Logistička regresija i primjene

Švoger, Marina

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Marina Švoger

LOGISTIČKA REGRESIJA I PRIMJENE

Diplomski rad

Voditelj rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, Veljača, 2022.

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

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

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Tati.

Sadržaj

Uvod

Jedna od najčešće korištenih statističkih metoda je regresijska analiza. Njome se želi odrediti odnos između jedne ili više varijabli odziva i jedne ili više varijabli poticaja. Ako varijable odziva mogu poprimiti konačan broj različitih vrijednosti, radi se o logističkoj regresiji. Odličan opis svih modela pa time i modela logističke regresije dao je britanski statističar George E. P. Box koji je rekao[1]: „...svi modeli su krivi: pitanje u praksi je koliko oni moraju biti krivi kako ne bi bili korisni." Dakle, nije pravo pitanje: „Je li model točan?" nego „Je li model dovoljno dobar da ga ima smisla koristiti u konkretnom slučaju?" Kako bismo dobili odgovor na to pitanje provodimo statističku analizu, ali i logički razmišljamo o tome koje varijable su od tolikog značaja da bi se svakako trebale nalaziti u modelu.

Logistička regresija ima brojne primjene u raznim područjima kao što su recimo medicina, ekonomija i inženjerstvo. U medicini se često koristi kako bi se utvrdila vjerojatnost da se razvije bolest ili da se dogodi smrtni ishod pod uvjetom da su nam dani podaci o određenim obilježjima kao što su recimo dob i spol pacijenta, njegove životne navike, druge eventualno povezane bolesti, rezultati nekih krvnih pretraga i slično. Konkretno, u ovom radu ćemo analizirati vjerojatnost da se dogodi smrt uzrokovana srčanim zatajenjem kod osoba koje su već imale jedno ili više težih srčanih zatajenja te vjerojatnost preživljavanja kod osoba koje imaju najčešći tip karcinoma jetre.

[1] Izjava u izvornom obliku na engleskom jeziku može se pronaći u [7].

Poglavlje 1

Logistička regresija

1.1 Regresijska analiza i linearna regresija

Regresijska analiza je skup postupaka i metoda pomoću kojih se nastoji ispitati i analizirati ovisnost jedne ili više zavisnih varijabli o jednoj ili više nezavisnih varijabli. Zavisne varijable koje želimo opisati ili procijeniti označavamo s $Y_1, Y_2, ..., Y_q$ i zovemo još i varijablama odziva ili odgovora dok nezavisne varijable označavamo s $x_1, x_2, ..., x_p$ i zovemo još varijablama poticaja ili kovarijatama. Ako imamo jednu varijablu poticaja tada govorimo o univarijabilnoj, a ako ih imamo više o multiploj ili multivarijabilnoj regresiji dok ako imamo jednu varijablu odziva govorimo o univarijatnoj, a ako ih imamo više o multivarijatnoj regresiji, no u ovom radu ćemo se baviti samo univarijatnom regresijom. [41]

Linearna regresija je vrsta regresijske analize u kojoj je povezanost između varijable odziva i varijabli poticaja opisana linearnom funkcijom. Ako imamo dodatna opažanja nezavisne varijable, model dobiven regresijskom analizom opaženih vrijednosti zavisne i nezavisnih varijabli možemo koristiti za procjenu odnosno predviđanje vrijednosti zavisne varijable. Također, u primjenama je vrlo važna činjenica da procijenjeni parametri govore koliko će se promijeniti vrijednost zavisne varijable ako se vrijednost jedne ili više nezavisnih varijabli promijeni za jednu jedinicu. [36]

Linearni regresijski model u slučaju jedne zavisne varijable ima sljedeći oblik:

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \epsilon$$

pri čemu je ϵ slučajna greška, a β_k, $k = 1, ..., p$ su parametri modela.

Za varijablu Y imamo slučajni uzorak duljine n što možemo zapisati u obliku

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \epsilon_i, \qquad i = 1, ..., n$$

pri čemu je x_{ki}, $k = 1, ..., p$, $i = 1, ..., n$ i-to opažanje k-te nezavisne varijable x_k, a ϵ_i, $i = 1, ..., n$ su slučajne greške.

U matričnom obliku to zapisujemo kao

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \epsilon$$

pri čemu je $\boldsymbol{Y} = (Y_1, ..., Y_n)^\tau$, $\boldsymbol{b} = (\beta_0, ..., \beta_p)^\tau$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^\tau$ i

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & ... & x_{np} \end{bmatrix}, \tag{1.1}$$

a $\boldsymbol{X}$ se još naziva i matricom dizajna. [28]

Osnovne pretpostavke koje moraju biti zadovoljene u modelu linearne regresije su:

1) veza između varijabli odziva i poticaja mora biti linearna
2) slučajne greške su međusobno nezavisne i normalno distribuirane s očekivanjem 0 i jednakom varijancom
3) opažanja su međusobno nezavisna.

Ako jedna ili više pretpostavki nisu zadovoljene, model koji dobijemo linearnom regresijom ne mora biti valjan. [28], [36]

Parametre modela najčešće procjenjujemo metodom najmanjih kvadrata pri čemu želimo minimizirati kvadratnu normu greške tj. želimo minimizirati $\|\epsilon\|_2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|_2$ po $\boldsymbol{b}$. Dobivamo da je procjenitelj dobiven metodom najmanjih kvadrata za $\boldsymbol{b}$

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}^\tau \boldsymbol{X})^{-1} \boldsymbol{X}^\tau \boldsymbol{Y}$$

uz uvjet da je $\boldsymbol{X}^\tau \boldsymbol{X}$ regularna. [28]

1.2 Razlike linearne i logističke regresije

Za razliku od linearne regresije gdje je varijabla odziva neprekidna, u praksi se često događa da ona može poprimiti konačan broj vrijednosti te tada govorimo o logističkoj regresiji. Ako zavisna varijabla može poprimiti samo dvije vrijednosti govorimo o binarnoj

ili dihotomnoj varijabli. Primjer takve varijable je varijabla *događaj* koja označava je li se neki događaj dogodio ili nije i ima Bernoullijevu razdiobu s parametrom p koji označava vjerojatnost da se taj događaj dogodio.

U primjenama je najčešće zavisna varijabla kategorijska, ali takva da može poprimiti više od dvije vrijednosti. Te vrijednosti mogu biti na neki način uređene pa tada govorimo o ordinalnim varijablama, ali i ne moraju biti pa takve varijable zovemo nominalnima. Primjer ordinalne varijable je varijabla *dobna kategorija* koja označava kojoj dobnoj kategoriji pripada osoba koja sudjeluje u nekom istraživanju pri čemu su moguće recimo sljedeće kategorije: $0-18$, $19-35$, $36-55$, $55-75$ i >75. Primjer nominalne varijable je npr. varijabla *državljanstvo* koja označava državljanstvo osobe.

Kod logističke regresije ne mora vrijediti pretpostavka o normalnoj distribuiranosti grešaka koja je morala biti zadovoljena kod linearne regresije. Ako imamo nezavisnu varijablu označenu s x te dihotomnu zavisnu varijablu koja može poprimiti vrijednosti $y = 0$ i $y = 1$ te ako označimo uvjetno očekivanje od Y uz dani x u oznaci $E(Y|x)$ s $\pi(Y|x)$ kao što je uobičajeno kad govorimo o logističkoj regresiji, tada vrijednost varijable Y uz dani x možemo izraziti kao

$$y = \pi(x) + \epsilon.$$

Tada ϵ može poprimiti samo dvije vrijednosti. Naime, ako je $y = 1$ tada je $\epsilon = 1 - \pi(x)$ s vjerojatnosti $\pi(x)$, a ako je $y = 0$ tada je $\epsilon = -\pi(x)$ s vjerojatnosti $1 - \pi(x)$. Dakle, greške imaju binomnu razdiobu s očekivanjem 0 i varijancom $\pi(x)[1 - \pi(x)]$. [24]

Nadalje, procjene parametara koje dobijemo u modelu više ne označavaju koliko se promijenila vrijednost zavisne varijable ako se vrijednost jedne ili više nezavisnih varijabli povećala za jednu jedinicu. Više govora o tome kako interpretiramo parametar β_1 na primjeru univarijabilne logističke regresije bit će u pododjeljku 1.6.1.

Važno je napomenuti i da se parametri modela kod logističke regresije najčešće procjenjuju metodom maksimalne vjerodostojnosti, a ne metodom najmanjih kvadrata kao kod linearne regresije jer se gube određena poželjna statistička svojstva kad imamo dihotomnu zavisnu varijablu. [24]

1.3 Model logističke regresije i logit funkcija

U svakoj regresijskoj analizi ključna vrijednost koja nas zanima je očekivana vrijednost zavisne varijable Y ako nam je dana vrijednost nezavisne varijable x. Kod linearne regresije pretpostavljamo da ta vrijednost može biti izražena kao linearna jednadžba po varijabli x

oblika $E(Y|x) = \beta_0 + \beta_1 x$. S obzirom da x poprima vrijednosti između $-\infty$ i $+\infty$, očito $E(Y|x)$ također može poprimiti bilo koju vrijednost.

Ako pretpostavimo da imamo po jednu zavisnu i nezavisnu varijablu, koristit ćemo sljedeći jednostavni logistički regresijski model:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{1.2}$$

Definiramo logit funkciju kao

$$\begin{aligned} g(x) &= \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \\ &= \beta_0 + \beta_1 x. \end{aligned} \tag{1.3}$$

Logit funkcija ima nekoliko korisnih svojstava linearnog regresijskog modela:
1) linearna je u svojim parametrima
2) može biti neprekidna
3) može poprimati vrijednosti od $-\infty$ do $+\infty$ ovisno o vrijednostima varijable x.

Ako imamo p nezavisnih varijabli $x_1, ..., x_p$ i vektor $\boldsymbol{x} = (x_1, ..., x_p)$ te s $\pi(\boldsymbol{x})$ označimo uvjetnu vjerojatnost da je $Y = 1$ uz dani $\boldsymbol{x}$ tj. $\pi(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x})$, multivarijabilni logistički model glasi

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}, \tag{1.4}$$

a pripadna logit funkcija

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p. \tag{1.5}$$

[24]

1.3.1 Metoda maksimalne vjerodostojnosti

Kao što je već rečeno, za procjenu parametara modela koristit ćemo metodu maksimalne vjerodostojnosti. Prvo moramo konstruirati funkciju vjerodostojnosti (eng. *likelihood function*) koja izražava vjerojatnost opaženih podataka kao funkciju nepoznatih parametara. Procjenitelji tih parametara metodom maksimalne vjerodostojnosti su tada vrijednosti koje maksimiziraju tu funkciju. [24]

Preciznije, imamo sljedeću definiciju:

Definicija 1.3.1. *Neka je $(x_1, ..., x_n)$ opaženi uzorak za slučajnu varijablu X s gustoćom $f(x|\theta)$, gdje je $\theta = (\theta_1, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$ nepoznati parametar.*

Definiramo funkciju vjerodostojnosti $L : \Theta \to \mathbb{R}$ *sa*

$$L(\theta) := f(x_1|\theta) \cdots f(x_n|\theta), \qquad \theta \in \Theta.$$

Vrijednost $\hat{\theta} = \hat{\theta}(x_1, ..., x_n) \in \Theta$ *za koju je*

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$

zovemo procjena metodom maksimalne vjerodostojnosti.
Statistika $\hat{\theta}(X_1, ..., X_n)$ *je procjenitelj metodom maksimalne vjerodostojnosti ili kraće MLE* (*eng.* maximum likelihood estimator). [30]

Promotrimo prvo slučaj univarijabilne regresije. Označimo s (x_i, y_i), $i = 1, ..., n$ n nezavisnih opažanja gdje y_i označava vrijednost zavisne varijable, a x_i vrijednost nezavisne varijable za *i-to* opažanje. Ako je Y dihotomna zavisna varijabla kodirana s 0 i 1 te ako je $\boldsymbol{\beta} = (\beta_0, \beta_1)^\tau$ onda izraz (1.2) za $\pi(x)$ daje uvjetnu vjerojatnost $P(Y = 1|x)$, a razlika $1 - \pi(x)$ daje uvjetnu vjerojatnost $P(Y = 0|x)$. Dakle, par (x_i, y_i) doprinosi funkciji vjerodostojnosti s

$$\pi(x_i)^{y_i}\Big[1 - \pi(x_i)\Big]^{1-y_i}. \tag{1.6}$$

S obzirom da je jedna od pretpostavki da su opažanja međusobno nezavisna, funkcija vjerodostojnosti se dobiva kao produkt izraza (1.6):

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\Big[1 - \pi(x_i)\Big]^{1-y_i}. \tag{1.7}$$

Funkcija $x \mapsto lnx$ je strogo rastuća pa je dovoljno izračunati logaritam izraza (1.7) kako bismo pronašli maksimum. Dobivamo izraz

$$\begin{aligned} L(\boldsymbol{\beta}) &= \ln\Big[l(\boldsymbol{\beta})\Big] \\ &= \sum_{i=1}^{n} \Big\{y_i \ln\Big[\pi(x_i)\Big] + (1 - y_i) \ln\Big[1 - \pi(x_i)\Big]\Big\} \end{aligned} \tag{1.8}$$

koji nazivamo log-vjerodostojnost (eng. *log likelihood*). Dalje, deriviramo $L(\boldsymbol{\beta})$ po β_0, a zatim i po β_1 te izjednačavamo te izraze s 0. Time dobivamo jednadžbe vjerodostojnosti (eng. *likelihood equations*):

$$\begin{aligned} \sum_{i=1}^{n} \Big[y_i - \pi(x_i)\Big] &= 0 \\ \sum_{i=1}^{n} x_i\Big[y_i - \pi(x_i)\Big] &= 0. \end{aligned} \tag{1.9}$$

Kod linearne regresije te su jednadžbe linearne po nepoznatim parametrima β_0 i β_1 pa se lako rješavaju dok su kod logističke regresije nelinearne pa se rješavaju posebnim metodama ili programima. Vrijednost $\hat{\boldsymbol{\beta}}$ koju dobijemo kao rješenje tih jednadžbi je procjena metodom maksimalne vjerodostojnosti koja je u definiciji 1.3.1 označena s $\hat{\theta}$.

Označimo s $\hat{\pi}(x_i)$ procjenu metodom maksimalne vjerodostojnosti za $\pi(x_i)$ čime dobivamo procjenu za uvjetnu vjerojatnost $P(Y = 1|x = x_i)$ koja predstavlja predviđene vrijednosti logističkog modela. Posljedica jednadžbi (1.9) je činjenica da je suma opaženih vrijednosti y jednaka sumi procijenjenih vrijednosti $\hat{\pi}(x_i)$, odnosno

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i).$$

U slučaju multivarijabilne regresije, označimo s $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, ..., n$ uzorak n nezavisnih opažanja. Želimo procijeniti vektor koeficijenata $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^\tau$. Postupak je sličan kao u univarijabilnom slučaju i funkcija vjerodostojnosti je slična kao u (1.7) samo što je $\pi(\boldsymbol{x})$ definiran kao u (1.4). Umjesto dvije jednadžbe vjerodostojnosti, jer funkciju log-vjerodostojnosti deriviramo po $\beta_0, \beta_1, ..., \beta_p$, dobivamo $p + 1$ jednadžbu vjerodostojnosti:

$$\begin{aligned} &\sum_{i=1}^{n} \Big[y_i - \pi(\boldsymbol{x}_i) \Big] = 0 \\ &\sum_{i=1}^{n} x_{ij} \Big[y_i - \pi(\boldsymbol{x}_i) \Big] = 0, \qquad j = 1, 2, ..., p. \end{aligned}$$

Rješavanjem tih jednadžbi posebnim programima dobivamo procjenu metodom maksimalne vjerodostojnosti $\hat{\boldsymbol{\beta}}$. [24]

1.4 Testiranje značajnosti koeficijenata

Nakon što smo procijenili parametre modela, želimo testirati jesu li oni statistički značajni, odnosno, želimo odgovoriti na pitanje je li model bolji ako uključimo određenu varijablu nego ako tu varijablu ne uključimo u model. Na to pitanje ćemo odgovoriti uspoređivanjem opaženih vrijednosti zavisne varijable s vrijednostima koje predviđaju model s tom varijablom i model bez te varijable.

Hipoteze koje testiramo su

$$\begin{aligned} H_0 : &\quad \beta_1 = 0 \\ H_1 : &\quad \beta_1 \neq 0 \end{aligned} \tag{1.10}$$

u slučaju univarijabilne regresije, odnosno

$$\begin{aligned} H_0 : &\quad \boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\tau = 0 \\ H_1 : &\quad \boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\tau \neq 0 \end{aligned} \tag{1.11}$$

u slučaju multivarijabilne regresije. [24]

Test omjera vjerodostojnosti

Ako imamo n opažanja saturirani model je onaj koji procjenjuje n parametara. Recimo, polinom prvog stupnja je primjer saturiranog modela ako imamo samo dva opažanja. Takvi modeli predstavljaju na neki način savršenu procjenu, a u slučaju regresije opažene vrijednosti zavisne varijable možemo zamišljati kao vrijednosti procijenjene saturiranim modelom. Saturirani modeli se najčešće koriste pri usporedbi dvaju modela te opažene i procijenjene vrijednosti stoga možemo usporediti pomoću funkcije vjerodostojnosti na sljedeći način:

$$\begin{aligned} D &= -2 \ln \left[\frac{\text{vjerodostojnost procijenjenog modela}}{\text{vjerodostojnost saturiranog modela}} \right] \\ &= -2 \ln \Big[\text{omjer vjerodostojnosti} \Big]. \end{aligned} \tag{1.12}$$

Ako se vrijednost dobivena u (1.12) koristi za testiranje hipoteza tada to nazivamo testom omjera vjerodostojnosti. U slučaju univarijabilne logističke regresije, koristeći izraz (1.8) dobivamo:

$$D = -2 \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\hat{\pi}(x_i)}{y_i} \right) + (1 - y_i) \ln \left(\frac{1 - \hat{\pi}(x_i)}{1 - y_i} \right) \right].$$

Statistika D naziva se devijancom i ima istu ulogu kao rezidualna suma kvadrata $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ u linearnoj regresiji pri čemu smo s y_i označili opaženu i-tu vrijednost, a s $\hat{y}_i$ predviđenu i-tu vrijednost zavisne varijable. Dodavanjem varijabli u model njena vrijednost se smanjuje te ona također ima i važnu ulogu u procjeni adekvatnosti modela.

Kako bismo odredili je li neka nezavisna varijabla značajna uspoređujemo devijancu modela bez te varijable s devijancom modela koji uključuje tu varijablu:

$$\begin{aligned} G &= D(\text{model bez varijable}) - D(\text{model s varijablom}) \\ &= -2 \ln \left[\frac{\text{vjerodostojnost modela bez varijable}}{\text{vjerodostojnost saturiranog modela}} \right] - \\ &\quad - \left(-2 \ln \left[\frac{\text{vjerodostojnost modela s varijablom}}{\text{vjerodostojnost saturiranog modela}} \right] \right) \\ &= -2 \ln \left[\frac{\text{vjerodostojnost modela bez varijable}}{\text{vjerodostojnost modela s varijablom}} \right]. \end{aligned} \tag{1.13}$$

Pod uvjetom da vrijedi hipoteza H_0 iz (1.10) vrijedi $G \sim \chi^2(1)$.

U slučaju multivarijabilne logističke regresije test značajnosti p koeficijenata provodimo na isti način. Koristimo statistiku G kao u (1.13) samo što se procijenjene vrijednosti $\hat{\pi}$ odnose na vektor $\hat{\boldsymbol{\beta}}$ s $p + 1$ koeficijentom. Pod uvjetom da vrijedi hipoteza H_0 iz (1.11) vrijedi $G \sim \chi^2(p)$. [24]

Waldov test

Waldov test je statistički ekvivalentan testu omjera vjerodostojnosti. U slučaju univarijabilnog modela Waldova testna statistika računa se kao

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\beta_1)} \tag{1.14}$$

pri čemu $\widehat{SE}(\beta_1)$ označava standardnu pogrešku koeficijenta β_1 definiranu kao

$$\widehat{SE}(\beta_1) = \sqrt{\frac{1}{n-2} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

pri čemu je n duljina uzorka, y_i vrijednost opažene i-te zavisne varijable, $\hat{y}_i$ predviđena vrijednost i-te zavisne varijable, x_i i-ta vrijednost nezavisne varijable, a $\bar{x}$ očekivanje nezavisne varijable ([38]). Pod pretpostavkom da vrijedi hipoteza H_0 iz (1.10) vrijedi $W \sim N(0, 1)$.

Multivarijabilni analogon Waldove statistike računa se kao

$$\begin{aligned} W &= \hat{\boldsymbol{\beta}}^{\tau} \left[\widehat{Var}(\hat{\boldsymbol{\beta}}) \right]^{-1} \hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}}^{\tau} (\boldsymbol{X}^{\tau} \boldsymbol{V} \boldsymbol{X}) \hat{\boldsymbol{\beta}} \end{aligned}$$

pri čemu je $\boldsymbol{X}$ kao u (1.1), a

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}(\boldsymbol{x_1})(1 - \hat{\pi}(\boldsymbol{x_1})) & 0 & \dots & 0 \\ 0 & \hat{\pi}(\boldsymbol{x_2})(1 - \hat{\pi}(\boldsymbol{x_2})) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}(\boldsymbol{x_n})(1 - \hat{\pi}(\boldsymbol{x_n})) \end{bmatrix}. \tag{1.15}$$

Ako vrijedi hipoteza H_0 iz (1.11) tada je $W \sim \chi^2(p + 1)$. [24]

1.5 Procjena pouzdanih intervala

Kod univarijabilne regresije procjene za pouzdane intervale za koeficijente β_0 i β_1 se dobivaju na temelju Waldovih statistika. $100(1-\alpha)\%$ pouzdani interval za koeficijent β_0 je

$$\left[\hat{\beta}_0 - z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_0),\, \hat{\beta}_0 + z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_0)\right] \tag{1.16}$$

pri čemu je $z_{1-\frac{\alpha}{2}}$ $(1-\frac{\alpha}{2})$-ti kvantil standardne normalne distribucije, a $\widehat{SE}(\hat{\beta}_0)$ procjena standardne pogreške temeljena na modelu. Analogno, $100(1-\alpha)\%$ pouzdani interval za koeficijent β_1 je

$$\left[\hat{\beta}_1 - z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1),\, \hat{\beta}_1 + z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1)\right], \tag{1.17}$$

a za logit funkciju

$$\left[\hat{g}(x) - z_{1-\frac{\alpha}{2}}\widehat{SE}\left[\hat{g}(x)\right],\, \hat{g}(x) + z_{1-\frac{\alpha}{2}}\widehat{SE}\left[\hat{g}(x)\right]\right]$$

pri čemu je

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x,$$

a

$$\begin{aligned}\widehat{SE}\left[\hat{g}(x)\right] &= \sqrt{\widehat{Var}\left[\hat{g}(x)\right]} \\ &= \sqrt{\widehat{Var}\left(\hat{\beta}_0\right) + x^2\widehat{Var}\left(\hat{\beta}_1\right) + 2x\widehat{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)}.\end{aligned}$$

Slično, $100(1-\alpha)\%$ pouzdani interval za vrijednosti koeficijenata procijenjenih modelom (1.2) je

$$\left[\frac{e^{\hat{g}(x)-z_{1-\frac{\alpha}{2}}\widehat{SE}\left[\hat{g}(x)\right]}}{1+e^{\hat{g}(x)-z_{1-\frac{\alpha}{2}}\widehat{SE}\left[\hat{g}(x)\right]}}, \frac{e^{\hat{g}(x)+z_{1-\frac{\alpha}{2}}\widehat{SE}\left[\hat{g}(x)\right]}}{1+e^{\hat{g}(x)+z_{1-\frac{\alpha}{2}}\widehat{SE}\left[\hat{g}(x)\right]}}\right].$$

Kod multivarijabilne regresije procjene za pouzdane intervale za koeficijente β_0, β_1, ..., β_p računaju se analogno kao u (1.16) i (1.17) dok je za logit funkciju situacija malo kompliciranija. Iz (1.5) slijedi da je

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

što pomoću vektorske notacije zapisujemo kao

$$\hat{g}(\boldsymbol{x}) = \boldsymbol{x}^\tau \hat{\boldsymbol{\beta}}$$

pri čemu je $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^\tau$ i $\boldsymbol{x}^\tau = (1, x_1, x_2, ..., x_p)$. Ako je $\boldsymbol{X}$ kao u (1.1) i $\boldsymbol{V}$ kao u (1.15) onda slijedi

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\tau \boldsymbol{V} \boldsymbol{X})^{-1}$$

iz čega slijedi

$$\begin{aligned}\widehat{Var}\Big[\Big(\hat{g}(\boldsymbol{x})\Big)\Big] &= \boldsymbol{x}^\tau \widehat{Var}\big(\hat{\boldsymbol{\beta}}\big)\boldsymbol{x} \\ &= \boldsymbol{x}^\tau (\boldsymbol{X}^\tau \boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}. \end{aligned} \tag{24}$$

1.6 Izgled (eng. *odds*) i omjer izgleda (eng. *odds ratio*)

Šansa ili izgled (eng. *odds*) definira se kao omjer vjerojatnosti da se neki događaj dogodio i vjerojatnosti da se nije dogodio. Preciznije, ako s p označimo vjerojatnost nekog događaja, tada izgled definiramo kao

$$odds = \frac{p}{1-p}. \tag{1.18}$$

Uočimo da je tada zapravo logit funkcija

$$g(x) = \ln(odds).$$

Iz izgleda možemo izračunati vjerojatnost koja se definira kao omjer broja povoljnih događaja i ukupnog broja događaja. Koristeći oznake kao u (1.18), vjerojatnost p nekog događaja možemo izraziti kao

$$p = \frac{odds}{1+odds}. \tag{1.19}$$

Za razliku od vjerojatnosti koja poprima vrijednosti od 0 do 1, izgled može poprimiti bilo koju nenegativnu vrijednost. [10]

Računanje izgleda je ponekad korisno kod uspoređivanja rezultata. Na primjer, uzmimo neku sportsku ekipu koja pobjeđuje u 40 od 100 odigranih utakmica. Tada je vjerojatnost pobjede ekipe $p = \frac{40}{100} = 0.4$ dok je šansa $odds = \frac{40}{60} = \frac{2}{3} = 0.6667$, odnosno ekipa pobjeđuje u dvije utakmice dok u tri gubi. Ako druga ekipa ima dvostruko veću vjerojatnost pobjede $p = \frac{80}{100} = 0.8$ tj. ako pobjeđuje u 80 od 100 utakmica tada ima šansu $odds = \frac{80}{20} = 4$ odnosno pobjeđuje u četiri utakmice dok u jednoj gubi. S druge strane, ako je šansa neke ekipe da pobijedi $odds = 3$, a druge ekipe $odds = 6$, to nipošto ne znači da druga ekipa ima vjerojatnost pobjede dvostruko veću od prve ekipe. Koristeći formulu (1.19), lako se izračuna da je vjerojatnost pobjede prve ekipe 0.75, a druge ekipe 0.8571.

Omjer šansi ili izgleda (eng. *odds ratio*) definira se kao omjer izgleda u jednoj grupi i izgleda u drugoj grupi. Preciznije, ako je x dihotomna varijabla koja može poprimiti samo vrijednosti 0 i 1, izgled da se događaj dogodio u grupi u kojoj je $x = 1$ je $\frac{\pi(1)}{1-\pi(1)}$, a izgled da se dogodio u grupi u kojoj je $x = 0$ je $\frac{\pi(0)}{1-\pi(0)}$. Tada omjer izgleda u oznaci *OR* definiramo

kao

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}. \qquad [24] \tag{1.20}$$

Omjer izgleda koristan je u primjenama. Npr. ako $x = 1$ označava da je pojedina osoba koja je sudjelovala u istraživanju pretila, a $x = 0$ da nije te $y = 1$ da ima visoki krvni tlak, a $y = 0$ da nema onda $\widehat{OR} = 2$ označava procjenu da se visoki tlak pojavljuje dvaput češće kod pretilih osoba u odnosu na osobe koje nisu pretile.

1.6.1 Interpretacija parametra β_1 u modelu univarijabilne logističke regresije

Pojam omjera izgleda potreban nam je kako bismo mogli interpretirati koeficijent smjera (eng. *slope coefficient*) β_1 u modelu univarijabilne logističke regresije (1.2). Prvo ćemo pogledati primjer kad je nezavisna varijabla dihotomna, zatim kad je kategorijska, ali takva da može poprimiti više od dvije vrijednosti te na kraju kad je kontinuirana.

Ako imamo dihotomnu nezavisnu varijablu koja je kodirana s 0 i 1 te ako je $x = 1$ i $y = 1$, uvrštavanjem vrijednosti x u formulu (1.2), dobijemo:

$$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}. \tag{1.21}$$

Slično, ako je $x = 1$ i $y = 0$ uvrštavanjem u istu formulu dobijemo:

$$1 - \pi(1) = \frac{1}{1 + e^{\beta_0+\beta_1}}. \tag{1.22}$$

Analogno napravimo za $x = 0$ i $y = 1$ te $x = 0$ i $y = 0$ te dobivene izraze zajedno s (1.21) i (1.22) uvrstimo u (1.20). Dobijemo:

$$\begin{aligned} OR &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) \Big/ \left(\frac{1}{1+e^{\beta_0}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{\beta_1}. \end{aligned} \tag{1.23}$$

Dakle, dobili smo vezu između omjera izgleda i koeficijenta regresije β_1 u logističkom regresijskom modelu s dihotomnom zavisnom varijablom koja je kodirana s 0 i 1. U tom

modelu taj koeficijent ujedno predstavlja i razliku u logit funkciji (1.3)

$$g(1) - g(0) = \beta_1,$$

odnosno označava promjenu u logit funkciji ako prijeđemo iz kategorije označene s 0 u kategoriju označenu s 1.

Zbog svoje interpretacije omjer izgleda je često parametar koji nas zanima kod logističke regresije. Osim njegove procjene $\widehat{OR}$, želimo izračunati i $100(1-\alpha)\%$ pouzdani interval za njegovu procjenu pomoću:

$$\left[e^{\hat{\beta}_1 - z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1)},\ e^{\hat{\beta}_1 + z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_1)}\right]. \tag{1.24}$$

Ako nezavisna varijabla nije kodirana pomoću 0 i 1 nego na drukčiji način, npr. s -1 i 1, na sličan način možemo dobiti vezu s koeficijentom regresije, međutim, ona ne mora uvijek biti ista kao u (1.23). Detaljnije se može naći u [24], ali ovdje se nećemo time baviti jer se svaka varijabla uvijek može kodirati pomoću 0 i 1 što nazivamo kodiranjem pomoću referentne ćelije (eng. *reference cell coding*).

Ako je zavisna varijabla kategorijska, ali može poprimiti više od dvije vrijednosti tj. ako može poprimiti $k, k > 2$ vrijednosti, kodiramo ju pomoću dizajn (eng. *design variables*) ili *dummy* varijabli. Pritom koristimo točno $k-1$ *dummy* varijablu. Najjednostavniji način je kodiranje pomoću referentne ćelije gdje sve *dummy* varijable postavimo na 0 za referentnu ćeliju, a za svaku od ostalih kategorija točno jedna *dummy* varijabla ima vrijednost 1 dok ostale imaju vrijednost 0.

Dakle, pretpostavimo da zavisna varijabla može poprimiti $k, k > 2$ vrijednosti. Tada model univarijabilne logističke regresije glasi

$$g(x) = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + ... + \beta_{k-1} d_{k-1} \tag{1.25}$$

pri čemu su d_1, d_2, ..., d_{k-1} *dummy* varijable. Ako želimo izračunati omjer izgleda između referentne kategorije i „prve" kategorije tj. kategorije koju označava 1 u varijabli d_1 i 0 u varijablama $d_2, ..., d_{k-1}$, računom kao u slučaju dihotomne nezavisne varijable i (1.23) dobivamo

$$OR(\text{referentna kategorija, kategorija}_1) = e^{\beta_1} \tag{1.26}$$

te je $100(1-\alpha)\%$ pouzdani interval za procjenu omjera izgleda isti kao u (1.24). Analogno, ako računamo omjer izgleda između referentne i „k-1" kategorije, slično kao u (1.26) dobivamo

$$OR(\text{referentna kategorija, kategorija}_{k-1}) = e^{\beta_{k-1}}$$

te $100(1-\alpha)\%$ pouzdani interval za procjenu omjera izgleda

$$\left[e^{\hat{\beta}_{k-1} - z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_{k-1})},\ e^{\hat{\beta}_{k-1} + z_{1-\frac{\alpha}{2}}\widehat{SE}(\hat{\beta}_{k-1})}\right].$$

U slučaju da je varijabla x kontinuirana i uz pretpostavku da je logit funkcija linearna u toj varijabli, koeficijent β_1 označava promjenu u logit funkciji kad se varijabla x promijeni za jednu jedinicu jer je

$$g(x+1) - g(x) = \beta_1.$$

Međutim, nekad pomak za jednu jedinicu nije praktičan u primjenama jer je prevelik ili premalen pa je potrebno promotriti što se događa s pomakom za c jedinica. Recimo, vrijednosti krvnog tlaka se najčešće izražavaju u mmHg, a povećanje ili smanjenje za 1 mm je najčešće premalo da bi se smatralo važnim pa gledamo pomak za recimo $c = 10$ mmHg. Sličnim računom kao u (1.23) dobijemo da je tada

$$OR = e^{c\beta_1}$$

te da je $100(1-\alpha)\%$ pouzdani interval za procjenu omjera izgleda

$$\left[e^{c\hat{\beta}_1 - z_{1-\frac{\alpha}{2}} c\widehat{SE}(\hat{\beta}_1)},\ e^{c\hat{\beta}_1 + z_{1-\frac{\alpha}{2}} c\widehat{SE}(\hat{\beta}_1)}\right]. \qquad [24]$$

1.6.2 Omjer izgleda u modelu multivarijabilne logističke regresije

Kad u logističkom modelu imamo više od jedne nezavisne varijable, koeficijente uz te varijable ne možemo više tako jednostavno interpretirati pomoću omjera izgleda, odnosno ne možemo omjer izgleda više procijeniti kao $\widehat{OR} = e^{\beta_1}$. Dodatno, situacija se komplicira kad je u modelu prisutna interakcija. Interakcija dviju ili više varijabli je situacija kad utjecaj jedne nezavisne varijable na zavisnu varijablu ovisi o drugoj ili više drugih nezavisnih varijabli.

U tom slučaju, kako bismo procijenili omjer izgleda, promotrimo primjer u kojem imamo dvije kovarijate od kojih je jedna dihotomna te njihovu interakciju. Ako dihotomnu varijablu označimo s F, a drugu nezavisnu varijablu s X te njihovu interakciju s $F \times X$, možemo primijeniti sljedeći postupak:

1) napišemo izraze za logit funkciju za vrijednosti $F = f_0$ i $F = f_1$ koje uspoređujemo
2) algebarski pojednostavnimo razliku između ta dva logita i izračunamo ju
3) dobivenu vrijednost eksponenciramo.

Logit model za $F = f$ i $X = x$ glasi

$$g(f, x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 f \times x,$$

a za $F = f_1$ i $F = f_0$ pri čemu je $X = x$

$$g(f_1, x) = \beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x$$
$$g(f_0, x) = \beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x.$$

Pojednostavljivanjem i računanjem razlike $g(f_1, x) - g(f_0, x)$ imamo

$$\begin{aligned} \ln\Big[OR(F = f_1, F = f_0, X = x)\Big] &= g(f_1, x) - g(f_0, x) \\ &= (\beta_0 + \beta_1 f_1 + \beta_2 x + \beta_3 f_1 \times x) - \\ &\quad - (\beta_0 + \beta_1 f_0 + \beta_2 x + \beta_3 f_0 \times x) \\ &= \beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0). \end{aligned} \tag{1.27}$$

Konačno, eksponenciranjem izraza (1.27) dobivamo

$$OR = e^{\beta_1(f_1 - f_0) + \beta_3 x(f_1 - f_0)}.$$

Primijetimo da se izraz (1.27) ne može pojednostavniti tako da ostane samo jedan koeficijent. [24]

1.7 Podaci koji nedostaju (eng. *missing data*)

Prije nego počnemo birati varijable koje će se nalaziti u modelu, moramo promotriti podatke. Može se dogoditi da nisu svi podaci dostupni (eng. *missing data*) zbog različitih razloga. Prema [29], bitno je koji su to razlozi jer se prema njima određuje na koji se način može riješiti problem nedostatka podataka. Ti razlozi se razvrstavaju u tri kategorije:

1) podaci nedostaju na potpuno slučajan način (MCAR, eng. *missing completely at random*)
2) podaci nedostaju na slučajan način (MAR, eng. *missing at random*)
3) podaci ne nedostaju na slučajan način (MNAR, eng. *missing not at random*).

Primjer podataka koji nedostaju na potpuno slučajan način je recimo kad se slučajno uništi uzorak krvi u nekom laboratoriju. Međutim, ako se vjerojatnost da je pojedini uzorak krvi uništen može predvidjeti iz ostalih dostupnih podataka o uzorku onda govorimo o podacima koji nedostaju na slučajan način. Recimo, ako je neki test kompliciraniji i češće se događa da se tijekom testiranja uzorak uništi onda na temelju toga što znamo da se na pojedinom uzorku trebao provesti taj test, možemo zaključiti da je veća vjerojatnost da će uzorak biti uništen. No, to se i dalje događa na slučajan način jer ne znamo točno koji uzorak će biti uništen. Primjer podataka koji ne nedostaju na slučajan način je kad nema podataka na dopinškom testu za određenu osobu jer ona namjerno nije bila dostupna za testiranje jer koristi doping.

Kako bismo dobili logistički regresijski model, svi podaci moraju biti dostupni. To možemo postići na više različitih načina, a većina ih podrazumijeva da podaci nedostaju ili na potpuno slučajan ili na slučajan način dok je situacija kompliciranija ako podaci ne nedostaju

na slučajan način te rezultati koje tada dobijemo ne moraju biti dobri. Među najčešće korištenim metodama za rješavanje problema nedostatka podataka su:

1) brisanje subjekata za koje nedostaje jedan ili više podataka (eng. *deletion methods*)
2) jednostruko nadomještanje podataka (eng. *single imputation methods*)
3) nadomještanje pomoću procjene metodom maksimalne vjerodostojnosti (eng. *maximum likelihood estimation method*)
4) višestruko nadomještanje (eng. *multiple imputations method*).

Brisanje subjekata za koje nedostaje barem jedan podatak je metoda koja pretpostavlja da podaci nedostaju na potpuno slučajan način te je to najjednostavnija metoda. Provodi se samo kad je u pitanju mali broj subjekata za koje podaci nisu dostupni jer se inače značajnije smanjuje veličina uzorka.

Metode jednostrukog nadomještanja podataka uključuju nadomještanje očekivanjem, nadomještanje regresijom te stohastičko nadomještanje regresijom. Nadomještanje očekivanjem podrazumijeva računanje očekivanja dostupnih vrijednosti te nadomještanje istima gdje je potrebno. Za nadomještanje regresijom potrebno je provesti linearnu regresijsku analizu s dostupnim podacima kao varijablama poticaja dok je varijabla odgovora varijabla za koju nedostaju podaci. Dakle, podaci koji nedostaju se zamijene procijenjenim vrijednostima dobivenog modela. Stohastičko nadomještanje regresijom je slično, ali uključuje i slučajnu grešku. Takvo nadomještanje daje dobre rezultate i za podatke koji nedostaju na potpuno slučajan i na slučajan način.

Metoda nadomještanja pomoću procjene metodom maksimalne vjerodostojnosti koristi sve dostupne podatke kako bi odredila parametre koji maksimiziraju vjerojatnost generiranja uzorka pritom koristeći funkciju log-vjerodostojnosti. Ta metoda ne mora davati dobre rezultate ako je uzorak mali te često zahtijeva korištenje posebnih programa.

Višestruko nadomještanje funkcionira tako da se generira m skupova podataka u kojima se vrijednosti koje nedostaju nadomjeste. Zatim se svaki skup podataka analizira posebno te se rezultati uklope u jedan konačan skup podataka poštujući Rubinova pravila koja se mogu naći u [23]. Ova metoda se smatra jednako dobrom kao nadomještanje pomoću procjene metodom maksimalne vjerodostojnosti, ali ne zahtijeva nikave posebne programe te se može primijeniti na bilo kakav tip podataka. [29]

1.8 Odabir odgovarajućeg modela

Nakon što smo riješili problem podataka koji nedostaju, cilj nam je odabrati varijable koje će ući u model tako da dobijemo najbolji model u smislu optimalnog broja varijabli, samih varijabli, ali i ukupne procjene modela. Postoje brojne statističke metode pomoću kojih pokušavamo na temelju testiranja i uspoređivanja odrediti koje varijable ćemo uključiti u model, ali te metode nisu jedini kriterij. Jednako je važno logički razmisliti koje varijable su neophodne za svaki pojedini problem. Te varijable svakako treba uključiti u model unatoč tome što će ponekad statističke metode možda sugerirati drukčije, a često se pri određivanju takvih varijabli konzultira struka.

Najčešći pristup je da odaberemo model koji sadrži najmanji mogući broj varijabli koje svejedno dobro opisuju podatke. Veća je vjerojatnost da su takvi modeli numerički stabilniji i da se mogu generalizirati. S druge strane, ako u modelu imamo previše varijabli, model može previše ovisiti o opaženim podacima što nije dobro. [24]

1.8.1 Odabir varijabli

Pri odabiru varijabli koje će se nalaziti u modelu najčešće se slijede sljedeći koraci:

1) univarijabilna analiza svake varijable
2) multivarijabilna analiza s odabranim varijablama iz koraka 1
3) provjera značajnosti svake odabrane varijable
4) provjera linearnosti logit funkcije za svaku odabranu neprekidnu varijablu
5) dodavanje interakcija.

Kod univarijabilne analize ako se radi o nominalnoj, ordinalnoj ili neprekidnoj varijabli koja ima mali broj cjelobrojnih vrijednosti, radimo kontingencijsku tablicu s dva retka za $y = 0$ i $y = 1$ te k stupaca za k kategorija nezavisne varijable. Pritom moramo paziti da nijedna ćelija nema vrijednost 0 što možemo riješiti npr. grupiranjem ćelija ili ako se radi o ordinalnoj varijabli, možemo tu varijablu modelirati kao da je neprekidna. Zatim koristimo χ^2-test omjera vjerodostojnosti s $k - 1$ stupnjem slobode koji je jednak vrijednosti testa omjera vjerodostojnosti za testiranje značajnosti koeficijenata za $k - 1$ *dummy* varijablu u univarijabilnom modelu koji sadrži jednu nezavisnu varijablu. Kod varijabli koje su statistički značajne na odabranoj razini značajnosti, poželjno je izračunati procjene omjera izgleda i pripadne procjene pouzdanih intervala za omjere izgleda.

Ako se radi o neprekidnim varijablama, prilagođavamo univarijabilni model da dobijemo procjenu koeficijenata, procjenu standardne greške, univarijabilnu Waldovu statistiku i test omjera vjerodostojnosti za značajnost koeficijenta uz pripadne p-vrijednosti.

Nakon što smo napravili univarijabilnu analizu, biramo varijable koje ćemo uključiti u multivarijabilnu analizu. Varijable kojima je p-vrijednost < 0.25 dolaze u obzir. Preporuka je da se uzme ta granica jer ako se uzme premala granica (npr. p-vrijednost < 0.05) može se dogoditi da ispadne da neke varijable nisu statistički značajne, a znamo da bi logički trebale biti. Također, može se dogoditi da zanemarimo neku varijablu koja sama po sebi nije dovoljno statistički značajna, ali zajedno s drugim varijablama to postaje. Ako se uzme prevelika granica, ispast će da je upitno velik broj varijabli statistički značajan i imat ćemo model s nepotrebno puno nezavisnih varijabli.

Ako nemamo puno nezavisnih varijabli, možemo odmah prijeći na multivarijabilnu analizu bez obzira na rezultate dobivene u koraku 1. Također, možemo koristiti metodu odabira najboljeg podskupa (eng. *best subsets selection*) pri čemu po određenom kriteriju uspoređujemo sve modele s jednom varijablom pa sve modele s dvije pa sve s tri, itd. kako bismo našli najbolji model. Ta se metoda ne koristi toliko često u logističkoj regresiji nego je puno korištenija *stepwise* metoda koja ima dvije glavne verzije:

- *forward* odabir s testom za *backward* eliminaciju
- *backward* odabir s testom za *forward* eliminaciju.

Nakon što smo za odabrane varijable nekom od navedenih metoda prilagodili multivarijabilni model, za svaku od varijabli u modelu provjeravamo značajnost. To uključuje provjeru Waldove statistike te usporedbu svakog procijenjenog koeficijenta iz tog modela s procijenjenim koeficijentom u modelu univarijabilne logističke regresije koji sadrži samo tu varijablu. Zatim varijable koje ne doprinose modelu mičemo iz modela, procjenjujemo novi model te ga uspoređujemo sa starim pomoću testa omjera vjerodostojnosti. Procijenjene koeficijente varijabli koje su ostale u modelu moramo još usporediti s procijenjenim koeficijentima u punom modelu jer velike promjene mogu značiti da neka varijabla ili više njih koje smo maknuli iz modela utječu na neku od varijabli koja je ostala u modelu. Ponavljamo postupak brisanja i dodavanja varijabli dok nismo sigurni da su sve značajne varijable u modelu, a sve varijable koje nisu u modelu nisu značajne.

Sada kada smo dobili multivarijabilni model za koji mislimo da sadrži sve potrebne varijable, za svaku neprekidnu varijablu moramo provjeriti linearnu povezanost s logit funkcijom. To se može napraviti pomoću *dummy* varijabli tako da umjesto neprekidne varijable prilagođavamo model s kategorijskom varijablom s nekoliko kategorija, preciznije, najčešće s četiri kategorije koje se odrede pomoću deskriptivne statistike. Drugi načini su pomoću frakcijskih polinoma (eng. *fractional polynomials*) ([24]), grafički pomoću *scatter plota* ili pomoću Box-Tidwellovog testa ([8]). Procedura za Box-Tidwellov test je sljedeća:

1) u modelu ostavimo samo neprekidne varijable
2) za svaku neprekidnu varijablu, u model dodamo interakciju *varijabla* $\times$ ln(*varijabla*)
3) provjeravamo statističku značajnost dodanih interakcija prema pripadnoj p-vrijednosti:

- ako je p-vrijednost > 0.05 interakcija nije statistički značajna što znači da je nezavisna varijabla linearna u logitu
- ako je p-vrijednost ≤ 0.05 interakcija je statistički značajna pa nezavisna varijabla nije linearna u logitu i trebamo provesti neku transformaciju nad varijablom. [8]

Dodajemo interakcije tako da razmišljamo koja bi nezavisna varijabla mogla utjecati na koju, odnosno koja varijabla bi mogla imati značajno drukčije vrijednosti po kategorijama ili vrijednostima druge. Na primjer, ako je zavisna dihotomna varijabla Y kodirana s 0 ako osoba nije imala srčani udar, a s 1 ako je, promatramo model u kojem su kovarijate *spol* i *pušenje* također dihotomne varijable pri čemu su žene kodirane s 0, muškarci s 1 te nepušači s 0, a pušači s 1. Smatramo da bi mogle postojati razlike u vjerojatnosti da osoba doživi srčani udar po spolu među pušačima te među nepušačima, odnosno da bi mogle postojati razlike da muška osoba doživi srčani udar u ovisnosti o tome puši li ili ne te da ženska osoba doživi srčani udar u ovisnosti o tome puši li ili ne. Zato u model dodajemo interakciju $spol \times pušenje$. Tada imamo logit funkciju $g(\boldsymbol{x}) = \beta_0 + \beta_1 \cdot spol + \beta_2 \cdot pušenje + \beta_3 \cdot spol \times pušenje$, a logistički model je $\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1+e^{g(\boldsymbol{x})}}$ pri čemu je $\boldsymbol{x} = (spol, pušenje)$ i $\pi(\boldsymbol{x}) = P(Y = srčani\, udar|\boldsymbol{x})$.

Interakcije uključujemo jednu po jednu u model te zatim pomoću testa omjera vjerodostojnosti provjeravamo njihovu značajnost. Smatra se da bi interakcije trebale doprinositi modelu na tradicionalnim razinama značajnosti, recimo na razini značajnosti od 5%. Konačnu odluku o tome hoćemo li uključiti interakciju u model donosimo na temelju statističke analize, ali i logičkog promišljanja. [24]

1.8.2 *Stepwise* procedura

Glavna ideja *stepwise* procedure je odabir ili eliminacija varijabli temeljena na algoritmu koji u svakom koraku provjerava značajnost varijabli prema kojoj varijable uključuje u model ili isključuje iz njega. Kod logističke regresije značajnost varijabli provjerava se χ^2-testom omjera vjerodostojnosti. Prilikom *forward* odabira testira se trenutni model u odnosu na model u kojem je dodana statistički najznačajnija varijabla, a prilikom *backward* eliminacije testira se sadašnji model u odnosu na model u kojem je eliminirana statistički najmanje značajna varijabla. U svakom koraku statistički najznačajnija varijabla je ona koja najviše promijeni log-vjerodostojnost u odnosu na model bez te varijable, odnosno ona koja dovede do najveće vrijednosti statistike G iz (1.13). [24], [10]

Algoritam *forward* odabira s testom za *backward* eliminaciju opisujemo po koracima:

Korak 0.

Pretpostavimo da imamo p mogućih nezavisnih varijabli za koje smo procijenili da bi

mogle biti logički značajne. Korak 0 započinje prilagodbom modela koji ima samo slobodni koeficijent β_0 i računanjem njegove log-vjerodostojnosti koju ćemo označiti s L_0. Tada prilagođavamo svaki od p mogućih univarijabilnih modela i računamo njihove log-vjerodostojnosti. Vrijednost log-vjerodostojnosti modela koji sadrži varijablu x_j u koraku 0 označavamo s $L_j^{(0)}$, a statistiku G koja označava vrijednost testa omjera vjerodostojnosti za model koji sadrži varijablu x_j u odnosu na model koji sadrži samo slobodni koeficijent označavamo s $G_j^{(0)}$ te zbog (1.13) vrijedi

$$G_j^{(0)} = -2(L_0 - L_j^{(0)}).$$

Pripadnu p-vrijednost označimo s $p_j^{(0)}$, a određujemo ju kao repnu vjerojatnost $P\left[\chi^2(v) > G_j^{(0)}\right] = p_j^{(0)}$ pri čemu je $v = 1$ ako je x_j neprekidna, a $v = k - 1$ ako je kategorijska s k kategorija.

Najznačajnija varijabla je ona s najmanjom p-vrijednosti te je ta varijabla kandidat za ulazak u model u koraku 1. Ako tu varijablu označimo s x_{e_1} onda je

$$p_{e_1}^{(0)} = \min(p_j^{(0)}).$$

Važno je napomenuti da to što je ta varijabla najznačajnija ne znači nužno da je ona i statistički značajna. Recimo, ako je $p_{e_1}^{(0)} = 0.8$, nema smisla dodati x_{e_1} u model i nastaviti dalje jer iako je najznačajnija među svim varijablama, nije statistički značajno povezana sa zavisnom varijablom dok ako je $p_{e_1}^{(0)} = 0.001$ želimo tu varijablu dodati u model u sljedećem koraku. Općenitije, ako s p_E označimo odabranu graničnu p-vrijednost temeljem koje procjenjujemo je li neka varijabla statistički dovoljno značajna za ulazak u model, vrijedi da ako je $p_{e_1}^{(0)} < p_E$ algoritam nastavlja s korakom 1 dok u protivnom staje ne dodajući nijednu varijablu u model. Najčešće se uzima p_E između 0.15 i 0.20, a ponekad čak i više.

Korak 1.

Korak 1 započinje prilagodbom modela koji sadrži varijablu x_{e_1}. Označimo s $L_{e_1}^{(1)}$ log-vjerodostojnost tog modela. Želimo odrediti značajnost preostale $p-1$ varijable pa prilagođavamo $p - 1$ logističkih modela koji sadrže varijable x_{e_1} i x_j, $j = 1, 2, ..., p$, $j \neq e_1$. Za model koji sadrži te dvije varijable označimo s $L_{e_1,j}^{(1)}$ log-vjerodostojnost te uz analognu oznaku kao u koraku 0 vrijedi

$$G_j^{(1)} = -2(L_{e_1}^{(1)} - L_{e_1,j}^{(1)})$$

uz pripadnu p-vrijednost $p_j^{(1)}$. Neka je varijabla s najmanjom p-vrijednosti u koraku 1 x_{e_2} pri čemu je

$$p_{e_2}^{(1)} = \min(p_j^{(1)}).$$

Ta varijabla je kandidat za ulazak u model u sljedećem koraku. Ako je $p^{(1)}_{e_2} < p_E$ algoritam nastavlja na korak 2, a u protivnom staje te u modelu imamo samo jednu varijablu.

Korak 2.

Analogno kao u koraku 1, korak 2 započinje prilagodbom modela koji sadrži varijable x_{e_1} i x_{e_2}. Moguće je da nakon što je x_{e_2} uključena u model x_{e_1} više nije značajna pa moramo procesom *backward* eliminacije utvrditi treba li ju izbaciti iz modela. Općenito, to se radi tako da se prilagođava model koji briše jednu od varijabli dodanih u prethodnim koracima i procjenjuje se značajnost izbrisane varijable.

Označimo s $L^{(2)}_{-e_j}$ log-vjerodostojnost modela u kojem je x_{e_j} izbrisana, s $G^{(2)}_{-e_j}$ test omjera vjerodostojnosti tog modela u usporedbi s punim modelom u koraku 2 te s $p^{(2)}_{-e_j}$ pripadnu p-vrijednost. Tada vrijedi

$$G^{(2)}_{-e_j} = -2(L^{(2)}_{-e_j} - L^{(2)}_{e_1,e_2}).$$

Kao potencijalnog kandidata kojeg će izbrisati iz modela, algoritam bira varijablu koja ako je izbrisana rezultira najvećom p-vrijednosti pripadnog χ^2-testa omjera vjerodostojnosti koji uspoređuje model s i bez te varijable. Ako tu varijablu označimo s x_{r_2} onda vrijedi

$$p^{(2)}_{r_2} = \max(p^{(2)}_{-e_j}, p^{(2)}_{e_2}).$$

Algoritam uspoređuje tu p-vrijednost s unaprijed odabranom p-vrijednosti koju smo označili s p_R. Mora vrijediti $p_R > p_E$ kako bismo osigurali da algoritam ne ubacuje pa izbacuje istu varijablu u dva neposredna koraka. Preporuča se uzeti p_R između 0.15 i 0.20, a ako ne želimo izbacivati puno varijabli nakon što su jednom ušle u model, možemo uzeti i puno veću p-vrijednost p_R. Ako vrijedi $p^{(2)}_{r_2} > p_R$ varijabla x_{r_2} se briše iz modela, a u protivnom ona ostaje u modelu. U oba slučaja, algoritam nastavlja dalje s *forward* selekcijom.

Prilagođava se $p-2$ logističkih modela koji sadrže varijable x_{e_1}, x_{e_2} i x_j, $j = 1, 2, ..., p$, $j \neq e_1, e_2$ te se računa log-vjerodostojnost svakog modela kao i test omjera vjerodostojnosti tog modela u odnosu na model koji sadrži samo x_{e_1} i x_{e_2} te se računaju pripadne p-vrijednosti. Ako je x_{e_3} varijabla s minimalnom p-vrijednosti, vrijedi

$$p^{(2)}_{e_3} = \min(p^{(2)}_j).$$

Ako je $p^{(2)}_{e_3} < p_E$ algoritam nastavlja na korak 3 dok u protivnom staje.

Korak 3.

Korak 3 je u potpunosti analogan koraku 2. Program prilagođava model koji sadrži i varijablu odabranu u koraku 2, provjerava treba li neku varijablu eliminirati te zatim provodi selekciju. Proces nastavlja tako do zadnjeg koraka koje smo označili sa S.

Korak S.

Korak S se događa kad vrijedi jedno od sljedećeg:

1) sve varijable su ušle u model
2) sve varijable u modelu imaju p-vrijednosti dobivene pri provjeri za eliminaciju $< p_R$ i sve varijable koje nisu u modelu imaju p-vrijednosti pri provjeri za selekciju $> p_E$.

U ovom trenutku model sadrži sve varijable koje su značajne na temelju p_E i p_R vrijednosti. [24]

Stepwise metoda je često korištena jer je brza i jeftina, ali problem je što uzima u obzir samo određeni broj modela. Također, kao rezultat daje jedan najbolji model, a pri odabiru modela u praksi najčešće ne postoji jedinstveni najbolji model. Zato je korisno doći do nekoliko modela koji se smatraju dobrima pa ih zatim međusobno usporediti. To možemo recimo tako da u *stepwise* metodi ne krenemo u koraku 0 od modela koji sadrži samo slobodni koeficijent nego od modela koji sadrži jednu ili više varijabli za koje znamo da su značajne za model. [10]

1.8.3 Usporedba dvaju modela

Dva modela možemo usporediti na više načina, a najčešće se to radi usporedbom vrijednosti R^2, prilagođenih R^2 (eng. *adjusted* R^2) i Akaikeovog informacijskog kriterija. Te vrijednosti za jedan odabrani model same po sebi nisu dovoljne da bismo u apsolutnom smislu zaključili je li model adekvatan, ali mogu biti korisne pri odlučivanju koji je model između dva ili više njih bolji. Također, pri uspoređivanju važno je obratiti pažnju na smislenost i interpretabilnost dobivenog modela te se na kraju ne moramo nužno odlučiti recimo za model koji ima najveći prilagođeni R^2 (ili najmanji AIC) ako postoji drugi model čiji prilagođeni R^2 (odnosno AIC) ima malo manju vrijednost (odnosno malo veću vrijednost), ali taj model ima više smisla. [10]

U praksi se često događa da želimo usporediti određeni model u odnosu na neki njegov podmodel u kojem smo izbacili jednu ili više varijabli. Drugim riječima, zanima nas je li potreban prošireni model ili je dovoljan podmodel. To najlakše možemo testirati pomoću testa omjera vjerodostojnosti. [24]

R^2

R^2 mjeri koliko je ukupne varijabilnosti objašnjeno modelom. Vrijednosti mogu biti između 0 i 1, a što je veća vrijednost, to je model bolji, iako kod logističke regresije te vrijednosti najčešće nisu toliko velike kao kod linearne regresije. Zbog toga ova statistika ima brojne inačice prilagođene za logističku regresiju koje se koriste u praksi.

Važno je naglasiti da se R^2 statistika može koristiti samo pri uspoređivanju dvaju modela s istim brojem nezavisnih varijabli jer će veći modeli uvijek imati veći R^2.

Ako s $D(X)$ označimo statistiku korištenu u (1.12) za testiranje omjera vjerodostojnosti modela X i saturiranog modela, a s $D(X_0)$ za testiranje omjera vjerodostojnosti modela koji sadrži samo slobodni koeficijent i saturiranog modela onda R^2 definiramo s

$$R^2 = \frac{D(X_0) - D(X)}{D(X_0)}. \qquad [10]$$

Prilagođeni R^2

Prilagođeni R^2 se može koristiti i pri uspoređivanju modela s različitim brojem nezavisnih varijabli jer ne „nagrađuje" automatski model s većim brojem varijabli. Veći prilagođeni R^2 znači da taj model zaista bolje opisuje podatke nego onaj s manjim prilagođenim R^2 no to ne mora nužno značiti da ima više nezavisnih varijabli što je slučaj ako imamo model s većim R^2 u odnosu na onaj s manjim R^2.

Uz oznake kao maloprije pri čemu je X model s p nezavisnih varijabli definiramo prilagođeni R^2 kao

$$adjusted\, R^2 = 1 - \frac{\frac{D(X)}{n-p-1}}{\frac{D(X_0)}{n-1}}. \qquad [10]$$

Akaikeov informacijski kriterij

Akaikeov informacijski kriterij (AIC, eng. *Akaike information criterion*) je još jedna mjera korisna pri uspoređivanju dvaju modela. Model se smatra boljim ako je AIC definiran na sljedeći način:

$$AIC = D(X) + 2(p + 1),$$

što manji. [10]

Test omjera vjerodostojnosti

Pretpostavimo da imamo trenutni multivarijabilni logistički model s p nezavisnih varijabli i neki njegov podmodel s r nezavisnih varijabli. Hipoteze koje testiramo su:

$$\begin{aligned} H_0 &: \text{ podmodel je dovoljan} \\ H_1 &: \text{ trenutni veći model je potreban.} \end{aligned} \qquad (1.28)$$

Uočimo da je to isto kao da testiramo hipoteze iz (1.11) samo ne želimo testirati jesu li svi koeficijenti $\beta_i = 0,\ i = 1, ...,\ p$ nego je li njih $r - p$ jednako 0. Računamo testnu statistiku

$$G = -2 \ln \left[\frac{\text{vjerodostojnost podmodela}}{\text{vjerodostojnost trenutnog većeg modela}} \right]$$

koja pod uvjetom da vrijedi hipoteza H_0 iz (1.28) ima $\chi^2(p - r)$ distribuciju. Ako je p-vrijednost dovoljno mala da odbacimo hipotezu H_0 zaključujemo da je trenutni veći model potreban. [34]

1.9 Procjena adekvatnosti modela (eng. *goodness-of-fit*)

Nakon što smo, na temelju statističke analize i logičkog razmišljanja, odabrali varijable za koje mislimo da bi trebale biti u modelu, preostaje procijeniti koliko je dobiveni model dobar, odnosno, koliko dobro opisuje zavisnu varijablu. Ako s $\boldsymbol{y} = (y_1, ..., y_n)^\tau$ označimo opaženi uzorak, a procijenjene vrijednosti s $\hat{\boldsymbol{y}} = (\hat{y_1}, ..., \hat{y_n})^\tau$ onda smatramo da je model dobar ako:

1) su $\boldsymbol{y}$ i $\hat{\boldsymbol{y}}$ ukupno vrlo malo udaljeni
2) je doprinos svakog para $(y_i, \hat{y_i}),\ i = 1, ..., n$ ukupnoj udaljenosti nesustavan i vrlo mali u odnosu na strukturu grešaka modela.

Prvo se bavimo statistikama koje računaju ukupnu razliku ili udaljenost između opaženih i procijenjenih vrijednosti. Ako dobijemo da je vrijednost neke takve statistike vrlo velika, tada znamo da model nije dobar. [24]

Parove $(y_i, \hat{y_i}),\ i = 1, ...,\ n$ možemo grafički prikazati tako da su na x-osi opažene, a na y-osi procijenjene vrijednosti. Ako je Y dihotomna varijabla kodirana s 0 i 1 onda bi za većinu procijenjenih vrijednosti oko 0 trebalo biti više opaženih vrijednosti jednakih 0, za većinu procijenjenih vrijednosti oko 1 više opaženih vrijednosti oko 1, a npr. za procijenjene vrijednosti oko 0.5 trebalo bi biti podjednako opaženih vrijednosti 0 i 1. [10]

Pretpostavimo sada da model sadrži p nezavisnih varijabli i označimo $\boldsymbol{x} = (x_1, ..., x_p)^\tau$. Neka J označava broj različitih opaženih vrijednosti koje može poprimiti $\boldsymbol{x}$. Na primjer, ako imamo u modelu samo varijable *spol* i *pušač* koje mogu poprimiti samo vrijednosti „muško" i „žensko", odnosno „da" i „ne", onda je $J \in \langle 1, 4 \rangle$, a ako u modelu imamo samo varijable *težina* i *visina* onda može biti $J \in \langle 1, n \rangle$. Pretpostavimo da je $J \approx n$ jer je to najčešće slučaj čim imamo bar jednu neprekidnu varijablu u modelu. Dalje, označimo s m_j broj subjekata takvih da je $\boldsymbol{x} = \boldsymbol{x}_j,\ j = 1, ..., J$. Slijedi $\sum_{j=1}^{J} m_j = n$. [24]

Testiramo hipoteze:

H_0 : model je adekvatan
H_1 : model nije adekvatan/neki drugi model je adekvatan.

Pearsonova χ^2-statistika

Uz oznake kao ranije, za procijenjene vrijednosti vrijedi

$$\hat{y}_j = m_j \hat{\pi}(\boldsymbol{x}_j)$$

iz čega slijedi da je Pearsonov rezidual

$$\begin{aligned} r(y_j, \hat{\pi}(\boldsymbol{x}_j)) &= \frac{\left(y_j - \hat{y}_j\right)}{\sqrt{\hat{y}_j\left(1 - \hat{\pi}(\boldsymbol{x}_j)\right)}} \\ &= \frac{\left(y_j - m_j\hat{\pi}(\boldsymbol{x}_j)\right)}{\sqrt{m_j\hat{\pi}(\boldsymbol{x}_j)\left(1 - \hat{\pi}(\boldsymbol{x}_j)\right)}} \end{aligned}$$

te Pearsonova χ^2-statistika

$$X^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}(\boldsymbol{x}_j))^2.$$

Ako je $J \approx n$ nailazimo na problem pri određivanju distribucije ove statistike, ali to možemo riješiti tako da zamislimo da imamo $2 \times J$ tablicu u kojoj prvi redak odgovara situaciji kad je $y = 1$, a drugi kad je $y = 0$, a u J stupaca se nalazi J različitih opaženih vrijednosti. Grupiramo podatke tako da dobijemo fiksan broj stupaca te zatim računamo opažene i očekivane frekvencije. Fiksiranjem broja stupaca očekivane frekvencije postaju velike ako je n dovoljno velik, a tada je i svaki m_j dovoljno velik te vrijedi $J < n$. Tada, pod pretpostavkom da je prilagođeni model točan vrijedi $X^2 \sim \chi^2(J - p - 1)$. [24]

Devijanca

Devijantni rezidual definira se kao

$$d(y_j, \hat{\pi}(\boldsymbol{x}_j)) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}(\boldsymbol{x}_j)}\right) + (m_j - y_j)\ln\left(\frac{(m_j - y_j)}{m_j(1 - \hat{\pi}(\boldsymbol{x}_j))}\right)\right]\right\}^{\frac{1}{2}}$$

pri čemu je predznak + ili - isti kao predznak izraza $y_j - m_j\hat{\pi}(\boldsymbol{x}_j)$.

Devijanca je tada

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}(\boldsymbol{x}_j))^2.$$

Kao i kod Pearsonove χ^2-statistike, pod pretpostavkom da je model točan i da je $J < n$ vrijedi $D \sim \chi^2(J - p - 1)$. [24]

Površina ispod ROC krivulje

Pojmove specifičnost i senzitivnost ili osjetljivost najbolje možemo opisati na medicinskom primjeru testa kojim želimo dijagnosticirati je li prisutna određena bolest. Analognim zaključivanjem jasno je što znače ti pojmovi na primjeru bilo kojeg testa ili modela. Promotrimo sljedeću tablicu:

	prisutna bolest	nije prisutna bolest
pozitivan test	TP	FP
negativan test	FN	TN

Tablica 1.1: Specifičnost i senzitivnost

pri čemu smo s TP (eng. *true positive*) označili stvarno pozitivne, dakle, one kojima je test pozitivan i imaju bolest, s FP (eng. *false positive*) lažno pozitivne (pozitivan test, ali nemaju bolest), s FN (eng. *false negative*) lažno negativne (negativan test, ali imaju bolest) te s TN (eng. *true negative*) stvarno negativne (negativan test i nemaju bolest).

Sada specifičnost (eng. *specificity, true positive rate*) možemo definirati kao sposobnost testa da točno prepozna stvarno pozitivne (TP), odnosno, kao vjerojatnost da je test pozitivan kad je prisutna bolest:

$$specifičnost = \frac{TP}{TP + FN}.$$

Slično, senzitivnost (eng. *sensitivity, true negative rate*) definiramo kao sposobnost testa da točno prepozna stvarno negativne (TN), odnosno, kao vjerojatnost da je test negativan kad zaista nije prisutna bolest:

$$senzitivnost = \frac{TN}{TN + FP}. \quad [32]$$

Želimo i da senzitivnost bude što veća, ali i da specifičnost bude što veća međutim, ta dva zahtjeva su često u suprotnosti. Senzitivnost i specifičnost najčešće grafički prikazujujemo tako da na x-os stavimo vrijednost razlike $1 - specifičnost = \frac{FN}{TP+FN}$ koja označava udio

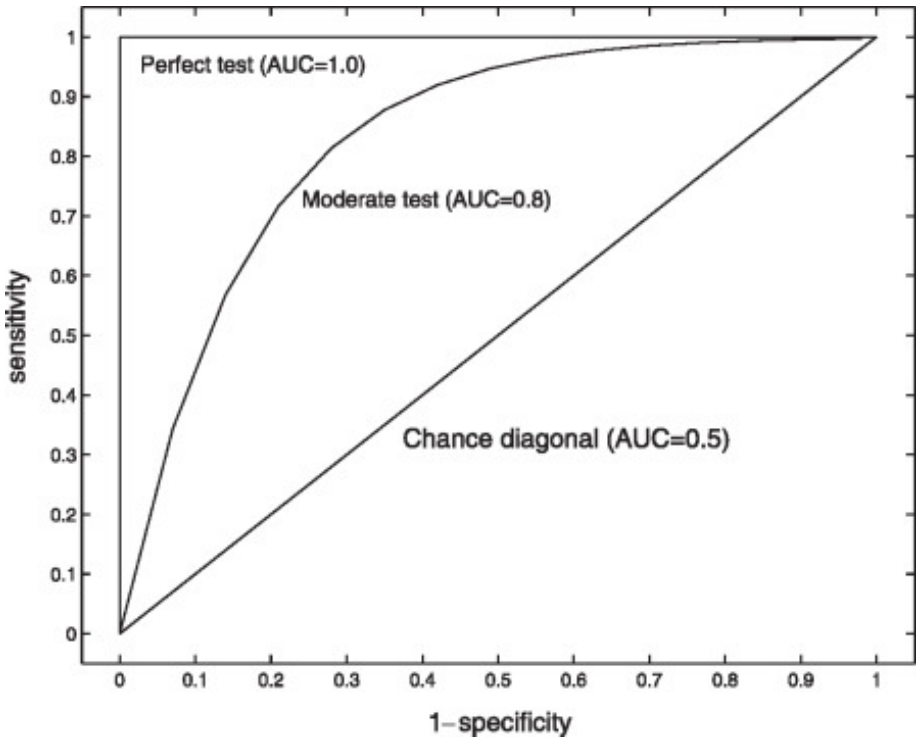
lažno negativnih među bolesnima (eng. *false positive rate*), a na y-os senzitivnost te crtamo ROC krivulju (eng. *receiver operating characteristic*). ROC krivulja u logističkoj regresiji služi kako bi se odredila najbolja granica za predviđanje vrijedi li za novo opažanje ishod $y = 1$ ili $y = 0$. Ta granica, npr. 0.5, služi da se svako opažanje s predviđenom vrijednosti većom ili jednakom od nje klasificira kao $y = 1$, a s manjom kao $y = 0$. Naravno, to ne znači da za to opažanje stvarno vrijedi ishod 1, odnosno 0 nego samo da je klasifikacija takva. Površina ispod ROC krivulje, koju često označavamo s AUC (eng. *area under the curve*), je veličina koja mjeri koliko dobro model logističke regresije klasificira pozitivne i negativne ishode za sve moguće granice.

Površina ispod ROC krivulje može biti od 0 do 1 i označava sposobnost razlikovanja ishoda 0 i 1 klasifikacijskog modela. Poželjno je da je ona što veća, a općenito vrijedi pravilo:

$AUC = 0.5$	model nije dobar
$AUC \in [0.7, 0.8\rangle$	model je prihvatljiv
$AUC \in [0.8, 0.9\rangle$	model je odličan
$AUC \geq 0.9$	model je izvanredan.

$AUC = 0.5$ je kao da model slučajnim odabirom klasificira subjekte za koje procjenjuje da je $y = 1$, odnosno za koje da je $y = 0$. Taj model nam ništa ne govori – možemo umjesto toga samo baciti simetričan novčić pa na temelju toga koja strana padne donijeti klasifikacijsku odluku. [24]

Sljedeća slika prikazuje tri ROC krivulje s pripadnim AUC vrijednostima. Kao što smo već rekli i kao što je označeno, veća površina ispod krivulje znači da je model bolji.



Slika 1.1: ROC krivulje i AUC vrijednosti tri različita testa

preuzeto s https://www.sciencedirect.com/topics/nursing-and-health-professions/receiver-operating-characteristic

Poglavlje 2

Primjene

Nakon što smo objasnili što je logistički regresijski model, što kada imamo nepotpune podatke, kako se biraju varijable koje ulaze u model i kako se procjenjuje adekvatnost modela, sada ćemo to detaljnije pokazati na dva primjera. Oba primjera se odnose na podatke s [12]. Prvi skup podataka odnosi se na zatajenje srca i može se naći na [21] te je korišten u članku [9] dok se drugi može naći na [19] i u članku [35].

2.1 Zatajenje srca

2.1.1 Opis problema i podataka

Prema [9] kardiovaskularne bolesti su uzrok smrti otprilike 17 milijuna ljudi godišnje, a najčešće među njima su infarkt miokarda i zatajenje srca (eng. *heart failure*) koje se događa kad srce ne može pumpati krvi koliko je tijelu potrebno. Zatajenje srca najčešće se dijeli na dvije grupe: sistoličko i dijastoličko zatajenje pri čemu do sistoličkog zatajenja dolazi kad je ejekcijska frakcija, odnosno postotak krvi koji izlazi iz srca pri jednoj kontrakciji, manji od 40%. Težina srčanog zatajenja značajno varira i obično se klasificira prema sustavu NYHA-a (*New York Heart Association*) na stupnjeve I–IV, a što je veći stupanj stanje je ozbiljnije.

Između travnja i prosinca 2015. u Pakistanu su prikupljani podaci o 299 pacijenata koji su imali sistoličko zatajenje srca te su već prije imali srčana zatajenja stupnja III ili IV. Promatrana obilježja se većinom odnose na dob, spol, krvnu sliku, životne navike i neke druge bolesti. Imamo ukupno 13 varijabli od kojih je 6 kvalitativnih, a 7 kvantitativnih. Nema vrijednosti koje nedostaju. Pacijenti su praćeni kroz različito dug period, a prosječno 130 dana. U trenutku prestanka praćenja, smrtni ishod se nije dogodio u 203 slučaja (kodirano s 0), a u 66 je (kodirano s 1). To će biti zavisna varijala modela logističke regresije.

Kvalitativne varijable su dihotomne i kodirane s 0 i 1:

- **anemija** (*anaemia*)
  - označava ima li osoba smanjen broj eritrocita ili hemoglobina [5]
- **visoki krvni tlak** (*high blood pressure*)
  - označava ima li osoba visoki krvni tlak
- **dijabetes** (*diabetes*)
  - označava ima li osoba dijabetes
- **spol** (*sex*)
  - označava spol osobe
  - 0 – žene
    1 – muškarci
- **pušenje** (*smoking*)
  - označava puši li osoba
- **smrtni ishod** (*death event*)
  - označava je li osoba umrla tijekom perioda praćenja.

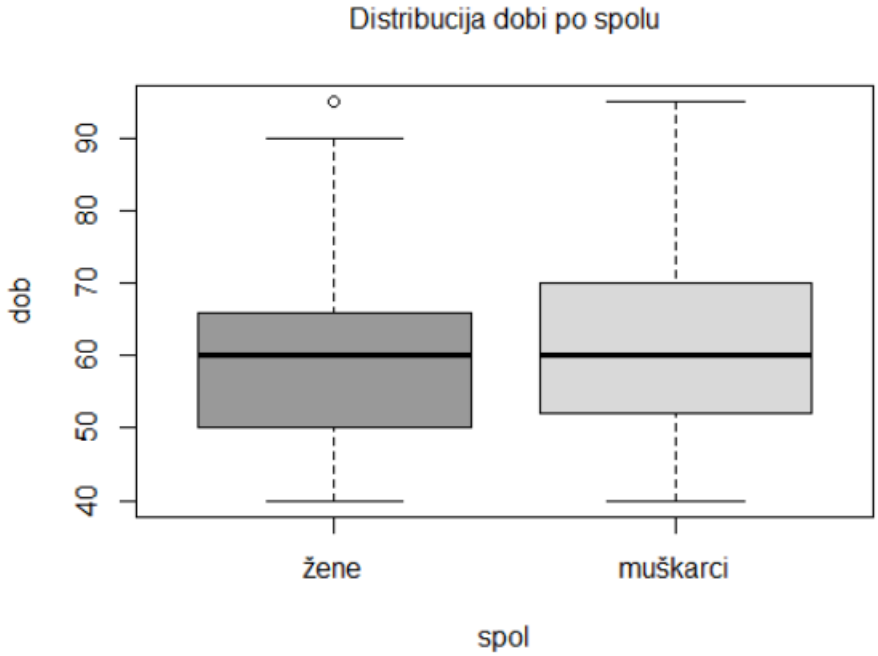
Kvantitativne varijable su sljedeće:

- **dob** (*age*)
  - označava dob pacijenta u godinama
- **kreatin fosfokinaza** (CPK, *creatinine phosphokinase*)
  - označava količinu enzima CPK u krvi u μg/L
- **ejekcijska frakcija** (*ejection fraction*)
  - označava postotak krvi koja napušta srce pri svakoj kontrakciji
  - normalno bi trebala biti između 50% i 75%
- **trombociti** (*platelets*)
  - označava količinu trombocita u mL krvi
- **kreatinin u serumu** (*serum creatinine*)
  - označava razinu kreatinina u krvi u mg/dL
- **natrij u serumu** (*serum sodium*)
  - označava količinu natrija u krvi u mEq/L
- **vrijeme praćenja** (*time, follow-up period*)
  - označava vrijeme praćenja osobe u danima. [9]

2.1.2 Odabir modela

Deskriptivna statistika

Promotrimo prvo distribuciju pacijenata po dobi i spolu.



Slika 2.1: Distribucija pacijenata po dobi i spolu

Vidimo da su muškarci koji su sudjelovali u ovom istraživanju ukupno nešto stariji nego žene, iako je medijan jednak i iznosi 60 godina. Kod muškaraca je nešto veća varijabilnost te ih najviše spada u kategoriju između 52 i 70 godina. Najviše žena je dobi između 50 i 66 godina. Imamo nekoliko žena i muškaraca koji imaju 40 godina te su oni najmlađi koji su sudjelovali u ovom istraživanju. Najstariji su muškarac i žena s 95 godina, međutim, kod žena je ta vrijednost dobi *outlier* odnosno, značajno je veća u odnosu na ostale žene.

Promotrimo prvo aritmetičku sredinu i standardnu devijaciju te karakterističnu petorku za svaku kvantitativnu varijablu uzorka, a zatim za svaku kvalitativnu varijablu pa tako i za zavisnu odredimo frekvenciju i postotak po kategorijama. Rezultati su dani u sljedećim tablicama:

varijabla	aritmetička sredina	standardna devijacija	mini-mum	donji kvartil	medijan	gornji kvartil	maksi-mum
dob	60.8339	11.8948	40	51	60	70	95
CPK	581.840	970.286	23.0	116.5	250.0	582.0	7861.0
ejekcijska frakcija	38.0836	11.8348	14	30	38	45	80
trombociti	263358	97804.2	25100	212500	262000	303500	850000
kreatinin	1.39388	1.03451	0.5	0.9	1.1	1.4	9.4
natrij	136.625	4.41248	113	134	137	140	148
praćenje	130.261	77.6142	4	73	115	203	285

Tablica 2.1: Deskriptivna analiza kvantitativnih varijabli

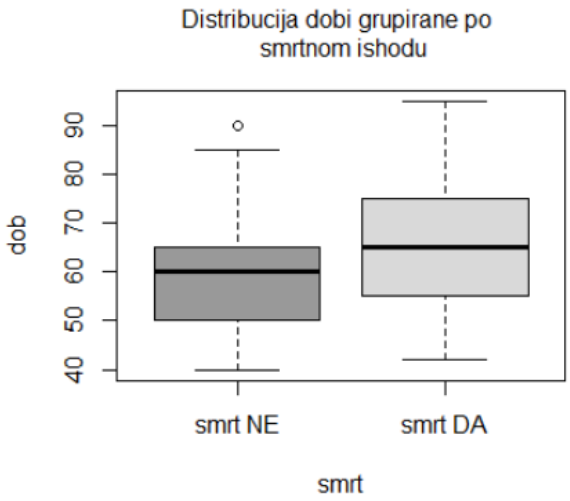
varijabla	kategorija	frekvencija	%
anemija	nema	170	56.86%
	ima	129	43.14%
dijabetes	nema	174	58.19%
	ima	125	41.81%
visoki tlak	nema	194	64.88%
	ima	105	35.12%
spol	žene	105	35.12%
	muškarci	194	64.88%
pušenje	ne	203	67.89%
	da	96	32.11%

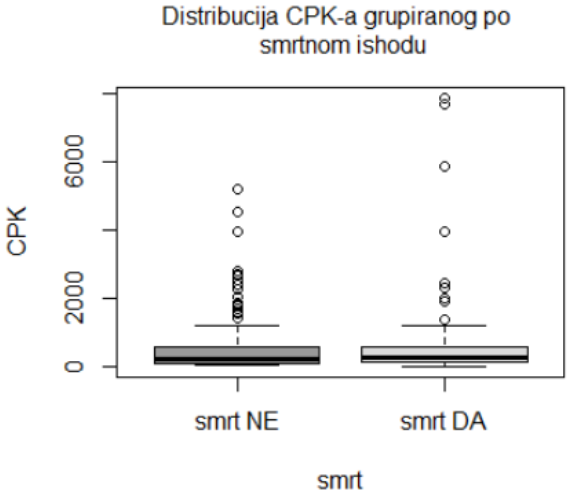
Tablica 2.2: Deskriptivna analiza kvalitativnih varijabli

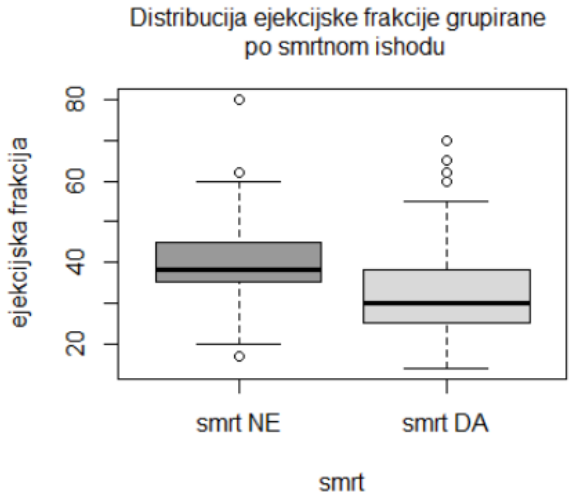
varijabla	kategorija	frekvencija	%
smrt	ne	203	67.89%
	da	96	32.11%

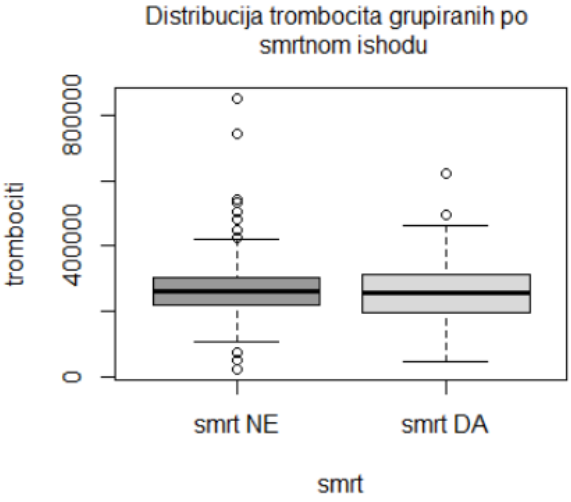
Tablica 2.3: Deskriptivna analiza zavisne varijable

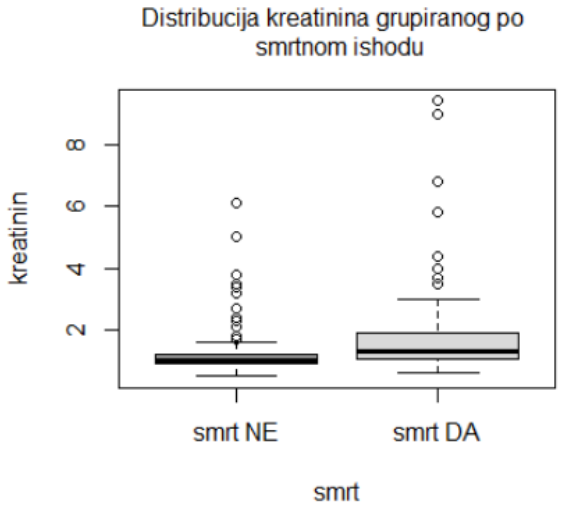
S obzirom da nam je cilj odrediti vezu između zavisne varijable *smrt* i navedenih nezavisnih varijabli, možemo grafički prikazati distribuciju svake kvantitativne varijable grupirane s obzirom na to je li se dogodio smrtni ishod:

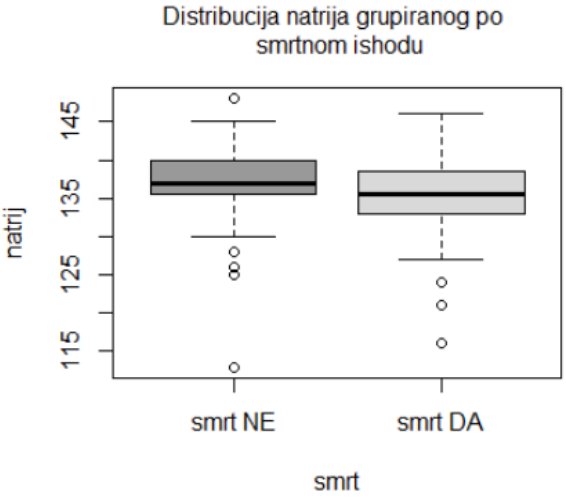

Distribucija dobi grupirane po smrtnom ishodu
dob
40
50
60
70
80
90
smrt NE
smrt DA
smrt

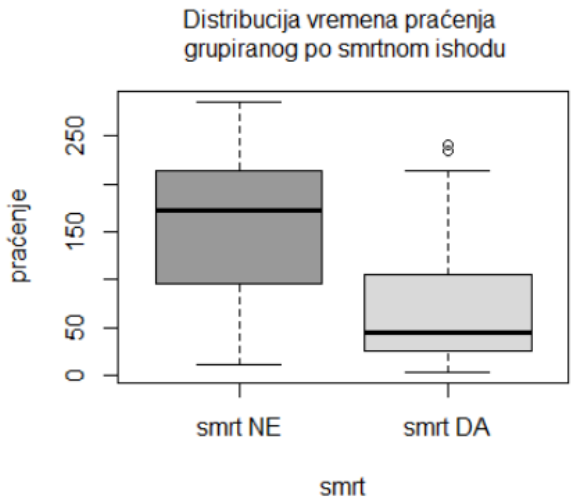

Distribucija CPK-a grupiranog po smrtnom ishodu
CPK
0
2000
6000
smrt NE
smrt DA
smrt


Distribucija ejekcijske frakcije grupirane po smrtnom ishodu
ejekcijska frakcija
20
40
60
80
smrt NE
smrt DA
smrt


Distribucija trombocita grupiranih po smrtnom ishodu
trombociti
0
400000
800000
smrt NE
smrt DA
smrt


Distribucija kreatinina grupiranog po smrtnom ishodu
kreatinin
2
4
6
8
smrt NE
smrt DA
smrt


Distribucija natrija grupiranog po smrtnom ishodu
natrij
115
125
135
145
smrt NE
smrt DA
smrt



Slika 2.2: Distribucija svake pojedine kvantitativne varijable grupirane prema smrtnom ishodu

Vidimo da je medijan u obje grupe skoro jednak na grafičkim prikazima distribucija *CPK-a* i *trombocita*, ali se ni kod ostalih varijabli ne razlikuje toliko puno osim kod varijable *praćenje* gdje je medijan puno veći kod osoba koje nisu umrle. Kad se dogodila smrt, veća varijabilnost je kod varijabli *dob*, *ejekcijska frakcija* i *kreatinin* nego kad se nije dogodila smrt dok je za varijablu *praćenje* situacija obrnuta. Kod ostalih varijabli je ta razlika vrlo mala. Kod varijable *dob* imamo samo jedan *outlier*, kod varijable *praćenje* dva, a kod ostalih varijabli ih imamo više, pogotovo kod varijabli *CPK* i *kreatinin*. Također, kod te dvije varijable kao i kod *praćenja* i *dobi* oni se nalaze samo s gornje strane što znači da imamo vrijednosti koje su značajnije povišene čak i do nekoliko puta u odnosu na prosjek kao što je slučaj kod *CPK-a* i *kreatinina*. Kod *dobi* imamo samo jednu povišenu vrijednost kad se nije dogodila smrt, a kod *praćenja* su obje kad se dogodila smrt. Suprotno, kod *natrija* imamo *outliere* s donje strane uz iznimku jednog *outliera* s gornje strane kad se nije dogodila smrt. Kod varijabli *ejekcijska frakcija* i *trombociti* kad se nije dogodila smrt imamo vrijednosti koje više odstupaju od ostalih podataka i s gornje i s donje strane, a kad se dogodila smrt te vrijednosti odstupaju samo s gornje strane.

Univarijabilna logistička regresija

Rezultati univarijabilne logističke regresije za kvantitativne varijable dani su u sljedećoj tablici pri čemu su vrijednosti Waldovog testa izračunate po formuli (1.14) u R-u označene kao *z value*:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
dob	0.04695	0.01107	4.241	$2.2 \cdot 10^{-5}$	[0.02525, 0.06865]
CPK	0.00013	0.00012	1.065	0.28700	[-0.00011, 0.00037]
ejekcijska frakcija	-0.05620	0.01258	-4.468	$7.9 \cdot 10^{-6}$	[-0.08085, -0.03155]
trombociti	$-1.1 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$	-0.847	0.39700	[-0.000004, 0.000001]
kreatinin	0.82420	0.19720	4.180	$2.9 \cdot 10^{-5}$	[0.43775, 1.21062]
natrij	-0.09639	0.02989	-3.224	0.00126	[-0.15498, -0.03779]
praćenje	-0.01995	0.00256	-7.804	$6.0 \cdot 10^{-15}$	[-0.02495, -0.01494]

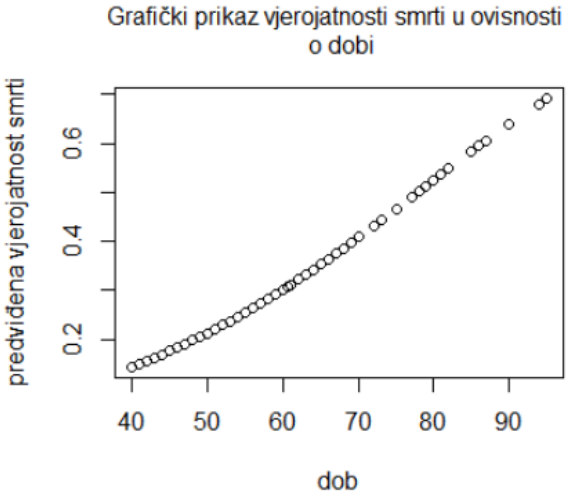
Tablica 2.4: Univarijabilna logistička regresija s kvantitativnim varijablama

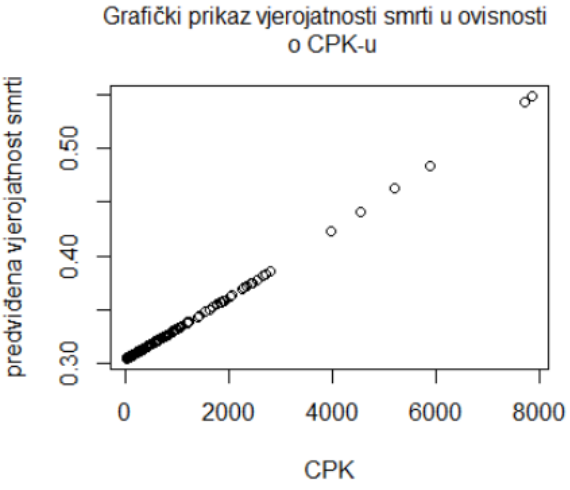
Testovima omjera vjerodostojnosti računamo značajnost procijenjenih koeficijenata u univarijabilnom modelu u odnosu na model bez varijabli. Vrijednost log-vjerodostojnosti modela bez varijabli je -187.67, a ostale vrijednosti su dane u sljedećoj tablici:

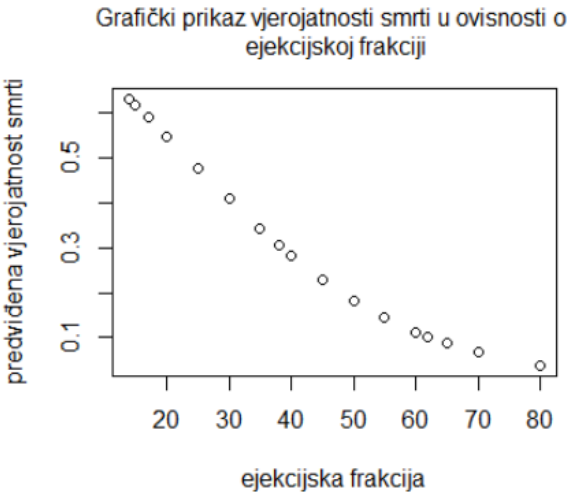
varijabla u modelu	log-vjerodostojnost modela univarijabilne regresije	testna statistika	p-vrijednost
dob	-178.00	19.356	$1.09 \cdot 10^{-5}$
CPK	-187.12	1.118	0.29030
ejekcijska frakcija	-175.98	23.381	$1.33 \cdot 10^{-6}$
trombociti	-187.31	0.738	0.39040
kreatinin	-173.63	28.097	$1.15 \cdot 10^{-7}$
natrij	-182.01	11.327	0.00076
praćenje	-139.54	96.275	$< 2.2 \cdot 10^{-16}$

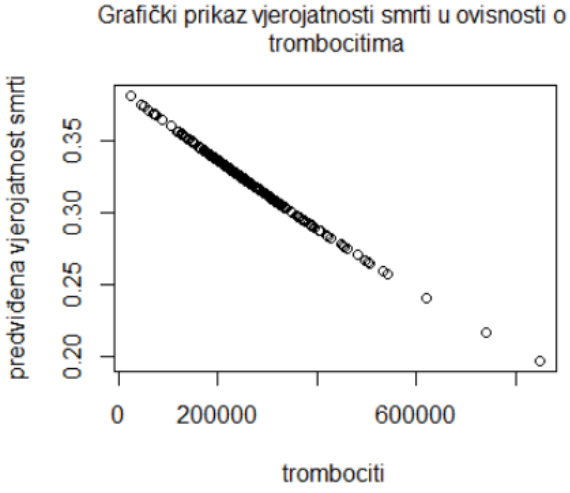
Tablica 2.5: Testovi omjera vjerodostojnosti za kvantitativne varijable

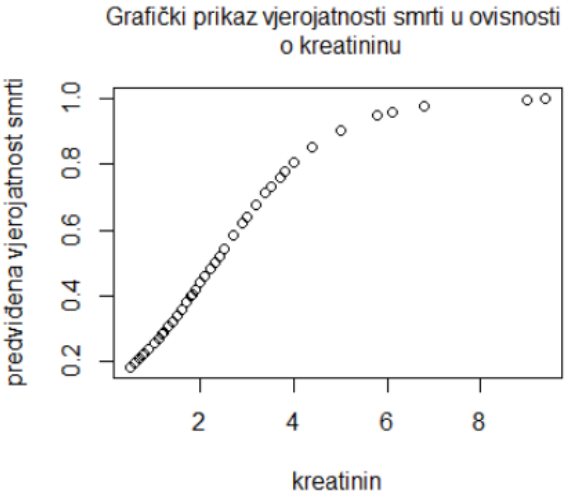
Gledamo koje varijable imaju p-vrijednost manju od 0.25 kao što je preporučeno u pododjeljku 1.8.1. To su varijable *dob*, *ejekcijska frakcija*, *kreatinin*, *natrij* i *praćenje* te ćemo njih uzimati u obzir pri daljnoj analizi. Vjerojatnost smrti procijenjene u svakom pojedinom modelu možemo grafički prikazati:

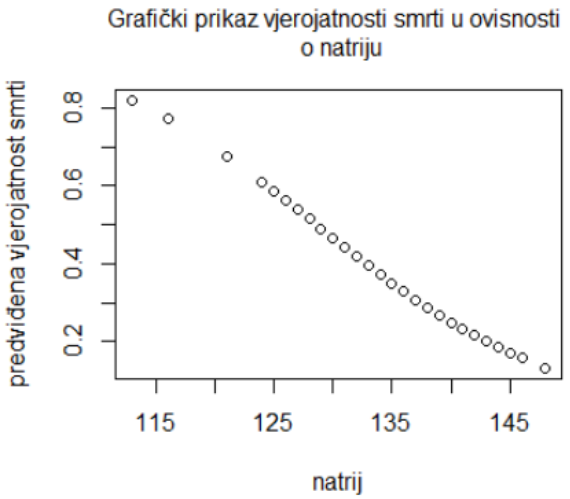

Grafički prikaz vjerojatnosti smrti u ovisnosti o dobi
predviđena vjerojatnost smrti
0.2
0.4
0.6
40
50
60
70
80
90
dob

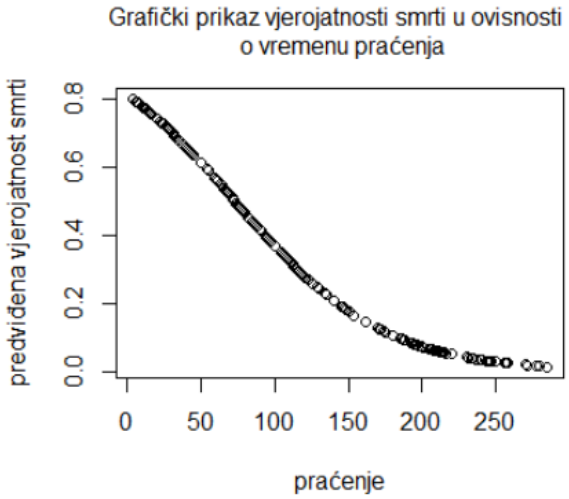

Grafički prikaz vjerojatnosti smrti u ovisnosti o CPK-u
predviđena vjerojatnost smrti
0.30
0.40
0.50
0
2000
4000
6000
8000
CPK


Grafički prikaz vjerojatnosti smrti u ovisnosti o ejekcijskoj frakciji
predviđena vjerojatnost smrti
0.1
0.3
0.5
20
30
40
50
60
70
80
ejekcijska frakcija


Grafički prikaz vjerojatnosti smrti u ovisnosti o trombocitima
predviđena vjerojatnost smrti
0.20
0.25
0.30
0.35
0
200000
600000
trombociti


Grafički prikaz vjerojatnosti smrti u ovisnosti o kreatininu
predviđena vjerojatnost smrti
0.2
0.4
0.6
0.8
1.0
2
4
6
8
kreatinin


Grafički prikaz vjerojatnosti smrti u ovisnosti o natriju
predviđena vjerojatnost smrti
0.2
0.4
0.6
0.8
115
125
135
145
natrij



Slika 2.3: Grafički prikaz vjerojatnosti smrti u ovisnosti o kvantitativnim varijablama

Iz grafičkih prikaza se također može naslutiti koje varijable su značajne. Npr. vidimo da varijabla *CPK* nije značajna u univarijabilnom modelu jer se vrijednost *CPK-a* mora puno više povećati da bi se vjerojatnost smrti značajnije promijenila dok je varijabla *dob* značajna jer se vjerojatnost smrti puno više promijeni i za relativno manji porast dobi. Iz negativnih predznaka procijenjenih koeficijenta iz tablice 2.4 kao i iz padajućih krivulja na slici 2.3, vidimo da se vjerojatnost smrti smanjuje povećanjem *ejekcijske frakcije*, *trombocita*, *natrija* ili *vremena praćenja*, a iz pozitivnih koeficijenata i rastućih krivulja možemo zaključiti da vjerojatnost smrti raste povećanjem *dobi*, *CPK-a* ili *kreatinina*.

Iz tablice 2.4 možemo izračunati procjene omjera izgleda i pripadne procjene 95% pouzdanih intervala za varijable koje su ispale statistički značajne:

varijabla	procjena omjera izgleda	procjena 95% pouzdanog intervala za omjer izgleda
dob	1.04807	[1.02557, 1.07106]
ejekcijska frakcija	0.94535	[0.92233, 0.96894]
kreatinin	2.28002	[1.54922, 3.35557]
natrij	0.90811	[0.85643, 0.96291]
praćenje	0.98025	[0.97535, 0.98517]

Tablica 2.6: Procjena omjera izgleda za statistički značajne kvantitativne varijable

Vidimo da se, gledajući modele univarijabilne logističke regresije, kad se:
- *dob* poveća za 1 godinu, omjer izgleda da se dogodi smrt poveća 1.04807 puta
- *ejekcijska frakcija* poveća za 1%, omjer izgleda da se dogodi smrt smanji 0.94535 puta
- *kreatinin* poveća za 1 mg/dL, omjer izgleda da se dogodi smrt poveća 2.28002 puta
- *natrij* poveća za 1 mEq/L, omjer izgleda da se dogodi smrt smanji 0.90811 puta
- *praćenje* poveća za 1 dan, omjer izgleda da se dogodi smrt smanji 0.98025 puta.

Zatim, za kvalitativne varijable radimo kontingencijske tablice:

		nema anemiju	ima anemiju	suma
smrt NE	frekvencija	120	83	203
	%	40.1%	27.8%	67.9%
	% u redu	59.1%	40.9%	-
	% u stupcu	70.6%	64.3%	-
smrt DA	frekvencija	50	46	96
	%	16.7%	15.4%	32.1%
	% u redu	52.1%	47.9%	-
	% u stupcu	29.4%	35.7%	-
suma	frekvencija	170	129	299
	%	56.9%	43.1%	100.0%

Tablica 2.7: Kontingencijska tablica za varijable *anemija* i *smrt*

		nema dijabetes	ima dijabetes	suma
smrt NE	frekvencija	118	85	203
	%	39.5%	28.4%	67.9%
	% u redu	58.1%	41.9%	-
	% u stupcu	67.8%	68.0%	-
smrt DA	frekvencija	56	40	96
	%	18.7%	13.4%	32.1%
	% u redu	58.3%	41.7%	-
	% u stupcu	32.2%	32.0%	-
suma	frekvencija	174	125	299
	%	58.2%	41.8%	100.0%

Tablica 2.8: Kontingencijska tablica za varijable *dijabetes* i *smrt*

		nema visoki tlak	ima visoki tlak	suma
smrt NE	frekvencija	137	66	203
	%	45.8%	22.1%	67.9%
	% u redu	67.5%	32.5%	-
	% u stupcu	70.6%	62.9%	-
smrt DA	frekvencija	57	39	96
	%	19.1%	13.0%	32.1%
	% u redu	59.4%	40.6%	-
	% u stupcu	29.4%	37.1%	-
suma	frekvencija	194	105	299
	%	64.9%	35.1%	100.0%

Tablica 2.9: Kontingencijska tablica za varijable *visoki tlak* i *smrt*

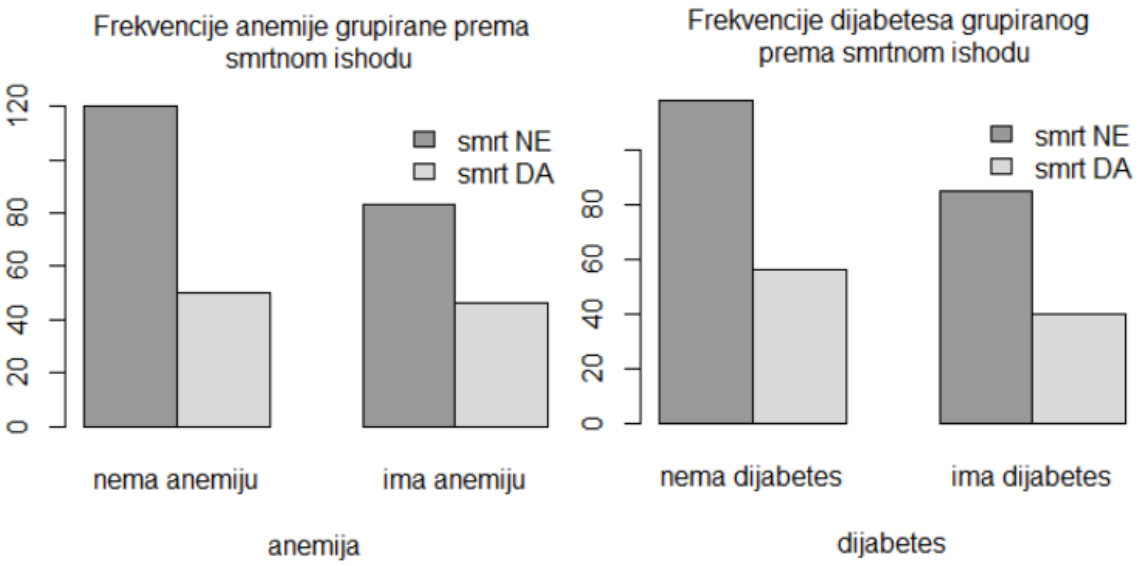
		žene	muškarci	suma
smrt NE	frekvencija	71	132	203
	%	23.7%	44.1%	67.9%
	% u redu	35.0%	65.0%	-
	% u stupcu	67.6%	68.0%	-
smrt DA	frekvencija	34	62	96
	%	11.4%	20.7%	32.1%
	% u redu	35.4%	64.6%	-
	% u stupcu	32.4%	32.0%	-
suma	frekvencija	105	194	299
	%	35.1%	64.9%	100.0%

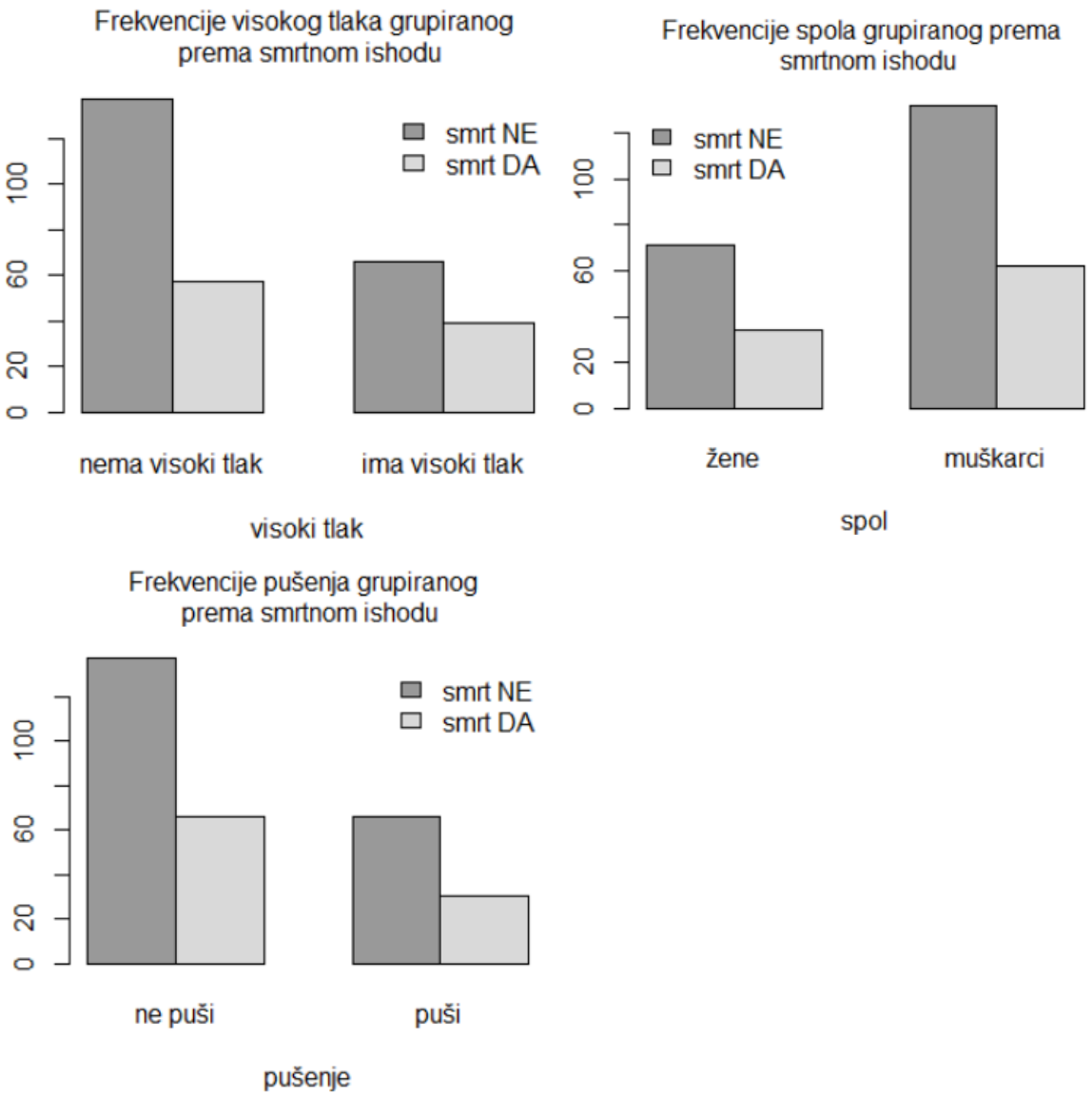
Tablica 2.10: Kontingencijska tablica za varijable *spol* i *smrt*

		ne puši	puši	suma
smrt NE	frekvencija	137	66	203
	%	45.8%	22.1%	67.9%
	% u redu	67.5%	32.5%	-
	% u stupcu	67.5%	68.8%	-
smrt DA	frekvencija	66	30	96
	%	22.1%	10.0%	32.1%
	% u redu	68.8%	31.2%	-
	% u stupcu	32.5%	31.2%	-
suma	frekvencija	203	96	299
	%	67.9%	32.1%	100.0%

Tablica 2.11: Kontingencijska tablica za varijable *pušenje* i *smrt*

Uočavamo da ni u jednoj ćeliji nemamo 0. Dobivene frekvencije možemo grafički prikazati:





Slika 2.4: Grafički prikaz frekvencija kvalitativnih varijabli grupiranih prema smrtnom ishodu

Osim što imamo više podataka o osobama koje nemaju anemiju, vidimo da je proporcionalno puno manja razlika u broju umrlih i preživjelih kod osoba koje imaju anemiju. Slična situacija je i za *visoki tlak* dok za ostale varijable to nije toliko izraženo gledajući proporcionalno. Više podataka imamo za osobe koje nemaju visoki tlak nego koje imaju, koje nemaju dijabetes nego koje imaju, koje ne puše nego koje puše te za muškarce u odnosu na žene.

Računamo procjene koeficijenata, standardne greške, procjene 95% pouzdanog intervala te vrijednosti Waldove statistike uz pripadne p-vrijednosti te rezultate prikazujemo u sljedećoj tablici:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
anemija1	0.28530	0.24920	1.145	0.252	[-0.20323, 0.77377]
dijabetes1	-0.00844	0.25119	-0.034	0.973	[-0.50076, 0.48388]
visoki tlak1	0.35080	0.25620	1.369	0.171	[-0.15130, 0.85297]
spol1	-0.01935	0.25923	-0.075	0.941	[-0.52743, 0.48874]
pušenje1	-0.05813	0.26634	-0.218	0.827	[-0.58014, 0.46388]

Tablica 2.12: Univarijabilna logistička regresija s kvalitativnim varijablama

Kvalitativne varijable su modelirane pomoću *dummy* varijabli kao u (1.25) na način da su nazvane odgovarajućim imenom varijable i oznakom kategorije.

Kao i za kvantitativne varijable, testovima omjera vjerodostojnosti računamo značajnost procijenjenih koeficijenata u univarijabilnim modelima u odnosu na model bez varijabli. Rezultati su dani u sljedećoj tablici:

varijabla u modelu	log-vjerodostojnost modela univarijabilne regresije	testna statistika	p-vrijednost
anemija	-187.02	1.3086	0.2527
dijabetes	-187.67	0.0011	0.9732
visoki tlak	-186.74	1.8630	0.1723
spol	-187.67	0.0056	0.9405
pušenje	-187.65	0.0478	0.8270

Tablica 2.13: Testovi omjera vjerodostojnosti za kvalitativne varijable

Vidimo da je statistički značajna još samo varijabla *visoki tlak* kad gledamo p-vrijednosti manje od 0.25. Za nju procjena omjera izgleda iznosi 1.42026 što znači da se omjer izgleda

da se dogodi smrt poveća 1.42026 puta kod osobe koja ima visoki tlak. Pripadna procjena 95% pouzdanog intervala je [0.85959, 2.34660]. Primijetimo da taj interval sadrži 1. To je zato što ta varijabla nije statistički značajna na razini značajnosti od 5%.

Dakle, univarijabilnom regresijom smo dobili da su značajne varijable *dob*, *ejekcijska frakcija*, *kreatinin*, *natrij*, *praćenje* i *visoki tlak* te su to varijable s kojima nastavljamo daljnju analizu.

Multivarijabilna logistička regresija *stepwise* metodom

Krećemo od praznog modela i uzimajući u obzir sve prethodno navedene statistički značajne varijable tražimo najbolji model. Dobijemo sljedeće rezultate:

Korak 0.

Algoritam kreće od modela bez varijabli i gleda koju bi varijablu prvu dodao u model. AIC trenutnog modela je 377.35.

moguće dodati ili izbaciti	varijabla	stupnjevi slobode	devijanca	AIC	test omjera vjerodostojnosti	p-vrijednost
+	praćenje	1	279.07	283.07	96.275	$< 2.2 \cdot 10^{-16}$
+	kreatinin	1	347.25	351.25	28.097	$1.15 \cdot 10^{-7}$
+	ejekcijska frakcija	1	351.97	355.97	23.381	$1.33 \cdot 10^{-6}$
+	dob	1	355.99	359.99	19.356	$1.09 \cdot 10^{-5}$
+	natrij	1	364.02	368.02	11.327	0.00076
	bez varijabli		375.35	377.35		
+	visoki tlak	1	373.49	377.49	1.863	0.17228

Tablica 2.14: *Stepwise* metoda – korak 0.

Korak 1.

Varijabla *praćenje* je statistički najznačajnija pa prva ulazi u model. Algoritam traži koju od preostalih varijabli treba sljedeću dodati u model. AIC trenutnog modela je 283.07.

moguće dodati ili izbaciti	varijabla	stupnjevi slobode	devijanca	AIC	test omjera vjerodostojnosti	p-vrijednost
+	ejekcijska frakcija	1	256.08	262.08	22.990	$1.63 \cdot 10^{-6}$

+	kreatinin	1	259.64	265.64	19.434	$1.04 \cdot 10^{-5}$
+	natrij	1	269.83	275.83	9.242	0.00237
+	dob	1	271.46	277.46	7.610	0.00581
	bez varijabli		279.07	283.07		
+	visoki tlak	1	278.96	284.96	0.117	0.73280
-	praćenje	1	375.35	377.35	96.275	$< 2.2 \cdot 10^{-16}$

Tablica 2.15: *Stepwise* metoda – korak 1.

Korak 2.

Varijabla *ejekcijska frakcija* se kao statistički najznačajnija ubacuje u model dok istovremeno varijabla *praćenje* i dalje ostaje u modelu jer je i dalje statistički značajna. Algoritam gleda bi li izbacio neku od njih, a nakon toga i koju od preostalih varijabli bi dodao u model. AIC trenutnog modela je 262.08.

moguće dodati ili izbaciti	varijabla	stupnjevi slobode	devijanca	AIC	test omjera vjerodostojnosti	p-vrijednost
+	kreatinin	1	235.41	243.41	20.670	$5.46 \cdot 10^{-6}$
+	dob	1	244.51	252.51	11.575	0.00067
+	natrij	1	249.73	257.73	6.354	0.01171
	bez varijabli		256.08	262.08		
+	visoki tlak	1	255.93	263.93	0.157	0.69180
-	ejekcijska frakcija	1	279.07	283.07	22.990	$1.63 \cdot 10^{-6}$
-	praćenje	1	351.97	355.97	95.884	$< 2.2 \cdot 10^{-16}$

Tablica 2.16: *Stepwise* metoda – korak 2.

Korak 3.

Kreatinin sljedeći ulazi u model, a nijedna od prethodno dodanih varijabli ne izlazi. Traži se koja će sljedeća varijabla ući u model. AIC modela je 243.41.

moguće dodati ili izbaciti	varijabla	stupnjevi slobode	devijanca	AIC	test omjera vjerodostojnosti	p-vrijednost
+	dob	1	226.30	236.30	9.113	0.00254
+	natrij	1	232.02	242.02	3.398	0.06529
	bez varijabli		235.41	243.41		

+	visoki tlak	1	235.41	245.41	0.006	0.94008
-	kreatinin	1	256.08	262.08	20.670	$5.46 \cdot 10^{-6}$
-	ejekcijska frakcija	1	259.64	265.64	24.226	$8.57 \cdot 10^{-7}$
-	praćenje	1	324.32	330.32	88.907	$< 2.2 \cdot 10^{-16}$

Tablica 2.17: *Stepwise* metoda – korak 3.

Korak 4.

Varijabla *dob* se sljedeća dodaje u model, a nijedna od prethodno dodanih varijabli se i dalje ne izbacuje te se zatim traži koja varijabla sljedeća ulazi u model. AIC trenutnog modela je 236.3.

moguće dodati ili izbaciti	varijabla	stupnjevi slobode	devijanca	AIC	test omjera vjerodostojnosti	p-vrijednost
+	natrij	1	223.49	235.49	2.815	0.09338
	bez varijabli		226.30	236.30		
+	visoki tlak	1	226.27	238.27	0.035	0.85167
-	dob	1	235.41	243.41	9.113	0.00254
-	kreatinin	1	244.51	252.51	18.207	$1.98 \cdot 10^{-5}$
-	ejekcijska frakcija	1	254.62	262.62	28.320	$1.03 \cdot 10^{-7}$
-	praćenje	1	305.28	313.28	78.981	$< 2.2 \cdot 10^{-16}$

Tablica 2.18: *Stepwise* metoda – korak 4.

Korak 5.

Natrij se zatim dodaje u model dok su sve prethodno dodane varijable i dalje ostale statistički značajne. Jedina varijabla koja se još potencijalno može dodati u model je *visoki tlak*. AIC trenutnog modela je 235.49.

moguće dodati ili izbaciti	varijabla	stupnjevi slobode	devijanca	AIC	test omjera vjerodostojnosti	p-vrijednost
	bez varijabli		223.49	235.49		
-	natrij	1	226.30	236.30	2.815	0.09338
+	visoki tlak	1	223.46	237.46	0.027	0.86877
-	dob	1	232.02	242.02	8.530	0.00349

-	kreatinin	1	239.56	249.56	16.077	$6.08 \cdot 10^{-5}$
-	ejekcijska frakcija	1	249.83	259.83	26.341	$2.86 \cdot 10^{-7}$
-	praćenje	1	303.09	313.09	79.603	$< 2.2 \cdot 10^{-16}$

Tablica 2.19: *Stepwise* metoda – korak 5.

Sve varijable koje su trenutno u modelu su statistički značajne, a varijablu *visoki tlak* ne treba dodati pa metoda ovdje staje i kao konačan model daje onaj u kojem su nezavisne varijable *praćenje*, *ejekcijska frakcija*, *kreatinin*, *dob* i *natrij* te su ujedno tim redom i ulazile u model.

Da smo umjesto *stepwise* metode s varijablama koje su ispale statistički značajne proveli *stepwise* metodu sa svim nezavisnim varijablama, dobili bismo da iste varijable istim redoslijedom ulaze u model te da model staje nakon jednakog broja koraka.

Provedimo multivarijabilnu regresiju za dobiveni model iz koraka 5:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
slobodni koeficijent	9.49303	5.40577	1.756	0.07907	[-1.10208, 20.08814]
praćenje	-0.02090	0.00292	-7.166	$7.74 \cdot 10^{-13}$	[-0.02661, -0.01518]
ejekcijska frakcija	-0.07343	0.01579	-4.652	$3.29 \cdot 10^{-6}$	[-0.10437, -0.04249]
kreatinin	0.68599	0.17404	3.941	$8.10 \cdot 10^{-5}$	[0.34487, 1.02711]
dob	0.04247	0.01503	2.825	0.00472	[0.01301, 0.07192]
natrij	-0.06456	0.03838	-1.682	0.09254	[-0.13978, 0.01066]

Tablica 2.20: Multivarijabilna logistička regresija s varijablama dobivenim *stepwise* metodom

Provjera značajnosti varijabli

Za svaku varijablu iz tablice 2.20 uspoređujemo procijenjeni koeficijent iz tog multivarijabilnog modela s procijenjenim koeficijentom u univarijabilnom modelu koji sadrži samo tu varijablu:

varijabla	procjena koeficijenta u multivarijabilnom modelu	procjena koeficijenta u univarijabilnom modelu
praćenje	-0.02090	-0.01995
ejekcijska frakcija	-0.07343	-0.05620
kreatinin	0.68599	0.82420
dob	0.04247	0.04695
natrij	-0.06456	-0.09639

Tablica 2.21: Usporedba procjena koeficijenata iz multivarijabilnog modela nakon *stepwise* metode i iz odgovarajućih univarijabilnih modela

Ako bismo gledali koje su varijable statistički značajne na razini značajnosti od 5% u multivarijabilnom logističkom modelu, zaključili bismo da su to sve varijable osim *natrija* pa možemo pomoću testa omjera vjerodostojnosti testirati je li potreban prošireni model koji sadrži i *natrij* ili je dovoljan podmodel bez *natrija*. Dobijemo da je vrijednost testne statistike 2.81510, a pripadna p-vrijednost 0.09338. Stoga na razini značajnosti od 5% zaključujemo da je dovoljan podmodel. Radimo multivarijabilnu logističku regresiju s preostalim varijablama:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
slobodni koeficijent	0.60447	1.03611	0.583	0.55962	[-1.42627, 2.63521]
praćenje	-0.02061	0.00288	-7.153	$8.48 \cdot 10^{-13}$	[-0.02626, -0.01496]
ejekcijska frakcija	-0.07480	0.01556	-4.809	$1.52 \cdot 10^{-6}$	[-0.10529, -0.04432]
kreatinin	0.71979	0.17460	4.123	$3.74 \cdot 10^{-5}$	[0.37758, 1.06199]
dob	0.04333	0.01487	2.913	0.00358	[0.01418, 0.07247]

Tablica 2.22: Multivarijabilna logistička regresija nakon izbacivanja *natrija*

Gledajući vrijednosti Waldovih testova i pripadne p-vrijednosti vidimo da su sve varijable iz tablice 2.22 statistički značajne na razini značajnosti 5%. Ponovno uspoređujemo procjene koeficijenata iz tog modela s procjenama koeficijenata iz odgovarajućih univarijabilnih modela. Rezultati su u sljedećoj tablici:

varijabla	procjena koeficijenta u multivarijabilnom modelu	procjena koeficijenta u univarijabilnom modelu
praćenje	-0.02061	-0.01995
ejekcijska frakcija	-0.07480	-0.05620
kreatinin	0.71979	0.82420
dob	0.04333	0.04695

Tablica 2.23: Usporedba procjena koeficijenata iz multivarijabilnog modela nakon izbacivanja *natrija* i iz odgovarajućih univarijabilnih modela

Potrebno je još usporediti procijenjene koeficijente iz trenutnog multivarijabilnog modela s procjenama koeficijenata iz punog multivarijabilnog modela:

varijabla	procjena koeficijenta u multivarijabilnom modelu	procjena koeficijenta u punom modelu
praćenje	-0.02061	-0.02104
ejekcijska frakcija	-0.07480	-0.07666
kreatinin	0.71979	0.66610
dob	0.04333	0.04742

Tablica 2.24: Usporedba procjena koeficijenata iz multivarijabilnog modela nakon izbacivanja *natrija* i iz punog multivarijabilnog modela

Zaključujemo da možemo sve varijable ostaviti u modelu i smatramo da se procjene koeficijenata u tablici 2.24 ne razlikuju previše u multivarijabilnom modelu u odnosu na one u punom modelu. Dakle, daljnju analizu nastavljamo s varijablama *praćenje*, *ejekcijska frakcija*, *kreatinin* i *dob*.

Provjera linearnosti logit funkcije za neprekidne varijable

Sve nezavisne varijable koje imamo u modelu su neprekidne. Kako bismo provjerili linearnost logit funkcije koristimo Box-Tidwellov test te u model dodajemo interakcije *praćenje* $\times$ ln(*praćenje*), *ejekcijska frakcija* $\times$ ln(*ejekcijska frakcija*), *kreatinin* $\times$ ln(*kreatinin*) i *dob* $\times$ ln(*dob*). Računamo procjene koeficijenata, pripadne standardne greške, Waldove testne statistike i pripadne p-vrijednosti:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost
slobodni koeficijent	-215.51308	298.34329	-0.722	0.4701
praćenje	-0.07230	0.20553	-0.352	0.7250

ln(praćenje)	-1.08658	2.38507	-0.456	0.6487
ejekcijska frakcija	2.16795	3.78300	0.573	0.5666
ln(ejekcijska frakcija)	-21.70047	24.84833	-0.873	0.3825
kreatinin	-9.98803	6.10725	-1.635	0.1020
ln(kreatinin)	9.51094	4.93154	1.929	0.0538
dob	-14.59333	13.97605	-1.044	0.2964
ln(dob)	136.21150	140.45666	0.970	0.3322
praćenje × ln(praćenje)	0.01115	0.03166	0.352	0.7248
ejekcijska frakcija × ln(ejekcijska frakcija)	-0.35727	0.67007	-0.533	0.5939
kreatinin × ln(kreatinin)	3.25717	2.00695	1.623	0.1046
dob × ln(dob)	2.41968	2.27903	1.062	0.2884

Tablica 2.25: Box-Tidwellov test

Vidimo da nijedna interakcija nije statistički značajna jer su sve pripadne p-vrijednosti veće od 0.05. Dakle, vrijedi linearnost logit funkcije za sve neprekidne varijable pa ne moramo raditi nikakve transformacije.

Interakcije

Nakon detaljnijeg istraživanja što su to točno kreatinin i ejekcijska frakcija, zaključili smo da bi bilo dobro ispitati jesu li statistički značajne interakcije *dob* × *kreatinin* i *dob* × *ejekcijska frakcija*. Navedene interakcije dodajemo u model jednu po jednu te zatim testom omjera vjerodostojnosti testiramo je li bolji prošireni model s interakcijom ili njegov podmodel bez interakcije. Prošireni model označimo s 1, a podmodel s 2 te se u njemu nalaze varijable *praćenje*, *ejekcijska frakcija*, *kreatinin* i *dob*. Log-vjerodostojnost modela 2 iznosi -113.15, a ostali rezultati su dani u sljedećoj tablici pri čemu je u prvom stupcu označena interakcija koja je dodana u model 1 u kojem se već nalaze sve varijable koje su u modelu 2:

dodana interakcija	log-vjerodostojnost modela s interakcijom	testna statistika	p-vrijednost
dob × kreatinin	-112.80	0.7082	0.4000
dob × ejekcijska frakcija	-112.29	1.7190	0.1898

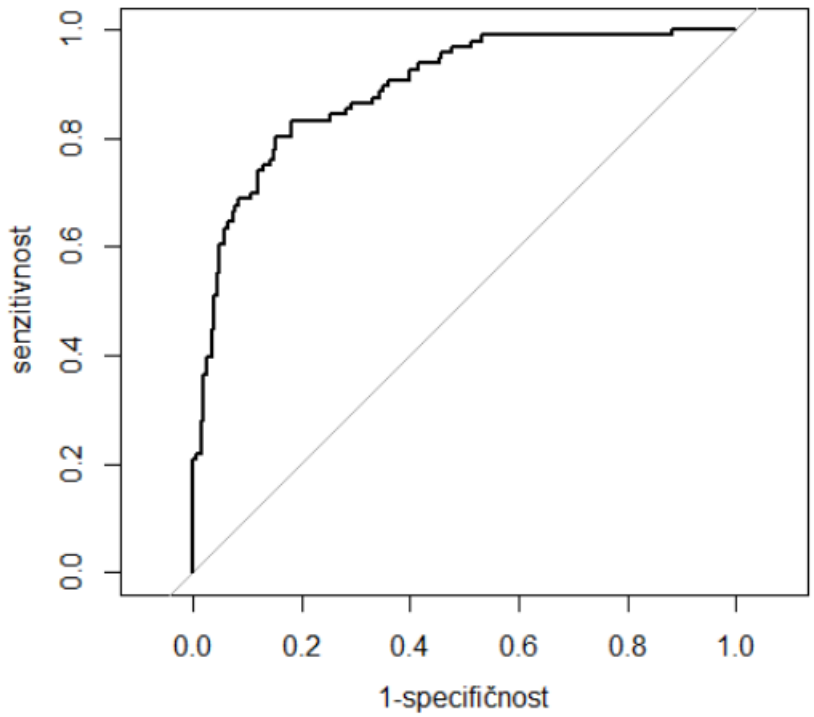
Tablica 2.26: Testovi omjera vjerodostojnosti za testiranje modela s interakcijama u odnosu na model bez njih

Na temelju dobivenih p-vrijednosti zaključujemo da ne odbacujemo nulte hipoteze da je dovoljan model bez interakcije. Dakle, nijednu interakciju nećemo uključiti u model.

2.1.3 Procjena adekvatnosti modela

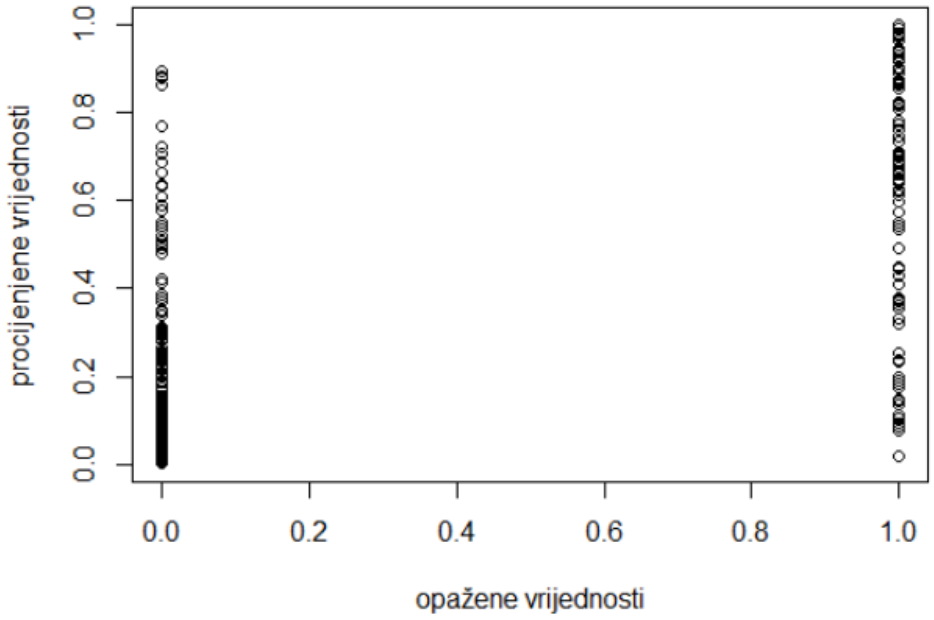
Primijetimo da su sve opažene vrijednosti varijable $\boldsymbol{x}$ = (*praćenje, ejekcijska frakcija, kreatinin, dob*) različite pa je $J = n$ što znači da Pearsonova χ^2-statistika i devijanca nemaju $\chi^2(J - p - 1)$ distribuciju. Stoga ćemo procijeniti adekvatnost modela pomoću površine ispod ROC krivulje te pomoću grafa opaženih i procijenjenih vrijednosti.

Površina ispod ROC krivulje je 0.8912 što znači da je model odličan, a sama ROC krivulja izgleda ovako:



Slika 2.5: ROC krivulja

Na grafu opaženih i procijenjenih vrijednosti



Slika 2.6: Grafički prikaz opaženih i procijenjenih vrijednosti

vidimo da se za opaženu vrijednost 0 više procijenjenih vrijednosti nalazi bliže 0, a za opaženu vrijednost 1 više oko 1.

Iz svega navedenog zaključujemo da model vrlo dobro opisuje podatke.

2.1.4 Zaključak

Nakon analize podataka po koracima iz pododjeljka 1.8.1 zaključili smo da je najbolji model onaj u kojem su nezavisne varijable *praćenje*, *ejekcijska frakcija*, *kreatinin* i *dob*. Preciznije, ako kao ranije označimo s $\boldsymbol{x} = (praćenje,\ ejekcijska\ frakcija,\ kreatinin,\ dob)$ i $\pi(\boldsymbol{x}) = P(smrt = 1|\boldsymbol{x})$ tada iz tablice 2.22 možemo iščitati da je pripadna logit funkcija

$$\begin{aligned} g(\boldsymbol{x}) = 0.60447 - 0.02061 \cdot praćenje - 0.07480 \cdot ejekcijska\,frakcija + \\ + 0.71979 \cdot kreatinin + 0.04333 \cdot dob, \end{aligned} \tag{2.1}$$

a model multivarijabilne logističke regresije

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}$$

uz g($\boldsymbol{x}$) iz (2.1).

2.2 Karcinom jetre

2.2.1 Opis problema i podataka

Prema [35] karcinom jetre je šesti najčešće dijagnosticirani karcinom, a hepatocelularni karcinom (HCC, eng. *hepatocellular carcinoma*) je najčešća vrsta karcinoma jetre. U Portugalu, gdje su podaci prikupljani, to je sedmi vodeći uzrok smrti povezanih s karcinomima. Prikupljeni su podaci o 165 pacijenata kojima je dijagnosticiran hepatocelularni karcinom te se oni odnose na način života, faktore rizika, rezultate laboratorijskih pretraga i slično. Točnije, prikupljeni su podaci o 49 različitih obilježja koja su prema smjernicama EASL–EORTC-a (*European Association for the Study of the Liver – European Organisation for Research and Treatment of Cancer*) smatrana najvažnijima. Od toga su 23 obilježja kvantitativna, a 26 je kvalitativnih. Međutim, nije bilo moguće kod svakog pacijenta prikupiti svaku traženu vrijednost.

Nakon perioda od godinu dana prikupljen je još podatak o tome je li svaki pojedini pacijent živ. To će biti zavisna varijabla *preživljavanje* u modelu logističke regresije te ćemo ju kodirati s 0 ako pacijent više nije živ, a s 1 ako je. Ukupno imamo 102 preživjela pacijenta i 63 pacijenta koji nisu preživjeli.

Kvalitativne varijable koje su dihotomne i kodirane s 0 i 1 su sljedeće:

- **spol** (*gender*)
  - označava spol osobe
  - 0 – žene
    1 – muškarci
- **simptomi** (*symptoms*)
  - označava je li osoba primijetila nekakve simptome bolesti
- **alkohol** (*alcohol*)
  - označava pije li osoba alkohol
- **hepatitis B površinski antigen** (HBsAg, *hepatitis B surface antigen*)
  - pozitivan test označava da je osoba akutno ili kronično zaražena hepatitisom B [18]
- **hepatitis B rani antigen** (HBeAg, *hepatitis B early antigen*)
  - pozitivan test označava da je aktivna replikacija virusa što znači da je osoba trenutno zarazna [17]
- **hepatitis B antitijelo jezgre** (HBcAb, *hepatitis B core antibody*)

  - pozitivan test označava da je osoba bila i/ili je trenutno zaražena hepatitisom B [18]

- **hepatitis C antitijelo virusa** (HCVAb, *hepatitis C virus antibody*)
  - negativan test označava da osoba trenutno nije zaražena hepatitisom C
  - pozitivan test označava da je osoba bila i/ili je trenutno zaražena hepatitisom C [20]

- **ciroza** (*cirrhosis*)
  - označava ima li osoba cirozu jetre

- **endemske države** (*endemic countries*)
  - označava je li osoba bila u nekoj endemskoj državi

- **pušenje** (*smoking*)
  - označava puši li osoba

- **dijabetes** (*diabetes*)
  - označava ima li osoba dijabetes

- **pretilost** (*obesity*)
  - označava je li osoba pretila

- **hemokromatoza** (*hemochromatosis*)
  - označava ima li osoba stanje kad se u organizmu nakuplja previše željeza [22]

- **arterijska hipertenzija** (AHT, *arterial hypertension*)
  - označava ima li osoba visoki arterijski krvni tlak

- **kronična bubrežna insuficijencija** (CRI, *chronical renal insufficiency*)
  - označava ima li osoba lošu funkciju bubrega [27]

- **HIV** (*human immunodeficiency virus*)
  - označava ima li osoba virus koji uzrokuje AIDS

- **nealkoholni steatohepatitis** (NASH, *nonalcoholic steatohepatitis*)
  - označava ima li osoba upalu i oštećenje jetre uzrokovano nakupljanjem masnoća u jetri [31]

- **variksi jednjaka** (*esophageal varices*)
  - označava ima li osoba povećanje vena jednjaka [45]

- **splenomegalija** (*splenomegaly*)

  - označava ima li osoba stanje povećane slezene [37]

- **portalna hipertenzija** (PHT, *portal hypertension*)
  - označava ima li osoba povišeni tlak u portalnoj veni koja prenosi krv iz probavnih organa do jetre [33]

- **tromboza portalne vene** (PVT, *portal vein thrombosis*)
  - označava ima li osoba blokadu ili suženje portalne vene uzrokovano krvnim ugruškom [40]

- **metastaze na jetri** (*liver metastasis*)
  - označava ima li osoba metastaze karcinoma na jetri

- **radiološko obilježje** (*radiological hallmark*)
  - označava vidi li se karcinom jetre na radiološkim snimkama osobe.

Kvalitativne ordinalne varijable su sljedeće:

- **status općeg tjelesnog stanja pacijenta** (PS, *performance status*)
  - označava koliko bolest utječe na svakodnevni život osobe opisujući mogućnost da se osoba brine sama za sebe, fizički kreće i bude aktivna
  - 0 – osoba je aktivna bez ograničenja
    1 – osoba je ograničenih sposobnosti, ali može obavljati lagane poslove
    2 – osoba je sposobna brinuti se za sebe, ali ne i obavljati ikakve poslove, u pokretu je više od 50% vremena kad je budna
    3 – osoba je sposobna za ograničenu brigu o sebi, u krevetu je ili stolcu više od 50% vremena kad je budna
    4 – osoba je nemoćna brinuti se o sebi, potpuno je u krevetu ili stolcu
    5 – smrt [39]

- **stupanj encefalopatije** (*encephalopathy degree*)
  - označava poremećaj živčanog sustava uzrokovan lošom funkcijom jetre nakon što se isključe neurološke bolesti [13]
  - 1 – nema
    2 – stupanj I ili II
    3 – stupanj III ili IV

- **stupanj ascitesa** (*ascites degree*)
  - označava stupanj nakupljanja tekućine u trbušnoj šupljini [6]
  - 1 – nema
    2 – malo
    3 – umjereno ili puno.

Kvantitativne varijable su sljedeće:

- **dob u trenutku dijagnoze** (*age of diagnosis*)
  - označava dob osobe u godinama u trenutku dijagnoze
- **grami alkohola dnevno** (*grams of alcohol per day*)
  - označava količinu čistog alkohola u gramima koju osoba popije dnevno
- **broj kutija cigareta godišnje** (*packs of cigarettes per year*)
  - označava koliko kutija cigareta godišnje osoba popuši
- **internacionalni normalizirani omjer** (INR, *international normalised ratio*)
  - označava mjeru izvedenu iz protrombinskog vremena, odnosno vremena zgrušavanja krvi [25]
- **alfa-fetoprotein** (AFP, *alpha-fetoprotein*)
  - povišena vrijednost označava veliku vjerojatnost prisutnosti karcinoma jetre ili nekog drugog tumora [1]
  - mjereno u ng/mL
- **hemoglobin** (*haemoglobin*)
  - označava koliko osoba ima hemoglobina u krvi, proteina koji prenosi kisik
  - mjereno u g/dL
- **prosječni volumen eritrocita** (MCV, *mean corpuscular volumen*)
  - označava prosječni volumen eritrocita u fl u krvi osobe
- **leukociti** (*leukocytes*)
  - označava koliko osoba ima leukocita u krvi
  - mjereno u G/L
- **trombociti** (*platelets*)
  - označava koliko osoba ima trombocita u krvi
  - mjereno u G/L
- **albumin** (*albumin*)
  - snižena vrijednost označava da osoba može imati bolest jetre [2]
  - mjereno u mg/dL
- **ukupni bilirubin** (*total bilirubin*)
  - povišena vrijednost označava da osoba možda ima bolest jetre [42]
  - mjereno u mg/dL

- **alanin transaminaza** (ALT, *alanine transaminase*)
  - povišena vrijednost označava da osoba može imati bolest jetre [4]
  - mjereno u U/L
- **aspartat transaminaza** (AST, *aspartate transaminase*)
  - povišena vrijednost označava da osoba može imati bolest jetre [4]
  - mjereno u U/L
- **gama-glutamiltransferaza** (GGT, *gamma-glutamyl transferase*)
  - inače se nalazi većinom u jetri, a povišena vrijednost u krvi označava da osoba može imati bolest jetre [16]
  - mjereno u U/L
- **alkalna fosfataza** (ALP, *alkaline phosphatase*)
  - povišena vrijednost označava da osoba može imati bolest jetre ili kostiju [3]
  - mjereno u U/L
- **ukupni proteini** (TP, *total proteins*)
  - označava ukupnu količinu proteina u krvi osobe u g/dL
  - povišena vrijednost označava da osoba može, između ostalog, imati upalu ili infekciju kao npr. hepatitis B ili C te HIV
  - snižena vrijednost označava da osoba može imati bolest jetre, između ostalog [43]
- **kreatinin** (*creatinine*)
  - označava koliko dobro funkcioniraju bubrezi [26]
  - mjereno u mg/dL
- **broj tumorskih čvorova** (*number of nodules*)
  - označava koliko osoba ima tumorskih čvorova u jetri
- **najveća dimenzija čvora** (*major dimension of nodule*)
  - označava najveću dimenziju u cm nađenog tumorskog čvora u jetri
- **direktni bilirubin** (*direct bilirubin*)
  - povišena vrijednost može označavati, između ostalog, da osoba ima probleme s jetrom ili hepatitis [11]
  - mjereno u mg/dL
- **željezo** (*iron*)
  - označava količinu željeza u μg/dL u krvi osobe

- **zasićenost kisikom** (*oxygen saturation*)
  - označava postotak zasićenost krvi kisikom

- **feritin** (*ferritin*)
  - označava koliko osoba pohranjuje željeza u ng/mL u tijelu
  - povišena vrijednost označava da osoba može imati bolest jetre. [15] [35]

2.2.2 Odabir modela

Podaci koji nedostaju

Među prikupljenim podacima ima vrijednosti koje nedostaju. Nedostaje 10.22% podataka kod nezavisnih varijabli dok je vrijednost zavisne varijable dostupna za svakog pacijenta. Pretpostavka je da podaci nedostaju na slučajan način jer kod nekih osoba nisu rađene sve krvne ili druge medicinske pretrage, odnosno osobe nisu odgovorile na određena pitanja o svojim životnim navikama i slično.

Za početak analiziramo koliko podataka nedostaje za svako od obilježja od interesa. Varijable su poredane počevši od onih za koje imamo najveći broja podataka koji nedostaju:

redni broj varijable	varijabla	broj pacijenata za koje nema podataka	% pacijenata za koje nema podataka
1.	zasićenost kisikom	80	48.48%
2.	feritin	80	48.48%
3.	željezo	79	47.88%
4.	cigarete	53	32.12%
5.	variksi	52	31.52%
6.	alkohol u gramima	48	29.09%
7.	direktni bilirubin	44	26.67%
8.	pušenje	41	24.85%
9.	HBeAg	39	23.64%
10.	endemska država	39	23.64%
11.	HBcAb	24	14.55%
12.	hemokromatoza	23	13.94%
13.	NASH	22	13.33%
14.	najveća dimenzija čvora	20	12.12%
15.	simptomi	18	10.91%
16.	HBsAg	17	10.30%

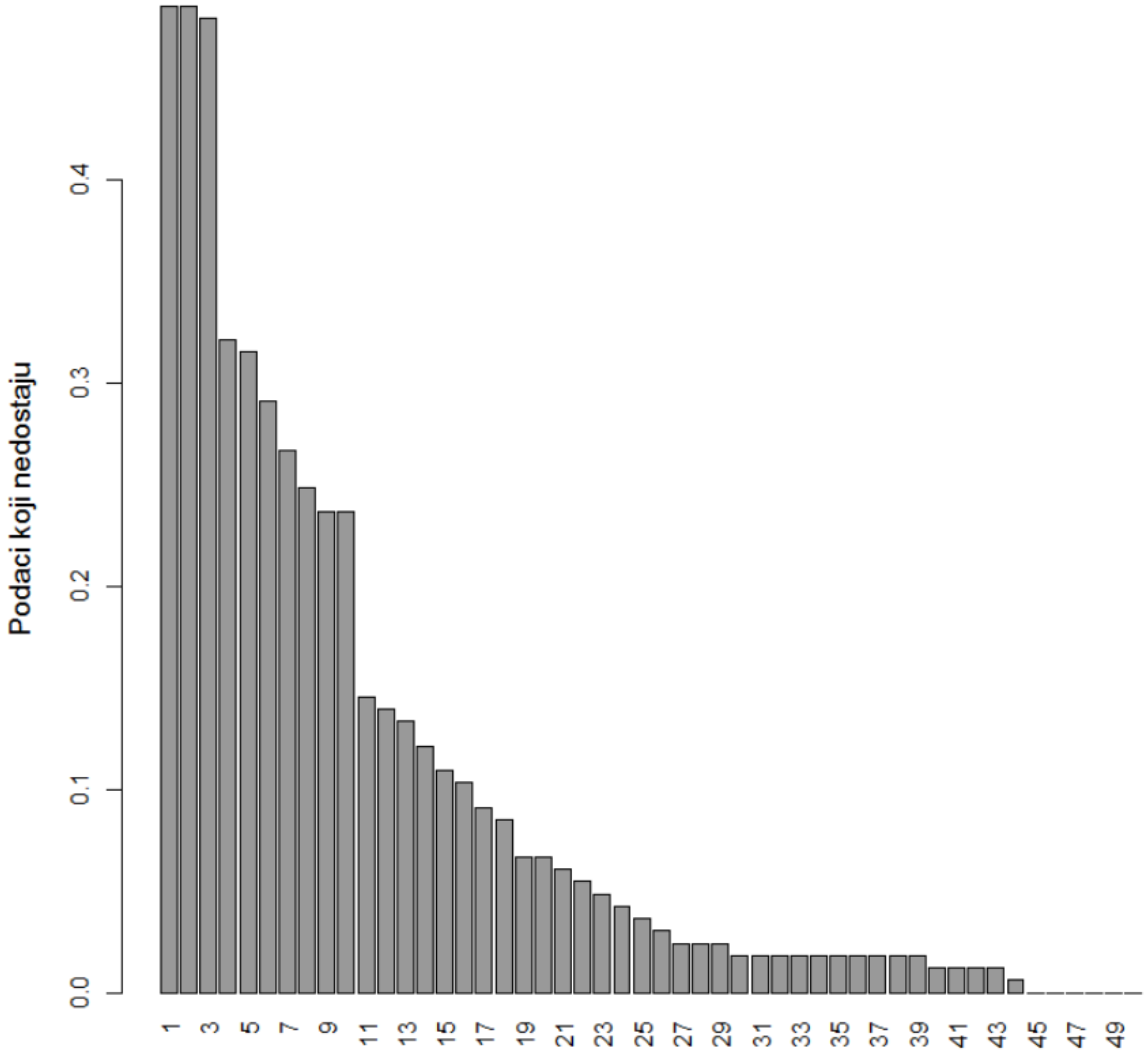
17.	splenomegalija	15	9.09%
18.	HIV	14	8.48%
19.	portalna hipertenzija	11	6.67%
20.	ukupni proteini	11	6.67%
21.	pretilost	10	6.06%
22.	HCVAb	9	5.45%
23.	AFP	8	4.85%
24.	kreatinin	7	4.24%
25.	albumin	6	3.64%
26.	ukupni bilirubin	5	3.03%
27.	metastaze	4	2.42%
28.	INR	4	2.42%
29.	ALT	4	2.42%
30.	dijabetes	3	1.82%
31.	arterijska hipertenzija	3	1.82%
32.	tromboza portalne vene	3	1.82%
33.	hemoglobin	3	1.82%
34.	prosječni volumen eritrocita	3	1.82%
35.	leukociti	3	1.82%
36.	trombociti	3	1.82%
37.	AST	3	1.82%
38.	GGT	3	1.82%
39.	ALP	3	1.82%
40.	kronična bubrežna insuficijencija	2	1.21%
41.	radiološko obilježje	2	1.21%
42.	stupanj ascitesa	2	1.21%
43.	broj čvorova	2	1.21%
44.	stupanj encefalopatije	1	0.61%
45.	spol	0	0.00%
46.	alkohol	0	0.00%
47.	ciroza	0	0.00%
48.	dob	0	0.00%
49.	stupanj stanja pacijenta	0	0.00%
50.	preživljavanje	0	0.00%

Tablica 2.27: Podaci koji nedostaju

Vidimo da za jako puno pacijenata nema podataka za *zasićenost kisikom*, *feritin* i *željezo*, a

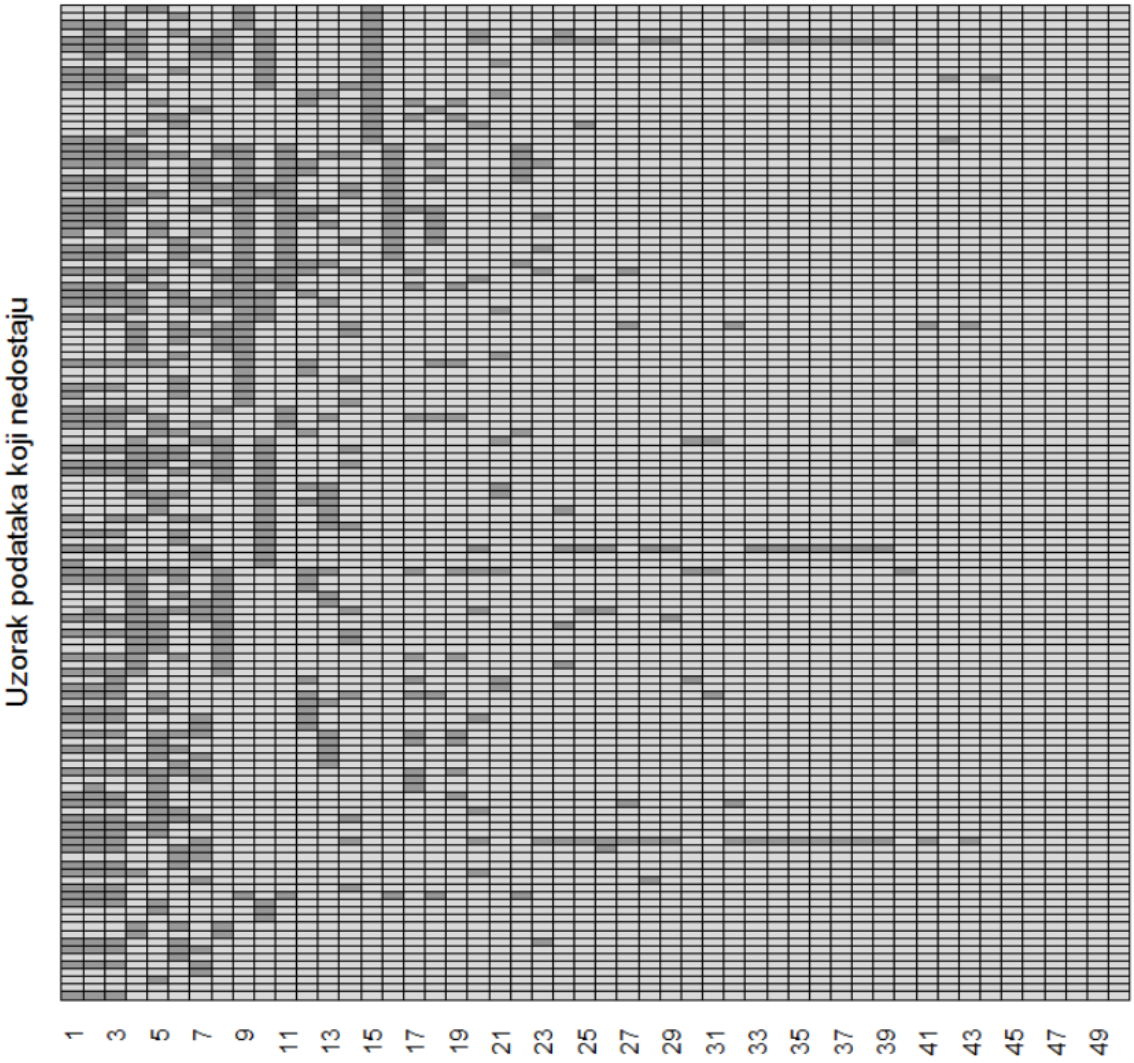
za varijable *spol*, *alkohol*, *ciroza*, *dob*, *stupanj stanja pacijenta* i *preživljavanje* su dostupni podaci za sve pacijente.

Dobivene rezultate možemo grafički prikazati stupčastim dijagramom pri čemu je na y-osi prikazan decimalan broj između 0 i 1 umjesto postotka, a stupci su označeni brojevima kao u prethodnoj tablici:



Slika 2.7: Grafički prikaz podataka koji nedostaju

Možemo još proučiti i uzorak pojavljivanja nedostatka podataka po pacijentima. U sljedećem grafičkom prikazu svaki redak predstavlja jednog pacijenta, a svaki stupac jednu varijablu poredano kao u tablici 2.27, odnosno po varijablama počevši od onih za koje imamo najveći broja podataka koji nedostaju do onih za koje ne nedostaju podaci. Tamnije siva polja predstavljaju podatke koji nedostaju, a svjetlije siva podatke koji su prikupljeni.



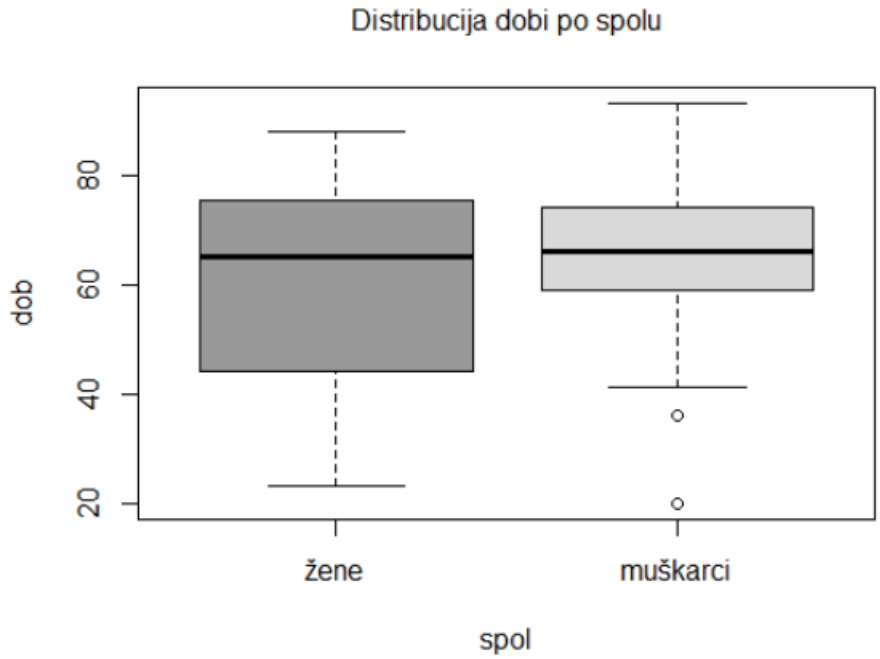
Slika 2.8: Prikaz uzorka podataka koji nedostaju

S obzirom da za samo 8 od 165 pacijenata imamo dostupne sve podatke, nije dobro rješenje

da sve ostale pacijente eliminiramo iz analize nego nadomještamo podatke koristeći višestruko nadomještanje pomoću pet skupova podataka kao što je opisano u pododjeljku 1.8.1.

Deskriptivna statistika

Kao i u primjeru o srčanom zatajenju, prvo možemo promotriti distribuciju pacijenata po dobi i spolu. Uočimo, za svakog pacijenta imamo dostupne podatke o dobi i spolu.



Slika 2.9: Distribucija pacijenata po dobi i spolu

Prosječna dob žena i muškaraca je otprilike jednaka, ali je kod žena dosta veća varijabilnost dobi. Ukupno imamo više žena koje su mlađe u odnosu na muškarce. Kod muškaraca dob dva najmlađa muškarca statistički značajno odstupa od dobi ostalih muškaraca dok kod žena nema statistički značajnih odstupanja dobi. Najstarija kao i najmlađa osoba u istraživanju su muškarci.

Analiziramo kvantitativne varijable osim varijable *broj čvorova* jer ona, iako kvantitativna, poprima šest različitih vrijednosti pa ju možemo lakše analizirati promatrajući frekvencije. U sljedećoj su tablici dani aritmetička sredina, standardna devijacija i karakteristična petorka uzorka:

varijabla	aritmetička sredina	standardna devijacija	minimum	donji kvartil	medijan	gornji kvartil	maksimum
dob	64.6909	13.3195	20	57	66	74	93
alkohol u gramima	81.1709	77.9892	0	0	80	100	500
cigarete	22.1570	44.9204	0.0	0.0	3.5	39.0	510.0
INR	1.42339	0.47968	0.84	1.17	1.30	1.53	4.82
AFP	18911.1	145628	1.2	5.2	37.2	622.8	$1.8 \cdot 10^6$
hemoglobin	12.8914	2.14495	5.00	11.40	13.08	14.60	18.70
prosječni volumen eritrocita	95.1176	8.42113	69.5	89.8	94.9	100.7	119.6
leukociti	1467.52	2899.41	2.2	5.1	7.2	18.7	13000.0
trombociti	113260	106993	1.7	255.4	92600	$1.7 \cdot 10^5$	$4.6 \cdot 10^5$
albumin	3.44164	0.68333	1.90	3.02	3.40	4.04	4.90
ukupni bilirubin	3.07487	5.43149	0.30	0.82	1.40	2.96	40.50
ALT	67.7067	57.8726	11.0	31.0	50.0	78.4	420.0
AST	96.0958	86.9315	17.0	46.0	71.4	110.0	553.0
GGT	267.167	258.457	23.0	90.8	179.0	345.2	1575.0
ALP	212.051	167.316	1.3	108.0	162.2	261.4	980.0
ukupni proteini	9.20752	12.2423	3.90	6.34	7.06	7.58	102.00
kreatinin	1.13972	0.97149	0.20	0.70	0.85	1.10	7.60
najveća dimenzija čvora	7.10449	5.24551	1.50	3.00	5.50	9.02	22.00
direktni bilirubin	1.71148	3.66282	0.10	0.35	0.75	1.42	29.30
željezo	91.9315	57.4083	0.0	43.6	88.6	137.0	224.0
zasićenost kisikom	39.3234	28.6770	0.0	16.6	33.0	59.2	126.0
feritin	469.454	491.455	0.0	80.0	290.0	748.2	2230.0

Tablica 2.28: Deskriptivna analiza kvantitativnih varijabli osim *broja čvorova*

Za varijable *broj čvorova*, zavisnu varijablu *preživljavanje* te za sve kvalitativne varija-

ble računamo frekvenciju i postotak po kategorijama. Rezultati su zapisani u sljedećim tablicama:

varijabla	kategorija	frekvencija	%
broj čvorova	0	1	0.61%
	1	68	41.21%
	2	25	15.15%
	3	11	6.67%
	4	2	1.21%
	5	58	35.15%
spol	žene	32	19.39%
	muškarci	133	80.61%
simptomi	nema	66	40.00%
	ima	99	60.00%
alkohol	ne pije	43	26.06%
	pije	122	73.94%
HBsAg	ne	137	83.03%
	da	28	16.97%
HBeAg	ne	145	87.88%
	da	20	12.12%
HBcAb	ne	108	65.45%
	da	57	34.55%
HCVAb	ne	125	75.76%
	da	40	24.24%
ciroza	nema	16	9.70%
	ima	149	90.30%
endemska država	ne	132	80.00%
	da	33	20.00%
pušenje	ne	79	47.88%
	da	86	52.12%
dijabetes	nema	106	64.24%
	ima	59	35.76%
pretilost	ne	140	84.85%
	da	25	15.15%
hemokromatoza	nema	148	89.70%
	ima	17	10.30%
arterijska hipertenzija	nema	103	62.42%
	ima	62	37.58%

kronična bubrežna insuficijencija	nema	144	87.27%
	ima	21	12.73%
HIV	nema	155	93.94%
	ima	10	6.06%
NASH	nema	149	90.30%
	ima	16	9.70%
variksi	nema	73	44.24%
	ima	92	55.76%
splenomegalija	nema	77	46.67%
	ima	88	53.33%
portalna hipertenzija	nema	53	32.12%
	ima	112	67.88%
tromboza portalne vene	nema	128	77.58%
	ima	37	22.42%
metastaze	nema	128	77.58%
	ima	37	22.42%
radiološko obilježje	ne	53	32.12%
	da	112	67.88%
stupanj stanja pacijenta	0	80	48.48%
	1	30	18.18%
	2	32	19.39%
	3	18	10.91%
	4	5	3.03%
	5	0	0.00%
stupanj encefalopatije	1	143	86.67%
	2	18	10.91%
	3	4	2.42%
stupanj ascitesa	1	110	66.67%
	2	36	21.82%
	3	18	10.91%

Tablica 2.29: Deskriptivna analiza kvalitativnih varijabli i varijable *broj čvorova*

varijabla	kategorija	frekvencija	%
preživljavanje	ne	63	38.20%
	da	102	61.80%

Tablica 2.30: Deskriptivna analiza zavisne varijable

Univarijabilna logistička regresija

Rezultate univarijabilne analize za kvantitativne varijable prikazujemo u sljedećoj tablici. Pritom, u stupcu Waldov test dana je multivarijabilna Waldova testna statistika kao što je opisano u [44].

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
dob	-0.02393	0.01291	-1.85263	0.06577	[-0.04943, 0.00158]
alkohol u gramima	-0.00231	0.00250	-0.92441	0.35724	[-0.00725, 0.00264]
cigarete	-0.00364	0.00424	-0.85835	0.39260	[-0.01204, 0.00476]
INR	-0.98767	0.42652	-2.31565	0.02187	[-1.83013, -0.14521]
AFP	$2.8 \cdot 10^{-8}$	$1.1 \cdot 10^{-6}$	0.02470	0.98032	$[-2.2 \cdot 10^{-6}, 2.2 \cdot 10^{-6}]$
hemoglobin	0.30091	0.08463	3.55543	0.00050	[0.13375, 0.46807]
prosječni volumen eritrocita	0.01031	0.01928	0.53491	0.59346	[-0.02777, 0.04840]
leukociti	$-7.1 \cdot 10^{-5}$	0.00005	-1.28877	0.19936	$[-0.00018, 3.8 \cdot 10^{-5}]$
trombociti	$-2.9 \cdot 10^{-6}$	$1.5 \cdot 10^{-6}$	-1.89514	0.05990	$[-5.9 \cdot 10^{-6}, 1.2 \cdot 10^{-7}]$
albumin	0.91993	0.26247	3.50490	0.00060	[0.40145, 1.43841]
ukupni bilirubin	-0.10926	0.04776	-2.28796	0.02348	[-0.20359, -0.01493]
ALT	-0.00022	0.00281	-0.07686	0.93883	[-0.00577, 0.00534]
AST	-0.00459	0.00202	-2.27455	0.02428	[-0.00858, -0.00060]
GGT	-0.00117	0.00064	-1.83654	0.06816	$[-0.00244, 8.9 \cdot 10^{-5}]$
ALP	-0.00405	0.00121	-3.35165	0.00101	[-0.00644, -0.00166]
ukupni proteini	-0.00499	0.01382	-0.36119	0.71846	[-0.03229, 0.02231]
kreatinin	-0.23106	0.17634	-1.31030	0.19204	[-0.57943, 0.11730]
broj čvorova	-0.10137	0.08968	-1.13033	0.26004	[-0.27848, 0.07575]
najveća dimenzija čvora	-0.07770	0.03405	-2.28166	0.02401	[-0.14503, -0.01038]
direktni bilirubin	-0.23088	0.11275	-2.04779	0.04282	[-0.45417, -0.00759]

željezo	0.01160	0.00446	2.60192	0.01100	[0.00273, 0.02047]
zasićenost kisikom	0.00302	0.00760	0.39709	0.69235	[-0.01210, 0.01814]
feritin	-0.00174	0.00064	-2.72130	0.00796	[-0.00301, -0.00047]

Tablica 2.31: Univarijabilna logistička regresija s kvantitativnim varijablama

Testom omjera vjerodostojnosti računamo značajnost procijenjenih koeficijenata u odgovarajućem univarijabilnom modelu u odnosu na model bez varijabli. Testna statistika se računa kako je navedeno u [44], a za svaku od kvantitativnih varijabli njena vrijednost dana je u sljedećoj tablici:

varijabla u univarijabilnom modelu	testna statistika	p-vrijednost
dob	3.63511	0.05657
alkohol u gramima	0.87397	0.34986
cigarete	0.87088	0.35071
INR	6.85439	0.00884
AFP	0.00061	0.98025
hemoglobin	14.27105	0.00016
prosječni volumen eritrocita	0.28694	0.59219
leukociti	1.66005	0.19760
trombociti	3.65980	0.05574
albumin	13.56671	0.00023
ukupni bilirubin	8.74151	0.00311
ALT	0.00589	0.93880
AST	5.84347	0.01563
GGT	3.52588	0.06042
ALP	14.47233	0.00014
ukupni proteini	0.12949	0.71896
kreatinin	1.82334	0.17692
broj čvorova	1.27950	0.25799
najveća dimenzija čvora	5.34440	0.02079
direktni bilirubin	10.32143	0.00131
željezo	7.66505	0.00563
zasićenost kisikom	0.15848	0.69056
feritin	9.68623	0.00186

Tablica 2.32: Testovi omjera vjerodostojnosti za kvantitativne varijable

Kao i u prethodnom primjeru o zatajenju srca, u obzir za daljna razmatranja uzet ćemo varijable koje imaju p-vrijednost manju od 0.25, a to su *dob*, *INR*, *hemoglobin*, *leukociti*, *trombociti*, *albumin*, *ukupni bilirubin*, *AST*, *GGT*, *ALP*, *kreatinin*, *najveća dimenzija čvora*, *direktni bilirubin*, *željezo* i *feritin*. Iz pozitivnih predznaka pripadnih procijenjenih koeficijenata iz tablice 2.31 zaključujemo da se vjerojatnost preživljavanja statistički značajno povećava ako se varijable *hemoglobin*, *albumin* ili *željezo* povećaju za jednu jedinicu dok se vjerojatnost preživljavanja statistički značajno smanjuje ako se varijable *dob*, *INR*, *leukociti*, *trombociti*, *ukupni bilirubin*, *AST*, *GGT*, *ALP*, *kreatinin*, *najveća dimenzija čvora*, *direktni bilirubin* ili *feritin* povećaju za jednu jedinicu.

Za statistički značajne kvantitativne varijable možemo izračunati procjene omjera izgleda i pripadne procjene 95% pouzdanih intervala:

varijabla	procjena omjera izgleda	procjena 95% pouzdanog intervala za omjer izgleda
dob	0.97636	[0.95177, 1.00158]
INR	0.37244	[0.16039, 0.86484]
hemoglobin	1.35109	[1.14311, 1.59691]
leukociti	0.99993	[0.99982, 1.00004]
trombociti	0.999997	[0.999994, 1.000000]
albumin	2.50911	[1.49399, 4.21398]
ukupni bilirubin	0.89650	[0.81580, 0.98518]
AST	0.99542	[0.99146, 0.99940]
GGT	0.99883	[0.99757, 1.00009]
ALP	0.99596	[0.99358, 0.99834]
kreatinin	0.79369	[0.56022, 1.12446]
najveća dimenzija čvora	0.92524	[0.86500, 0.98968]
direktni bilirubin	0.79383	[0.63497, 0.99244]
željezo	1.01167	[1.00274, 1.02068]
feritin	0.99826	[0.99699, 0.99953]

Tablica 2.33: Procjena omjera izgleda za statistički značajne kvantitativne varijable

Gledajući one modele univarijabilne logističke regresije u kojima je kvantitativna varijabla ispala statistički značajnom, kad se:

- *dob* poveća za 1 godinu, omjer izgleda da osoba preživi se smanji 0.97636 puta
- *INR* poveća za 1, omjer izgleda da osoba preživi se smanji 0.37244 puta
- *hemoglobin* poveća za 1 g/dL, omjer izgleda da osoba preživi se poveća 1.35109 puta

- *leukociti* povećaju za 1 G/L, omjer izgleda da osoba preživi se smanji 0.99993 puta, odnosno kad se *leukociti* povećaju za 1000 G/L, omjer izgleda da osoba preživi se smanji 0.93185 puta
- *trombociti* povećaju za 1 G/L, omjer izgleda da osoba preživi se smanji 0.999997 puta, odnosno kad se *trombociti* povećaju za 1000 G/L, omjer izgleda da osoba preživi se smanji 0.99712 puta
- *albumin* poveća za 1 mg/dL, omjer izgleda da osoba preživi se poveća 2.50911 puta
- *ukupni bilirubin* poveća za 1 mg/dL, omjer izgleda da osoba preživi se smanji 0.89650 puta
- *AST* poveća za 1 U/L, omjer izgleda da osoba preživi se smanji 0.99542 puta
- *GGT* poveća za 1 U/L, omjer izgleda da osoba preživi se smanji 0.99883 puta
- *ALP* poveća za 1 U/L, omjer izgleda da osoba preživi se smanji 0.99596 puta
- *kreatinin* poveća za 1 mg/dL, omjer izgleda da osoba preživi se smanji 0.79369 puta
- *najveća dimenzija čvora* poveća za 1 cm, omjer izgleda da osoba preživi se smanji 0.92524 puta
- *direktni bilirubin* poveća za 1 mg/dL, omjer izgleda da osoba preživi se smanji 0.793833 puta
- *željezo* poveća za 1 μg/dL, omjer izgleda da osoba preživi se poveća 1.01167 puta
- *feritin* poveća za 1 ng/mL, omjer izgleda da osoba preživi se smanji 0.99826 puta.

Kontingencijske tablice bismo trebali napraviti za svaku kvalitativnu varijablu i za svaki od pet skupova podataka posebno osim ako za tu kvalitativnu varijablu nema vrijednosti koje nedostaju u originalnim podacima. To je, kao što smo ranije već zaključili, slučaj kod varijabli *spol*, *alkohol*, *ciroza* i *stupanj stanja pacijenta* te je za svaku od tih varijabli dovoljno napraviti po jednu kontingencijsku tablicu za originalne podatke. Dakle, s obzirom da imamo 22 nezavisne kvalitativne varijable kojima smo morali nadomještati vrijednosti, 4 kvalitativne varijable kojima nismo i 5 skupova podataka, ukupno moramo napraviti 114 kontingencijskih tablica pa ih zbog opširnosti ovdje ne navodimo. Međutim, uočavamo da u kontingencijskoj tablici za varijablu *stupanj stanja pacijenta* imamo 0 u ćeliji kad je osoba preživjela, a *stupanj stanja pacijenta* joj iznosi 5. Kao što je rečeno u potpoglavlju 1.8.1, s obzirom da je *stupanj stanja pacijenta* ordinalna varijabla, modeliramo ju kao neprekidnu. Slična je situacija recimo za prvi skup podataka i za varijablu *stupanj encefalopatije* u ćeliji tablice kad je je osoba preživjela, a *stupanj encefalopatije* je 3 pa *stupanj encefalopatije* također modeliramo kao neprekidnu varijablu.

Rezultati univarijabilne logističke regresije za *stupanj stanja pacijenta* i *stupanj encefalopatije* prikazani su u sljedećoj tablici:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
stupanj stanja pacijenta	-0.69984	0.15255	-4.58753	$8.97 \cdot 10^{-6}$	[-1.00110, -0.39857]
stupanj encefalo-patije	-0.69320	0.38076	-1.82056	0.07054	[-1.44516, 0.05877]

Tablica 2.34: Univarijabilna logistička regresija sa *stupnjem stanja pacijenta* i *stupnjem encefalopatije*

Dalje, kao i ranije, provodimo odgovarajuće testove omjera vjerodostojnosti:

varijabla u univarijabilnom modelu	testna statistika	p-vrijednost
stupanj stanja pacijenta	24.11950	$9.05 \cdot 10^{-7}$
stupanj encefalopatije	3.45945	0.06289

Tablica 2.35: Testovi omjera vjerodostojnosti za *stupanj stanja pacijenta* i *stupanj encefalopatije*

Zaključujemo da su i *stupanj stanja pacijenta* i *stupanj encefalopatije* statistički značajne varijable. Za *stupanj stanja pacijenta* procjena omjera izgleda iznosi 0.49667 što znači da se omjer izgleda da osoba preživi smanji 0.49667 puta kad prijeđemo iz niže kategorije u prvu sljedeću višu kategoriju. Pripadna procjena 95% pouzdanog intervala iznosi [0.36748, 0.67128]. Za *stupanj encefalopatije* omjer izgleda da osoba preživi se smanji 0.49997 puta kad prijeđemo iz niže u prvu sljedeću višu kategoriju, a pripadna procjena 95% pouzdanog intervala iznosi [0.23571, 1.06053].

Rezultati univarijabilne logističke regresije za ostale kvalitativne varijable dani su u sljedećim tablicama:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
spol1	-0.20342	0.41244	-0.49322	0.62253	[-1.01791, 0.61106]
simptomi1	-1.35650	0.38788	-3.49722	0.00063	[-2.12322, -0.58978]
alkohol1	-0.19129	0.36977	-0.51733	0.60564	[-0.92151, 0.53893]
HBsAg1	0.22884	0.56890	0.40225	0.68809	[-0.89563, 1.35331]
HBeAg1	-15.14143	882.743	-0.01715	0.98634	[-1762.62, 1732.33]
HBcAb1	0.23639	0.40435	0.58462	0.55976	[-0.56317, 1.03595]

HCVAb1	-0.56309	0.39344	-1.43122	0.15442	[-1.34040, 0.21421]
ciroza1	0.25593	0.53158	0.48146	0.63085	[-0.79384, 1.30570]
endemska država1	0.96644	0.81303	1.18868	0.23687	[-0.64304, 2.57592]
pušenje1	0.32850	0.37310	0.88046	0.38037	[-0.41021, 1.06722]
dijabetes1	-0.48009	0.33662	-1.42619	0.15579	[-1.14495, 0.18477]
pretilost1	0.15144	0.50106	0.30225	0.76288	[-0.83855, 1.14143]
hemokromatoza1	-0.43899	0.78553	-0.55885	0.57717	[-1.99221, 1.11423]
arterijska hipertenzija1	0.29463	0.34036	0.86563	0.38800	[-0.37761, 0.96686]
kronična bubrežna insuficijencija1	-0.52952	0.47956	-1.10417	0.27119	[-1.47665, 0.41761]
HIV1	0.13884	1.23659	0.11227	0.91076	[-2.30495, 2.58262]
NASH1	0.59962	0.83557	0.71762	0.47420	[-1.05245, 2.25170]
variksi1	0.23684	0.41420	0.57181	0.56863	[-0.58408, 1.05776]
splenomegalija1	-0.12430	0.33969	-0.36592	0.71496	[-0.79564, 0.54704]
portalna hipertenzija1	-0.19189	0.36854	-0.52068	0.60336	[-0.92009, 0.53631]
tromboza portalne vene1	-1.02962	0.38730	-2.65847	0.00866	[-1.79457, -0.26467]
metastaze1	-1.20576	0.39198	-3.07607	0.00247	[-1.97999, -0.43152]
radiološko obilježje1	0.10697	0.34377	0.31117	0.75608	[-0.57198, 0.78592]

Tablica 2.36: Univarijabilna logistička regresija s kvalitativnim varijablama osim *stupnja stanja pacijenta*, *stupnja encefalopatije* i *stupnja ascitesa*

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
stupanj ascitesa2	-0.83423	0.39316	-2.12187	0.03541	[-1.61075, -0.05771]
stupanj ascitesa3	-1.52737	0.54172	-2.81949	0.00543	[-2.59732, -0.45743]

Tablica 2.37: Univarijabilna logistička regresija sa *stupnjem ascitesa*

Kao i u primjeru sa zastojem srca, kvalitativne varijable su modelirane pomoću *dummy* varijabli kao u (1.25) samo su umjesto oznaka d_k korišteni nazivi odgovarajuće varijable i odgovarajuća kategorijska vrijednost koju varijabla poprima. Dalje, provodimo testove omjera vjerodostojnosti kako bismo izračunali značajnost procijenjenih koeficijenata u odgovarajućim univarijabilnim modelima u odnosu na model bez varijabli:

varijabla u univarijabilnom modelu	testna statistika	p-vrijednost
spol	0.24636	0.61965
simptomi	13.54310	0.00023
alkohol	0.27012	0.60325
HBsAg	0.16533	0.68430
HBeAg	2.02922	0.15430
HBcAb	0.34651	0.55609
HCVAb	2.02974	0.15425
ciroza	0.22942	0.63196
endemska država	1.64454	0.19970
pušenje	0.77800	0.37775
dijabetes	2.03152	0.15407
pretilost	0.09235	0.76121
hemokromatoza	0.30461	0.58100
arterijska hipertenzija	0.75688	0.38431
kronična bubrežna insuficijencija	1.21079	0.27118
HIV	0.01279	0.90997
NASH	0.55877	0.45476
variksi	0.32566	0.56822
splenomegalija	0.13414	0.71418
portalna hipertenzija	0.27318	0.60121
tromboza portalne vene	7.20902	0.00725
metastaze	9.78256	0.00176

radiološko obilježje	0.09660	0.75595
stupanj ascitesa	5.49979	0.00409

Tablica 2.38: Testovi omjera vjerodostojnosti za kvalitativne varijable osim za *stupanj stanja pacijenta* i *stupanj encefalopatije*

Varijable koje imaju p-vrijednost manju od 0.25 su *simptomi*, *HCVAb*, *endemska država*, *dijabetes*, *tromboza portalne vene*, *metastaze* i *stupanj ascitesa*. Vjerojatnost preživljavanja se statistički značajno povećava kod osoba koje su bile u nekoj endemskoj državi dok se statistički značajno smanjuje kod osoba koje imaju simptome, pozitivan test na površinski antigen hepatitisa B, pozitivan test na antitijelo virusa hepatitisa C, dijabetes, trombozu portalne vene ili metastaze kao i prelaskom u viši stupanj ascitesa u odnosu na referentnu kategoriju, a to je najniži mogući stupanj.

Računamo procjene omjera izgleda i pripadne procjene 95% pouzdanih intervala za statistički značajne varijable:

varijabla	procjena omjera izgleda	procjena 95% pouzdanog intervala za omjer izgleda
simptomi1	0.25756	[0.11965, 0.55445]
HCVAb1	0.56944	[0.26174, 1.23889]
endemska država1	2.62857	[0.52569, 13.14348]
dijabetes1	0.61873	[0.31824, 1.20294]
tromboza portalne vene1	0.35714	[0.16620, 0.76746]
metastaze1	0.29947	[0.13807, 0.64952]

Tablica 2.39: Procjena omjera izgleda za statistički značajne kvalitativne varijable osim *stupnja stanja pacijenta*, *stupnja encefalopatije* i *stupnja ascitesa*

varijabla	procjena omjera izgleda	procjena 95% pouzdanog intervala za omjer izgleda
stupanj ascitesa2	0.43422	[0.19974, 0.94393]
stupanj ascitesa3	0.21711	[0.07447, 0.63291]

Tablica 2.40: Procjena omjera izgleda za *stupanj ascitesa*

Gledajući one modele univarijabilne logističke regresije u kojima je kvalitativna varijabla ispala statistički značajnom, omjer izgleda da osoba preživi se:

- smanji 0.25756 puta ako osoba ima simptome
- smanji 0.56944 puta ako osoba ima antitijela na virus hepatitis C
- poveća 2.62857 puta ako je osoba posjetila endemsku državu
- smanji 0.61873 puta ako osoba ima dijabetes
- smanji 0.35714 puta ako osoba ima trombozu portalne vene
- smanji 0.29947 puta ako osoba ima metastaze
- smanji 0.43422 puta ako je stupanj ascitesa 2 u odnosu na osobu čiji je stupanj ascitesa 1
- smanji 0.21711 puta ako je stupanj ascitesa 3 u odnosu na osobu čiji je stupanj ascitesa 1.

Dakle, univarijabilnom analizom dobijemo da su statistički značajne varijable *dob*, *INR*, *hemoglobin*, *leukociti*, *trombociti*, *albumin*, *ukupni bilirubin*, *AST*, *GGT*, *ALP*, *kreatinin*, *najveća dimenzija čvora*, *direktni bilirubin*, *željezo*, *feritin*, *simptomi*, *HCVAb*, *endemska država*, *dijabetes*, *tromboza portalne vene*, *metastaze*, *stupanj stanja pacijenta*, *stupanj encefalopatije* i *stupanj ascitesa*.

Multivarijabilna logistička regresija *stepwise* metodom

Kad imamo skup podataka u kojem nedostaju neki podaci i koristimo metodu višestrukog nadomještanja m puta, tada *stepwise* metodu provodimo na svakom od m skupova podataka posebno i zatim gledamo koliko puta od 0 do m je svaka varijabla odabrana. Varijable koje su odabrane barem $\frac{m}{2}$ puta biramo za daljnju analizu. Počinjemo *stepwise* metodu od praznog modela i uzimajući u obzir sve prethodno navedene statistički značajne varijable tražimo najbolji model. U sljedećim tablicama smo zapisali samo koje nezavisne varijable su u modelu u trenutnom koraku te koliki je trenutni AIC za svaki pojedini skup podataka [44]:

korak	varijable u modelu	AIC
0	bez varijabli	221.4326
1	stupanj stanja pacijenta	199.3131
2	stupanj stanja pacijenta i feritin	184.4084
3	stupanj stanja pacijenta, feritin i ALP	176.3939
4	stupanj stanja pacijenta, feritin, ALP i HCVAb	172.6920
5	stupanj stanja pacijenta, feritin, ALP, HCVAb i dob	168.7997
6	stupanj stanja pacijenta, feritin, ALP, HCVAb, dob i hemoglobin	164.6515
7	stupanj stanja pacijenta, feritin, ALP, HCVAb, dob, hemoglobin i tromboza portalne vene	162.2844
8	stupanj stanja pacijenta, feritin, ALP, HCVAb, dob, hemoglobin, tromboza portalne vene i simptomi	160.9376

9	stupanj stanja pacijenta, feritin, ALP, HCVAb, dob, hemoglobin, tromboza portalne vene, simptomi i GGT	160.7665

Tablica 2.41: *Stepwise* metoda za prvi nadomješteni skup podataka

korak	varijable u modelu	AIC
0	bez varijabli	221.4326
1	stupanj stanja pacijenta	199.3131
2	stupanj stanja pacijenta i ALP	191.8329
3	stupanj stanja pacijenta, ALP i HCVAb	186.6110
4	stupanj stanja pacijenta, ALP, HCVAb i dob	182.2805
5	stupanj stanja pacijenta, ALP, HCVAb, dob i simptomi	179.0352
6	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi i hemoglobin	176.5429
7	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, hemoglobin i tromboza portalne vene	174.4303
8	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, hemoglobin, tromboza portalne vene i feritin	173.2059
9	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, hemoglobin, tromboza portalne vene, feritin i GGT	173.1702

Tablica 2.42: *Stepwise* metoda za drugi nadomješteni skup podataka

korak	varijable u modelu	AIC
0	bez varijabli	221.4326
1	stupanj stanja pacijenta	199.3131
2	stupanj stanja pacijenta i simptomi	190.6949
3	stupanj stanja pacijenta, simptomi i HCVAb	185.2548
4	stupanj stanja pacijenta, simptomi, HCVAb i ALP	180.8506
5	stupanj stanja pacijenta, simptomi, HCVAb, ALP i dob	176.4855
6	stupanj stanja pacijenta, simptomi, HCVAb, ALP, dob i INR	173.0690
7	stupanj stanja pacijenta, simptomi, HCVAb, ALP, dob, INR i AST	172.0747
8	stupanj stanja pacijenta, simptomi, HCVAb, ALP, dob, INR, AST i hemoglobin	170.6816

9	stupanj stanja pacijenta, simptomi, HCVAb, ALP, dob, INR, AST, hemoglobin i najveća dimenzija čvora	169.4526

Tablica 2.43: *Stepwise* metoda za treći nadomješteni skup podataka

korak	varijable u modelu	AIC
0	bez varijabli	221.4326
1	stupanj stanja pacijenta	199.3131
2	stupanj stanja pacijenta i simptomi	190.6949
3	stupanj stanja pacijenta, simptomi i feritin	184.3435
4	stupanj stanja pacijenta, simptomi, feritin i ALP	179.0932
5	stupanj stanja pacijenta, simptomi, feritin, ALP i dob	174.3840
6	stupanj stanja pacijenta, simptomi, feritin, ALP, dob i HCVAb	170.0580
7	stupanj stanja pacijenta, simptomi, feritin, ALP, dob, HCVAb i hemoglobin	166.6612
8	stupanj stanja pacijenta, simptomi, feritin, ALP, dob, HCVAb, hemoglobin i tromboza portalne vene	163.6467
9	stupanj stanja pacijenta, simptomi, feritin, ALP, dob, HCVAb, hemoglobin, tromboza portalne vene i GGT	163.1546
10	stupanj stanja pacijenta, simptomi, feritin, ALP, dob, HCVAb, hemoglobin, tromboza portalne vene, GGT i najveća dimenzija čvora	161.7885
11	stupanj stanja pacijenta, simptomi, feritin, ALP, dob, HCVAb, hemoglobin, tromboza portalne vene, GGT, najveća dimenzija čvora i INR	161.2422
12	simptomi, feritin, ALP, dob, HCVAb, hemoglobin, tromboza portalne vene, GGT, najveća dimenzija čvora i INR	160.7374
13	simptomi, feritin, ALP, dob, HCVAb, hemoglobin, tromboza portalne vene, GGT, najveća dimenzija čvora, INR i endemska država	160.5447

Tablica 2.44: *Stepwise* metoda za četvrti nadomješteni skup podataka

korak	varijable u modelu	AIC
0	bez varijabli	221.4326
1	stupanj stanja pacijenta	199.3131

2	stupanj stanja pacijenta i ALP	190.5379
3	stupanj stanja pacijenta, ALP i HCVAb	184.4566
4	stupanj stanja pacijenta, ALP, HCVAb i dob	180.5708
5	stupanj stanja pacijenta, ALP, HCVAb, dob i simptomi	176.9076
6	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi i INR	173.4940
7	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, INR i AST	172.1509
8	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, INR, AST i željezo	170.0375
9	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, INR, AST, željezo i dijabetes	168.7223
10	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, INR, AST, željezo, dijabetes i najveća dimenzija čvora	168.1834
11	stupanj stanja pacijenta, ALP, HCVAb, dob, simptomi, INR, AST, željezo, dijabetes i najveća dimenzija čvora i endemska država	167.9516

Tablica 2.45: *Stepwise* metoda za peti nadomješteni skup podataka

Sada gledamo za svaku od varijabli za koliko skupova podataka se nalazi u modelu u završnom koraku *stepwise* procedure:

varijabla	broj skupova podataka za koje se nalazi u modelu u završnom koraku
dob	5
INR	3
hemoglobin	4
leukociti	0
trombociti	0
albumin	0
ukupni bilirubin	0
AST	2
GGT	3
ALP	5
kreatinin	0
najveća dimenzija čvora	3
direktni bilirubin	0
željezo	1
feritin	3

simptomi	5
HCVAb	5
endemska država	2
dijabetes	1
tromboza portalne vene	3
metastaze	0
stupanj stanja pacijenta	4
stupanj encefalopatije	0
stupanj ascitesa	0

Tablica 2.46: Broj skupova podataka za koje se svaka pojedina varijabla našla u modelu u završnom koraku *stepwise* metode

Dakle, varijable koje smo dobili u završnom koraku modela *stepwise* metodom u 3 ili više skupova podataka su *dob*, *INR*, *hemoglobin*, *GGT*, *ALP*, *najveća dimenzija čvora*, *feritin*, *simptomi*, *HCVAb*, *tromboza portalne vene* i *stupanj stanja pacijenta*.

Provedimo multivarijabilnu analizu koristeći navedene varijable:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
slobodni koeficijent	5.49627	2.42195	2.26936	0.02494	[0.70350, 10.28904]
dob	-0.05681	0.01961	-2.89652	0.00446	[-0.09564, -0.01799]
INR	-0.83102	0.63182	-1.31528	0.19148	[-2.08481, 0.42278]
hemoglobin	0.24925	0.11916	2.09175	0.03867	[0.01321, 0.48529]
GGT	0.00146	0.00108	1.35541	0.17797	[-0.00067, 0.00359]
ALP	-0.00480	0.00188	-2.54535	0.01208	[-0.00852, -0.00107]
najveća dimenzija čvora	-0.06852	0.04430	-1.54685	0.12531	[-0.15650, 0.01945]
feritin	-0.00102	0.00082	-1.24262	0.25050	[-0.00294, 0.00089]
simptomi1	-0.85850	0.47756	-1.79768	0.07457	[-1.80337, 0.08638]
HCVAb1	-1.77181	0.55251	-3.20685	0.00166	[-2.86412, -0.67951]
tromboza portalne vene1	-0.98255	0.55181	-1.78059	0.07731	[-2.07420, 0.10910]

stupanj stanja pacijenta	-0.37859	0.22207	-1.70481	0.09174	[-0.81987, 0.06270]

Tablica 2.47: Multivarijabilna logistička regresija s varijablama dobivenim nakon *stepwise* metode

Provjera značajnosti varijabli

Gledajući tablicu 2.47 vidimo da na razini značajnosti od 10% nisu statistički značajne varijable *INR*, *GGT*, *najveća dimenzija čvora* i *feritin*. Želimo ih izbaciti pa pomoću testa omjera vjerodostojnosti testiramo je li potreban prošireni model koji sadrži i te četiri varijable ili je dovoljan podmodel koji ih ne sadrži. Dobijemo da je vrijednost testne statistike 1.72853, a odgovarajuća p-vrijednost 0.15232 iz čega zaključujemo da je dovoljan model bez tih varijabli. Ako pokušamo iz modela iz tablice 2.47 izbaciti i varijable koje nisu statistički značajne na razini značajnosti od 5%, a to su još *simptomi*, *tromboza portalne vene* i *stupanj stanja pacijenta*, test omjera vjerodostojnosti daje vrijednost testne statistike 3.72557 i p-vrijednost 0.00068 iz čega zaključujemo da su te varijable ipak potrebne u modelu.

Multivarijabilna analiza za dobiveni model dana je u sljedećoj tablici:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
slobodni koeficijent	2.73448	1.79058	1.52714	0.12881	[-0.80326, 6.27222]
dob	-0.04348	0.01655	-2.62679	0.00949	[-0.07617, -0.01078]
hemoglobin	0.24283	0.10465	2.32049	0.02164	[0.03609, 0.44957]
ALP	-0.00319	0.00135	-2.36945	0.01909	[-0.00586, -0.00053]
simptomi1	-0.97585	0.44981	-2.16949	0.03193	[-1.86605, -0.08565]
HCVAb1	-1.78025	0.52568	-3.38657	0.00090	[-2.81876, -0.74173]
tromboza portalne vene1	-0.95053	0.48585	-1.95643	0.05222	[-1.91028, 0.00923]
stupanj stanja pacijenta	-0.45221	0.18683	-2.42051	0.01666	[-0.82127, -0.08315]

Tablica 2.48: Multivarijabilna logistička regresija za model bez *INR-a*, *GGT-a*, *najveće dimenzije čvora* i *feritina*

Sada su sve varijable statistički značajne na razini značajnosti od 5% osim *tromboze portalne vene*. Testom omjera vjerodostojnosti testiramo je li potreban model s varijablama iz tablice 2.48 ili je dovoljan podmodel bez *tromboze portalne vene* i dobijemo da je vrijednost testne statistike 3.79770 te da je pripadna p-vrijednost 0.05136 iz čega možemo zaključiti da je dovoljan model bez te varijable. Ponovno procjenjujemo multivarijabilni regresijski model:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
slobodni koeficijent	2.88656	1.78772	1.61466	0.10843	[-0.64502, 6.41814]
dob	-0.03991	0.01632	-2.44593	0.01556	[-0.07214, -0.00768]
hemoglobin	0.20358	0.10074	2.02073	0.04504	[0.00456, 0.40259]
ALP	-0.00330	0.00133	-2.47487	0.01442	[-0.00593, -0.00066]
simptomi1	-1.04762	0.44144	-2.37318	0.01909	[-1.92091, -0.17433]
HCVAb1	-1.58747	0.50699	-3.13119	0.00208	[-2.58899, -0.58596]
stupanj stanja pacijenta	-0.55416	0.18048	-3.07057	0.00252	[-0.91067, -0.19766]

Tablica 2.49: Multivarijabilna logistička regresija za model bez *INR-a*, *GGT-a*, *najveće dimenzije čvora*, *feritina* i *tromboze portalne vene*

Preostaje još usporediti procijenjene koeficijente iz trenutnog multivarijabilnog modela s onima iz odgovarajućeg univarijabilnog modela te zatim s onima iz punog multivarijabilnog modela. Rezultati su u sljedećim tablicama:

varijabla	procjena koeficijenta u multivarijabilnom modelu	procjena koeficijenta u univarijabilnom modelu
dob	-0.03991	-0.02393
hemoglobin	0.20358	0.30091
ALP	-0.00330	-0.00405
simptomi1	-1.04762	-1.35650
HCVAb1	-1.58747	-0.56309
stupanj stanja pacijenta	-0.55416	-0.69984

Tablica 2.50: Usporedba procjena koeficijenata iz multivarijabilnog modela s varijablama iz tablice 2.49 i iz odgovarajućih univarijabilnih modela

varijabla	procjena koeficijenta u multivarijabilnom modelu	procjena koeficijenta u punom modelu
dob	-0.03991	-0.15588
hemoglobin	0.20358	0.29785
ALP	-0.00330	-0.00943
simptomi1	-1.04762	-2.40855
HCVAb1	-1.58747	-2.34481
stupanj stanja pacijenta	-0.55416	-1.05123

Tablica 2.51: Usporedba procjena koeficijenata iz multivarijabilnog modela s varijablama iz tablice 2.49 i iz punog multivarijabilnog modela

Zaključujemo da daljnju analizu nastavljamo s varijablama *dob*, *hemoglobin*, *ALP*, *simptomi*, *HCVAb* i *stupanj stanja pacijenta*.

Provjera linearnosti logit funkcije za neprekidne varijable

Neprekidne nezavisne varijable koje imamo u modelu su *dob*, *hemoglobin* i *ALP*. Kako bismo provjerili linearnost logit funkcije ponovno koristimo Box-Tidwellov test te u model dodajemo interakcije *dob* $\times$ ln(*dob*), *hemoglobin* $\times$ ln(*hemoglobin*) i *ALP* $\times$ ln(*ALP*). U sljedećoj tablici dane su procjene koeficijenata, pripadne standardne greške, Waldove testne statistike i pripadne p-vrijednosti:

varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost
slobodni koeficijent	85.93311	75.19854	1.14275	0.25494
dob	5.80426	4.41548	1.31452	0.19065
ln(dob)	-47.91101	37.55661	-1.27570	0.20401
hemoglobin	4.10559	22.73138	0.18061	0.85691
ln(hemoglobin)	-4.76695	55.44420	-0.08598	0.93160
ALP	-0.07464	0.05324	-1.40177	0.16301
ln(ALP)	0.02877	1.66886	0.01724	0.98627
dob $\times$ ln(dob)	-0.98437	0.74120	-1.32808	0.18615
hemoglobin $\times$ ln(hemoglobin)	-0.98330	5.18258	-0.18973	0.84977
ALP $\times$ ln(ALP)	0.01027	0.00702	1.46355	0.14537

Tablica 2.52: Box-Tidwellov test

Vidimo da nijedna interakcija nije statistički značajna jer su sve pripadne p-vrijednosti veće od 0.05 pa zaključujemo da vrijedi linearnost logit funkcije za sve neprekidne varijable.

Interakcije

Interakcije za koje želimo ispitati jesu li statistički značajne su *dob* × *hemoglobin*, *dob* × *ALP*, *dob* × *simptomi*, *dob* × *HCVAb*, *dob* × *stupanj stanja pacijenta*, *hemoglobin* × *ALP*, *simptomi* × *HCVAb* i *simptomi* × *stupanj stanja pacijenta*. Navedene interakcije dodajemo u model jednu po jednu te zatim testom omjera vjerodostojnosti testiramo je li bolji prošireni model s varijablama iz tablice 2.49 i interakcijom ili njegov podmodel s istim varijablama samo bez interakcije.

dodana interakcija	testna statistika	p-vrijednost
dob × hemoglobin	0.71487	0.39786
dob × ALP	2.27165	0.13177
dob × simptomi	1.04039	0.30785
dob × HCVAb	1.11047	0.29198
dob × stupanj stanja pacijenta	0.08466	0.77108
hemoglobin × ALP	3.92606	0.04769
simptomi × HCVAb	0.66595	0.41457
simptomi × stupanj stanja pacijenta	0.00235	0.96135

Tablica 2.53: Testovi omjera vjerodostojnosti za testiranje modela s interakcijama u odnosu na model bez njih

Uočimo da je na razini značajnosti od 5% statistički značajna jedino interakcija *hemoglobin* × *ALP* pa ju uključujemo u konačan model.

Radimo multivarijabilnu analizu dobivenog modela:

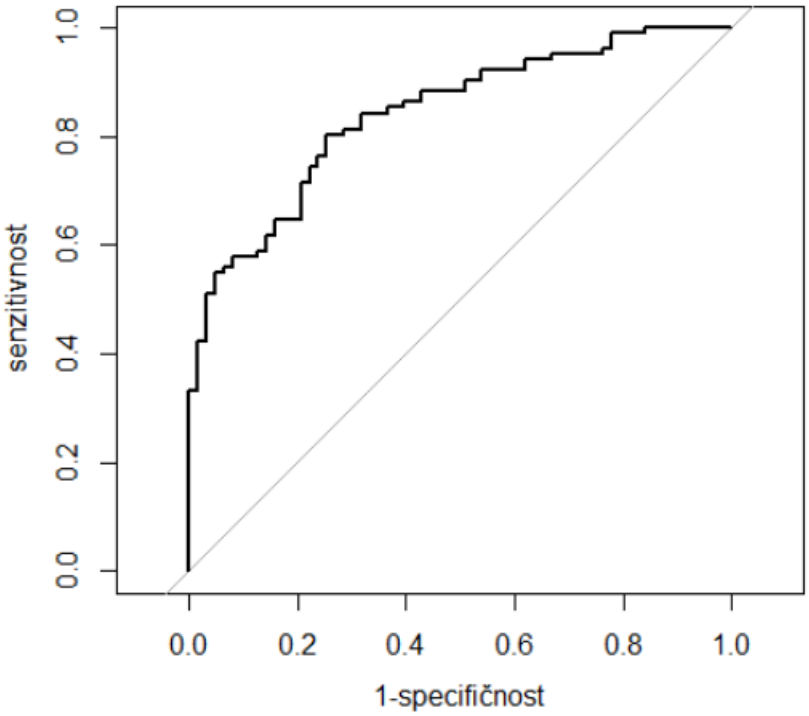
varijabla	procjena koeficijenta	standardna greška	Waldov test	p-vrijednost	procjena 95% pouzdanog intervala
slobodni koeficijent	-0.39517	2.44605	-0.16155	0.87187	[-5.22761, 4.43728]
dob	-0.03630	0.01640	-2.21318	0.02836	[-0.06871, -0.00390]
hemo-globin	0.45357	0.16735	2.71037	0.00749	[0.12297, 0.78417]
ALP	0.01246	0.00812	1.53391	0.12717	[-0.00359, 0.02852]
simptomi1	-0.96093	0.44828	-2.14360	0.03395	[-1.84792, -0.07394]
HCVAb1	-1.56913	0.51571	-3.04265	0.00276	[-2.58789, -0.55036]
stupanj stanja pacijenta	-0.53527	0.18248	-2.93329	0.00387	[-0.89576, -0.17478]

hemo-globin × ALP	-0.00134	0.00069	-1.93526	0.05485	[-0.00271, 0.00003]

Tablica 2.54: Multivarijabilna logistička regresija konačnog modela

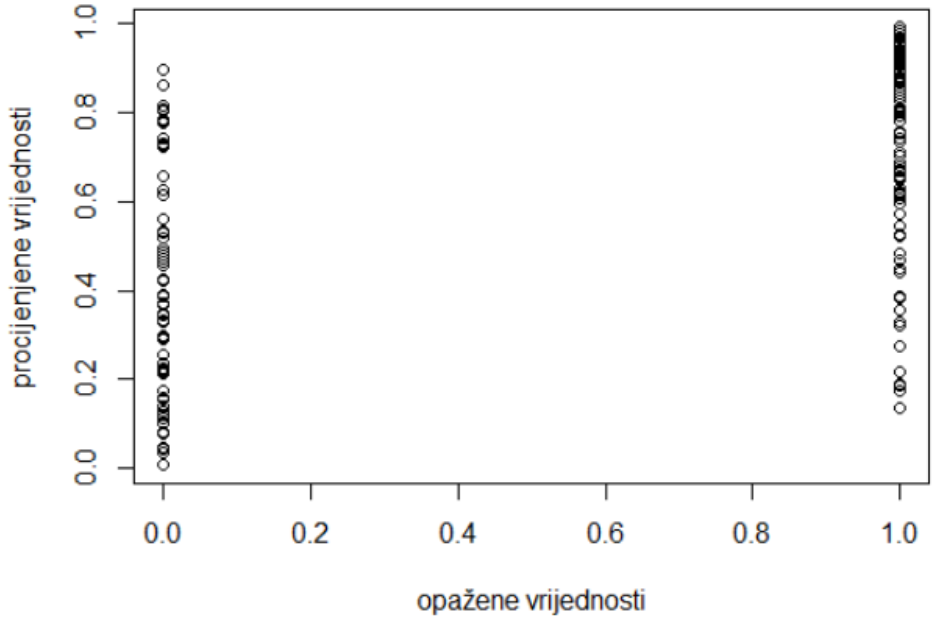
2.2.3 Procjena adekvatnosti modela

Koristeći Rubinova pravila iz [23] kako bismo izračunali procijenjene vrijednosti procjenjujemo adekvatnost modela. Površina ispod ROC krivulje je 0.8405 što znači da je model odličan. Na sljedećoj slici prikazujemo ROC krivulju:



Slika 2.10: ROC krivulja

Na grafu opaženih i procijenjenih vrijednosti



Slika 2.11: Grafički prikaz opaženih i procijenjenih vrijednosti

vidimo da se za opaženu vrijednost 0 nešto više procijenjenih vrijednosti nalazi bliže 0, a za opaženu vrijednost 1 više oko 1.

Iz svega navedenog zaključujemo da model prilično dobro opisuje podatke.

2.2.4 Zaključak

Nakon detaljne analize zaključujemo da je najbolji model onaj u kojem su nezavisne varijable *dob*, *hemoglobin*, *ALP*, *simptomi*, *antitijelo virusa hepatitisa C*, *stupanj stanja pacijenta* i interakcija *hemoglobin* $\times$ *ALP*. Ako označimo s $\boldsymbol{x}$ = (*dob, hemoglobin, ALP, simptomi, HCVAb, stupanj stanja pacijenta, hemoglobin* $\times$ *ALP*) i $\pi(\boldsymbol{x}) = P(preživljavanje = 1|\boldsymbol{x})$ tada iz tablice 2.54 možemo iščitati da je pripadna logit funkcija

$$\begin{aligned} g(\boldsymbol{x}) = &-0.39517 - 0.03630 \cdot dob + 0.45357 \cdot hemoglobin + 0.01246 \cdot ALP - \\ &- 0.96093 \cdot simptomi - 1.56913 \cdot HCVAb - 0.53527 \cdot stupanj\,stanja \\ &pacijenta - 0.00134 \cdot hemoglobin \times ALP \end{aligned} \tag{2.2}$$

a model multivarijabilne logističke regresije

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}$$

uz g($\boldsymbol{x}$) iz (2.2).

Dodatak A

Kod u R-u

A.1 Zatajenje srca

Učitavanje dodatnih paketa i podataka te mijenjanje imena stupaca u tablici s podacima:

```
library(DescTools)
library(data.table)
library(lmtest)
library(pROC)

podaci<-read.csv
    ("heart_failure_clinical_records_dataset.csv")
setnames(podaci,old=c(1:13),new=c("dob","anemija","CPK",
    "dijabetes","ejekcijska","visoki_tlak","trombociti",
    "kreatinin","natrij","spol","pusenje","pracenje",
    "smrt"))
```

Pretvaranje kvalitativnih varijabli u faktore:

```
indeksi<-c(2,4,6,10,11)
podaci[,indeksi]<-lapply(podaci[,indeksi],factor)
```

Grafički prikaz distribucije dobi po spolu:

```
par(font.main=1,cex.main=1)
boxplot(podaci$dob~podaci$spol,main="Distribucija dobi po
    spolu",xlab="spol",ylab="dob",col=c("gray60","gray85"),
    names=c("zene","muskarci"))
```

Deskriptivna statistika:

```
kvantitativne<-subset(podaci,select=c(1,3,5,7,8,9,12))
apply(kvantitativne,2,function(x) list(mean(x),sd(x),
    quantile(x)))
kvalitativne<-subset(podaci,select=c(2,4,6,10,11,13))
apply(kvalitativne,2,function(x) list(data.frame(table(x)),
    data.frame(prop.table(table(x))*100)))
```

Distribucija svake pojedine kvalitativne varijable grupirane prema smrtnom ishodu:

```
boxplot(podaci$dob~podaci$smrt,main="Distribucija dobi
    grupirane po \n smrtnom ishodu",xlab="smrt",ylab="dob",
    col=c("gray60","gray85"),names=c("smrt NE","smrt DA"))
boxplot(podaci$CPK~podaci$smrt,main="Distribucija CPK-a
    grupiranog po \n smrtnom ishodu",xlab="smrt",ylab="CPK",
    col=c("gray60","gray85"),names=c("smrt NE","smrt DA"))
boxplot(podaci$ejekcijska~podaci$smrt,main="Distribucija
    ejekcijske frakcije grupirane \n po smrtnom ishodu",
    xlab="smrt",ylab="ejekcijska frakcija",col=c("gray60",
    "gray85"),names=c("smrt NE","smrt DA"))
boxplot(podaci$trombociti~podaci$smrt,main="Distribucija
    trombocita grupiranih po \n smrtnom ishodu",xlab="smrt",
    ylab="trombociti",col=c("gray60","gray85"),names=c(
    "smrt NE","smrt DA"))
boxplot(podaci$kreatinin~podaci$smrt,main="Distribucija
    kreatinina grupiranog po \n smrtnom ishodu",xlab="smrt",
    ylab="kreatinin",col=c("gray60","gray85"),names=c(
    "smrt NE","smrt DA"))
boxplot(podaci$natrij~podaci$smrt,main="Distribucija natrija
    grupiranog po \n smrtnom ishodu",xlab="smrt",ylab=
    "natrij",col=c("gray60","gray85"),names=c("smrt NE",
    "smrt DA"))
boxplot(podaci$pracenje~podaci$smrt,main="Distribucija
    vremena pracenja \n grupiranog po smrtnom ishodu",xlab=
    "smrt",ylab="pracenje",col=c("gray60","gray85"),names=c(
    "smrt NE","smrt DA"))
```

Univarijabilna logistička regresija i test omjera vjerodostojnosti za značajnost koeficijenta za svaku od kvantitativnih varijabli te procjena omjera izgleda i procjena 95% pouzdanog intervala za omjer izgleda za varijable koje su ispale statistički značajne:

```
model_dob<-glm(smrt~dob,data=podaci,family="binomial")
summary(model_dob)
round(confint.default(model_dob,level=0.95),5)
lrtest(model_dob)
round(exp(coefficients(model_dob)[2]),5)
round(exp(confint.default(model_dob,level=0.95)),5)

model_CPK<-glm(smrt~CPK,data=podaci,family="binomial")
summary(model_CPK)
round(confint.default(model_CPK,level=0.95),5)
lrtest(model_CPK)

model_ejekcijska<-glm(smrt~ejekcijska,data=podaci,family=
    "binomial")
summary(model_ejekcijska)
round(confint.default(model_ejekcijska,level=0.95),5)
lrtest(model_ejekcijska)
round(exp(coefficients(model_ejekcijska)[2]),5)
round(exp(confint.default(model_ejekcijska,level=0.95)),5)

model_trombociti<-glm(smrt~trombociti,data=podaci,family=
    "binomial")
summary(model_trombociti)
round(confint.default(model_trombociti,level=0.95),6)
lrtest(model_trombociti)

model_kreatinin<-glm(smrt~kreatinin,data=podaci,family=
    "binomial")
summary(model_kreatinin)
round(confint.default(model_kreatinin,level=0.95),5)
lrtest(model_kreatinin)
round(exp(coefficients(model_kreatinin)[2]),5)
round(exp(confint.default(model_kreatinin,level=0.95)),5)

model_natrij<-glm(smrt~natrij,data=podaci,family="binomial")
summary(model_natrij)
round(confint.default(model_natrij,level=0.95),5)
lrtest(model_natrij)
round(exp(coefficients(model_natrij)[2]),5)
```

```
round(exp(confint.default(model_natrij,level=0.95)),5)

model_pracenje<-glm(smrt~pracenje,data=podaci,family=
    "binomial")
summary(model_pracenje)
round(confint.default(model_pracenje,level=0.95),5)
lrtest(model_pracenje)
round(exp(coefficients(model_pracenje)[2]),5)
round(exp(confint.default(model_pracenje,level=0.95)),5)
```

Grafički prikaz vjerojatnosti smrti u ovisnosti o kvantitativnim varijablama:

```
plot(podaci$dob,predict(model_dob,type="response"),xlab=
    "dob",ylab="predvidena vjerojatnost smrti",main=
    "Graficki prikaz vjerojatnosti smrti u ovisnosti \n o
    dobi")
plot(podaci$CPK,predict(model_CPK,type="response"),xlab=
    "CPK",ylab="predvidena vjerojatnost smrti",main=
    "Graficki prikaz vjerojatnosti smrti u ovisnosti
    \n o CPK-u")
plot(podaci$ejekcijska,predict(model_ejekcijska,type=
    "response"),xlab="ejekcijska frakcija",ylab="predvidena
    vjerojatnost smrti",main="Graficki prikaz vjerojatnosti
    smrti u ovisnosti o \n ejekcijskoj frakciji")
plot(podaci$trombociti,predict(model_trombociti,type=
    "response"),xlab="trombociti",ylab="predvidena
    vjerojatnost smrti",main="Graficki prikaz vjerojatnosti
    smrti u ovisnosti o \n trombocitima")
plot(podaci$kreatinin,predict(model_kreatinin,type=
    "response"),xlab="kreatinin",ylab="predvidena
    vjerojatnost smrti",main="Graficki prikaz vjerojatnosti
    smrti u ovisnosti \n o kreatininu")
plot(podaci$natrij,predict(model_natrij,type="response"),
    xlab="natrij",ylab="predvidena vjerojatnost smrti",main=
    "Graficki prikaz vjerojatnosti smrti u ovisnosti \n o
    natriju")
plot(podaci$pracenje,predict(model_pracenje,type=
    "response"),xlab="pracenje",ylab="predvidena
    vjerojatnost smrti",main="Graficki prikaz
    vjerojatnosti smrti u ovisnosti\n o vremenu pracenja")
```

Kontingencijske tablice za kvalitativne nezavisne varijable:

```
apply(kvalitativne,2,function(x) Desc(table(podaci$smrt,x)))
```

Grafički prikazi frekvencija kvalitativnih varijabli grupiranih prema smrtnom ishodu:

```
barplot(table(podaci$smrt,podaci$anemija),beside=TRUE,xlab=
    "anemija",names.arg=c("nema anemiju","ima anemiju"),col=
    c("gray60","gray85"),main="Frekvencije anemije grupirane
    prema \n smrtnom ishodu")
legend("topright",legend=c("smrt NE","smrt DA"),fill=c(
    "gray60","gray85"),bty="n")
barplot(table(podaci$smrt,podaci$dijabetes),beside=TRUE,
    xlab="dijabetes",names.arg=c("nema dijabetes","ima
    dijabetes"),col=c("gray60","gray85"),main="Frekvencije
    dijabetesa grupiranog \n prema smrtnom ishodu")
legend("topright",legend=c("smrt NE","smrt DA"),fill=c(
    "gray60","gray85"),bty="n")
barplot(table(podaci$smrt,podaci$visoki_tlak),beside=TRUE,
    xlab="visoki tlak",names.arg=c("nema visoki tlak","ima
    visoki tlak"),col=c("gray60","gray85"),main="Frekvencije
    visokog tlaka grupiranog \n prema smrtnom ishodu")
legend("topright",legend=c("smrt NE","smrt DA"),fill=c(
    "gray60","gray85"),bty="n")
barplot(table(podaci$smrt,podaci$spol),beside=TRUE,xlab=
    "spol",names.arg=c("zene","muskarci"),col=c("gray60",
    "gray85"),main="Frekvencije spola grupiranog prema \n
    smrtnom ishodu")
legend("topleft",legend=c("smrt NE","smrt DA"),fill=c(
    "gray60","gray85"),bty="n")
barplot(table(podaci$smrt,podaci$pusenje),beside=TRUE,xlab=
    "pusenje",names.arg=c("ne pusi","pusi"),col=c("gray60",
    "gray85"),main="Frekvencije pusenja grupiranog \n prema
    smrtnom ishodu")
legend("topright",legend=c("smrt NE","smrt DA"),fill=c(
    "gray60","gray85"),bty="n")
```

Univarijabilna logistička regresija i test omjera vjerodostojnosti za značajnost koeficijenta za svaku kvalitativnu varijablu te procjena omjera izgleda i procjena pripadnog 95% pouzdanog intervala za omjer izgleda za varijablu koja je ispala statistički značajna:

```
model_anemija<-glm(smrt~anemija,data=podaci,family=
```

```
    "binomial")
summary(model_anemija)
round(confint.default(model_anemija, level=0.95),5)
lrtest(model_anemija)

model_dijabetes<-glm(smrt~dijabetes, data=podaci, family=
    "binomial")
summary(model_dijabetes)
round(confint.default(model_dijabetes, level=0.95),5)
lrtest(model_dijabetes)

model_visoki_tlak<-glm(smrt~visoki_tlak, data=podaci, family=
    "binomial")
summary(model_visoki_tlak)
round(confint.default(model_visoki_tlak, level=0.95),5)
lrtest(model_visoki_tlak)
round(exp(coefficients(model_visoki_tlak)[2]),5)
round(exp(confint.default(model_visoki_tlak, level=0.95)),5)

model_spol<-glm(smrt~spol, data=podaci, family="binomial")
summary(model_spol)
round(confint.default(model_spol, level=0.95),5)
lrtest(model_spol)

model_pusenje<-glm(smrt~pusenje, data=podaci, family=
    "binomial")
summary(model_pusenje)
round(confint.default(model_pusenje, level=0.95),5)
lrtest(model_pusenje)
```

Stepwise metoda i multivarijabilna logistička regresija s varijablama dobivenim u završnom koraku *stepwise* metode:

```
model_bez<-glm(smrt~1, data=podaci, family="binomial")
model_odabrane<-glm(podaci$smrt~., data=
    podaci[c(1,5,6,8,9,12)], family="binomial")
model_stepwise<-step(model_bez, data=podaci, scope=list(lower=
    model_bez, upper=model_odabrane), direction="both", trace=
    TRUE, test="LRT")
summary(model_stepwise)
```

```
round(confint.default(model_stepwise,level=0.95),5)
```

Testiranje je li potreban model s *natrijem* ili je dovoljan podmodel bez *natrija*:

```
model_bez_natrija<-glm(smrt~pracenje+ejekcijska+kreatinin+
    dob,data=podaci,family="binomial")
summary(model_bez_natrija)
round(confint.default(model_bez_natrija,level=0.95),5)
lrtest(model_bez_natrija,model_stepwise)
```

Procjene koeficijenata punog multivarijabilnog modela:

```
model_puni<-glm(podaci$smrt~.,data=podaci[-13],family=
    "binomial")
summary(model_puni)
```

Provjera linearnosti logit funkcije za neprekidne nezavisne varijable:

```
model_linearnost<-glm(smrt~pracenje+pracenje*log(pracenje)+
    ejekcijska+ejekcijska*log(ejekcijska)+kreatinin+
    kreatinin*log(kreatinin)+dob*log(dob),data=podaci,
    family="binomial")
summary(model_linearnost)
```

Provjera jesu li interakcije statistički značajne:

```
model_interakcije1<-glm(smrt~pracenje+ejekcijska+kreatinin+
    dob+pracenje+dob*kreatinin,data=podaci,family=
    "binomial")
lrtest(model_interakcije1,model_bez_natrija)

model_interakcije2<-glm(smrt~pracenje+ejekcijska+kreatinin+
    dob+dob*ejekcijska,data=podaci,family="binomial")
lrtest(model_interakcije2,model_bez_natrija)
```

Procjena adekvatnosti modela:

```
procijenjene_vjerojatnosti<-predict(model_bez_natrija,
    type="response")
roc_krivulja<-roc(podaci$smrt,procijenjene_vjerojatnosti,
    direction="<",print.auc=TRUE)
plot(roc_krivulja,legacy.axes=TRUE,xlab="1-specificnost",
    ylab="senzitivnost")
plot(podaci$smrt,procijenjene_vjerojatnosti,xlab="opazene
    vrijednosti",ylab="procijenjene vrijednosti")
```

A.2 Karcinom jetre

Učitavanje dodatnih paketa te podataka i mijenjanje imena stupaca u tablici s učitanim podacima:

```
library(DescTools)
library(data.table)
library(VIM)
library(mice)
library(pROC)

podaci<-read.table("hcc-data.txt",sep=",")
setnames(podaci,old=c(1:50),new=c("spol","simptomi",
    "alkohol","HBsAg","HBeAg","HBcAb","HCVAb","ciroza",
    "drzava","pusenje","dijabetes","pretilost",
    "hemokromatoza","AHT","CRI","HIV","NASH","variksi",
    "splenomegalija","PHT","PVT","metastaze","obiljezje",
    "dob","alkohol_grami","cigarete","PS","encefalopatija",
    "ascites","INR","AFP","hemoglobin","MCV","leukociti",
    "trombociti","albumin","ukupni_bilirubin","ALT","AST",
    "GGT","ALP","TP","kreatinin","broj_cvorova","dimenzija",
    "direktni_bilirubin","zeljezo","zasicenost","feritin",
    "prezivljavanje"))
```

Pretvaranje podataka u numeričke vrijednosti kako bismo upitnike iz originalnih podataka pretvorili u NA (eng. *not available)* da bi ih R-ove funkcije mogle prepoznati kao podatke koji nedostaju:

```
podaci[]<-lapply(podaci,function(x) as.numeric(
    as.character(x)))
```

Definiranje kvalitativnih varijabli kao faktora:

```
indeksi<-c(1:23,27:29)
podaci[,indeksi]<-lapply(podaci[,indeksi],factor)
```

Koliko podataka od ukupne količine podataka danih nezavisnim varijablama nedostaje:

```
nedostatak<-sum(is.na(podaci))
ukupno<-dim(podaci)[1]*(dim(podaci)[2]-1)
postotak_ukupni<-(nedostatak/ukupno)*100
```

Koliko podataka nedostaje u svakoj varijabli:

```
broji<-sort(sapply(podaci,function(y) sum(length(which
    (is.na(y))))),decreasing=TRUE)
postotak<-round((sort(sapply(podaci,function(y) sum(length
    (which(is.na(y))))),decreasing=TRUE)/165)*100,2)
```

Grafički prikazi podataka koji nedostaju:

```
stupci<-c("45","15","46","16","9","11","22","47","10","8",
    "30","21","12","31","40","18","13","5","17","19","32",
    "27","41","48","6","4","49","44","42","28","23","33",
    "34","35","36","25","26","29","37","38","39","20","24",
    "43","14","7","3","1","2","50")
missing_vrijednosti<-aggr(podaci,col=c("gray85","gray60"),
    sortVars=TRUE,labels=stupci,ylab=c("Podaci koji
    nedostaju","Uzorak podataka koji nedostaju"))
```

Nadomještanje podataka koji nedostaju $m = 5$ puta i spremanje nadomještenih podataka u varijablu *dopunjeni_podaci*:

```
dopunjeni<-mice(podaci,m=5,seed=123,method=c("","polyreg",
    "","polyreg","polyreg","polyreg","polyreg","","polyreg",
    "polyreg","polyreg","polyreg","polyreg","polyreg",
    "polyreg","polyreg","polyreg","polyreg","polyreg",
    "polyreg","polyreg","polyreg","polyreg","","pmm","pmm",
    "","polr","polr","pmm","pmm","pmm","pmm","pmm","pmm",
    "pmm","pmm","pmm","pmm","pmm","pmm","pmm","pmm","pmm",
    "pmm","pmm","pmm","pmm","pmm",""))
dopunjeni_podaci<-complete(dopunjeni,"long")
```

Grafički prikaz distribucije dobi po spolu:

```
par(font.main=1,cex.main=1)
boxplot(podaci$dob~podaci$spol,main="Distribucija dobi po
    spolu",col=c("gray60","gray85"),names=c("zene",
    "muskarci"),xlab="spol",ylab="dob")
```

Deskriptivna analiza kvantitativnih varijabli osim varijable *broj čvorova*:

```
mean(podaci$dob)
sd(podaci$dob)
quantile(podaci$dob)

mean(with(dopunjeni_podaci,tapply(alkohol_grami,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(alkohol_grami,.imp,sd)))
```

```
kvartili_alkohol_grami<-with(dopunjeni_podaci,tapply(
    alkohol_grami,.imp,quantile))
(kvartili_alkohol_grami[[1]]+kvartili_alkohol_grami[[2]]+
    kvartili_alkohol_grami[[3]]+kvartili_alkohol_grami[[4]]+
    kvartili_alkohol_grami[[5]])/5

mean(with(dopunjeni_podaci,tapply(cigarete,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(cigarete,.imp,sd)))
kvartili_cigarete<-with(dopunjeni_podaci,tapply(cigarete,
    .imp,quantile))
(kvartili_cigarete[[1]]+kvartili_cigarete[[2]]+
    kvartili_cigarete[[3]]+kvartili_cigarete[[4]]+
    kvartili_cigarete[[5]])/5

mean(with(dopunjeni_podaci,tapply(INR,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(INR,.imp,sd)))
kvartili_INR<-with(dopunjeni_podaci,tapply(INR,.imp,
    quantile))
(kvartili_INR[[1]]+kvartili_INR[[2]]+kvartili_INR[[3]]+
    kvartili_INR[[4]]+kvartili_INR[[5]])/5

mean(with(dopunjeni_podaci,tapply(AFP,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(AFP,.imp,sd)))
kvartili_AFP<-with(dopunjeni_podaci,tapply(AFP,.imp,
    quantile))
(kvartili_AFP[[1]]+kvartili_AFP[[2]]+kvartili_AFP[[3]]+
    kvartili_AFP[[4]]+kvartili_AFP[[5]])/5

mean(with(dopunjeni_podaci,tapply(hemoglobin,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(hemoglobin,.imp,sd)))
kvartili_hemoglobin<-with(dopunjeni_podaci,tapply(
    hemoglobin,.imp,quantile))
(kvartili_hemoglobin[[1]]+kvartili_hemoglobin[[2]]+
kvartili_hemoglobin[[3]]+kvartili_hemoglobin[[4]]+
    kvartili_hemoglobin[[5]])/5

mean(with(dopunjeni_podaci,tapply(MCV,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(MCV,.imp,sd)))
kvartili_MCV<-with(dopunjeni_podaci,tapply(MCV,.imp,
```

```
    quantile))
(kvartili_MCV[[1]]+kvartili_MCV[[2]]+kvartili_MCV[[3]]+
    kvartili_MCV[[4]]+kvartili_MCV[[5]])/5

mean(with(dopunjeni_podaci,tapply(leukociti,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(leukociti,.imp,sd)))
kvartili_leukociti<-with(dopunjeni_podaci,tapply(leukociti,
    .imp,quantile))
(kvartili_leukociti[[1]]+kvartili_leukociti[[2]]+
    kvartili_leukociti[[3]]+kvartili_leukociti[[4]]+
    kvartili_leukociti[[5]])/5

mean(with(dopunjeni_podaci,tapply(trombociti,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(trombociti,.imp,sd)))
kvartili_trombociti<-with(dopunjeni_podaci,tapply(
    trombociti,.imp,quantile))
(kvartili_trombociti[[1]]+kvartili_trombociti[[2]]+
    kvartili_trombociti[[3]]+kvartili_trombociti[[4]]+
    kvartili_trombociti[[5]])/5

mean(with(dopunjeni_podaci,tapply(albumin,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(albumin,.imp,sd)))
kvartili_albumin<-with(dopunjeni_podaci,tapply(albumin,.imp,
    quantile))
(kvartili_albumin[[1]]+kvartili_albumin[[2]]+
    kvartili_albumin[[3]]+kvartili_albumin[[4]]+
    kvartili_albumin[[5]])/5

mean(with(dopunjeni_podaci,tapply(ukupni_bilirubin,.imp,
    mean)))
mean(with(dopunjeni_podaci,tapply(ukupni_bilirubin,.imp,
    sd)))
kvartili_ukupni_bilirubin<-with(dopunjeni_podaci,tapply(
    ukupni_bilirubin,.imp,quantile))
(kvartili_ukupni_bilirubin[[1]]+
    kvartili_ukupni_bilirubin[[2]]+
    kvartili_ukupni_bilirubin[[3]]+
    kvartili_ukupni_bilirubin[[4]]+
    kvartili_ukupni_bilirubin[[5]])/5
```

```
mean(with(dopunjeni_podaci,tapply(ALT,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(ALT,.imp,sd)))
kvartili_ALT<-with(dopunjeni_podaci,tapply(ALT,.imp,
    quantile))
(kvartili_ALT[[1]]+kvartili_ALT[[2]]+kvartili_ALT[[3]]+
    kvartili_ALT[[4]]+kvartili_ALT[[5]])/5

mean(with(dopunjeni_podaci,tapply(AST,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(AST,.imp,sd)))
kvartili_AST<-with(dopunjeni_podaci,tapply(AST,.imp,
    quantile))
(kvartili_AST[[1]]+kvartili_AST[[2]]+kvartili_AST[[3]]+
    kvartili_AST[[4]]+kvartili_AST[[5]])/5

mean(with(dopunjeni_podaci,tapply(GGT,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(GGT,.imp,sd)))
kvartili_GGT<-with(dopunjeni_podaci,tapply(GGT,.imp,
    quantile))
(kvartili_GGT[[1]]+kvartili_GGT[[2]]+kvartili_GGT[[3]]+
    kvartili_GGT[[4]]+kvartili_GGT[[5]])/5

mean(with(dopunjeni_podaci,tapply(ALP,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(ALP,.imp,sd)))
kvartili_ALP<-with(dopunjeni_podaci,tapply(ALP,.imp,
    quantile))
(kvartili_ALP[[1]]+kvartili_ALP[[2]]+kvartili_ALP[[3]]+
    kvartili_ALP[[4]]+kvartili_ALP[[5]])/5

mean(with(dopunjeni_podaci,tapply(TP,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(TP,.imp,sd)))
kvartili_TP<-with(dopunjeni_podaci,tapply(TP,.imp,quantile))
(kvartili_TP[[1]]+kvartili_TP[[2]]+kvartili_TP[[3]]+
    kvartili_TP[[4]]+kvartili_TP[[5]])/5

mean(with(dopunjeni_podaci,tapply(kreatinin,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(kreatinin,.imp,sd)))
kvartili_kreatinin<-with(dopunjeni_podaci,tapply(kreatinin,
    .imp,quantile))
(kvartili_kreatinin[[1]]+kvartili_kreatinin[[2]]+
```

```
      kvartili_kreatinin[[3]]+kvartili_kreatinin[[4]]+
      kvartili_kreatinin[[5]])/5

mean(with(dopunjeni_podaci,tapply(dimenzija,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(dimenzija,.imp,sd)))
kvartili_dimenzija<-with(dopunjeni_podaci,tapply(dimenzija,
      .imp,quantile))
(kvartili_dimenzija[[1]]+kvartili_dimenzija[[2]]+
      kvartili_dimenzija[[3]]+kvartili_dimenzija[[4]]+
      kvartili_dimenzija[[5]])/5

mean(with(dopunjeni_podaci,tapply(direktni_bilirubin,.imp,
     mean)))
mean(with(dopunjeni_podaci,tapply(direktni_bilirubin,.imp,
     sd)))
kvartili_direktni_bilirubin<-with(dopunjeni_podaci,tapply(
      direktni_bilirubin,.imp,quantile))
(kvartili_direktni_bilirubin[[1]]+
      kvartili_direktni_bilirubin[[2]]+
      kvartili_direktni_bilirubin[[3]]+
      kvartili_direktni_bilirubin[[4]]+
      kvartili_direktni_bilirubin[[5]])/5

mean(with(dopunjeni_podaci,tapply(zeljezo,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(zeljezo,.imp,sd)))
kvartili_zeljezo<-with(dopunjeni_podaci,tapply(zeljezo,.imp,
      quantile))
(kvartili_zeljezo[[1]]+kvartili_zeljezo[[2]]+
      kvartili_zeljezo[[3]]+kvartili_zeljezo[[4]]+
      kvartili_zeljezo[[5]])/5

mean(with(dopunjeni_podaci,tapply(zasicenost,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(zasicenost,.imp,sd)))
kvartili_zasicenost<-with(dopunjeni_podaci,tapply(
      zasicenost,.imp,quantile))
(kvartili_zasicenost[[1]]+kvartili_zasicenost[[2]]+
      kvartili_zasicenost[[3]]+kvartili_zasicenost[[4]]+
      kvartili_zasicenost[[5]])/5
```

```
mean(with(dopunjeni_podaci,tapply(feritin,.imp,mean)))
mean(with(dopunjeni_podaci,tapply(feritin,.imp,sd)))
kvartili_feritin<-with(dopunjeni_podaci,tapply(feritin,.imp,
    quantile))
(kvartili_feritin[[1]]+kvartili_feritin[[2]]+
    kvartili_feritin[[3]]+kvartili_feritin[[4]]+
    kvartili_feritin[[5]])/5
```

Deskriptivna analiza kvalitativnih varijabli, zavisne varijable i varijable *broj čvorova*:

```
frekvencije_broj_cvorova<-with(dopunjeni_podaci,tapply(
    broj_cvorova,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_broj_cvorova<-round((
    frekvencije_broj_cvorova[[1]][2]+
    frekvencije_broj_cvorova[[2]][2]+
    frekvencije_broj_cvorova[[3]][2]+
    frekvencije_broj_cvorova[[4]][2]+
    frekvencije_broj_cvorova[[5]][2])/5,0)
round((ukupna_frekvencija_broj_cvorova/165)*100,2)

data.frame(table(podaci$spol))
data.frame(prop.table(table(podaci$spol))*100)

frekvencije_simptomi<-with(dopunjeni_podaci,tapply(simptomi,
    .imp,function(x) data.frame(table(x))))
ukupna_frekvencija_simptomi<-round((
    frekvencije_simptomi[[1]][2]+
    frekvencije_simptomi[[2]][2]+
    frekvencije_simptomi[[3]][2]+
    frekvencije_simptomi[[4]][2]+
    frekvencije_simptomi[[5]][2])/5,0)
round((ukupna_frekvencija_simptomi/165)*100,2)

data.frame(table(podaci$alkohol))
data.frame(prop.table(table(podaci$alkohol))*100)

frekvencije_HBsAg<-with(dopunjeni_podaci,tapply(HBsAg,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_HBsAg<-round((frekvencije_HBsAg[[1]][2]+
    frekvencije_HBsAg[[2]][2]+frekvencije_HBsAg[[3]][2]+
```

```
    frekvencije_HBsAg[[4]][2]+frekvencije_HBsAg[[5]][2])
    /5,0)
round((ukupna_frekvencija_HBsAg/165)*100,2)

frekvencije_HBeAg<-with(dopunjeni_podaci,tapply(HBeAg,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_HBeAg<-round((frekvencije_HBeAg[[1]][2]+
    frekvencije_HBeAg[[2]][2]+frekvencije_HBeAg[[3]][2]+
    frekvencije_HBeAg[[4]][2]+frekvencije_HBeAg[[5]][2])
    /5,0)
round((ukupna_frekvencija_HBeAg/165)*100,2)

frekvencije_HBcAb<-with(dopunjeni_podaci,tapply(HBcAb,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_HBcAb<-round((frekvencije_HBcAb[[1]][2]+
    frekvencije_HBcAb[[2]][2]+frekvencije_HBcAb[[3]][2]+
    frekvencije_HBcAb[[4]][2]+frekvencije_HBcAb[[5]][2])
    /5,0)
round((ukupna_frekvencija_HBcAb/165)*100,2)

frekvencije_HCVAb<-with(dopunjeni_podaci,tapply(HCVAb,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_HCVAb<-round((frekvencije_HCVAb[[1]][2]+
    frekvencije_HCVAb[[2]][2]+frekvencije_HCVAb[[3]][2]+
    frekvencije_HCVAb[[4]][2]+frekvencije_HCVAb[[5]][2])
    /5,0)
round((ukupna_frekvencija_HCVAb/165)*100,2)

data.frame(table(podaci$ciroza))
data.frame(prop.table(table(podaci$ciroza))*100)

frekvencije_drzava<-with(dopunjeni_podaci,tapply(drzava,
    .imp,function(x) data.frame(table(x))))
ukupna_frekvencija_drzava<-round((frekvencije_drzava[[1]][2]
    +frekvencije_drzava[[2]][2]+frekvencije_drzava[[3]][2]+
    frekvencije_drzava[[4]][2]+frekvencije_drzava[[5]][2])
    /5,0)
round((ukupna_frekvencija_drzava/165)*100,2)
```

```
frekvencije_pusenje<-with(dopunjeni_podaci,tapply(pusenje,
    .imp,function(x) data.frame(table(x))))
ukupna_frekvencija_pusenje<-round((
    frekvencije_pusenje[[1]][2]+frekvencije_pusenje[[2]][2]+
    frekvencije_pusenje[[3]][2]+frekvencije_pusenje[[4]][2]+
    frekvencije_pusenje[[5]][2])/5,0)
round((ukupna_frekvencija_pusenje/165)*100,2)

frekvencije_dijabetes<-with(dopunjeni_podaci,tapply(
    dijabetes,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_dijabetes<-round((
    frekvencije_dijabetes[[1]][2]+
    frekvencije_dijabetes[[2]][2]+
    frekvencije_dijabetes[[3]][2]+
    frekvencije_dijabetes[[4]][2]+
    frekvencije_dijabetes[[5]][2])/5,0)
round((ukupna_frekvencija_dijabetes/165)*100,2)

frekvencije_pretilost<-with(dopunjeni_podaci,tapply(
    pretilost,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_pretilost<-round((
    frekvencije_pretilost[[1]][2]+
    frekvencije_pretilost[[2]][2]+
    frekvencije_pretilost[[3]][2]+
    frekvencije_pretilost[[4]][2]+
    frekvencije_pretilost[[5]][2])/5,0)
round((ukupna_frekvencija_pretilost/165)*100,2)

frekvencije_hemokromatoza<-with(dopunjeni_podaci,tapply(
    hemokromatoza,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_hemokromatoza<-round((
    frekvencije_hemokromatoza[[1]][2]+
    frekvencije_hemokromatoza[[2]][2]+
    frekvencije_hemokromatoza[[3]][2]+
    frekvencije_hemokromatoza[[4]][2]+
    frekvencije_hemokromatoza[[5]][2])/5,0)
round((ukupna_frekvencija_hemokromatoza/165)*100,2)

frekvencije_AHT<-with(dopunjeni_podaci,tapply(AHT,.imp,
```

```
    function(x) data.frame(table(x))))
ukupna_frekvencija_AHT<-round((frekvencije_AHT[[1]][2]+
    frekvencije_AHT[[2]][2]+frekvencije_AHT[[3]][2]+
    frekvencije_AHT[[4]][2]+frekvencije_AHT[[5]][2])/5,0)
round((ukupna_frekvencija_AHT/165)*100,2)

frekvencije_CRI<-with(dopunjeni_podaci,tapply(CRI,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_CRI<-round((frekvencije_CRI[[1]][2]+
    frekvencije_CRI[[2]][2]+frekvencije_CRI[[3]][2]+
    frekvencije_CRI[[4]][2]+frekvencije_CRI[[5]][2])/5,0)
round((ukupna_frekvencija_CRI/165)*100,2)

frekvencije_HIV<-with(dopunjeni_podaci,tapply(HIV,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_HIV<-round((frekvencije_HIV[[1]][2]+
    frekvencije_HIV[[2]][2]+frekvencije_HIV[[3]][2]+
    frekvencije_HIV[[4]][2]+frekvencije_HIV[[5]][2])/5,0)
round((ukupna_frekvencija_HIV/165)*100,2)

frekvencije_NASH<-with(dopunjeni_podaci,tapply(NASH,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_NASH<-round((frekvencije_NASH[[1]][2]+
   frekvencije_NASH[[2]][2]+frekvencije_NASH[[3]][2]+
   frekvencije_NASH[[4]][2]+frekvencije_NASH[[5]][2])/5,0)
round((ukupna_frekvencija_NASH/165)*100,2)

frekvencije_variksi<-with(dopunjeni_podaci,tapply(variksi,
    .imp,function(x) data.frame(table(x))))
ukupna_frekvencija_variksi<-round((
    frekvencije_variksi[[1]][2]+frekvencije_variksi[[2]][2]+
    frekvencije_variksi[[3]][2]+frekvencije_variksi[[4]][2]+
    frekvencije_variksi[[5]][2])/5,0)
round((ukupna_frekvencija_variksi/165)*100,2)

frekvencije_splenomegalija<-with(dopunjeni_podaci,tapply(
    splenomegalija,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_splenomegalija<-round((
    frekvencije_splenomegalija[[1]][2]+
```

```
    frekvencije_splenomegalija[[2]][2]+
    frekvencije_splenomegalija[[3]][2]+
    frekvencije_splenomegalija[[4]][2]+
    frekvencije_splenomegalija[[5]][2])/5,0)
round((ukupna_frekvencija_splenomegalija/165)*100,2)

frekvencije_PHT<-with(dopunjeni_podaci,tapply(PHT,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_PHT<-round((frekvencije_PHT[[1]][2]+
    frekvencije_PHT[[2]][2]+frekvencije_PHT[[3]][2]+
    frekvencije_PHT[[4]][2]+frekvencije_PHT[[5]][2])/5,0)
round((ukupna_frekvencija_PHT/165)*100,2)

frekvencije_PVT<-with(dopunjeni_podaci,tapply(PVT,.imp,
    function(x) data.frame(table(x))))
ukupna_frekvencija_PVT<-round((frekvencije_PVT[[1]][2]+
    frekvencije_PVT[[2]][2]+frekvencije_PVT[[3]][2]+
    frekvencije_PVT[[4]][2]+frekvencije_PVT[[5]][2])/5,0)
round((ukupna_frekvencija_PVT/165)*100,2)

frekvencije_metastaze<-with(dopunjeni_podaci,tapply(
    metastaze,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_metastaze<-round((
    frekvencije_metastaze[[1]][2]+
    frekvencije_metastaze[[2]][2]+
    frekvencije_metastaze[[3]][2]+
    frekvencije_metastaze[[4]][2]+
    frekvencije_metastaze[[5]][2])/5,0)
round((ukupna_frekvencija_metastaze/165)*100,2)

frekvencije_obiljezje<-with(dopunjeni_podaci,tapply(
    obiljezje,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_obiljezje<-round((
    frekvencije_obiljezje[[1]][2]+
    frekvencije_obiljezje[[2]][2]+
    frekvencije_obiljezje[[3]][2]+
    frekvencije_obiljezje[[4]][2]+
    frekvencije_obiljezje[[5]][2])/5,0)
round((ukupna_frekvencija_obiljezje/165)*100,2)
```

```
data.frame(table(podaci$PS))
data.frame(prop.table(table(podaci$PS))*100)

frekvencije_encefalopatija<-with(dopunjeni_podaci,tapply(
    encefalopatija,.imp,function(x) data.frame(table(x))))
ukupna_frekvencija_encefalopatija<-round((
    frekvencije_encefalopatija[[1]][2]+
    frekvencije_encefalopatija[[2]][2]+
    frekvencije_encefalopatija[[3]][2]+
    frekvencije_encefalopatija[[4]][2]+
    frekvencije_encefalopatija[[5]][2])/5,0)
round((ukupna_frekvencija_encefalopatija/165)*100,2)

frekvencije_ascites<-with(dopunjeni_podaci,tapply(ascites,
    .imp,function(x) data.frame(table(x))))
ukupna_frekvencija_ascites<-round((
    frekvencije_ascites[[1]][2]+frekvencije_ascites[[2]][2]+
    frekvencije_ascites[[3]][2]+frekvencije_ascites[[4]][2]+
    frekvencije_ascites[[5]][2])/5,0)
round((ukupna_frekvencija_ascites/165)*100,2)

Desc(podaci$prezivljavanje)
```

Modeliranje *stupnja stanja pacijenta* i *stupnja encefalopatije* kao neprekidnih varijabli:

```
podaci$PS<-as.numeric(levels(podaci$PS))[podaci$PS]
podaci$encefalopatija<-as.numeric(levels(podaci$
    encefalopatija))[podaci$encefalopatija]

dopunjeni$data$PS<-as.numeric(levels(dopunjeni$data$PS))
    [dopunjeni$data$PS]
dopunjeni_podaci$PS<-as.numeric(levels(dopunjeni_podaci$PS))
    [dopunjeni_podaci$PS]

dopunjeni$data$encefalopatija<-as.numeric(levels(dopunjeni$
    data$encefalopatija))[dopunjeni$data$encefalopatija]
dopunjeni[["imp"]][["encefalopatija"]][["1"]]<-as.numeric(
    levels(dopunjeni[["imp"]][["encefalopatija"]][["1"]]))
    [dopunjeni[["imp"]][["encefalopatija"]][["1"]]]
dopunjeni[["imp"]][["encefalopatija"]][["2"]]<-as.numeric(
```

```
    levels(dopunjeni[["imp"]][["encefalopatija"]][["2"]]))
    [dopunjeni[["imp"]][["encefalopatija"]][["2"]]]
dopunjeni[["imp"]][["encefalopatija"]][["3"]]<-as.numeric(
    levels(dopunjeni[["imp"]][["encefalopatija"]][["3"]]))
    [dopunjeni[["imp"]][["encefalopatija"]][["3"]]]
dopunjeni[["imp"]][["encefalopatija"]][["4"]]<-as.numeric(
    levels(dopunjeni[["imp"]][["encefalopatija"]][["4"]]))
    [dopunjeni[["imp"]][["encefalopatija"]][["4"]]]
dopunjeni[["imp"]][["encefalopatija"]][["5"]]<-as.numeric(
    levels(dopunjeni[["imp"]][["encefalopatija"]][["5"]]))
    [dopunjeni[["imp"]][["encefalopatija"]][["5"]]]
dopunjeni_podaci$encefalopatija<-as.numeric(levels(
    dopunjeni_podaci$encefalopatija))[dopunjeni_podaci$
    encefalopatija]
```

Univarijabilna logistička regresija, odgovarajući testovi omjera vjerodostojnosti za značajnost koeficijenata, a zatim i procjene omjera izgleda te procjene 95% pouzdanog intervala za omjer izgleda za varijable koje su ispale značajne:

```
modeli_dob<-with(dopunjeni,exp=glm(prezivljavanje~dob,data=
    podaci,family="binomial"))
model_dob<-pool(modeli_dob)
summary(model_dob,conf.int=TRUE)
summary(model_dob,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_dob)

modeli_alkohol_grami<-with(dopunjeni,exp=glm(prezivljavanje~
    alkohol_grami,data=podaci,family="binomial"))
model_alkohol_grami<-pool(modeli_alkohol_grami)
summary(model_alkohol_grami,conf.int=TRUE)
D3(modeli_alkohol_grami)

modeli_cigarete<-with(dopunjeni,exp=glm(prezivljavanje~
    cigarete,data=podaci,family="binomial"))
model_cigarete<-pool(modeli_cigarete)
summary(model_cigarete,conf.int=TRUE)
D3(modeli_cigarete)

modeli_INR<-with(dopunjeni,exp=glm(prezivljavanje~INR,data=
    podaci,family="binomial"))
```

```
model_INR<-pool(modeli_INR)
summary(model_INR, conf.int=TRUE)
summary(model_INR, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_INR)

modeli_AFP<-with(dopunjeni, exp=glm(prezivljavanje~AFP, data=
    podaci, family="binomial"))
model_AFP<-pool(modeli_AFP)
summary(model_AFP, conf.int=TRUE)
D3(modeli_AFP)

modeli_hemoglobin<-with(dopunjeni, exp=glm(prezivljavanje~
    hemoglobin, data=podaci, family="binomial"))
model_hemoglobin<-pool(modeli_hemoglobin)
summary(model_hemoglobin, conf.int=TRUE)
summary(model_hemoglobin, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_hemoglobin)

modeli_MCV<-with(dopunjeni, exp=glm(prezivljavanje~MCV, data=
    podaci, family="binomial"))
model_MCV<-pool(modeli_MCV)
summary(model_MCV, conf.int=TRUE)
D3(modeli_MCV)

modeli_leukociti<-with(dopunjeni, exp=glm(prezivljavanje~
    leukociti, data=podaci, family="binomial"))
model_leukociti<-pool(modeli_leukociti)
summary(model_leukociti, conf.int=TRUE)
summary(model_leukociti, conf.int=TRUE, exponentiate=TRUE)
exp(model_leukociti$pooled$estimate[2]*1000)
D3(modeli_leukociti)

modeli_trombociti<-with(dopunjeni, exp=glm(prezivljavanje~
    trombociti, data=podaci, family="binomial"))
model_trombociti<-pool(modeli_trombociti)
summary(model_trombociti, conf.int=TRUE)
summary(model_trombociti, conf.int=TRUE, exponentiate=TRUE)
exp(model_trombociti$pooled$estimate[2]*1000)
D3(modeli_trombociti)
```

```
modeli_albumin<-with(dopunjeni,exp=glm(preživljavanje~
    albumin,data=podaci,family="binomial"))
model_albumin<-pool(modeli_albumin)
summary(model_albumin,conf.int=TRUE)
summary(model_albumin,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_albumin)

modeli_ukupni_bilirubin<-with(dopunjeni,exp=glm(
    preživljavanje~ukupni_bilirubin,data=podaci,family=
    "binomial"))
model_ukupni_bilirubin<-pool(modeli_ukupni_bilirubin)
summary(model_ukupni_bilirubin,conf.int=TRUE)
summary(model_ukupni_bilirubin,conf.int=TRUE,exponentiate=
    TRUE)
D3(modeli_ukupni_bilirubin)

modeli_ALT<-with(dopunjeni,exp=glm(preživljavanje~ALT,data=
    podaci,family="binomial"))
model_ALT<-pool(modeli_ALT)
summary(model_ALT,conf.int=TRUE)
D3(modeli_ALT)

modeli_AST<-with(dopunjeni,exp=glm(preživljavanje~AST,data=
    podaci,family="binomial"))
model_AST<-pool(modeli_AST)
summary(model_AST,conf.int=TRUE)
summary(model_AST,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_AST)

modeli_GGT<-with(dopunjeni,exp=glm(preživljavanje~GGT,data=
    podaci,family="binomial"))
model_GGT<-pool(modeli_GGT)
summary(model_GGT,conf.int=TRUE)
summary(model_GGT,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_GGT)

modeli_ALP<-with(dopunjeni,exp=glm(preživljavanje~ALP,data=
    podaci,family="binomial"))
model_ALP<-pool(modeli_ALP)
```

```
summary(model_ALP, conf.int=TRUE)
summary(model_ALP, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_ALP)

modeli_TP<-with(dopunjeni, exp=glm(prezivljavanje~TP, data=
    podaci, family="binomial"))
model_TP<-pool(modeli_TP)
summary(model_TP, conf.int=TRUE)
D3(modeli_TP)

modeli_kreatinin<-with(dopunjeni, exp=glm(prezivljavanje~
    kreatinin, data=podaci, family="binomial"))
model_kreatinin<-pool(modeli_kreatinin)
summary(model_kreatinin, conf.int=TRUE)
summary(model_kreatinin, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_kreatinin)

modeli_broj_cvorova<-with(dopunjeni, exp=glm(prezivljavanje~
    broj_cvorova, data=podaci, family="binomial"))
model_broj_cvorova<-pool(modeli_broj_cvorova)
summary(model_broj_cvorova, conf.int=TRUE)
D3(modeli_broj_cvorova)

modeli_dimenzija<-with(dopunjeni, exp=glm(prezivljavanje~
    dimenzija, data=podaci, family="binomial"))
model_dimenzija<-pool(modeli_dimenzija)
summary(model_dimenzija, conf.int=TRUE)
summary(model_dimenzija, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_dimenzija)

modeli_direktni_bilirubin<-with(dopunjeni, exp=glm(
    prezivljavanje~direktni_bilirubin, data=podaci, family=
    "binomial"))
model_direktni_bilirubin<-pool(modeli_direktni_bilirubin)
summary(model_direktni_bilirubin, conf.int=TRUE)
summary(model_direktni_bilirubin, conf.int=TRUE, exponentiate=
    TRUE)
D3(modeli_direktni_bilirubin)
```

```
modeli_zeljezo<-with(dopunjeni,exp=glm(prezivljavanje~
    zeljezo,data=podaci,family="binomial"))
model_zeljezo<-pool(modeli_zeljezo)
summary(model_zeljezo,conf.int=TRUE)
summary(model_zeljezo,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_zeljezo)

modeli_zasicenost<-with(dopunjeni,exp=glm(prezivljavanje~
    zasicenost,data=podaci,family="binomial"))
model_zasicenost<-pool(modeli_zasicenost)
summary(model_zasicenost,conf.int=TRUE)
summary(model_zasicenost,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_zasicenost)

modeli_feritin<-with(dopunjeni,exp=glm(prezivljavanje~
    feritin,data=podaci,family="binomial"))
model_feritin<-pool(modeli_feritin)
summary(model_feritin,conf.int=TRUE)
summary(model_feritin,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_feritin)

modeli_spol<-with(dopunjeni,exp=glm(prezivljavanje~spol,
    data=podaci,family="binomial"))
model_spol<-pool(modeli_spol)
summary(model_spol,conf.int=TRUE)
D3(modeli_spol)

modeli_simptomi<-with(dopunjeni,exp=glm(prezivljavanje~
    simptomi,data=podaci,family="binomial"))
model_simptomi<-pool(modeli_simptomi)
summary(model_simptomi,conf.int=TRUE)
summary(model_simptomi,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_simptomi)

modeli_alkohol<-with(dopunjeni,exp=glm(prezivljavanje~
    alkohol,data=podaci,family="binomial"))
model_alkohol<-pool(modeli_alkohol)
summary(model_alkohol,conf.int=TRUE)
D3(modeli_alkohol)
```

```
modeli_HBsAg<-with(dopunjeni,exp=glm(prezivljavanje~HBsAg,
    data=podaci,family="binomial"))
model_HBsAg<-pool(modeli_HBsAg)
summary(model_HBsAg,conf.int=TRUE)
D3(modeli_HBsAg)

modeli_HBeAg<-with(dopunjeni,exp=glm(prezivljavanje~HBeAg,
    data=podaci,family="binomial"))
model_HBeAg<-pool(modeli_HBeAg)
summary(model_HBeAg,conf.int=TRUE)
D3(modeli_HBeAg)

modeli_HBcAb<-with(dopunjeni,exp=glm(prezivljavanje~HBcAb,
    data=podaci,family="binomial"))
model_HBcAb<-pool(modeli_HBcAb)
summary(model_HBcAb,conf.int=TRUE)
D3(modeli_HBcAb)

modeli_HCVAb<-with(dopunjeni,exp=glm(prezivljavanje~HCVAb,
    data=podaci,family="binomial"))
model_HCVAb<-pool(modeli_HCVAb)
summary(model_HCVAb,conf.int=TRUE)
summary(model_HCVAb,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_HCVAb)

modeli_ciroza<-with(dopunjeni,exp=glm(prezivljavanje~ciroza,
    data=podaci,family="binomial"))
model_ciroza<-pool(modeli_ciroza)
summary(model_ciroza,conf.int=TRUE)
D3(modeli_ciroza)

modeli_drzava<-with(dopunjeni,exp=glm(prezivljavanje~drzava,
    data=podaci,family="binomial"))
model_drzava<-pool(modeli_drzava)
summary(model_drzava,conf.int=TRUE)
summary(model_drzava,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_drzava)

modeli_pusenje<-with(dopunjeni,exp=glm(prezivljavanje~
```

```
    pusenje, data=podaci, family="binomial"))
model_pusenje<-pool(modeli_pusenje)
summary(model_pusenje, conf.int=TRUE)
D3(modeli_pusenje)

modeli_dijabetes<-with(dopunjeni, exp=glm(prezivljavanje~
    dijabetes, data=podaci, family="binomial"))
model_dijabetes<-pool(modeli_dijabetes)
summary(model_dijabetes, conf.int=TRUE)
summary(model_dijabetes, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_dijabetes)

modeli_pretilost<-with(dopunjeni, exp=glm(prezivljavanje~
    pretilost, data=podaci, family="binomial"))
model_pretilost<-pool(modeli_pretilost)
summary(model_pretilost, conf.int=TRUE)
D3(modeli_pretilost)

modeli_hemokromatoza<-with(dopunjeni, exp=glm(prezivljavanje~
    hemokromatoza, data=podaci, family="binomial"))
model_hemokromatoza<-pool(modeli_hemokromatoza)
summary(model_hemokromatoza, conf.int=TRUE)
D3(modeli_hemokromatoza)

modeli_AHT<-with(dopunjeni, exp=glm(prezivljavanje~AHT, data=
    podaci, family="binomial"))
model_AHT<-pool(modeli_AHT)
summary(model_AHT, conf.int=TRUE)
D3(modeli_AHT)

modeli_CRI<-with(dopunjeni, exp=glm(prezivljavanje~CRI, data=
    podaci, family="binomial"))
model_CRI<-pool(modeli_CRI)
summary(model_CRI, conf.int=TRUE)
D3(modeli_CRI)

modeli_HIV<-with(dopunjeni, exp=glm(prezivljavanje~HIV, data=
    podaci, family="binomial"))
model_HIV<-pool(modeli_HIV)
```

```
summary(model_HIV, conf.int=TRUE)
D3(modeli_HIV)

modeli_NASH<-with(dopunjeni, exp=glm(prezivljavanje~NASH,
    data=podaci, family="binomial"))
model_NASH<-pool(modeli_NASH)
summary(model_NASH, conf.int=TRUE)
D3(modeli_NASH)

modeli_variksi<-with(dopunjeni, exp=glm(prezivljavanje~
    variksi, data=podaci, family="binomial"))
model_variksi<-pool(modeli_variksi)
summary(model_variksi, conf.int=TRUE)
D3(modeli_variksi)

modeli_splenomegalija<-with(dopunjeni, exp=glm(prezivljavanje
    ~splenomegalija, data=podaci, family="binomial"))
model_splenomegalija<-pool(modeli_splenomegalija)
summary(model_splenomegalija, conf.int=TRUE)
D3(modeli_splenomegalija)

modeli_PHT<-with(dopunjeni, exp=glm(prezivljavanje~PHT, data=
    podaci, family="binomial"))
model_PHT<-pool(modeli_PHT)
summary(model_PHT, conf.int=TRUE)
D3(modeli_PHT)

modeli_PVT<-with(dopunjeni, exp=glm(prezivljavanje~PVT, data=
    podaci, family="binomial"))
model_PVT<-pool(modeli_PVT)
summary(model_PVT, conf.int=TRUE)
summary(model_PVT, conf.int=TRUE, exponentiate=TRUE)
D3(modeli_PVT)

modeli_metastaze<-with(dopunjeni, exp=glm(prezivljavanje~
    metastaze, data=podaci, family="binomial"))
model_metastaze<-pool(modeli_metastaze)
summary(model_metastaze, conf.int=TRUE)
summary(model_metastaze, conf.int=TRUE, exponentiate=TRUE)
```

```
D3(modeli_metastaze)

modeli_obiljezje<-with(dopunjeni,exp=glm(prezivljavanje~
    obiljezje,data=podaci,family="binomial"))
model_obiljezje<-pool(modeli_obiljezje)
summary(model_obiljezje,conf.int=TRUE)
D3(modeli_obiljezje)

modeli_PS<-with(dopunjeni,exp=glm(prezivljavanje~PS,data=
    podaci,family="binomial"))
model_PS<-pool(modeli_PS)
summary(model_PS,conf.int=TRUE)
summary(model_PS,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_PS)

modeli_encefalopatija<-with(dopunjeni,exp=glm(prezivljavanje
    ~encefalopatija,data=podaci,family="binomial"))
model_encefalopatija<-pool(modeli_encefalopatija)
summary(model_encefalopatija,conf.int=TRUE)
summary(model_encefalopatija,conf.int=TRUE,exponentiate=
    TRUE)
D3(modeli_encefalopatija)

modeli_ascites<-with(dopunjeni,exp=glm(prezivljavanje~
    ascites,data=podaci,family="binomial"))
model_ascites<-pool(modeli_ascites)
summary(model_ascites,conf.int=TRUE)
summary(model_ascites,conf.int=TRUE,exponentiate=TRUE)
D3(modeli_ascites)
```

Stepwise analiza:

```
svi_i_bez<-list(upper=~+dob+INR+hemoglobin+leukociti+
    trombociti+albumin+ukupni_bilirubin+AST+GGT+ALP+
    kreatinin+dimenzija+direktni_bilirubin+zeljezo+
    feritin+simptomi+HCVAb+drzava+dijabetes+PVT+metastaze+
    PS+encefalopatija+ascites,lower=~1)
izraz<-expression(model_bez<-glm(prezivljavanje~1,family=
    "binomial"),model<-step(model_bez,scope=svi_i_bez,
    direction="both",test="LRT"))
```

```
modeli_stepwise<-with(dopunjeni,izraz)
modeli_stepwise$analyses[[1]]$anova
modeli_stepwise$analyses[[2]]$anova
modeli_stepwise$analyses[[3]]$anova
modeli_stepwise$analyses[[4]]$anova
modeli_stepwise$analyses[[5]]$anova
```

Koliko puta se koja varijabla pojavila u završnom koraku *stepwise* metode:

```
formule<-lapply(modeli_stepwise$analyses,formula)
varijable<-lapply(formule,terms)
koliko<-unlist(lapply(varijable,labels))
table(koliko)
```

Multivarijabilna logistička regresija s varijablama koje su dobivene u završnom koraku *stepwise* metode barem 3 puta:

```
modeli_odabrane_stepwise<-with(dopunjeni,glm(prezivljavanje~
    dob+INR+hemoglobin+GGT+ALP+dimenzija+feritin+simptomi+
    HCVAb+PVT+PS,family="binomial"))
model_odabrane_stepwise<-pool(modeli_odabrane_stepwise)
summary(model_odabrane_stepwise,conf.int=TRUE)
```

Test omjera vjerodostojnosti za testiranje je li dovoljan model bez varijabli *INR*, *GGT*, *najveća dimenzija čvora* i *feritin* ili je potreban model s tim varijablama:

```
modeli_odabrane_znacajnost<-with(dopunjeni,glm(
    prezivljavanje~dob+hemoglobin+ALP+simptomi+HCVAb+PVT+PS,
    family="binomial"))
model_odabrane_znacajnost<-pool(modeli_odabrane_znacajnost)
D3(modeli_odabrane_stepwise,modeli_odabrane_znacajnost)
summary(model_odabrane_znacajnost,conf.int=TRUE)
```

Test omjera vjerodostojnosti za testiranje je li dovoljan model i bez varijabli *simptomi*, *tromboza portalne vene* te *stupanj stanja pacijenta* ili je potreban model s njima:

```
modeli_odabrane_znacajnost2<-with(dopunjeni,glm(
    prezivljavanje~dob+hemoglobin+ALP+HCVAb,family=
    "binomial"))
model_odabrane_znacajnost2<-pool(
    modeli_odabrane_znacajnost2)
D3(modeli_odabrane_stepwise,modeli_odabrane_znacajnost2)
```

Test omjera vjerodostojnosti za testiranje je li varijabla *tromboza portalne vene* potrebna u modelu:

```
modeli_odabrane_znacajnost3<-with(dopunjeni,glm(
    preživljavanje~dob+hemoglobin+ALP+simptomi+HCVAb+PS,
    family="binomial"))
model_odabrane_znacajnost3<-pool(
    modeli_odabrane_znacajnost3)
D3(modeli_odabrane_znacajnost,modeli_odabrane_znacajnost3)
summary(model_odabrane_znacajnost3,conf.int=TRUE)
```

Procjene koeficijenata u punom multivarijabilnom modelu:

```
modeli_puni<-with(dopunjeni,glm(preživljavanje~.,data=
    dopunjeni_podaci,family="binomial"))
model_puni<-pool(modeli_puni)
summary(model_puni)
```

Provjera linearnosti logit funkcije za neprekidne nezavisne varijable:

```
modeli_linearnost<-with(dopunjeni,glm(preživljavanje~dob+
    dob*log(dob)+hemoglobin+hemoglobin*log(hemoglobin)+ALP+
    ALP*log(ALP),family="binomial"))
model_linearnost<-pool(modeli_linearnost)
summary(model_linearnost)
```

Za svaku interakciju posebno test je li statistički značajno bolji model s njom nego bez nje:

```
modeli_interakcije1<-with(dopunjeni,glm(preživljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+dob*hemoglobin,family=
    "binomial"))
model_interakcije1<-pool(modeli_interakcije1)
D3(modeli_interakcije1,modeli_odabrane_znacajnost3)

modeli_interakcije2<-with(dopunjeni,glm(preživljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+dob*ALP,family=
    "binomial"))
model_interakcije2<-pool(modeli_interakcije2)
D3(modeli_interakcije2,modeli_odabrane_znacajnost3)

modeli_interakcije3<-with(dopunjeni,glm(preživljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+dob*simptomi,family=
    "binomial"))
```

```
model_interakcije3<-pool(modeli_interakcije3)
D3(modeli_interakcije3 ,modeli_odabrane_znacajnost3)

modeli_interakcije4<-with(dopunjeni ,glm(prezivljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+dob*HCVAb, family=
    "binomial"))
model_interakcije4<-pool(modeli_interakcije4)
D3(modeli_interakcije4 ,modeli_odabrane_znacajnost3)

modeli_interakcije5<-with(dopunjeni ,glm(prezivljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+dob*PS, family=
    "binomial"))
model_interakcije5<-pool(modeli_interakcije5)
D3(modeli_interakcije5 ,modeli_odabrane_znacajnost3)

modeli_interakcije6<-with(dopunjeni ,glm(prezivljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+hemoglobin*ALP, family=
    "binomial"))
model_interakcije6<-pool(modeli_interakcije6)
D3(modeli_interakcije6 ,modeli_odabrane_znacajnost3)

modeli_interakcije7<-with(dopunjeni ,glm(prezivljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+simptomi*HCVAb, family=
    "binomial"))
model_interakcije7<-pool(modeli_interakcije7)
D3(modeli_interakcije7 ,modeli_odabrane_znacajnost3)

modeli_interakcije8<-with(dopunjeni ,glm(prezivljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+simptomi*PS, family=
    "binomial"))
model_interakcije8<-pool(modeli_interakcije8)
D3(modeli_interakcije8 ,modeli_odabrane_znacajnost3)
```

Konačni model:

```
modeli_konacni<-with(dopunjeni ,glm(prezivljavanje~dob+
    hemoglobin+ALP+simptomi+HCVAb+PS+hemoglobin*ALP,
    family="binomial"))
model_konacni<-pool(modeli_konacni)
summary(model_konacni ,conf.int=TRUE)
```

Procjena adekvatnosti modela:

```
procijenjene<-(predict(modeli_konacni$analyses[[1]],type=
    "response")+predict(modeli_konacni$analyses[[2]],type=
    "response")+predict(modeli_konacni$analyses[[3]],type=
    "response")+predict(modeli_konacni$analyses[[4]],type=
    "response")+predict(modeli_konacni$analyses[[5]],type=
    "response"))/5
roc_krivulja<-roc(podaci$prezivljavanje,procijenjene,
    direction="<",print.auc=TRUE)
plot(roc_krivulja,legacy.axes=TRUE,xlab="1-specificnost",
    ylab="senzitivnost")
plot(podaci$prezivljavanje,procijenjene,xlab="opazene
    vrijednosti",ylab="procijenjene vrijednosti")
```

Bibliografija

[1] http://breyer.hr/pretrage/sve-pretrage/afp, posjećena u studenom 2021.

[2] https://poliklinika-labplus.hr/albumini/, posjećena u studenom 2021.

[3] https://medlineplus.gov/lab-tests/alkaline-phosphatase/, posjećena u studenom 2021.

[4] https://poliklinika-aviva.hr/usluge/ast-alt/, posjećena u studenom 2021.

[5] https://www.plivazdravlje.hr/bolest-clanak/bolest/287/Anemija.html, posjećena u siječnju 2022.

[6] http://www.msd-prirucnici.placebo.hr/msd-prirucnik/bolesti-jetre-i-zuci/pristup-jetrenom-bolesniku/ascites, posjećena u studenom 2021.

[7] Box, G.E.P. i N.R. Draper: *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 74, 1987.

[8] https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290, posjećena u prosincu 2021.

[9] Chicco, D. i G. Jurman: *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*. BMC Medical Informatics and Decision Making, Springer, svezak 20, članak 16, 2020.

[10] Christensen, R.: *Log-linear Models and Logistic Regression*. Springer Texts in Statistics. Springer, Second Edition, 1997.

[11] https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=bilirubin_direct, posjećena u studenom 2021.

[12] Dua, D. i C. Graff: *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017. http://archive.ics.uci.edu/ml, posjećena u rujnu 2021.

[13] https://my.clevelandclinic.org/health/diseases/21220-hepatic-encephalopathy, posjećena u studenom 2021.

[14] Everitt, B. S. i Hothorn T.: *A Handbook of Statistical Analyses Using R*. Taylor and Francis Group, LLC, Second Edition, 2010.

[15] https://www.mayoclinic.org/tests-procedures/ferritin-test/about/pac-20384928, posjećena u studenom 2021.

[16] https://medlineplus.gov/lab-tests/gamma-glutamyl-transferase-ggt-test/, posjećena u studenom 2021.

[17] https://www.mayocliniclabs.com/test-catalog/Clinical+and+Interpretive/8311, posjećena u studenom 2021.

[18] https://www.hepb.org/prevention-and-diagnosis/diagnosis/hbv-blood-tests/, posjećena u studenom 2021.

[19] *HCC Survival Data Set*. https://archive.ics.uci.edu/ml/datasets/HCC+Survival. posjećena u rujnu 2021.

[20] https://www.cdc.gov/hepatitis/hcv/HepatitisCTesting.htm, posjećena u studenom 2021.

[21] *Heart failure clinical records Data Set*. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records. posjećena u rujnu 2021.

[22] https://my.clevelandclinic.org/health/diseases/14971-hemochromatosis-iron-overload, posjećena u studenom 2021.

[23] Heymans, M. W. i I. Eekhout: *Applied Misssing Data Analysis with SPSS and (R)Studio*. First Draft, 2019. https://bookdown.org/mwheymans/bookmi/, posjećena u siječnju 2022.

[24] Hosmer, D. W. i S. Lemeshow: *Applied Logistic Regression*. Wiley Series in Probability and Statistics - Texts and References Section. Wiley, Second Edition, 2000.

[25] https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=international_normalized_ratio, posjećena u studenom 2021.

[26] https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646, posjećena u studenom 2021.

[27] https://health.ucdavis.edu/vascular/diseases/renal_insufficiency.html, posjećena u studenom 2021.

[28] *Linearni modeli, Statistički praktikum 2, Druge vježbe*. Mrežne stranice PMF-MO, ograničen pristup – objavljivano tijekom akademske godine 2020./2021.

[29] Meeyai, S.: *Logistic Regression with Missing Data: A Comparison of Handling Methods and Effects of Percent Missing Values*. Journal of Traffic and Logistics Engineering, svezak 4, 2016. http://www.jtle.net/uploadfile/2016/1108/20161108041850649.pdf, posjećena u prosincu 2021.

[30] Mimica, A. i M. Ninčević: *Statistika primjeri i zadaci*. Mrežne stranice PMF-MO, https://web.math.pmf.unizg.hr/nastava/stat/files/vjezbe_novo.pdf, 2010. ograničen pristup, posjećena u listopadu 2021.

[31] https://stanfordhealthcare.org/medical-conditions/liver-kidneys-and-urinary-system/nonalcoholic-steatohepatitis-nash.html, posjećena u studenom 2021.

[32] Parikh, R., A. Mathai, S. Parikh, G. C. Sekhar i R. Thomas: *Understanding and using sensitivity, specificity and predictive values*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062/, posjećena u studenom 2021.

[33] https://my.clevelandclinic.org/health/diseases/4912-portal-hypertension, posjećena u studenom 2021.

[34] https://online.stat.psu.edu/stat504/lesson/6/6.2/6.2.3, posjećena u prosincu 2021.

[35] Santos, M. S., P. H. Abreu, P. J. García-Laencina, A. Simão i A. Carvalho: *A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients*. Journal of biomedical informatics, svezak 58, 49-59, 2015.

[36] Slijepčević, S.: *Statistika, bilješke s predavanja*. Nastale tijekom slušanja kolegija akademske godine 2018./2019.

[37] https://www.healthline.com/health/splenomegaly, posjećena u studenom 2021.

[38] https://www.statology.org/standard-error-of-regression-slope/, posjećena u siječnju 2022.

[39] https://ecog-acrin.org/resources/ecog-performance-status, posjećena u studenom 2021.

[40] https://www.healthline.com/health/portal-vein-thrombosis, posjećena u studenom 2021.

[41] Tsai, A. C.: *Achieving Consensus on Terminology Describing Multivariable Analyses*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679183/, posjećena u studenom 2021.

[42] https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=total_bilirubin_blood, posjećena u studenom 2021.

[43] https://www.healthline.com/health/total-protein#results, posjećena u studenom 2021.

[44] Van Buuren, S.: *Flexible Imputation of Missing Data*. Chapman Hall/CRC, Second Edition, 2018. https://stefvanbuuren.name/fimd/, posjećena u siječnju 2022.

[45] https://my.clevelandclinic.org/health/diseases/15429-esophageal-varices, posjećena u studenom 2021.

Sažetak

U prvom poglavlju ovog rada objašnjeni su osnovni pojmovi vezani za logističku regresiju kao i razlike između linearne te logističke regresije. Detaljno je definiran pojam logit funkcije kao i metoda maksimalne vjerodostojnosti kojom se procjenjuju parametri modela. Navedeno je nekoliko testova za testiranje značajnosti koeficijenata te kako se može dobiti procjena pouzdanih intervala za vrijednosti koeficijenata. Objašnjen je pojam omjera izgleda koji je temeljan za razumijevanje interpretacije parametara modela. Zatim je rečeno što ako nisu svi podaci dostupni. Po koracima je naveden postupak koji se najčešće slijedi kako bismo na kraju dobili odgovarajući model te je detaljno opisana *stepwise* metoda odabira varijabli koje će se nalaziti u modelu. Navedeni su i načini kako možemo usporediti dva modela te kako možemo procijeniti adekvatnost modela.

U drugom poglavlju ovog rada prikazani su rezultati dobiveni provođenjem opisanih koraka za odabir varijabli koje će biti u modelu na dva primjera iz područja medicine. Pritom je korišten programski jezik R. U oba primjera kao konačan rezultat dobiveni su modeli koji su prema iznosu površine ispod ROC krivulje ocijenjeni odličnima. U primjeru o srčanom zatajenju dobiveno je da vjerojatnost smrti ovisi o vremenu praćenja pacijenta, o njegovoj dobi te o vrijednostima kreatinina u krvi i ejekcijske frakcije. Vjerojatnost preživljavanja osoba koje imaju hepatocelularni karcinom može se modelirati u ovisnosti o dobi pacijenta, o tome ima li simptome karcinoma, o statusu njegovog općeg tjelesnog stanja, o tome je li bio i/ili je trenutno zaražen hepatitisom C te o vrijednostima hemoglobina i enzima alkalne fosfataze kao i o vrijednostima njihove međusobne interakcije.

Summary

In the first chapter of this master thesis we introduced the basic concepts of logistic regression as well as the differences between linear and logistic regression. Logit function and maximum likelihood method for estimating parameters of the model were described in detail. A few statistical tests for testing for the significance of the coefficients and confidence interval estimations of the coefficients were also provided. The concept of odds ratio which is fundamental in understanding how to interpret parameters of the model is also introduced. Additionally, we discussed what to do when there are missing values in the data. We introduced the most common procedure for variable selection step by step and described stepwise procedure in great detail. Lastly, some ways of comparing two models and goodness-of-fit statistics were suggested.

In the second chapter we demonstrated variable selection procedure on two medical examples using programming language R. In both examples the final models are considered to be excellent in terms of obtained value of the area under the ROC curve. In the first example regarding the heart failure data we can conclude that probability that patient died depends on their follow-up period, age, serum creatinine value and ejection fraction percentage whereas in the second example about hepatocellular carcinoma we can conclude that probability that patient survived is related to their age, performance status, haemoglobin and alkaline phosphatase values as well as its interaction value, whether they have symptoms of the disease and whether they are and/or were infected with hepatitis C.

Životopis

Rođena sam 8. lipnja 1996. godine u Zagrebu gdje sam i odrasla. Nakon završene Osnovne škole kralja Tomislava 2011. godine upisala sam „opći" program V. gimnazije te sam sve razrede završila s odličnim uspjehom. Obrazovanje sam nastavila 2015. godine upisujući preddiplomski sveučilišni studij Matematika na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu te sam stekla akademski naziv sveučilišne prvostupnice matematike (*univ. bacc. math.*). Nakon toga, 2019. godine sam na istom odsjeku upisala diplomski sveučilišni studij Matematička statistika.

Tečno govorim engleski, a tijekom osnovnoškolskog i srednjoškolskog obrazovanja učila sam i njemački te pohađala tečajeve španjolskog jezika u školi stranih jezika.