crijeva.

Ključne riječi: Analiza preživljenja, cenzuriranje, funkcija preživljenja, funkcija rizika, Kaplan-Meier procjenitelj, log-rang test, Coxov regresijski model proporcionalnog rizika.

Abstract. The paper introduces basic concepts of survival analysis such as censoring, survival function, hazard function etc.
We define the Kaplan-Meier estimator of the survival function and two nonparametric procedures for comparing two or more survival functions, log-rank test, and Wilcoxon test.
After the theoretical background for Cox regression model is developed, we define procedures for selecting variables in the model and introducing different definitions of residuals used in model diagnosis.
Ultimately, using the principles defined in the theoretical part, we build Cox's regression model for subjects with colon cancer surgery.

Key words: Survival analysis, censoring, survival function, hazard function, Kaplan-Meier estimator, log-rank test, Cox regression.

Životopis

Rođen sam 9.11.1992. u Osijeku, Hrvatska. 2007. godine završavam osnovnoškolsko obrazovanje u Semeljcima tijekom kojega redovito sudjelujem na općinskim, županijskim, regionalnim te državnim natjecanjima iz matematike. Iste godine upisujem III. gimnaziju u Osijeku. 2011. godine upisujem preddiplomski studij matematike na Odjelu za matematiku Sveučilišta Josipa Jurja Strossmayera u Osijeku. Tri godine kasnije uspješno završavam preddiplomski studij uz završni rad Eulerov teorem pod mentorstvom doc. dr. sc. Ivana Matića. U listopadu 2014. godine upisujem diplomski studij matematike, smjer financijska matematika i statistika. Krajem 2015. godine diplomskog studija odrađujem stručnu studentsku praksu u tvrtki Farmeron d.o.o. Suradnja s tvrtkom Farmeron i kolegom Miroslavom Jankovićem rezultira seminarskim radom Woodov i Milkbot laktacijski modeli koji je nagrađen Rektorovom nagradom.