Analiza ishoda liječenja pacijenata s Clostridium difficile infekcijom logističkom regresijom

Bogdanić, Matea

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Matea Bogdanić

ANALIZA ISHODA LIJEČENJA PACIJENATA S CLOSTRIDIUM DIFFICILE INFEKCIJOM LOGISTIČKOM REGRESIJOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, veljača 2019.

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. _____

Mojoj obitelji

Sadržaj

Uvod

Bakterija C. difficile je otkrivena godine 1935. (Hall i O'Tolle) kada je zbog teškog i kompliciranog uzgoja nazvana Bacillus difficilis. Otkrićem antibiotika i njihovom sve češćom upotrebom tijekom 20. stoljeća dolazi do porasta incidencije navedene bolesti. Intenzivno istraživanje „klindamicinskog kolitisa" je 1970-ih godina rezultiralo nepobitnim dokazivanjem uloge toksina C. difficile u patogenezi bolesti. Neželjene posljedice primjene antibiotika, ovisno o vrsti antibiotika, epidemiološkim okolnostima i populaciji iz koje bolesnik potiče, pojavljuju se u 25% do 50% osoba koje koriste antibiotsko liječenje. Procjenjuje se da je bakterija Clostridium difficile uzročnik oko 25% slučajeva postantimikrobne dijareje te je uzročnik gotovo svih teških oblika bolesti. Od tada pa do naših dana, zbog porasta incidencije teških oblika bolesti i sklonosti rekurentnom pojavljivanju, raste medicinski i ekonomski značaj infekcije uzrokovane s C. difficile.[3]

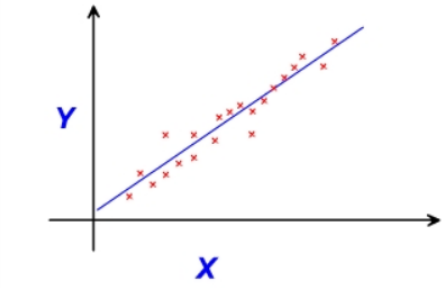
U ovom radu ćemo koristeći model logističke regresije sa dihotomnom zavisnom varijablom napraviti analizu ishoda liječenja pacijenata s C. difficile infekcijom. Podaci su prikupljeni iz Arhive za medicinsku dokumentaciju Klinike za infektivne bolesti "Dr. Fran Mihaljević". Etičko povjerenstvo Klinike za infektivne bolesti „Dr. Fran Mihaljević" je odobrilo ovo istraživanje, te korištenje njihove baze podataka. Zavisna varijabla je konačan ishod liječenja, dok su nezavisne varijable: dob, spol, McCabe score, klinička težina bolesti, trajanje hospitalizacije, ataka, temperatura, leukociti, pokretnost, kreatinin, konkomitantne infekcije, te JIM (boravak u jedinici intenzivne medicine zbog CDI). Podatke ćemo obrađivati u statističkom programu SAS 9.4.

Poglavlje 1

Logistička regresija

1.1 Linearna regresija

Regresijska analiza je metoda ispitivanja ovisnosti zavisne varijable (varijable odaziva) o jednoj ili više nezavisnih varijabli (varijabli poticaja). Glavni rezultat regresijske analize jest regresijski model - matematička jednadžba koja kvantificira tu povezanost. Ako je ta povezanost linearna, radi se o linearnoj regresiji.
Kod linearne regresije povezanost izmedu zavisne i nezavisne varijable opisana je jednadžbom pravca, koristeći metodu najmanjih kvadrata. Pravac koji najbolje opisuje povezanost tih varijabli odaberemo tako da iz skupa svih pravaca odaberemo onaj čija je suma odstupanja svake točke od pravca najmanja.[4]



Slika 1.1: Grafički prikaz univarijatne linearne regresije
izvor:https://tex.stackexchange.com

Uvedimo oznake:

- $x_1, x_2, ..., x_p$ - nezavisne varijable (varijable poticaja)
- y - zavisna varijabla (varijabla odaziva)
- ε - slučajna greška
- $\beta_0, \beta_1, ..., \beta_p$ - parametri modela

Linearni regresijski model izgleda ovako:

$$y = \beta_0 + \sum_{k=1}^{p} \beta_k x_k + \varepsilon. \tag{1.1}$$

Ovisno o p razlikujemo univarijatnu (jednostruku) linearnu regresiju ($p = 1$) i multivarijatnu (višestruku) linearnu regresiju ($p > 1$).
U primjeni imamo više opažanja pa model zapisujemo:

$$y = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} + \varepsilon_i, \qquad i = 1, 2, .., n, \tag{1.2}$$

gdje pretpostavljamo da su greške ε_i nezavisne s distribucijom $N(0, \sigma^2)$.

Uvedimo oznake te zapišimo model u matričnom obliku:

- $\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$
- $Y = (y_1, y_2, ..., y_n)^T$
- $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$
- $b = (\beta_0, \beta_1, ..., \beta_p)^T$

te zapis modela sada izgleda ovako:

$$Y = Xb + \varepsilon. \tag{1.3}$$

Naš je cilj minimizirati funkciju

$$L(b) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{11} - \cdots - \beta_k x_{ik})^2$$

da bi mogli procijeniti parametre modela b. Ono što zapravo radimo je minimiziramo sumu kvadratnih reziduala

$$L(b) = \sum_{i=1}^{n} \varepsilon_i^2,$$

pa dobivamo da je najbolja ocjena za b

$$\hat{b} = (X^T X)^{-1} X^T Y,$$

uz uvjet regularnosti matrice $X^T X$.

Procijenjene vrijednosti tada su jednake

$$\hat{Y} = X\hat{b} = X(X^T X)^{-1} X^T Y,$$

a ostatci

$$\varepsilon = Y - \hat{Y}.$$

Glavne pretpostavke koje opravdavaju korištenje linearnog regresijskog modela su:

- linearni odnos između varijabli poticaja i odaziva
- nekoreliranost varijable poticaja i greške
- nezavisnost grešaka
- homogenost grešaka
- normalna distribuiranost grešaka

Ukoliko neka od pretpostavki nije zadovoljena, naši rezultati mogu biti nevaljani.[4][5]

1.2 Logistička regresija - osnovni pojmovi

Regresijske metode se koriste u analizama podataka kada je potrebno opisati vezu između varijable odaziva (zavisne) i jedne ili više varijabli poticaja (nezavisnih varijabli). Osnovna ideja ovakve analize je naći najbolji odgovarajući, situaciji prikladan model koji bi opisao tu vezu. Nezavisne varijable se često nazivaju kovarijatama. Razlikujemo slučajeve u ovisnosti o vrsti zavisne varijable. Zavisna varijabla može biti kontinuirana (numerička) ili diskretna (kategorijska) - odnosno da poprima jednu, dvije ili više mogućih vrijednosti. Slučaj u kojem je zavisna varijabla kontinuirana se analizira metodom linearne regresije, dok se slučaj sa kategorijskom zavisnom varijablom analizira metodom logističke regresije.

Jedan od osnovnih slučajeva je kada zavisna varijabla poprima samo dvije vrijednosti - dihotomna varijabla odaziva. Povijesno gledano, kada se javila potreba za analizom takvih modela, različite funkcije su bile promatrane za korištenje. U tu svrhu je izabrana logistička distribucija iz dva osnovna razloga. Prvi razlog je matematičke prirode - ta funkcija je fleksibilna i lako se koristi, dok je drugi razlog da se lako interpretira.[1][2]

Forma logističkog regresijskog modela (logistička funkcija) koji koristimo je:

$$p(x) = \frac{1}{1 + e^{-x}} \tag{1.4}$$

gdje je $x \in \langle -\infty, \infty \rangle$, a $p(x) \in \langle 0, 1 \rangle$.

Inverzna funkcija logističke funkcije zove se logit i definira se:

$$\begin{aligned} logit(p(x)) =& \log\left[\frac{p(x)}{1-p(x)}\right] \\ =& \log(p(x)) - \log(1 - p(x)), \end{aligned} \tag{1.5}$$

gdje je $p \in \langle 0, 1 \rangle$, a $logit(p) \in \langle -\infty, \infty \rangle$.

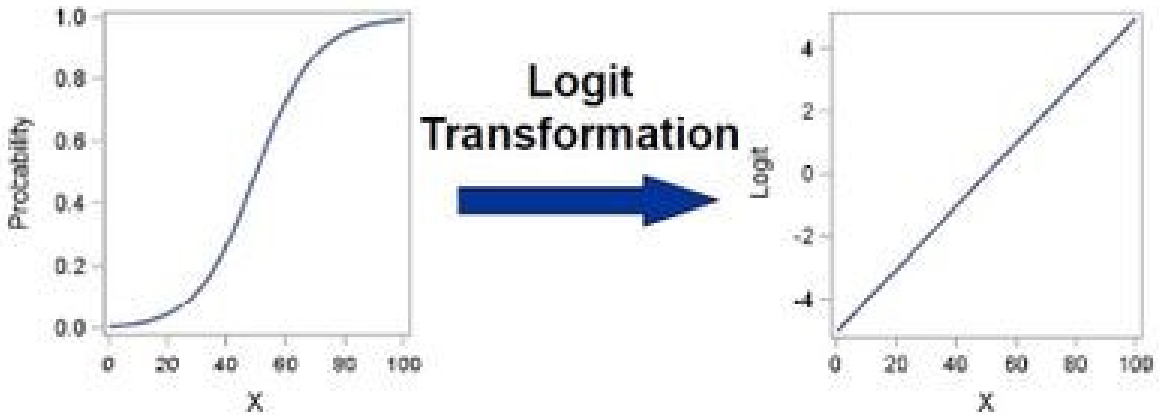
Promotrimo vjerojatnost da se neki događaj dogodi. Primijetimo da je vjerojatnost p funkcija koja poprima vrijednosti u intervalu $\langle 0, 1 \rangle$.
Izgled (engl. odds) nekog događaja je omjer očekivanog broja puta kada će se događaj dogoditi naspram očekivanog broja puta kada se događaj neće dogoditi. Veza između vjerojatnosti i izgleda je:

$$izgled = \frac{p(x)}{1 - p(x)}. \tag{1.6}$$

Tablica 1.1: Odnos vjerojatnosti, izgleda i log izgleda

vjerojatnost	izgled	log izgleda
0.100	0.111	-2.197
0.200	0.250	-1.386
0.300	0.428	-0.847
0.400	0.667	-0.405
0.500	1.000	0.000
0.600	1.500	0.405
0.700	2.333	0.847
0.800	4.000	1.386
0.900	9.000	2.197

Vidimo da *izgled* poprima vrijednosti u intervalu $\langle 0, \infty \rangle$, dok log izgleda poprima vrijednosti u intervalu $\langle -\infty, \infty \rangle$. Točnije, transformacijom $\frac{p(x)}{1-p(x)}$ mičemo gornju granicu, dok logaritmiranjem mičemo donju granicu. Iz tog razloga koristimo upravo log izgleda, kako bi izbjegli modeliranje varijable sa restrikcijama kao što su vjerojatnost i izgled. Logističkom regresijom modeliramo logit transformiranu vjerojatnost kao linearnu vezu sa prediktorskim varijablama. Na slici 1.2 vdimo kako izgleda opisana transformacija.



Slika 1.2: Grafički prikaz logit transformacije
izvor: https://communities.sas.com

Logistički regresijski model izgleda:

$$\begin{aligned} logit(p(x)) &= log(izgled(x)) \\ &= \beta_0 + \beta_1 x. \end{aligned} \tag{1.7}$$

Slijedi da je

$$\begin{aligned} izgled(x) &= \frac{p(x)}{1 - p(x)} \\ &= e^{\beta_0 + \beta_1 x}. \end{aligned} \tag{1.8}$$

Nadalje, slijedi

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}. \tag{1.9}$$

Model sa k prediktorskih varijabli izgleda:

$$\begin{aligned} logit(p(x)) &= log(izgled(x)) \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k. \end{aligned} \tag{1.10}$$

Bitna razlika logističkog i linearnog regresijskog modela je u distribuciji slučajne greške. U linearnom modelu greška ε slijedi normalnu distribuciju s očekivanjem nula i konstantnom varijancom. U slučaju dihotomne varijable odaziva, vrijednost varijable odaziva je $y = p(x) + \varepsilon$. Ovdje ε može poprimati dvije moguće vrijednosti. Ako je $y = 1$ tada je $\varepsilon = 1 - p(x)$ sa vjerojatnošću $p(x)$, a ako je $y = 0$ tada je $\varepsilon = -p(x)$ sa vjerojatnošću $1 - p(x)$. Dakle, u tom slučaju, slučajna greška ε slijedi binomnu distribuciju.[1]
Osnovna metoda procjene parametara logističkog modela zove se metoda maksimalne vjerodostojnosti (ML).

1.3 Metoda maksimalne vjerodostojnosti

Pretpostavimo da imamo uzorak od n nezavisnih observacija parova (x_i, y_i), $i = 1, 2, ..., n$ gdje y_i označava vrijednost dihotomne varijable odaziva a x_i vrijednost nezavisne varijable i-tog mjerenja. Pretpostavimo da varijabla odaziva može poprimati samo vrijendosti 1 ili 0, što bi predstavljalo prisutstvo ili odsustvo određene karakteristike. Potrebno je odrediti vrijednosti parametara β_0 i β_1. U linearnoj regresiji smo u tu svrhu koristili metodu najmanjih kvadrata - odabirom koeficijenata tako da suma kvadratnih odstupanja bude minimizirana. Dobiveni procjenitelji su imali dobra statistička svojstva. Metoda koja se u logističkom regresijskom modelu koristi za procjenu parametara modela naziva se metoda maksimalne vjerodostojnosti. U vrlo općem smislu ova metoda pridružuje vrijednosti nepoznatim parametrima tako da oni maksimiziraju vjerojatnost dobivanja promatranog skupa podataka. Da bi se ova metoda mogla primijeniti, prvo treba konstruirati funkciju koja se naziva vjerodostojnost (eng. likelihood function). Ova funkcija predstavlja vjerojatnost dobivanja promatranih podataka kao funkciju nepoznatih parametara. ML procjenitelji nepoznatih parametara su one vrijednosti koje maksimiziraju vjerodostojnost.[1]

Označimo sa $\hat{y}$ ML procjenu od y, sa $\hat{p(x)}$ ML procjenu od $p(x)$, te sa $\beta = (\beta_0, \beta_1)$ parametre univarijatnog logističkog modela.

Vjerodostojnost izgleda ovako:

$$l(\beta) = \prod_{i=1}^{n} [p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}] \tag{1.11}$$

pri čemu su (x_i, y_i), $i = 1, 2, \ldots, n$ naše observacije.

Zbog jednostavnosti korištenja, daljnji računi se rade sa funkcijom log-vjerodostojnosti, odnosno $log(l(\beta))$.
Funkcija log-vjerodostojnost koju označavamo sa $L(\beta)$ izgleda ovako:

$$L(\beta) = \sum_{i=1}^{n} [(y_i) ln(p(x_i)) + (1 - y_i) ln(1 - p(x_i))]. \tag{1.12}$$

Kako bi se pronašle tražene vrijednosti, dakle parametre $\beta = (\beta_0, \beta_1)$ koji maksimiziraju funkciju $L(\beta)$, funkciju $L(\beta)$ je potrebno derivirati po β_0, β_1 te izjednačiti s nulom. Jednadžbe koje na taj način dobijemo nazivaju se jednadžbe vjerodostojnosti (eng. likelihood equations):

$$\begin{aligned} \sum_{i=1}^{n} [y_i - p(x_i)] &= 0 \\ \sum_{i=1}^{n} x_i [y_i - p(x_i)] &= 0. \end{aligned} \tag{1.13}$$

Ove jednadžbe su nelinearne u parametrima β_0, β_1, pa se rješavaju iterativnim metodama koje su implementirane u statističkim programima.
Vrijednost β koja se dobije rješavanjem ovih jednadžbi zove se ML procjenitelj i označava sa $\hat{\beta}$.

1.4 Testiranje adekvatnosti modela i značajnosti parametara

Nakon procjene parametara, testira se značajnost varijabli u modelu. Ovaj dio uključuje formulaciju i testiranje statističke hipoteze u svrhu određivanja da li su nezavisne (prediktorske) varijable u našem modelu značajno povezane sa zavisnom varijablom. U ovom dijelu predstavljena je generalna ideja koja se može u detaljima razlikovati ovisno o kakvom se modelu radi. Zbog jednostavnosti, pretpostavimo da se radi o univarijatnom logističkom modelu. Pitanje od kojeg krećemo je: Da li model koji sadrži nezavisnu varijablu govori više o zavisnoj varijabli, nego model koji ne sadrži tu nezavisnu varijablu? Na ovo pitanje se odgovara usporedbom promatranih vrijednosti zavisne varijable sa procijenjenim vrijednostima svakog od ta dva modela - modela sa i bez nezavisne varijable. Matematička funkcija koja se koristi za usporedbu tih vrijednosti ovisi o određenom problemu. Ako su procijenjene vrijednosti s nezavisnom varijablom u modelu bolje, odnosno točnije, upućuje da bi nezavisna varijabla mogla biti značajna za model. U logističkoj regresijskoj analizi, usporedba promatranih i predviđenih vrijednosti zasniva se na funkciji log-vjerodostojnosti koju smo ranije definirali. Da bi se bolje razumjela ta usporedba, konceptualno razmišljamo o promatranim vrijednostima zavisne varijable kao predviđenim vrijednostima koje dobivamo iz satuiranog modela. Satuirani model je onaj koji sadrži onoliko parametara koliko ima observacija. Jednostavan primjer satuiranog modela je korištenje linearnog regresijskog modela kada imamo samo dvije observacije. Usporedba promatranih i dobivenih vrijednosti koristeći funkciju vjerodostojnosti bazira se na sljedećem izrazu:

$$D = -2\log\Big[\frac{\text{funkcija vjerodostojnosti modela}}{\text{funkcija vjerodostojnosti satuiranog modela}}\Big]. \tag{1.14}$$

Izraz unutar velikih zagrada naziva se omjer vjerodostojnosti (eng. likelihood ratio).
Da bi došli do poznate distribucije u svrhu testiranja hipoteza nužno je koristiti funkciju $-2log$(omjer vjerodostojnosti).
Takav test se zove test omjera vjerodostojnosti (eng.likelihood ratio test).

$$\begin{aligned} D &= -2\log\Big[\frac{\text{funkcija vjerodostojnosti modela}}{\text{funkcija vjerodostojnosti satuiranog modela}}\Big] \\ &= -2\sum_{i=1}^{n}\Big[y_i\log\frac{p(\hat{x_i})}{y_i} + (1-y_i)\log\frac{1-p(\hat{x_i})}{1-y_i}\Big]. \end{aligned} \tag{1.15}$$

Statistika D se naziva devijanca. Devijanca u logističkoj regresiji ima istu ulogu kao rezidualna suma kvadrata u linearnoj regresiji. Da bi procijenili značajnost nezavisne varijable, uspoređujemo vrijednost statistike D sa i bez nezavisne varijable u jednadžbi.

$$G = D(\text{model bez varijable}) - D(\text{model sa varijablom}) \tag{1.16}$$

$$G = -2\log\Big[\frac{\text{funkcija vjerodostojnosti bez varijable}}{\text{funkcija vjerodostojnosti sa varijablom}}\Big] \tag{1.17}$$

Uz pretpostavku $\beta_1 = 0$, statistika G slijedi χ^2 distribuciju sa jednim stupnjem slobode. Kod univarijatnog logističkog modela testiramo hipoteze:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Kod multivarijatnog logističkog modela testiramo hipoteze:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \text{barem jedan } \beta_i \neq 0, i = 1, 2, \ldots, k.$$

Za parametre modela kažemo da su statistički značajni ako se statistički značajno razlikuju od nule. Nakon što zadamo razinu značajnosti α, dobivenu p-vrijednost uspoređujemo sa razinom značajnosti.
Ukoliko je $p < \alpha$, odbacujemo H_0 u korist alternativne hipoteze H_1, što znači da je nezavisna varijabla statistički značajna na razini značajnosti α. U suprotnom ne možemo odbaciti H_0 u korist alternative, što znači da nezavisna varijabla nije statistički značajna na razini značajnosti α.

1.5 Interpretacija parametara logističkog regresijskog modela

U ovom dijelu ćemo se baviti interpretacijom koeficijenata logističkog regresijskog modela za situaciju kada je nezavisna varijabla nominalna i dihotomna. Pretpostavimo da nezavisna varijabla, x, može poprimati vrijednosti nula ili jedan. Razlika u logit funkciji (u oznaci g) za slučaj kada $x = 1$ i $x = 0$ je:

$$\begin{aligned} g(1) - g(0) &= [\beta_0 + \beta_1] - [\beta_0] \\ &= \beta_1. \end{aligned} \tag{1.18}$$

Prvi korak u interpretaciji efekta nezavisne varijable u modelu je izraziti željenu logit razliku u terminima modela. U ovom slučaju, logit razlika je jednaka β_1.
Da bi dalje interpretirali ovaj rezultat moramo uvesti mjeru povezanosti omjer izgleda (eng. odds ratio).
Ranije smo spomenuli da ćemo vrijednost zavisne varijable $Y = 1$ smatrati prisustvom neke određene karakteristike, a vrijednost $Y = 0$ odsustvom iste. Izgled da ta karakteristika bude prisutna u skupini za koju vrijedi $x = 1$ definirana je kao

$$\frac{p(1)}{1-p(1)}.$$

Slično, izgled da ta karakteristika bude prisutna u skupini za koju vrijedi $x = 0$ definirana je kao

$$\frac{p(0)}{1-p(0)}.$$

Omjer izgleda, u oznaci OR, definira se kao omjer izgleda za $x = 1$ naprema izgledima za $x = 0$, te je dan ovom jednadžbom:

$$OR = \frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}}. \tag{1.19}$$

Nadalje, dobivamo:

$$\begin{aligned} OR &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e_0^{\beta}}{1+e_0^{\beta}}\right)/\left(\frac{1}{1+e_0^{\beta}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e_0^{\beta}} \\ &= e^{(\beta_0+\beta_1)-\beta_0} \\ &= e_1^{\beta}. \end{aligned} \tag{1.20}$$

Dakle, za logističku regresiju s dihotomnom nezavisnom varijablom koja poprima vrijednosti 0 i 1, veza između OR i regresijskog koeficijenta je:

$$OR = e^{\beta_1}. \tag{1.21}$$

Paramentar β_0 ili intercept je očekivana vrijednost zavisne varijable y kada je $x = 0$. Za kontinuirane nezavisne varijable vrijedi:

$$g(x+1) - g(x) = \beta_1. \tag{1.22}$$

Parametar β_1 pokazuje promjenu u log izgledu kada se nezavisna varijabla x pomakne za 1. Analogno, pomak nezavisne varijable x za konstantu c uzrokuje pomak u log izgledu:

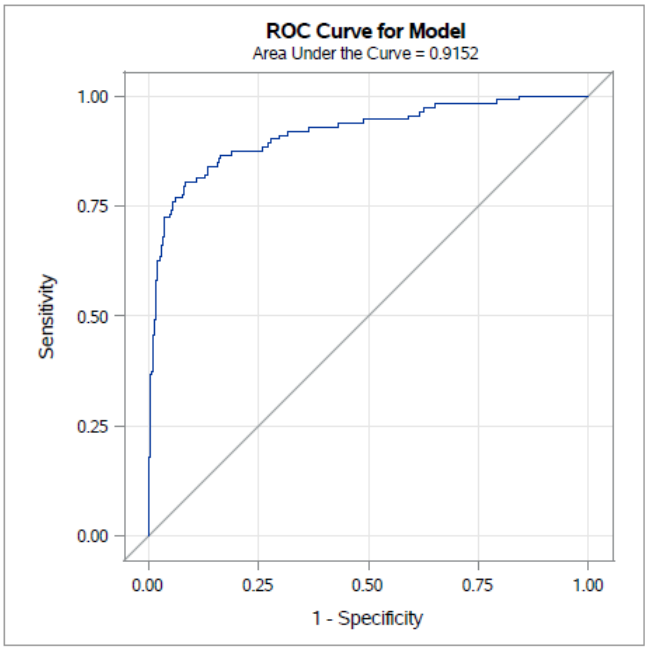
$$g(x+c) - g(x) = c\beta_1 \tag{1.23}$$

$$OR(c) = OR(x+c, x) = e^{c\beta_1} \tag{1.24}$$

Upravo ova jednostavna veza između regresijskog koeficijenta i OR je fundamentalni razlog zbog kojeg je logistička regresija izuzetno snažan analitički alat. OR je mjera povezanosti koja ima vrlo široku primjenu, posebno u epidemiologiji, jer aproksimira koliko je vjerojatno (ili nije vjerojatno) da neka karakteristika bude prisutna među određenim skupinama.

1.6 Roc krivulja

ROC (eng. Receiver Operating Characteristic) krivulja je grafički prikaz valjanosti dijagnostičkog testa. Valjanost dijagnostičkog testa je složeni pokazatelj i ima dvije komponente: osjetljivost i specifičnost. Osjetljivost testa je proporcija dobro detektiranih bolesnih osoba od sveukupnog broja bolesnih, a specifičnost testa je proporcija zdravih osoba koje su dobro detektirane kao zdrave, od ukupnog broja zdravih osoba. Bolji opis klasifikacijske točnosti dan je površinom ispod ROC krivulje. Ova krivulja, koja proizlazi iz teorije obrade signala, pokazuje kako prijemnik upravlja signalima koje prima u prisustvu buke. Grafički prikazuje vjerojatnost detektiranja stvarnog signala (osjetljivost) i lažnog signala (1- specifičnost) za sve moguće točke. Područje (površina) ispod ROC krivulje, koja poprima vrijednosti između 0 i 1, daje mjeru sposobnosti modela da razlikuje subjekte koji imaju ili nemaju prisustvo određene karakteristike (mjera točnosti). Uopćeno pravilo za tumačenje vrijednosti c, koja predstavlja površinu ispod ROC krivulje za određeni model je da kada je vrijednost c manja od 0.5 smatramo da taj model nema prediktivnu vrijednost, dok za vrijednost c veću od 0.8 smatramo da pripada modelu jako dobre prediktivne vrijednosti.[1]



Slika 1.3: primjer ROC krivulje (ispis iz SAS-a)

Poglavlje 2

Analiza ishoda liječenja pacijenata s Clostridium difficile infekcijom logističkom regresijom

2.1 Općenito o C. difficile i deskriptivna statistika

Primjena antibiotika može biti praćena mnogobrojnim neželjenim posljedicama od kojih su gastrointestinalne nuspojave jedne od češćih. Procjenjuje se da je bakterija Clostridium difficile uzročnik oko 25% slučajeva postantibiotskog proljeva (dijareje) te je uzročnik gotovo svih teških oblika bolesti. Sve češćom upotrebom antibiotika tijekom 20. stoljeća dolazi do porasta incidencije Clostridium difficile infekcije (CDI). Od tada pa do naših dana, zbog porasta incidencije teških oblika bolesti i lošijih ishoda liječenja, raste medicinski i ekonomski značaj CDI. Rizični čimbenici za obolijevanje od CDI se mogu podijeliti na primarne i sekundarne. Glavni primarni čimbenici opisani u literaturi su muški spol, dob iznad 65 godina, dob manja od jedne godine uz predležeće kronične bolesti, dugo trajanje hospitalizacije i antimikrobna terapija. Najznačajniji sekundarni rizični čimbenici su komorbiditeti i predležeća zdravstvena stanja, upalna bolest crijeva, imunodeficijencija i HIV, malnutricija, niska razina serumskih albumina, zloćudni tumori, cistična fibroza i dijabetes. Najznačajniji rizični čimbenik za razvoj bolesti je upotreba antibiotika širokog spektra. Infekcija uzrokovana s C. difficile se može očitovati širokim spektrom kliničkih stanja, od asimptomatske kolonizacije preko blagog proljeva do životno ugrožavajuće bolesti sa smrtnim ishodom liječenja. [3]

U ovom radu analizirani su bolesnici koji su hospitalno liječeni u Klinici za infektivne bolesti „Dr. Fran Mihaljević“ zbog CDI koja se definira kao prisutnost dijareje uz dokaz toksina bakretije c.difficile u uzorku stolice tijekom razdoblja od 1. siječnja 2013. do 31. prosinca 2017. godine. Radi se o retrospektivnom istraživanju koje je odobreno od strane Etičkog povjerenstva Klinike za infektivne bolesti „Dr. Fran Mihaljević“. Sve potrebne varijable prikupili su Nikolina Bogdanić, dr.med., Karlo Vidović, dr.med. te doc. dr. sc. Mirjana Balen Topić, specijalist infektolog, iz originalne arhivirane medicinske dokumentacije. Svrha istraživanja je bila utvrditi potencijalne rizične čimbenike za smrtni ishod liječenja CDI.
Podatci koje ćemo obraditi sastoje se od 1080 observacija, od kojih se svaka sastoji od 13 varijabli:

- **dob** bolesnika izražena u godinama

Basic Statistical Measures			
Location		Variability	
Mean	71.46204	Std Deviation	17.35708
Median	76.00000	Variance	301.26825
Mode	77.00000	Range	97.00000
		Interquartile Range	16.00000

Slika 2.1: Deskriptivna statistika za varijablu dob (ispis iz SAS-a)

- **spol** bolesnika - muškarci (0) i žene (1)

spol	Frequency	Percent
0	451	41.76
1	629	58.24

Slika 2.2: Tablica frekvencija za varijablu spol (ispis iz SAS-a)

- **trajanje hospitalizacije** izraženo u danima

Basic Statistical Measures			
Location		Variability	
Mean	16.22593	Std Deviation	17.92454
Median	11.00000	Variance	321.28904
Mode	10.00000	Range	235.00000
		Interquartile Range	9.00000

Slika 2.3: Deskriptivna statistika za varijablu trajanje hospitalizacije (ispis iz SAS-a)

- **McCabe score** - score prema kojem se kategoriziraju kronične predležeće bolesti na ljestvici od 0 do 3

McCabe	Frequency	Percent
0	48	4.44
1	573	53.06
2	411	38.06
3	48	4.44

Slika 2.4: Tablica frekvencija za varijablu McCabe score (ispis iz SAS-a)

- **klinička težina bolesti** - klinička definicija bolesti prema smjernicama na ljestvici od 1 do 4

klinicka_tezina	Frequency	Percent
1	151	13.98
2	425	39.35
3	381	35.28
4	123	11.39

Slika 2.5: Tablica frekvencija za varijablu klinička težina bolesti (ispis iz SAS-a)

- **ataka bolesti** - broj epizode bolesti

Basic Statistical Measures			
Location		Variability	
Mean	1.516667	Std Deviation	0.86734
Median	1.000000	Variance	0.75227
Mode	1.000000	Range	8.00000
		Interquartile Range	1.00000

Slika 2.6: Deskriptivna statistika za varijablu ataka (ispis iz SAS-a)

- **pokretnost** bolesnika - razlikujemo nepokretne (0) i pokretne (1) bolesnike

pokretnost	Frequency	Percent
0	583	53.98
1	497	46.02

Slika 2.7: Tablica frekvencija za varijablu pokretnost (ispis iz SAS-a)

- **tjelesna temperatura** bolesnika

Basic Statistical Measures			
Location		Variability	
Mean	37.85315	Std Deviation	0.85728
Median	38.00000	Variance	0.73493
Mode	38.00000	Range	7.30000
		Interquartile Range	1.20000

Slika 2.8: Deskriptivna statistika za varijablu tjelesna temperatura (ispis iz SAS-a)

- **broj leukocita** u krvi

Basic Statistical Measures			
Location		Variability	
Mean	15.27923	Std Deviation	8.69793
Median	13.50000	Variance	75.65390
Mode	8.90000	Range	94.30000
		Interquartile Range	8.40000

Slika 2.9: Deskriptivna statistika za varijablu broj leukocita u krvi (ispis iz SAS-a)

- **koncentracija kreatinina** u krvi - razlikujemo stanje kada je kreatinin bio povišen 1.5 puta u odnosu na vrijednosti prije razbolijevanja (1)

kreat	Frequency	Percent
0	886	82.04
1	194	17.96

Slika 2.10: Tablica frekvencija za varijablu koncentracija kreatinina u krvi (ispis iz SAS-a)

- **konkomitantne infekcije** - razlikujemo slučaj kada su se pojavile dodatne infekcije uz CDI (1)

konkomit_inf	Frequency	Percent
0	411	38.06
1	669	61.94

Slika 2.11: Tablica frekvencija za varijablu konkomitantne infekcije (ispis iz SAS-a)

- **liječenje u jedinici intenzivne medicine** - razlikujemo slučaj kada se bolesnik liječio u JIM-u zbog CDI infekcije (1)

JIM	Frequency	Percent
0	1058	97.96
1	22	2.04

Slika 2.12: Tablica frekvencija za varijablu liječenje u jedinici intenzivne medicine (ispis iz SAS-a)

- **ishod liječenja** - razlikujemo smrtni ishod (1)

ishod	Frequency	Percent
0	968	89.63
1	112	10.37

Slika 2.13: Tablica frekvencija za varijablu ishod liječenja (ispis iz SAS-a)

Zavisna varijabla u ovoj analizi je ishod liječenja, dok su ostale varijable nezavisne.

2.2 Univarijatna logistička regresija

Za svaku od 12 nezavisnih varijabli iz baze ćemo provesti univarijatnu logističku regresiju, te odrediti značajnost svake varijable. U tablici 2.1 nalaze se osnovne dobivene vrijednosti.

Tablica 2.1: Rezultati analize univarijatnih logističkih modela

varijabla	df	-2logL(intercept only)	-2logL(intercept and covariates)	likelihood ratio (χ^2)	konvergencija	p-vrijednost
dob	1	719.594	694.103	25.4911	zadovoljena	< .0001
trajanje	1	719.594	718.346	1.2480	zadovoljena	0.3064
McCabe score	1	719.594	658.390	61.2045	zadovoljena	<.0001
klinička tež	1	719.594	440.402	279.1925	zadovoljena	<.0001
ataka	1	719.594	711.932	7.6619	zadovoljena	0.0121
temperatura	1	719.594	709.022	10.5721	zadovoljena	0.0012
leukociti	1	719.594	663.425	56.1696	zadovoljena	<.0001
spol	1	719.594	719.532	0.0618	zadovoljena	0.8035
pokretnost	1	719.594	666.496	53.0978	zadovoljena	<.0001
kreatinin	1	719.594	679.541	40.0531	zadovoljena	<.0001
konkom inf	1	719.594	679.512	40.0826	zadovoljena	<.0001
JIM	1	719.594	697.035	22.5588	zadovoljena	<.0001

Iz tablice 2.1 se vidi da je kriterij kovergencije zadovoljen za sve varijable. Iz tablice 2.1 i izračunatih p-vrijednosti vidimo da su na razini značajnosti od 5% statistički značajne varijable dob, McCabe score, klinička težina bolesti, ataka, temperatura, leukociti, pokretnost, kreatinin, konkomitantne infekcije, te JIM (boravak u jedinici intenzivne medicine zbog CDI).

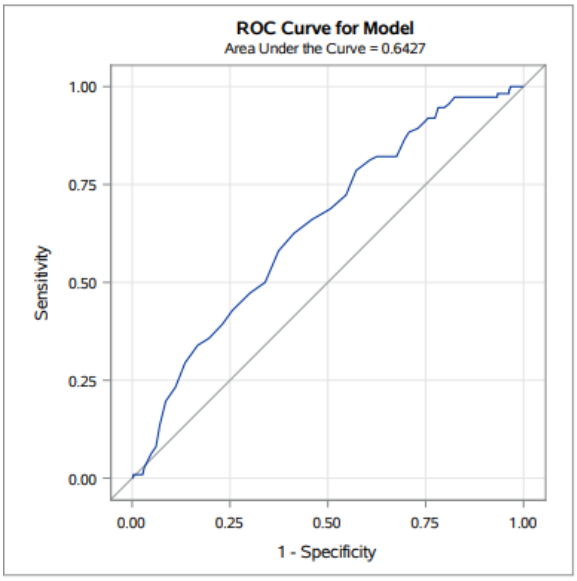
Tablica 2.2: Procjene omjera izgleda za univarijatne logističke modele

varijabla	procjena OR	95% pouzdani interval	c
dob	1.040	1.022 - 1.059	0.643
trajanje	0.993	0.979 - 1.007	0.635
McCabe score	3.343	2.441 - 4.578	0.697
klinička tež	14.243	9.399 - 21.584	0.883
ataka	0.678	0.500 - 0.918	0.556
temperatura	1.464	1.162 - 1.845	0.605
leukociti	1.076	1.054 - 1.097	0.709
spol	0.951	0.640 - 1.412	0.506
pokretnost	0.182	0.107 - 0.309	0.672
kreatinin	4.038	2.669 - 6.111	0.634
konkom inf	4.867	2.740 - 8.644	0.643
JIM	9.475	4.007 - 22.403	0.543

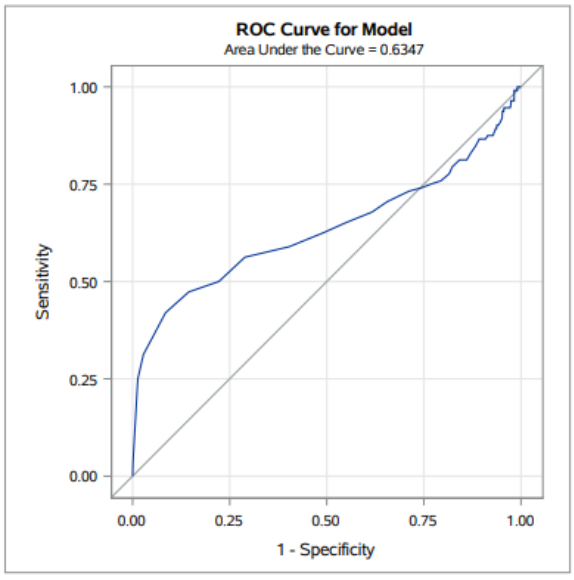
Drugi kriterij kojim možemo provjeriti statističku značajnost varijabli je 95% pouzdani interval. Ako 95% pouzdani interval ne sadrži jedinicu, onda je varijabla statistički značajna. Iz tablice 2.2 vidimo da su navedene varijable po tom kriteriju statistički značajne na razini značajnosti od 5%. Vrijednosti u tablici 2.2 pod "Procjena OR" predstavljaju omjer izgleda prelaska iz nesmrtnog ishoda u smrtni ishod, uz prelazak nezavisne varijable iz niže u višu kategoriju, odnosno uz pomak za 1.

- Povećanje dobi za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 1.040 puta
- Povećanje trajanja hospitalizacije za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 0.993 puta
- Povećanje McCabe score-a za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 3.343 puta
- Povećanje kliničke težine za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 14.243 puta
- Povećanje atake za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 0.678 puta
- Povećanje temperature za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 1.464 puta
- Povećanje leukocita za 1 jedinicu povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 1.076 puta
- Žene s obzirom na muškarce imaju manji omjer izgleda za smrtni ishod (OR=0.951)
- Pokretne osobe s obzirom na nepokretne imaju manji omjer rizika za smrtni ishod (OR=0.182)
- Osobe s povišenim kreatininom imaju veći omjer izgleda za smrtni ishod od osoba bez povišenog kreatinina (OR=4.038)
- Osobe s konkomitantnim infekcijama imaju veći omjer rizika za smrtni ishod od osoba bez konkomitantnih infekcija (OR=4.867)
- Osobe koje su se liječile u JIM-u zbog CDI imaju veći omjer rizika za smrtni ishod od osoba koje se nisu liječile u JIM-u zbog CDI (OR=9.475)

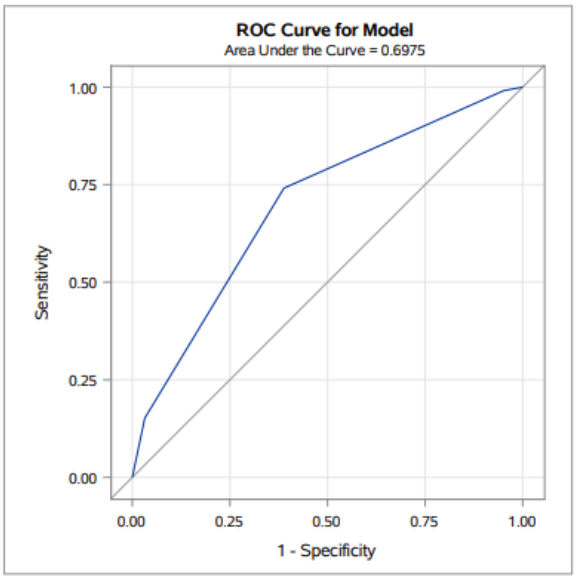
Podatak koji nam govori o prediktivnoj snazi modela je vrijednost c, koju iščitavamo iz tablice 2.2, za svaku varijablu. Varijabla klinička težina bolesti ima najveću c-vrijednost (0.883), slijedi je varijabla leukociti (0.709), dok najmanju c-vrijednost ima varijabla spol (0.506). Ostale varijable imaju c- vrijednost u intervalu od 0.6 do 0.7. ROC krivulje svakog pojedinog modela prikazane su narednim grafovima.



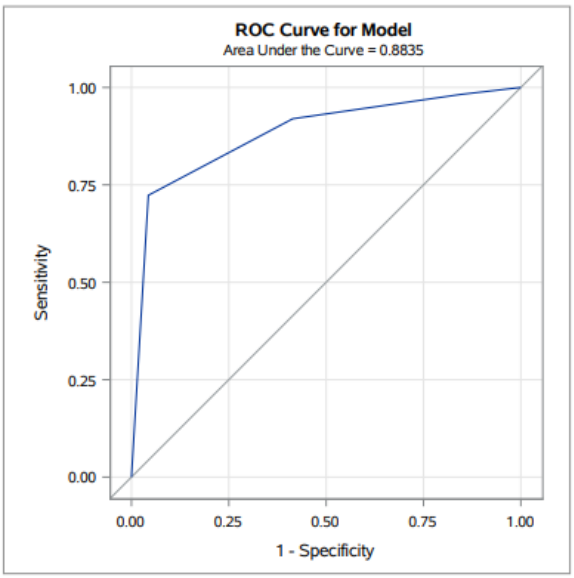
Slika 2.14: ROC krivulja za varijablu dob (ispis iz SAS-a)



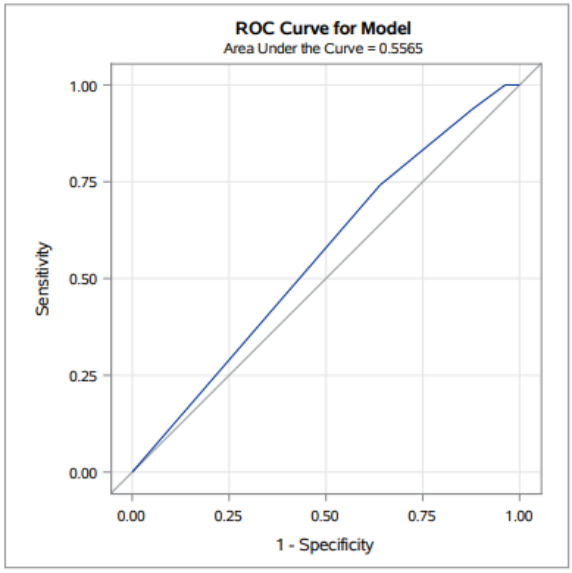
Slika 2.15: ROC krivulja za varijablu trajanje hospitalizacije (ispis iz SAS-a)



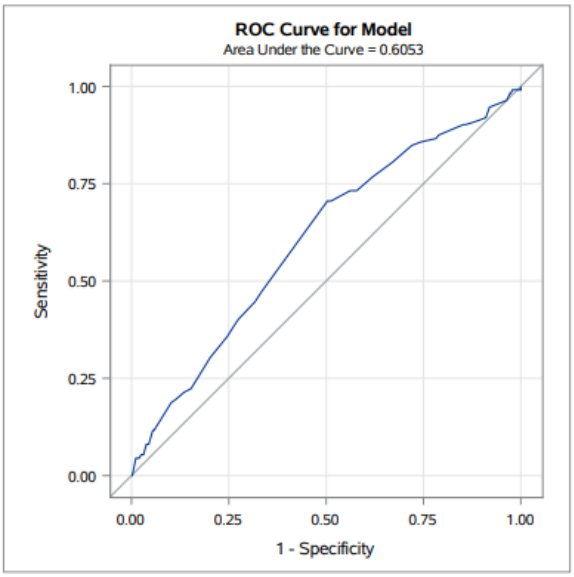
Slika 2.16: ROC krivulja za varijablu McCabe score (ispis iz SAS-a)



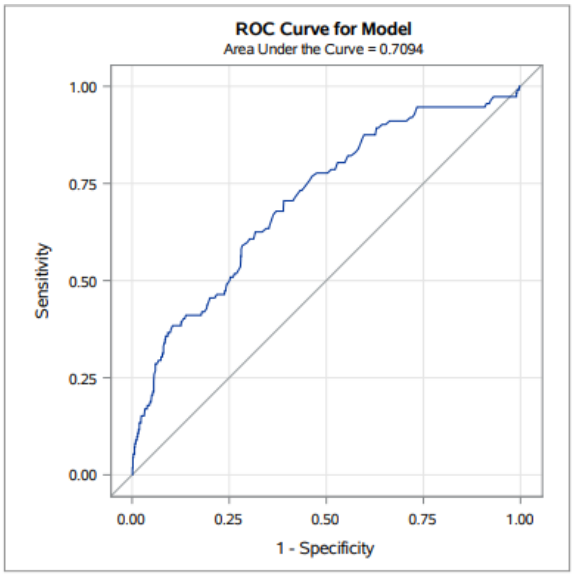
Slika 2.17: ROC krivulja za varijablu klinička težina bolesti (ispis iz SAS-a)



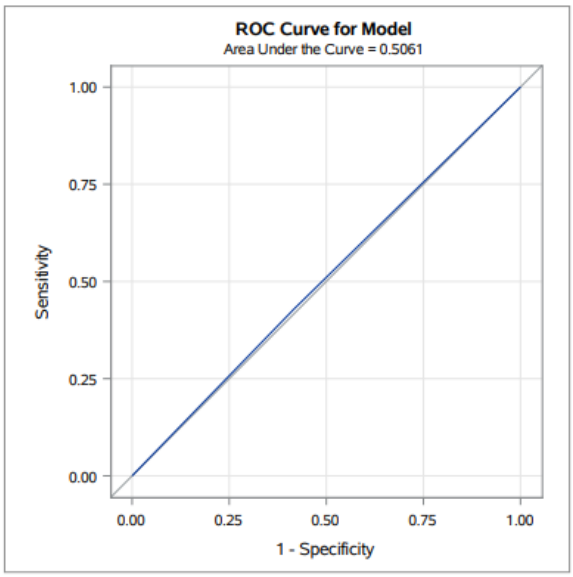
Slika 2.18: ROC krivulja za varijablu ataka bolesti (ispis iz SAS-a)



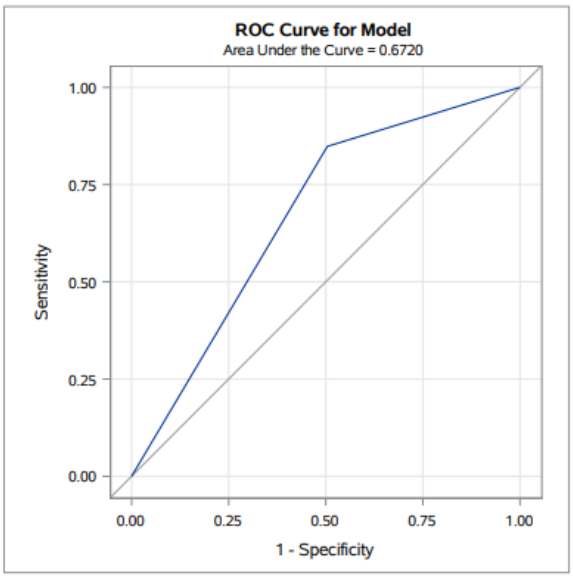
Slika 2.19: ROC krivulja za varijablu tjelesna temperatura (ispis iz SAS-a)



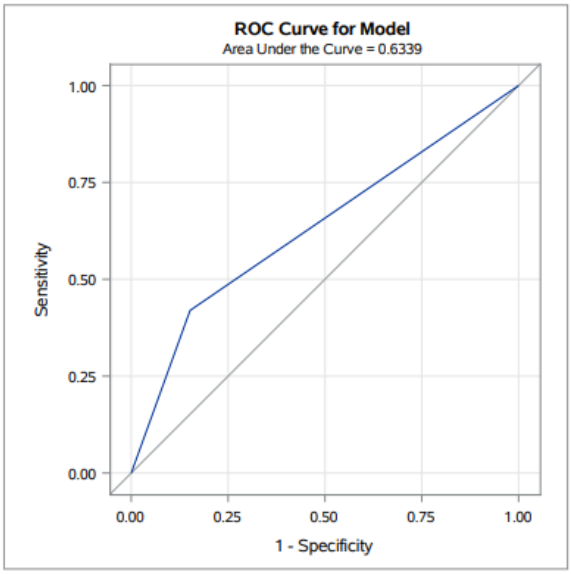
Slika 2.20: ROC krivulja za varijablu broj leukocita u krvi (ispis iz SAS-a)



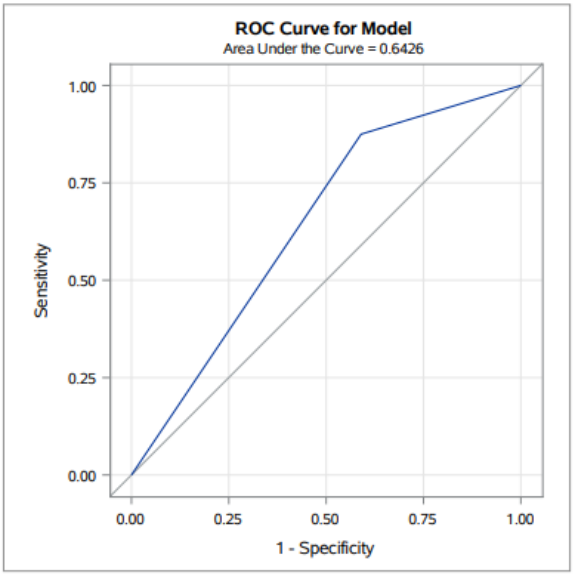
Slika 2.21: ROC krivulja za varijablu spol (ispis iz SAS-a)



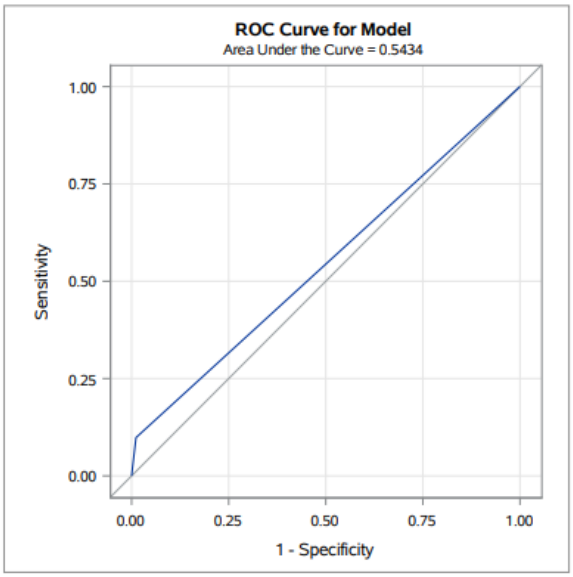
Slika 2.22: ROC krivulja za varijablu pokretnost (ispis iz SAS-a)



Slika 2.23: ROC krivulja za varijablu koncentracija kreatinina u krvi (ispis iz SAS-a)



Slika 2.24: ROC krivulja za varijablu konkomitantne infekcije (ispis iz SAS-a)



Slika 2.25: ROC krivulja za varijablu liječenje u jedinici intenzivne medicine (ispis iz SAS-a)

Tablica 2.3: Rezultati ML procjene parametara za univarijatne logističke modele

varijabla	procjena intercepta	procjena varijable
dob	-5.1245	0.0395
trajanje	-2.0423	-0.00743
McCabe score	-4.1152	1.2069
klinička tež	-10.2395	2.6563
ataka	-1.6053	-0.3888
temperatura	-16.6325	0.3813
leukociti	-3.4239	0.0729
spol	-2.1277	-0.0502
pokretnost	-1.6364	-1.7041
kreatinin	-2.5361	1.3959
konkom inf	-3.3449	1.5825
JIM	-2.2487	2.2487

Iz tablice 2.3 možemo iščitati kako glase jednadžbe ovih modela:

- Za dob: logit(p) = -5.1245 + 0.0395 × dob
- Za trajanje: logit(p) = -2.0423 - 0.00743 × trajanje
- Za McCabe score: logit(p) = -4.1152 + 1.2069 × McCabe score
- Za klinička težina: logit(p) = -10.2395 + 2.656 × klinička težina
- Za ataka: logit(p) = -1.6053 - 0.3888 × ataka
- Za temperatura: logit(p) = -16.6325 + 0.38135 × temperatura
- Za leukociti: logit(p) = -3.4239 + 0.0729 × leukociti
- Za spol: logit(p) = -2.1277 - 0.0502 × spol
- Za pokretnost: logit(p) = -1.6364 - 1.704 × pokretnost
- Za kreatinin: logit(p) = -2.5361 + 1.3959 × kreatinin
- Za konkomitantne infekcije: logit(p) = -3.3449 + 1.5825 × konkomitantne infekcije
- Za JIM: logit(p) = -2.2487 + 2.2487 × JIM

Iz ovih jednadžbi lako dođemo do vjerojatnosti smrtnog ishoda osobe s određenim karakteristikama. Odredimo sada vjerojatnost smrtnog ishoda za pacijenta koji ima 75 godina.

$$p(75) = \frac{e^{-5.1245+0.0395x75}}{1 + e^{-5.1245+0.0395x75}} = 0.10322$$

2.3 Multivarijatna logistička regresija

U ovom poglavlju ćemo provesti multivarijatnu logističku regresiju za sve nezavisne varijable iz baze koje su se univarijatnom analizom pokazale kao značajne. Nezavisne varijable koje ulaze u ovaj model su: dob, McCabe score, klinička težina bolesti, ataka, temperatura, leukociti, pokretnost, kreatinin, konkomitantne infekcije, te JIM (boravak u jedinici intenzivne medicine zbog CDI).

Tablica 2.4: Rezultati analize multivarijatnog logističkog modela

df	-2logL(intercept only)	-2logL(intercept and covariates)	likelihood ratio (χ^2)	konvergencija	c	p-vrijednost
10	719.594	397.289	322.3052	zadovoljena	0.915	< .0001

Iz tablice 2.4 vidimo da je multivarijatni logistički model značajan na razini značajnosti od 5%.

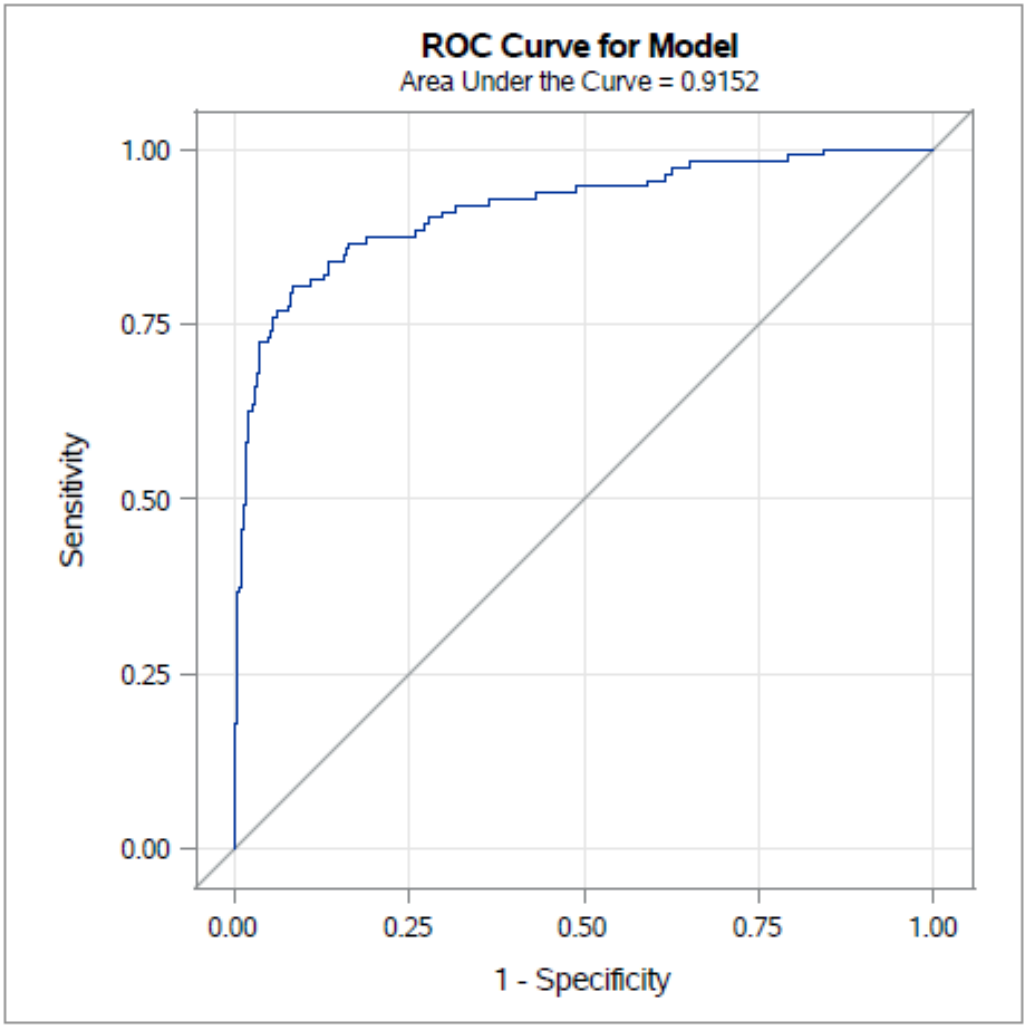
Tablica 2.5: Rezultati ML procjene parametara za multivarijatni logistički model

varijabla	procjena parametra	p-vrijednost
intercept	-10.6332	0.0858
dob	0.00910	0.4224
McCabe score	0.9065	<.0001
klinička tež	2.2669	<.0001
ataka	-0.3010	0.1410
temperatura	-0.0216	0.8916
leukociti	0.0138	0.2589
pokretnost	-0.6260	0.0667
kreatinin	0.3127	0.2574
konkom inf	0.6421	0.0826
JIM	0.3662	0.5185

Jednadžba ovog modela glasi:

logit(p) = -10.6332 + 0.00910 × dob + 0.9065 × McCabe score + 2.2669 × klinička težina - 0.3010 × ataka - 0.0216 × temperatura + 0.0138 × leukociti - 0.6260 × pokretnost + 0.3127 × kreatinin + 0.6421 × konkomitantne infekcije + 0.3662 × JIM

Iz tablice 2.5 vidimo koje su varijable statistički značajne na razini značajnosti od 5% u ovom modelu. To su varijable: McCabe score i klinička težina bolesti.
Iz tablice 2.4 i vrijednosti c vidimo da je prediktivna snaga smrtnog ishoda ovim modelom 91,5%, što je po uobičajenim kriterijima izuzetno jak model.



Slika 2.26: ROC krivulja za multivarijatni logistički model (ispis iz SAS-a)

Tablica 2.6: Procjene omjera izgleda za multivarijatni logistički model

varijabla	procjena OR	95% pouzdani interval
dob	1.009	0.987 - 1.032
McCabe score	2.476	1.620 - 3.783
klinička tež	9.649	6.197 - 15.025
ataka	0.740	0.496 - 1.105
temperatura	0.979	0.717 - 1.335
leukociti	1.014	0.990 - 1.039
pokretnost	0.535	0.274 - 1.044
kreatinin	1.367	0.796 - 2.349
konkom inf	1.900	0.920 - 3.923
JIM	1.442	0.475 - 4.383

Iz tablice 2.6 vidimo da intervali pouzdanosti za varijable McCabe score i klinička težina bolesti ne sadrže jedinicu, pa zaključujemo da su te varijable statistički značajne. Interpretacija omjera izgleda za statistički značajne varijable:

- Povećanje McCabe score-a za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 2.476 puta
- Povećanje kliničke težine za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 9.649 puta

2.4 Stepwise procedura

SAS procedura koja se koristi u ovom dijelu je stepwise procedura koja je kombinacija selekcije unaprijed (eng. forward) i eliminacije unatrag (eng. backward). Ona počinje kao selekcija unaprijed, međutim, varijabla koja se nađe u modelu ne mora i ostati u modelu.

Tablica 2.7: Rezultati stepwise analize multivarijatnog logističkog modela

varijabla	df	-2logL(intercept only)	-2logL(intercept and covariates)	likelihood ratio (χ^2)	konvergencija	p-vrijednost
klin tež	1	719.594	440.402	279.1925	zadovoljena	< .0001
McCabe	2	719.594	414.899	304.6952	zadovoljena	< .0001
pokretnost	3	719.594	408.443	311.1512	zadovoljena	< .0001
konkom inf	4	719.594	404.174	315.4206	zadovoljena	< .0001

Tablica 2.8: Rezultati ML procjene parametara za stepwise multivarijatni logistički model

varijabla	procjena parametra	p-vrijednost
intercept	-11.2579	< .0001
klinička tež	2.3906	< .0001
McCabe score	0.9218	< .0001
pokretnost	-0.6764	0.0407
konkom inf	0.7110	0.0460

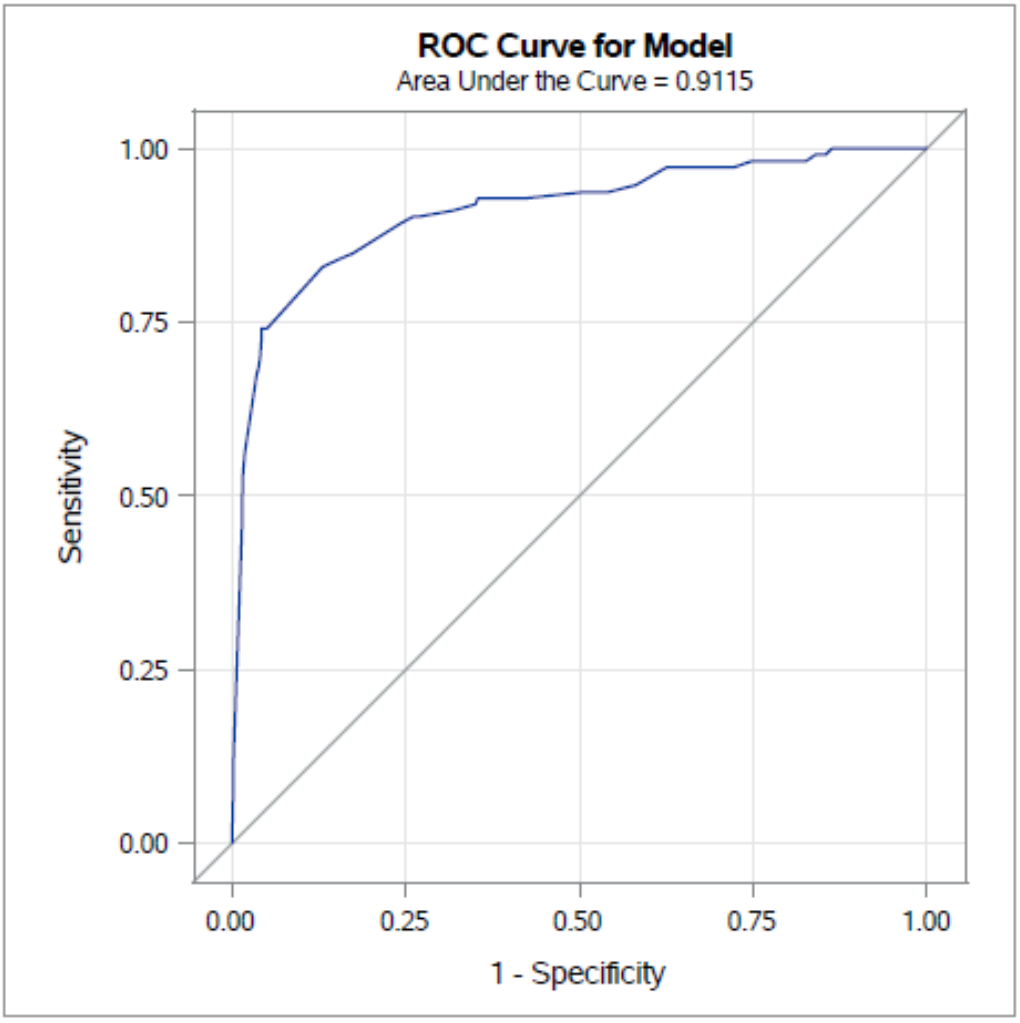
Jednadžba našeg modela izgleda ovako:

logit(p) = -11.2579 + 2.3906 $\times$ klinička težina + 0.9218 $\times$ McCabe score - 0.6764 $\times$ pokretnost + 0.7110 $\times$ konkomitantne infekcije

Tablica 2.9: Procjene omjera izgleda za stepwise multivarijatni logistički model

varijabla	procjena OR	95% pouzdani interval
klinička tež	10.920	7.163 - 16.648
McCabe score	2.514	1.664 - 3.799
pokretnost	0.508	0.266 - 0.972
konkom inf	2.036	1.013 - 4.093

- Povećanje McCabe score-a za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 2.514 puta
- Povećanje kliničke težine za 1 povećava omjer izgleda za prelazak nesmrtnog ishoda u smrtni ishod za 10.920 puta
- Pokretne osobe s obzirom na nepokretne imaju manji omjer izgleda za smrtni ishod (OR=0.508)
- Osobe s konkomitantnim infekcijama imaju veći omjer izgleda za smrtni ishod u odnosu na osobe bez konkomitantnih infekcija (OR=2.036)



Slika 2.27: ROC krivulja za stepwise logistički model (ispis iz SAS-a)

2.5 Usporedba modela

Usporedba univarijatnih modela, punog modela te modela dobivenog stepwise procedurom

Pri korištenju regresijskih modela, dostupna su dva temeljno različita pristupa kako istražiti učinak prediktorskih varijabli na varijablu odgovora: sve prediktorske varijable se mogu unijeti istovremeno u model, ili se mogu unositi postupno. Prva metoda (puni model) provodi se tako da se sve prediktorske varijable unose u isto vrijeme u model. Njihov

zajednički doprinos u objašnjavanju zavisne varijable utvrđuje se i sažima u jednom testu značajnosti punog modela. Druga metoda se provodi tako da se prediktorske varijable dodaju postupno (eng. stepwise) i / ili uklanjaju iz modela. Kada se varijable redom unose u model (eng. forward selection), početni model obuhvaća samo intercept i u svakom sljedećem koraku dodaje se varijabla koja dovodi do najvećeg (i značajnog) poboljšanja prikladnosti statističkog modela. U brisanju unatrag (eng. backward deletion), početni model je puni model koji uključuje sve varijable, a u svakom koraku isključena je varijabla koja dovodi do najmanjeg (ne značajnog) smanjenja prikladnosti modela. Moguć je i pristup koji započinje selekcijom prema naprijed, ali nakon uključivanja svake varijable testira na svakom koraku može li se uključena varijabla izostaviti iz modela bez značajnog smanjenja prikladnosti modela. Model koji dobijemo svakim od ovih postupaka trebao bi sadržavati podskup prediktorskih varijabli koje utječu i najbolje objašnjavaju zavisnu varijablu.[6]
Primjena stepwise procedura kritizirana je na više osnova (za pregled, vidi Wittingham et al. 2006). Stepwise metode često ne uspijevaju uključiti sve varijable koje imaju stvarni utjecaj na zavisnu varijablu, dok često uključuju i druge varijable koje ne utječu na zavisnu varijablu (Derksen i Keselman 1992). Posljedično, krajnji model općenito ne mora biti najbolji model (Miller 1984). Osim toga, stepwise procedure su obično nestabilne, što znači da samo neznatne promjene podataka mogu dovesti do velikih razlika u kranjim modelima, te je zato važna klinička pozadina istraživanja. [6]

Gledamo li rezultate univarijatnih logističkih modela, varijable dob, McCabe score, klinička težina, ataka, temperatura, leukociti, pokretnost, kreatinin, konkomitantne infekcije te liječenje u JIM-u tbog CDI su statistički značajne na razini značajnosti od 5%. Jedino varijable spol i trajanje nisu statistički značajne. Od svih statistički značajnih varijabli, najbolju predikcijsku snagu ima varijabla klinička težina bolesti, a najmanju liječenje u JIM-u zbog CDI.
U punom modelu su statistički značajne varijable bile samo McCabe score i klinička težina, što može ukazivati na kolinearnost nezavisnih varijabli. Prediktivna snaga ovog modela je 91.52%.
Model dobiven stepwise procedurom se malo razlikuje od punog modela. U njemu su na razini značajnosti od 5%, statistički značajne varijable: klinička težina, McCabe score, pokretnost i konkomitantne infekcije. Prediktivna snaga ovog modela je 91.15%.

Poglavlje 3

Dodatak

3.1 Kod u SAS-u

Zbog povjerljivosti podataka ne prilaže se baza podataka.

```
proc univariate data=cdiff;
var dob trajanje_hosp temp leuk;
run;
proc freq data=cdiff;
table god spol pokretnost kreat konkomit_inf JIM ishod
McCabe klinicka_tezina ataka/ chisq;run;

title "Univarijatna logisticka regresija-dob";
proc logistic data=cdiff descending;
model ishod=dob /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-trajanje hospitalizacije";
proc logistic data=cdiff descending;
model ishod=trajanje_hosp /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-McCabe";
proc logistic data=cdiff descending;
model ishod=McCabe /lackfit rsq outroc=rocgraf;
run;
```

```
title "Univarijatna logisticka regresija-klinicka tezina";
proc logistic data=cdiff descending;
model ishod=klinicka_tezina /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-ataka";
proc logistic data=cdiff descending;
model ishod=ataka /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-temperatura";
proc logistic data=cdiff descending;
model ishod=temp /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-leukociti";
proc logistic data=cdiff descending;
model ishod=leuk /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-spol";
proc logistic data=cdiff descending;
model ishod=spol/lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-pokretnost";
proc logistic data=cdiff descending;
model ishod=pokretnost/lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-kreatinin";
proc logistic data=cdiff descending;
model ishod=kreat/lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logisticka regresija-konkomitantne infekcije";
proc logistic data=cdiff descending;
model ishod=konkomit_inf/lackfit rsq outroc=rocgraf;
run;
```

```
title "Univarijatna logisticka regresija-JIM";
proc logistic data=cdiff descending;
model ishod=JIM/lackfit rsq outroc=rocgraf;
run;

title "Multivarijatna logisticka regresija";
proc logistic data=cdiff descending;
model ishod=dob McCabe klinicka_tezina ataka temp leuk
pokretnost kreat konkomit_inf JIM /lackfit rsq
outroc=rocgraf;
run;

title "Multivarijatna logisticka regresija-stepwise";
proc logistic data=cdiff descending;
model ishod=dob McCabe klinicka_tezina ataka temp leuk pokretnost
kreat konkomit_inf JIM / selection=stepwise;
run;

title "Stepwise ROC krivulja";
proc logistic data=cdiff descending;
model ishod=McCabe klinicka_tezina pokretnost
konkomit_inf/lackfit rsq outroc=rocgraf;
run;
```

Bibliografija

[1] D. W. Hosmer, S. Lemeshow, *Applied logistic regression (2nd ed.)*. New York, NY: John Wiley and Sons, 2000.

[2] M. H. Katz, *Multivariable Analysis: A Practical Guide for Clinicians*, Cambridge, 2006.

[3] N. Bogdanić, *Epidemiološka i klinička obilježja bolesnika hospitaliziranih zbog dijareje uzrokovane bakterijom Clostridium difficile*, diplomski rad, 2017.

[4] A. Jazbec, *Odabrane statističke metode u biomedicini*, PMF-MO, nastavni materijali, 2017.

[5] V. Wagner, *Statistički praktikum 2*, PMF-MO, nastavni materijali, 2017.

[6] Mundry, Roger and Charles Nunn. 2009. *Stepwise model fitting and statistical inference: turning noise into signal pollution*. American Naturalist 173(1): 119-123.

Sažetak

U ovom radu analizirali smo ishode liječenja pacijenata s Clostridium difficile infekcijom. Korištena je metoda logističke regresije na bazi podataka koja se sastoji od 1080 observacija i 13 varijabli. Istraživanje je odobreno od strane Etičkog povjerenstva Klinike za infektivne bolesti „Dr. Fran Mihaljević“. Sve potrebne varijable prikupili su Nikolina Bogdanić, dr.med., Karlo Vidović, dr.med. te doc. dr. sc. Mirjana Balen Topić, specijalist infektolog, iz originalne arhivirane medicinske dokumentacije. Za obradu podataka je korišten statistički program SAS. Iz provedene analize smo zaključili da su varijable koje imaju značajan utjecaj na krajnji ishod liječenja na razini značajnosti od 5% McCabe score, klinička težina bolesti, pokretnost i konkomitantne infekcije.

Summary

In this paper we have analyzed the outcomes of treatment of patients with Clostridium difficile infection. A logistic regression method was used on a database consisting of 1080 observations and 13 variables. The study was approved by the Ethics Committee of the Clinic for Infectious Diseases "Dr. Fran Mihaljević ". All necessary variables were collected by Nikolina Bogdanić, MD, Karlo Vidović, MD. and doc. dr. sc. Mirjana Balen Topić, an infectious deseases specialist, from the original archived medical documentation. The statistical program SAS was used for data processing. Analysis showed that the variables that significantly affect the ultimate outcome of treatment at the significance level of 5% are McCabe score, clinical severity of the disease, patient mobility and concurrent infections.

Životopis

Rodena sam 4. veljače 1994. godine u Splitu. Nakon završene prirodoslovno-matematičke gimnazije u Splitu, 2012. godine upisala sam Preddiplomski sveučilišni studij Matematika, smjer nastavnički na Prirodoslovno-matematičkom fakultetu Sveučilišta u Zagrebu. Nakon završenog preddiplomskog studija, 2016. godine upisala sam Diplomski sveučilišni studij Matematička statistika na istom fakultetu.