Analiza utjecaja promatranih varijabli na bolest srca logističkom regresijom

Barbarić, Andrea

Analiza utjecaja promatranih varijabli na bolest srca logističkom regresijom

Barbarić, Andrea

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Andrea Barbarić

ANALIZA UTJECAJA PROMATRANIH VARIJABLI NA BOLEST SRCA LOGISTIČKOM REGRESIJOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, ožujak 2021.

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. _____

Hvala mentorici, prof.dr.sc. Anamariji Jazbec na pomoći pri izradi ovog rada.
Hvala mojim sekama, mom Lovri i prijateljima koji su vjerovali u mene.
Posebno hvala mami i tati na neizmjernoj podršci i ljubavi tijekom svih godina studiranja.

Sadržaj

Uvod

U različitim područjima znanosti i industrije, poput ekonomije, zdravstva, biomedicine, kriminalistike, ekologije, inženjerstva i sl., često želimo utvrditi ovisi li neka odabrana veličina o drugim mjerenim veličinama. Ta je veza vrlo rijetko jasno definirana, stoga se pri njenom određivanju služimo raznim vjerojatnosnim i statističkim modelima. Metode regresijske analize postale su vrlo važna komponenta pri obradi i analizi podataka u svrhu određivanja te ovisnosti. Pri tome veličinu koja nas zanima modeliramo kao slučajnu varijablu te je nazivamo zavisna varijabla ili varijabla odziva, dok ostale mjerene veličine zovemo nezavisne varijable ili kovarijate.

U ovome radu, koristeći bazu koja sadrži podatke o pacijentima, točnije o njihovoj dobi, spolu, boli u prsima, tlaku, kolesterolu, šećeru, nalazu EKG-a itd., pomoću logističke regresije analizirat ćemo i procijeniti koje od prikupljenih varijabli i koliko statistički značajno utječu na dihotomnu varijablu bolest srca. Također, pomoću analiziranih varijabli želimo naći najbolji model za procjenu vjerojatnosti dobivanja bolesti srca.

U prvom poglavlju upoznat ćemo se s glavnim pojmovima vezanima za logistički regresijski model te uočiti razlike u odnosu na linearnu regresiju, dok ćemo u drugom poglavlju primijeniti opisanu logističku regresiju na konkretnom primjeru. Pri analizi ćemo koristiti statistički program SAS.

Poglavlje 1

Logistička regresija

1.1 Regresijska analiza

Regresijska analiza obuhvaća metode ispitivanja ovisnosti jedne (zavisne) varijable o jednoj ili više drugih (nezavisnih) varijabli. Pritom je glavni cilj objasniti koje ulazne (nezavisne) varijable, i u kojoj mjeri, utječu na izlaznu (zavisnu) varijablu, koliko pouzdano to možemo zaključiti te koliko je, u konačnici, naš model uspješan u objašnjavanju ponašanja izlazne varijable kao funkcije ulaznih varijabli. Također, zanima nas njegova uspješnost u predviđanju izlaznih vrijednosti za određene zadane ulazne vrijednosti. Rezultat regresijske analize je regresijski model, tj. matematička jednadžba koja objašnjava tu povezanost. Razlikujemo više vrsta regresije. Ako model ima samo jednu ulaznu vrijablu, govorimo o univarijatnoj regresiji. Pritom, ako je povezanost linearna, radi se o linearnoj regresiji i tada je njen rezultat jednadžba pravca, dok za nelinearnu univarijatnu regresiju dobivamo jednadžbu krivulje. S druge strane, ako model ima više ulaznih varijabli, govorimo o multivarijatnoj regresiji za koju opet razlikujemo linearnu (rezultat je jednadžba ravnine) i nelinearnu ovisnost (rezultat je jednadžba plohe). [5]

Općenito, jednadžba modela glasi:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon \,, \tag{1.1}$$

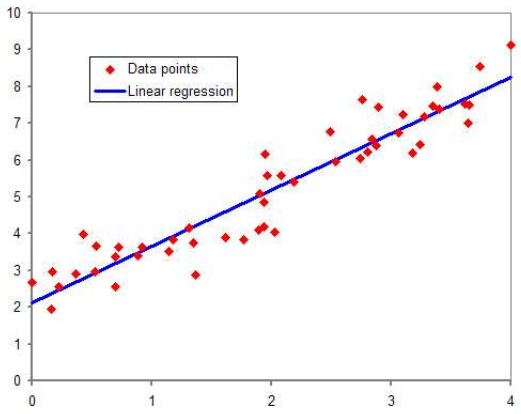
gdje su:

- Y – zavisna varijabla
- x_i – nezavisne varijable, $i = 1, \ldots, n,\ n \in \mathbb{N}$
- ε – slučajna greška
- β_i – parametri modela, $i = 1, \ldots, n,\ n \in \mathbb{N}$.

1.2 Usporedba linearne i logističke regresije

Vrste regresije razlikujemo i prema tipu zavisne varijable.

Linearna regresija vrsta je regresijske analize u kojoj je povezanost između zavisne i nezavisne varijable opisana jednadžbom pravca. Taj pravac određujemo metodom najmanjih kvadrata, tako da, iz skupa svih pravaca, odaberemo onaj čija je suma odstupanja svake točke od pravca najmanja. [4] [5]
Problem kod tog modela nastaje ako zavisna varijabla nije kontinuirana, već kategorijska. Tada se narušavaju pretpostavke o normalnosti i homogenosti slučajne greške, potrebne za korištenje linearnog regresijskog modela, stoga naša predviđanja mogu biti pogrešna. [7] Iz tih razloga razvila se logistička regresija.



Slika 1.1: Primjer grafa linearne regresije
izvor: `https://hr.wikipedia.org/wiki/Linearna_regresija` (2020. godina)

Logistička regresija vrsta je regresijske analize u kojoj je zavisna varijabla kategorijska. Kod nje nema pretpostavki o distribuciji nezavisnih varijabli. U njenoj najraširenijoj primjeni, zavisna varijabla je jednostavna dihotomna, tj. može poprimiti samo dvije vrijednosti (kategorije). Za razliku od linearnog modela, ovdje za procjenu parametara koristimo metodu maksimalne vjerodostojnosti. Ostale procedure (poput korelacije varijabli, stepwise procedure, procjene statističke značajnosti varijabli) odvijaju se kao i kod obične regresije. Više o tome svemu govorit ćemo u idućim poglavljima.

1.3 Logistički regresijski model

Logistička funkcija, prepoznatljiva po svojoj S-krivulji, dana je s $p : (-\infty, \infty) \to (0, 1)$

$$p(x) = \frac{1}{1 + e^{-x}} . \tag{1.2}$$

Njoj inverzna funkcija je funkcija logit, dana s $logit : (0, 1) \to (-\infty, \infty)$

$$logit(p(x)) = log\left[\frac{p(x)}{1 - p(x)}\right] = log(p(x)) - log(1 - p(x)) . \tag{1.3}$$

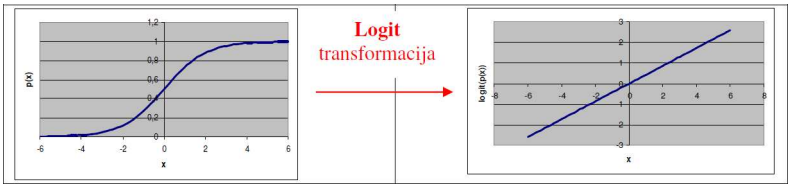
Logistički model procjenjuje vjerojatnost da će neka jedinica opažanja ući u jednu dihotomnu skupinu umjesto u drugu. Glavni cilj je linearizirati takav model jer je linearnu funkciju puno lakše interpretirati. No, znamo da je vjerojatnost funkcija ograničena u intervalu (0, 1), dok je linearna funkcija neograničena. Iz tog razloga, provodimo transformaciju u dva koraka.
Ako je $p(x)$ vjerojatnost nekog događaja, onda **izgled** ili **šansa** (eng. *odds*) tog događaja predstavlja omjer očekivanog broja puta kada će se on dogoditi i očekivanog broja puta kada se događaj neće dogoditi.

Prvi korak je transformirati vjerojatnost u izgled. Dakle,

$$odds(x) = \frac{p(x)}{1 - p(x)}. \tag{1.4}$$

Time smo maknuli gornju granicu intervala.

Drugi korak je izračunati vrijednost prirodnog logaritma dobivenog izgleda, čime mičemo donju granicu intervala.
Navedenu transformaciju vidimo na slici 1.2 [1] i nazivamo je *logit transformacija* logističke funkcije. Drugim riječima, logističkom regresijom modeliramo logit transformiranu vjerojatnost koja je u linearnoj vezi s kovarijatama.



Slika 1.2: Grafički prikaz logit transformacije logističke funkcije

[1] preuzeta iz [5]

Tablica 1.1: Odnos između vjerojatnosti, izgleda i log vrijednosti izgleda

Vjerojatnost	**Izgled**	**Log(izgled)**
0.00	0.00	/
0.10	0.11	-2.21
0.20	0.25	-1.39
0.30	0.43	-0.84
0.40	0.67	-0.40
0.50	1.00	0.00
0.60	1.50	0.41
0.70	2.33	0.85
0.80	4.00	1.39
0.90	9.00	2.20
1.00	/	/

Iz gornje tablice vidimo da logaritamska vrijednost izgleda poprima i pozitivne i negativne neograničene vrijednosti, tj. da smo navedenom transformacijom makli i gornju i donju granicu. [1] Također, ona je definirana za sve vjerojatnosti između 0 i 1, ali ne i za vrijednosti jednake 0 ili 1.

Dakle, univarijatni logistički regresijski model je oblika:

$$logit(p(x)) = log(odds(x)) = log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x. \tag{1.5}$$

Izrazimo $p(x)$ iz jednakosti (1.5):

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}. \tag{1.6}$$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}. \tag{1.7}$$

Slično, multivarijatni logistički regresijski model (za $x = (x_1, \ldots, x_k)$) je oblika :

$$logit(p(x)) = log(odds(x)) = log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{1.8}$$

Slijedi:

$$p(x_1, \ldots, x_k) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}. \tag{1.9}$$

1.4 Procjena parametara modela

Za razliku od linearne regresije u kojoj koristimo metodu najmanjih kvadrata, kod logističkog modela za procjenu nepoznatih parametara koristimo **metodu maksimalne vjerodostojnosti** (oznaka ML). Kod nje tražimo najmanje moguće odstupanje između opaženih (y) i prediktivnih ($\hat{y}$) vrijednosti koristeći iterativne računalne metode. Jednom kada nađemo najbolje rješenje, to odstupanje nazivamo *Deviance* ili *-2LogLikelihood*. [5]
Opišimo ukratko matematičku pozadinu te metode ([4]).

Neka je $\boldsymbol{X} = (X_1, \ldots, X_n)$ slučajni uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$. Na osnovu zadanog uzorka želimo procijeniti vrijednost parametra θ ili neke njegove funkcije.
Ako je $\boldsymbol{x} = (x_1, \ldots, x_n)$ realizacija tog slučajnog uzorka, tada je **vjerodostojnost** funkcija definirana s:

$$L : \Theta \to \mathbb{R}, \qquad L(\theta|x) = L(\theta) := f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{1.10}$$

Definicija 1.4.1. *Statistika $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$ je **procjenitelj maksimalne vjerodostojnosti** za parametar θ (**MLE**[2]) ako vrijedi*

$$L(\hat{\theta}|\boldsymbol{X}) = \max_{\theta\in\Theta} L(\theta|\boldsymbol{X}). \tag{1.11}$$

Umjesto maksimiziranja produkta (1.10), često se, radi jednostavnosti, maksimizira funkcija log-vjerodostojnost jer je logaritam rastuća funkcija.

Prikažimo navedeni postupak na našem logističkom modelu sa zavisnom dihotomnom varijablom. Pretpostavimo da imamo uzorak od n nezavisnih opažanja parova $(x_i, y_i), i = 1, \ldots, n$, pri čemu su X_i nezavisne varijable, a Y_i zavisne s vrijednostima u skupu $\{0, 1\}$ (0 predstavlja odsutnost, a 1 prisutnost promatranog obilježja). Neka je $\hat{y}$ ML procjena od y, a $\hat{p}$ ML procjena od $p(x) = \mathbb{P}(y = 1|x)$. Želimo procijeniti nepoznati parametar $\boldsymbol{\beta} = (\beta_0, \beta_1)$ (radi jednostavnosti prikaza gledamo univarijatni logistički model, analogni postupak vrijedi i za multivarijatni model). Funkcija vjerodostojnosti za taj model dana je s:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i} \right]. \tag{1.12}$$

Umjesto maksimiziranja $L(\boldsymbol{\beta})$, maksimizirat ćemo funkciju log-vjerodostojnosti $l(\boldsymbol{\beta})$

$$l(\boldsymbol{\beta}) = ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} \left[y_i ln(p(x_i)) + (1 - y_i) ln(1 - p(x_i)) \right]. \tag{1.13}$$

[2] eng. Maximum Likelihood Estimator

Da bi se pronašli procjenitelji maksimalne vjerodostojnosti, tj. vrijednosti za parametar $\boldsymbol{\beta}$ koji će maksimizirati funkciju $l(\boldsymbol{\beta})$, potrebno je parcijalno derivirati funkciju $l(\boldsymbol{\beta})$ po β_0 i β_1 te izjednačiti s nulom. Dobivamo jednadžbe:

$$\sum_{i=1}^{n} [y_i - p(x_i)] = 0 \tag{1.14}$$

$$\sum_{i=1}^{n} x_i [y_i - p(x_i)] = 0. \tag{1.15}$$

Jednadžbe se rješavaju iterativnim metodama te dobiveno rješenje nazivamo ML procjenitelj parametra $\boldsymbol{\beta}$, u oznaci $\hat{\boldsymbol{\beta}}$. [5]

1.5 Testiranje adekvatnosti modela (eng. *Goodness of fit*)

Nakon procjene parametara, želimo odabrati najpogodniji model, tj. onaj koji najbolje opisuje dane podatke. Osnovna mjera adekvatnosti modela je tzv. **devijanca** (oznaka D, eng. *Deviance*) iz prethodnog potpoglavlja. Ona je istovjetna sumi kvadrata reziduala kod linearne regresije. D računamo kao omjer vjerodostojnosti:

$$\begin{aligned} D &= -2ln\left[\frac{likelihood\ procijenjenog\ modela}{likelihood\ saturiranog\ modela}\right] \\ &= -2\sum_{i=1}^{n}\left[y_i ln(\frac{\hat{p}(x_i)}{y_i}) + (1-y_i)ln(\frac{1-\hat{p}(x_i)}{1-y_i})\right] \approx \chi^2 \end{aligned} \tag{1.16}$$

Saturirani model je onaj koji sadrži onoliko parametara koliko ima podataka.
Za svaku varijablu želimo testirati je li model bolji kada je ona uključena ili isključena iz njega. Pri tome koristimo G statistiku. Ukoliko dodamo varijablu koja pospješuje naš model, očekujemo da se D smanji, i to statistički značajno.

$$\begin{aligned} G &= D(model\ bez\ varijabli) - D(model\ s\ k\ varijabli) \\ &= -2l(0) - (-2l(k)) \\ &= -2ln\left[\frac{L(0)}{L(k)}\right] \approx \chi^2(k) \end{aligned} \tag{1.17}$$

Testiranje značajnosti parametara

Provođenjem statističkog testa utvrđujemo je li koeficijent uz pojedinu nezavisnu varijablu jednak ili različit od nule.

Kod univarijatnog logističkog modela testiramo:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{1.18}$$

Kod multivarijatnog logističkog modela s k nezavisnih varijabli testiramo:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_k = 0 \\ H_1 &: \textit{barem jedan od } \beta_i \neq 0,\ i = 1, \ldots, k \end{aligned} \tag{1.19}$$

Pri tome se koristi Waldov test. Ukoliko je parametar jednak 0, onda pripadna nezavisna varijabla nema utjecaj u modelu. Kažemo da je parametar statistički značajan ako se statistički značajno razlikuje od nule te u tom slučaju pripadna nezavisna varijabla statistički značajno pospješuje model.

1.6 Interpretacija parametara

Za jednostavnije tumačenje, bazirat ćemo se na univarijatni model logističke regresije

$$logit(p(x)) = ln(odds(x)) = log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x. \tag{1.20}$$

Koeficijent β_0 (*intercept*) nužno postoji u svakom modelu, međutim nema značenje u tumačenju istoga. On predstavlja vrijednost $ln(odds)$ kada je prediktorska varijabla jednaka 0. Preko jednadžbe (1.4) uveli smo pojam izgleda ili šanse. Za potrebe interpretacije koeficijenata, treba nam još i pojam **omjera šansi** (**eng. *odds ratio***).
Promotrimo:

$$g(x) := logit(p(x)) = ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x \tag{1.21}$$

$$g(x+1) = \beta_0 + \beta_1(x+1) \tag{1.22}$$

$$g(x+1) - g(x) = \beta_1 \tag{1.23}$$

$$logit(p(x+1)) - logit(p(x)) = \beta_1 \tag{1.24}$$

$$ln(odds(p(x+1))) - ln(odds(p(x))) = \beta_1 \tag{1.25}$$

$$ln\left[\frac{odds(p(x+1))}{odds(p(x))}\right] = \beta_1 \tag{1.26}$$

- Ako je nezavisna varijabla dihotomna s vrijednostima 0 ili 1, tada:

$$\begin{aligned} x = 1 &\rightarrow odds = \frac{p(1)}{1-p(1)} \\ x = 0 &\rightarrow odds = \frac{p(0)}{1-p(0)} \end{aligned} \tag{1.27}$$

$$\begin{aligned} g(1) - g(0) &= ln\left[\frac{odds(p(1))}{odds(p(0))}\right] \\ &= ln\left[\frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}}\right] \\ &= ln(odds\ ratio) = \beta_1 \\ \Rightarrow odds\ ratio(1,0) &= e^{\beta_1} \end{aligned} \tag{1.28}$$

Interpretacija: Prelaskom nezavisne varijable iz niže u višu kategoriju, povećava se omjer šansi za prelazak zavisne varijable iz niže u višu kategoriju za iznos e^{β_1}.

- Ako je nezavisna varijabla kontinuirana, tada:

$$g(x+1) - g(x) = \beta_1 \tag{1.29}$$

Parametar β_1 pokazuje promjenu u logaritmiranom izgledu (oddsu) kada se nezavisna varijabla x pomakne za 1. Međutim, kod kontinuiranih varijabli nam pomak za jednu jedinicu često nije od velikog značaja u istraživanjima, stoga nas češće zanima:

$$\begin{aligned} g(x+c) - g(x) &= c\beta_1,\ c\ konst. \\ \Rightarrow odds\ ratio(c) &= odds\ ratio(x+c,x) = e^{c\beta_1} \end{aligned} \tag{1.30}$$

Interpretacija: Povećanjem nezavisne varijable za iznos $c \in \mathbb{R}$, povećava se omjer šansi za prelazak zavisne varijable iz niže u višu kategoriju za iznos $e^{c\beta_1}$.

Npr. pretpostavimo da zavisna dihotomna varijabla Y označuje prisutnost (vrijednost 1) ili odsutnost (vrijednost 0) bolesti srca, a nezavisna varijabla X označuje bavi li se osoba (vrijednost 1) ili ne (vrijednost 0) redovitom tjelesnom aktivnošću. Ako je procijenjen *odds ratio* jednak 0.5, onda je upola manja vjerojatnost da će se srčana bolest razviti među onima koji vježbaju nego među nevježbačima unutar proučavane populacije.
Upravo je ta jednostavna veza između koeficijenata modela i omjera šansi jedan od glavnih razloga zašto se logistička regresija pokazala vrlo snažnim analitičkim alatom u raznim istraživanjima. [3]

1.7 ROC krivulja

ROC krivulja (eng. *Receiver Operating Characteristic curve*) grafički je prikaz koji se često koristi u svrhu ispitivanja valjanosti dijagnostičkog testa. Valjanost dijagnostičkog testa ima dvije komponente: **osjetljivost** i **specifičnost**. Osjetljivost testa predstavlja udio bolesnih osoba koje je test pravilno prepoznao kao ”pozitivne” (bolesne) u ukupnom broju

bolesnih, a specifičnost testa predstavlja udio zdravih ispitanika koji su pravilno prepoznati kao "negativni" (zdravi) u ukupnom broju zdravih. ROC krivuljom želi se prikazati odnos proporcija lažno pozitivnih (1 − specifičnost) i stvarno pozitivnih (osjetljivost).
Jedna od standardnih mjera za ocjenu dobrote prilagodbe modela je *c-statistika* koja predstavlja površinu ispod ROC krivulje. Ona nam dakle pokazuje koliko je naš model dobar klasifikator.
Statistiku c računamo prema sljedećoj formuli:

$$c = \frac{nc + 0.5(t - nc - nd)}{t}\,, \tag{1.31}$$

gdje je nc broj *concordant* parova, nd broj *discordant* parova, a t broj parova s jednakim vrijednostima odgovora. Kažemo da je par opservacija s različitim odgovorima podudaran (eng. *concordant*) ako opservacija s više rangiranim odgovorom (npr. 2, "događaj se nije dogodio") ima nižu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom (npr. 1, "događaj se dogodio"), a nepodudaran (eng. *discordant*) ako opservacija s više rangiranim odgovorom ima višu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom. Ako par opservacija nije ni podudaran ni nepodudaran, kažemo da ima jednak odgovor (eng. *tie*). [5]
c poprima vrijednosti od 0 do 1. Ukoliko je ona manja ili jednaka 0.5, smatramo da model nema dobru prediktivnu vrijednost. Što je c bliže vrijednosti 1, to je snaga prediktivnog modela veća (bolja). [3]

Poglavlje 2

Primjena logističke regresije na primjeru bolesti srca

2.1 Opis problema

Za potrebu analize podataka primjenom logističke regresije, koristit će se baza dostupna na `https://www.kaggle.com/cherngs/heart-disease-cleveland-uci`. Baza se sastoji od 297 opservacija i 14 atributa. Kako je navedeno, originalnu bazu (sa 76 atributa) kreirali su M.D. Andras Janosi s mađarskog Instituta za kardiologiju, M.D. William Steinbrunn sa Sveučilišne bolnice u Zürichu, M.D. Matthias Pfisterer sa Sveučilišne bolnice u Baselu te M.D., Ph.D. Robert Detrano s Klinike Cleveland. Međutim, u svim objavljenim radovima i eksperimentima iz područja strojnog učenja korišten je podskup od 14 atributa koji se koriste i u ovom radu.
Opservacije predstavljaju pacijente, a atributi podatke o pacijentima. Točnije, za 297 pacijenata imamo informacije o njihovoj dobi, spolu, tipu boli u prsima, izmjerenom tlaku, šećeru u krvi i ostalo. "Ciljni" atribut je dihotomna varijabla koja govori ima li pacijent srčanu bolest ili nema. Upravo na temelju tih podataka želimo procijeniti koje varijable i koliko statistički značajno utječu na pojavu bolesti srca kod osobe.

Navedimo popis svih varijabli, kratice koje će biti korištene u radu te njihov kratak opis:

- **dob**
  Dob pacijenta (izražena u godinama).
- **spol**
  Spol pacijenta. To je kategorijska dihotomna varijabla: 0−žensko, 1−muško.
- **tip boli u prsima**
  Vrsta boli u prsima koju pacijent osjeća u mirovanju i opisuje doktoru. To je katego-

rijska nominalna varijabla: 0−tipična angina, 1−atipična angina, 2−neanginalna bol, 3−asimptomatski (za detaljniji opis vidi [2]). U daljnjem tekstu koristimo oznaku **bol**.

- **krvni tlak u mirovanju**
  Izmjeren krvni tlak u mirovanju na prijemu u bolnicu. Mjeri se u milimetrima žive (mmHg). Kod zdravih osoba optimalna vrijednost krvnog tlaka trebala bi biti ispod 120/80 mmHg. U daljnjem tekstu koristimo oznaku **tlak**.

- **kolesterol**
  Razina kolesterola u krvi (mjerena u mg/dl). Povišena razina ukupnog i LDL-kolesterola te smanjena razina HDL-kolesterola povećavaju rizik za razvoj bolesti srca i krvnih žila. Preporučena vrijednost ukupnog kolesterola je $< 5,0 mmol/l$, tj. $< 193 mg/dL$.

- **šećer u krvi**
  Razina šećera u krvi. To je kategorijska dihotomna varijabla: 0−nema povišen šećer, 1−ima povišen šećer (ako je izmjerena razina $> 120 mg/dl$). U daljnjem tekstu koristimo oznaku **šećer**.

- **nalaz EKG-a**
  Zabilježena električna aktivnost srca. To je kategorijska ordinalna varijabla: 0−uredan nalaz, 1−uočena abnormalnost ST i/ili T vala (pod tim se podrazumijeva inverzija T-vala i/ili elevacija ili depresija ST-vala $> 0,05 mV$, za detalje vidi [6]), 2−moguća hipertrofija lijeve klijetke. U daljnjem tekstu koristimo oznaku **EKG**.

- **maksimalni puls**
  Maksimalni broj otkucaja srca dok je pacijent u pokretu, izmjeren tokom scintigrafije srca. Protok krvi kroz srčani mišić obično se ispituje ubrizgavanjem talija-201 u venu i snimanjem tijekom testa opterećenjem. U daljnjem tekstu koristimo oznaku **puls**.

- **angina izazvana tjelesnom aktivnošću**
  Kategorijska dihotomna varijabla koja pokazuje je li pacijent osjetio bol u prsima prilikom neke tjelesne aktivnosti: 0−ne, 1−da. U daljnjem tekstu koristimo oznaku **ex_angina**.

- **ST-depresija**
  Duljina depresije (spuštenosti) ST-segmenta. ST-depresija se izaziva fizičkom aktivnosti i mjeri se u odnosu na pacijenta u stanju mirovanja. Kod zdrave osobe, ST-segment je horizontalan. To je numerička varijabla. U daljnjem tekstu koristimo oznaku **STdepresija**.

- **vršni nagib ST-segmenta u pokretu**
  Kategorijska nominalna varijabla: 0−uzdignut, 1−ravan, 2−spušten (kod zdrave osobe je obično ravan). U daljnjem tekstu koristimo oznaku **STnagib**.

- **broj "obojenih"(vidljivih) glavnih krvnih žila kod pretrage fluoroskopije**
  Diskretna varijabla koja poprima vrijednosti od 0 do 3. Prilikom fluoroskopije se mogu uočiti neke promjene u protoku krvnih žila, ugrušci i slično. U daljnjem tekstu koristimo oznaku **krvne_zile**.

- **talij test**
  Kategorijska nominalna varijabla s vrijednostima 0, 1 i 2. Talij (točnije, njegov izotop talij-201) se ubrizgava u krv u maloj količini te se snimaju srce i krvne žile pacijenta u pokretu. Idealno bi bilo da postoji jednaka raspodjela talija u svim segmentima miokarda (0). Ako postoji oštećenje koje je fiksno, onda se neće moći jednako raspodijeliti (1), a ako je reverzibilno, onda će do jednake raspodjele doći nakon nekog vremena (2).

- **stanje**
  Kategorijska dihotomna varijabla koja označava ima li osoba bolest srca (1) ili nema (0). Pri analizi ćemo nju koristiti kao zavisnu varijablu. U daljnjem tekstu koristimo oznaku **bolest**.

2.2 Deskriptivna statistika

Deskriptivna ili opisna statistika grana je matematičke statistike koja se bavi predočavanjem i opisivanjem glavnih karakteristika izmjerenog ili zadanog skupa podataka (tablice, grafikoni, srednje vrijednosti, ...).
Tablica 2.1 prikazuje deskriptivnu statistiku svih nezavisnih numeričkih varijabli. Pritom *N* označava ukupan broj opservacija, *Mean* je aritmetička sredina podataka, *Std Dev* standardna devijacija te su još iskazani *medijan* te *minimalni* i *maksimalni* element vrijednosti pojedine varijable.
U tablicama 2.2 do 2.8 prikazana je deskriptivna statistika (tj. tablice frekvencija) svih nezavisnih kategorijskih varijabli, a u tablici 2.9 deskriptivna statistika zavisne varijable bolest koja je također kategorijska, točnije dihotomna. Pritom stupac *Frequency* označava frekvenciju, tj. broj pojavljivanja određene kategorije, a stupac *Cumulative Frequency* kumulativnu frekvenciju. Obje vrijednosti izražene su i u postotcima.

Tablica 2.1: Deskriptivna statistika nezavisnih numeričkih varijabli (SAS ispis)

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Median	Maximum
dob	297	54.5421	9.0497	29.0000	56.0000	77.0000
tlak	297	131.7	17.7628	94.0000	130.0	200.0
kolesterol	297	247.4	51.9976	126.0	243.0	564.0
puls	297	149.6	22.9416	71.0000	153.0	202.0
STdepresija	297	1.0556	1.1661	0	0.8000	6.2000
krvne_zile	297	0.6768	0.9390	0	0	3.0000

Tablica 2.2: Deskriptivna statistika nezavisne kategorijske dihotomne varijable **spol** (SAS ispis)

The FREQ Procedure

spol	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	96	32.32	96	32.32
1	201	67.68	297	100.00

Tablica 2.3: Deskriptivna statistika nezavisne kategorijske nominalne varijable **bol** (SAS ispis)

bol	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	23	7.74	23	7.74
1	49	16.50	72	24.24
2	83	27.95	155	52.19
3	142	47.81	297	100.00

Tablica 2.4: Deskriptivna statistika nezavisne kategorijske dihotomne varijable **šećer** (SAS ispis)

secer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	254	85.52	254	85.52
1	43	14.48	297	100.00

Tablica 2.5: Deskriptivna statistika nezavisne kategorijske ordinalne varijable **EKG** (SAS ispis)

EKG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	147	49.49	147	49.49
1	4	1.35	151	50.84
2	146	49.16	297	100.00

Tablica 2.6: Deskriptivna statistika nezavisne kategorijske dihotomne varijable **ex_angina** (SAS ispis)

ex_angina	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	200	67.34	200	67.34
1	97	32.66	297	100.00

Tablica 2.7: Deskriptivna statistika nezavisne kategorijske nominalne varijable **STnagib** (SAS ispis)

STnagib	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	139	46.80	139	46.80
1	137	46.13	276	92.93
2	21	7.07	297	100.00

Tablica 2.8: Deskriptivna statistika nezavisne kategorijske nominalne varijable **talij_test** (SAS ispis)

talij_test	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	164	55.22	164	55.22
1	18	6.06	182	61.28
2	115	38.72	297	100.00

Tablica 2.9: Deskriptivna statistika zavisne kategorijske dihotomne varijable **bolest** (SAS ispis)

bolest	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	160	53.87	160	53.87
1	137	46.13	297	100.00

2.3 Univarijatna logistička regresija

U ovome potpoglavlju provest ćemo univarijatnu logističku regresiju za svaku nezavisnu varijablu te na taj način odrediti koliko je statistički značajna svaka od varijabli pojedinačno. Kao što smo već ranije pokazali, univarijatni logistički model dan je jednadžbom (1.5), tj.

$$logit(p(x)) = \beta_0 + \beta_1 x\,. \tag{2.1}$$

Dakle, imamo 13 različitih modela jer je 13 nezavisnih varijabli. U svakom modelu zavisna varijabla je bolest. Za diskusiju ćemo koristiti razinu značajnosti $\alpha = 5\%$.

No, prije toga, potrebno je konstruirati pomoćne, tzv. *dummy* varijable. Naime, ukoliko je nezavisna varijabla kontinuirana, tada je veza između zavisne i nezavisne varijable linearna pa ne postoji problem kod procjene parametara $\beta_i,\ i = 0, 1$ (2.1). Međutim, ukoliko ovisnost zavisne o nezavisnoj varijabli nije linearna ili ukoliko razmak između kategorija varijable nije ekvidistantan, kao i kod svih nominalnih varijabli gdje uopće ne postoji uređaj između kategorija, trebalo bi koristiti *dummy* varijable kako bi se dobro procijenili parametri β_i. [5] To su binarne varijable, s vrijednostima 0 i 1. Dakle, za sve nezavisne kategorijske varijable s $k \geq 3$ kategorija, konstruira se $k - 1$ pomoćnih varijabli tako da se jedna (referentna) kategorija fiksira, a preostale kategorije modeliramo pomoću nje.
U našem slučaju, *dummy* varijable konstruiramo za 4 nezavisne kategorijske varijable: bol, EKG, STnagib i talij_test (*SAS kod* se može vidjeti u poglavlju 3).

Glavni rezultati dobiveni univarijatnom logističkom regresijom za sve nezavisne varijable prikazani su u tablicama 2.10 do 2.12.

Tablica 2.10: Rezultati analize univarijatnih logističkih modela provedene u SAS-u

Varijabla	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vr
dob	409.946	394.252	15.6945	<.0001
spol	409.946	386.118	23.8284	<.0001
bol_tip bol_atip bol_asimp	409.946	328.752	81.1946	<.0001
tlak	409.946	402.883	7.0635	0.0079
kolesterol	409.946	408.026	1.9206	0.1658
secer	409.946	409.944	0.0030	0.9565
EKG_abnorm EKG_hiper	409.946	400.282	9.6648	0.0080
puls	409.946	351.970	57.9761	<.0001
ex_angina	409.946	355.477	54.4697	<.0001
STdepresija	409.946	350.484	59.4625	<.0001
ST_uzdignut ST_spusten	409.946	365.161	44.7859	<.0001
krvne_zile	409.946	339.421	70.5254	<.0001
talij_fiks talij_reverz	409.946	323.387	86.5593	<.0001

Za sve varijable konvergencija procedure je zadovoljena. Model je tim bolji, što je devijanca manja, odnosno što je vrijednost u stupcu *Likelihood Ratio* veća. Prema izračunatim p-vrijednostima, iz tablice 2.10 možemo zaključiti da su svi modeli, osim onih koji sadrže varijable kolesterol i šećer, statistički značajni.

Tablica 2.11: Rezultati procjene parametara metodom maksimalne vjerodostojnosti

Varijabla	Procjena parametra	Stand.greška	Wald χ^2	p-vr	Procjena Intercepta
dob	0.0529	0.0138	14.6640	0.0001	-3.0512
spol	1.2737	0.2725	21.8484	<.0001	-1.0438
bol_tip	0.4573	0.5256	0.7570	0.3843	
bol_atip	-0.2076	0.4550	0.2081	0.6482	-1.2840
bol_asimp	2.2552	0.3260	47.8486	<.0001	
tlak	0.0177	0.00681	6.7947	0.0091	-2.4940
kolesterol	0.00313	0.00228	1.8859	0.1697	-0.9300
secer	0.0181	0.3306	0.0030	0.9565	-0.1578
EKG_abnorm	1.6131	1.1672	1.9099	0.1670	-0.5145
EKG_hiper	0.6792	0.2380	8.1456	0.0043	
puls	-0.0443	0.00663	44.7135	<.0001	6.4718
ex_angina	1.9454	0.2831	47.2112	<.0001	-0.7768
STdepresija	0.9160	0.1381	44.0119	<.0001	-1.0827
ST_uzdignut	-1.6686	0.2637	40.0308	<.0001	0.6174
ST_spusten	-0.3298	0.4759	0.4800	0.4884	
krvne_zile	1.2475	0.1776	49.3315	<.0001	-0.9324
talij_fiks	1.9264	0.5338	13.0258	0.0003	-1.2333
talij_reverz	2.4148	0.2886	69.9996	<.0001	

Prema dobivenim p-vrijednostima, iz tablice 2.11 zaključujemo kako varijable koje statistički značajno utječu na vjerojatnost da osoba ima bolest srca su: dob, spol, bol_asimp, tlak, EKG_hiper, puls, ex_angina, ST_depresija, ST_uzdignut, krvne_zile, talij_fiks i talij_reverz.
Također, iz gornje tablice možemo iščitati jednadžbe univarijatnih logističkih modela za pojedine varijable:

- za dob: $logit(p) = -3.0512 + 0.0529 \times dob$
- za spol: $logit(p) = -1.0438 + 1.2737 \times spol$
- za bol: $logit(p) = -1.2840 + 0.4573 \times bol\_tip - 0.2076 \times bol\_atip + 2.2552 \times bol\_asimp$
- za tlak: $logit(p) = -2.4940 + 0.0177 \times tlak$
- za kolesterol: $logit(p) = -0.9300 + 0.00313 \times kolesterol$
- za šećer: $logit(p) = -0.1578 + 0.0181 \times secer$
- za EKG: $logit(p) = -0.5145 + 1.6131 \times EKG\_abnorm + 0.6792 \times EKG\_hiper$

- za puls: $logit(p) = 6.4718 - 0.0443 \times puls$
- za ex_angina: $logit(p) = -0.7768 + 1.9454 \times ex\_angina$
- za STdepresija: $logit(p) = -1.0827 + 0.9160 \times STdepresija$
- za STnagib: $logit(p) = 0.6174 - 1.6686 \times ST\_uzdignut - 0.3298 \times ST\_spusten$
- za krvne_zile: $logit(p) = -0.9324 + 1.2475 \times krvne\_zile$
- za talij_test: $logit(p) = -1.2333 + 1.9264 \times talij\_fiks + 2.4148 \times talij\_reverz$

Pomoću tih jednadžbi možemo izračunati vjerojatnost da osoba s određenim karakteristikama razvije bolest srca. Naime, prema jednadžbi (1.7) imamo:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{2.2}$$

Primjer 1: Vjerojatnost da muška osoba ima bolest srca je

$$p(1) = \frac{e^{-1.0438+1.2737\times 1}}{1 + e^{-1.0438+1.2737\times 1}} = 0.5572 \tag{2.3}$$

Primjer 2: Vjerojatnost da osoba koja ima depresiju ST segmenta jednaku 1.7 razvije bolest srca iznosi

$$p(1.7) = \frac{e^{-1.0827+0.9160\times 1.7}}{1 + e^{-1.0827+0.9160\times 1.7}} = 0.6164 \tag{2.4}$$

Tablica 2.12: Rezultati procjene omjera šansi univarijatnih modela

Varijabla	OR procjena	95% pouzdani interval	c
dob	**1.054**	**1.026 - 1.083**	**0.639**
spol	**3.574**	**2.095 - 6.097**	**0.631**
bol_tip	1.580	0.564 - 4.426	
bol_atip	0.813	0.333 - 1.982	**0.764**
bol_asimp	**9.537**	**5.034 - 18.068**	
tlak	**1.018**	**1.004 - 1.032**	**0.576**
kolesterol	1.003	0.999 - 1.008	0.567
secer	1.018	0.533 - 1.947	0.501
EKG_abnorm	5.018	0.509 - 49.440	**0.590**
EKG_hiper	**1.972**	**1.237 - 3.144**	
puls	**0.957**	**0.944 - 0.969**	**0.748**
ex_angina	**6.997**	**4.017 - 12.187**	**0.698**
STdepresija	**2.499**	**1.907 - 3.276**	**0.734**
ST_uzdignut	**0.189**	**0.112 - 0.316**	**0.696**
ST_spusten	0.719	0.283 - 1.828	
krvne_zile	**3.481**	**2.458 - 4.931**	**0.752**
talij_fiks	**6.865**	**2.412 - 19.542**	**0.766**
talij_reverz	**11.187**	**6.354 - 19.697**	

Još jedan način kojim možemo provjeriti statističku značajnost varijabli je 95% pouzdani interval za procjenu omjera šansi. Ukoliko on sadrži jedinicu, varijabla nije statistički značajna, a ako ne sadrži jedinicu, varijabla jest statistički značajna. Vodeći se time, iz tablice 2.12 ponavljamo prethodni zaključak o statističkoj značajnosti varijabli dob, spol, bol_asimp, tlak, EKG_hiper, puls, ex_angina, ST_depresija, ST_uzdignut, krvne_zile, talij_fiks i talij_reverz. (*Opaska*: one su u tablici 2.12 istaknute podebljanim slovima.)

Vrijednosti u stupcu *OR procjena* predstavljaju omjere šansi prelaska zavisne varijable iz kategorije "osoba nema bolest srca" u kategoriju "osoba ima bolest srca", prilikom prelaska nezavisne varijable iz niže u višu kategoriju (ako se radi o kategorijskoj varijabli), odnosno prilikom njenog pomaka za 1 (ako se radi o numeričkoj varijabli). Dakle, na temelju podataka iz tablice 2.12 možemo izvesti sljedeću interpretaciju:

- Povećanjem dobi osobe za 1 godinu, povećava joj se izgled da razvije bolest srca za 5.4% (tj. 1.054 puta), i to statistički značajno.
- Muškarci, u odnosu na žene, imaju 3.574 puta (tj. za 257.4%) veći omjer šansi da razviju bolest srca, i taj rezultat je isto statistički značajan.

- Osobe s tipičnom anginom imaju 1.580 puta (tj. za 58%) veći omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli, međutim taj rezultat nije statistički značajan.
- Osobe s atipičnom anginom imaju manji omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli (OR=0.813), ali to također nije statistički značajno.
- Asimptomatske osobe imaju 9.537 puta veći omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli. To je statistički značajan rezultat.
- Povećanje tlaka za 1 jedinicu povećava omjer šansi da osoba razvije bolest srca 1.018 puta (statistički značajno).
- Povećanje kolesterola za 1 jedinicu povećava omjer šansi da osoba razvije bolest srca 1.003 puta (nije statistički značajno).
- Osobe s povišenom razinom šećera u krvi imaju 1.018 puta veći omjer šansi da razviju bolest srca u odnosu na one koje nemaju povišen šećer (nije statistički značajno).
- Osobe s uočenom abnormalnosti ST i/ili T vala imaju 5.018 puta veći omjer šansi da razviju bolest srca u odnosu na one s urednim EKG nalazom (nije statistički značajno).
- Osobe s mogućom hipertrofijom lijeve klijetke imaju 1.972 puta veći omjer šansi da razviju bolest srca u odnosu na one s urednim EKG nalazom (statistički značajan rezultat).
- Povećanje pulsa za 1 statistički značajno povećava omjer šansi da osoba razvije bolest srca 0.957 puta (dakle osobe s višim pulsom imaju manji omjer šansi za razvoj bolesti).
- Osobe koje su osjetile bol u prsima prilikom tjelesne aktivnosti imaju 6.997 puta veći omjer šansi da razviju bolest srca u odnosu na one koje pritom ne osjete bol u prsima (i to je statistički značajno).
- Povećanje ST-depresije za 1 povećava omjer šansi da osoba razvije bolest srca 2.499 puta (statistički značajno).
- Osobe s uzdignutim ST-segmentom u pokretu imaju manji omjer šansi da razviju bolest srca u odnosu na osobe s ravnim ST-segmentom (OR=0.189) (statistički značajan rezultat).

- Osobe sa spuštenim ST-segmentom u pokretu imaju manji omjer šansi da razviju bolest srca u odnosu na osobe s ravnim ST-segmentom (OR=0.719), ali to nije statistički značajno.

- Povećanje broja vidljivih krvnih žila prilikom fluoroskopije za 1 (statistički značajno) povećava omjer šansi da osoba razvije bolest srca 3.481 puta.

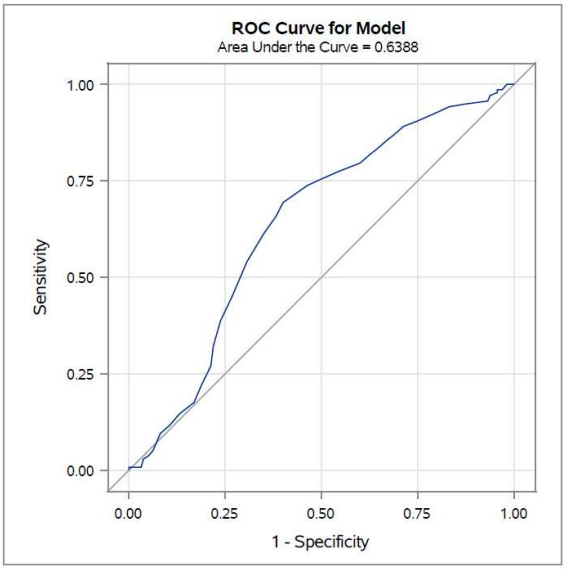
- Osobe s nemogućnosti jednake raspodjele talija imaju 6.865 puta veći omjer šansi da razviju bolest srca u odnosu na osobe s normalnim rezultatima talij testa (statistički značajan rezultat).

- Osobe s mogućnosti naknadne jednake raspodjele talija imaju 11.187 puta veći omjer šansi da razviju bolest srca u odnosu na osobe s normalnim rezultatima talij testa (statistički značajan rezultat).

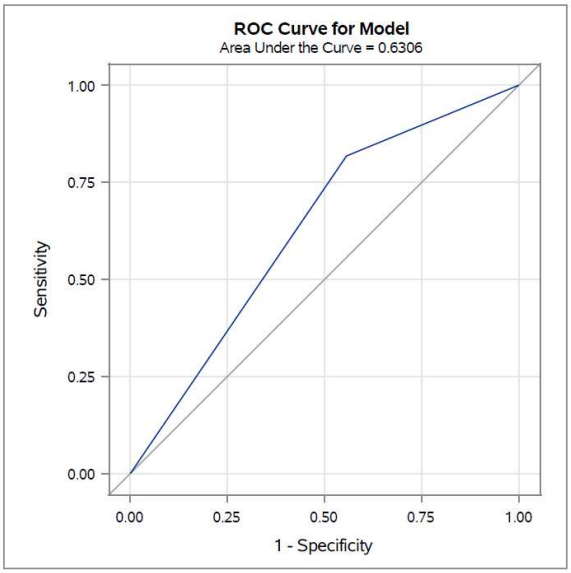
Ovdje uočavamo zanimljiv primjer za varijablu tlak. Naime, kao što smo već rekli u potpoglavlju 1.6, često nam je od većeg interesa promatrati što se događa prilikom pomaka kontinuirane varijable za nekoliko jedinica, umjesto za 1 jedinicu.
Npr. Povećanjem tlaka za 10 mmHg, povećava se omjer šansi da osoba razvije bolest srca $e^{10\times 0.0177} = 1.194$ puta, tj. za 19.4% (to smo izračunali pomoću jednadžbe (1.30)).

Treći pokazatelj koji govori kakav je model klasifikator, tj. koliko dobro razdjeljuje zdrave i bolesne, jest površina ispod ROC-krivulje o kojoj smo više rekli u potpoglavlju 1.7. Nju predstavlja vrijednost c iz zadnjeg stupca tablice 2.12. Vidimo da najveću prediktivnu snagu imaju varijable bol i talij test. Model s varijablom šećer nema dobru prediktivnu vrijednost jer je c=0.501, a prate je kolesterol (c=0.567) i tlak (c=0.576).

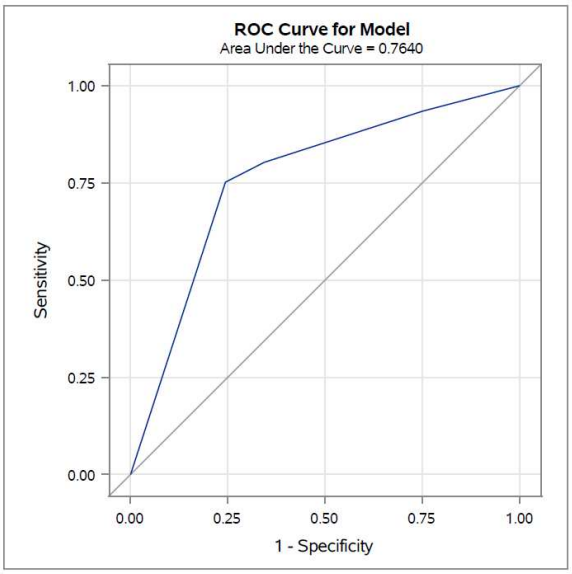
Prikažimo grafički pripadne ROC-krivulje za sve univarijatne modele:



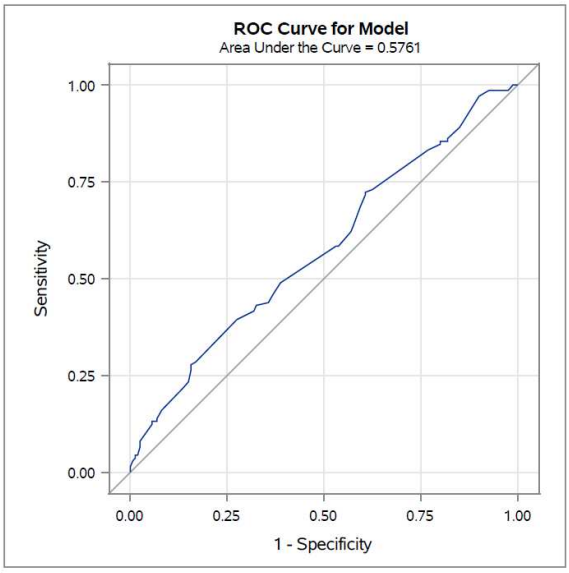
Slika 2.1: ROC krivulja za varijablu **dob**
(SAS ispis)



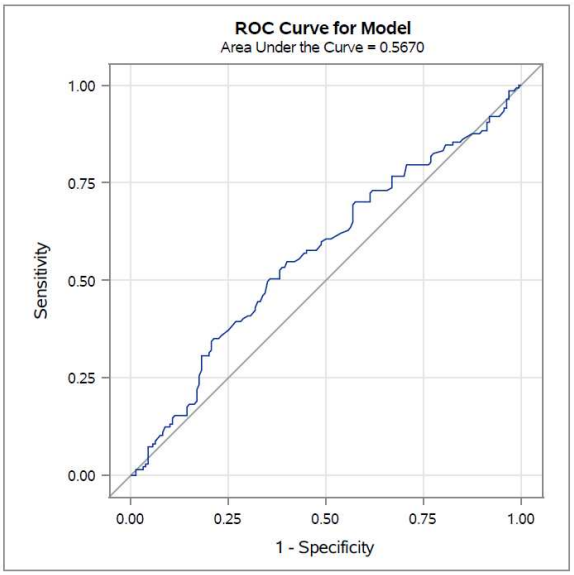
Slika 2.2: ROC krivulja za varijablu **spol**
(SAS ispis)



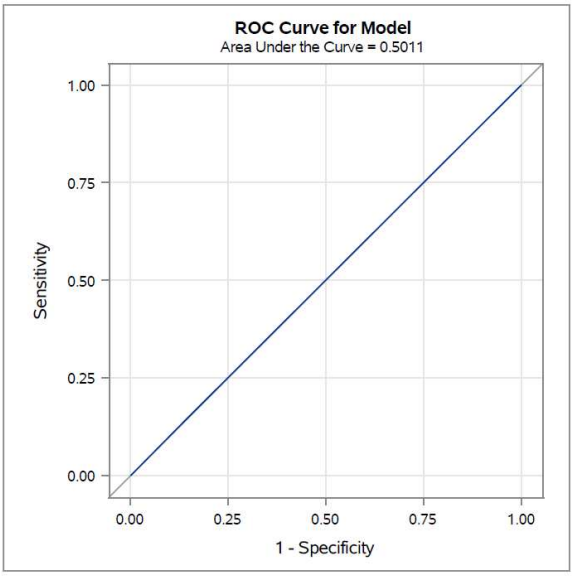
Slika 2.3: ROC krivulja za varijablu **bol**
(SAS ispis)



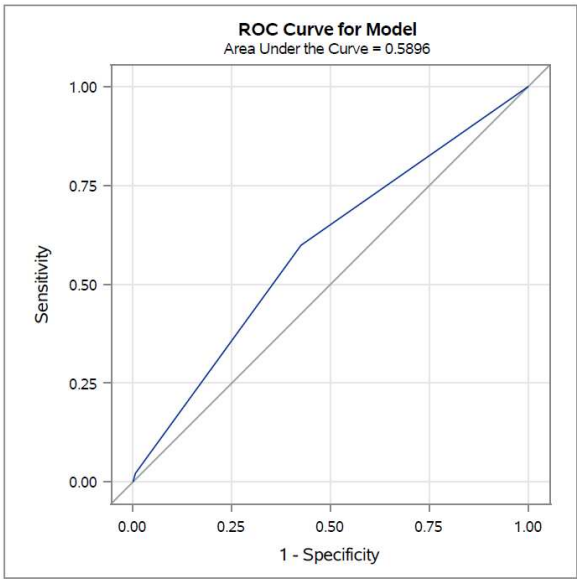
Slika 2.4: ROC krivulja za varijablu **tlak** (SAS ispis)



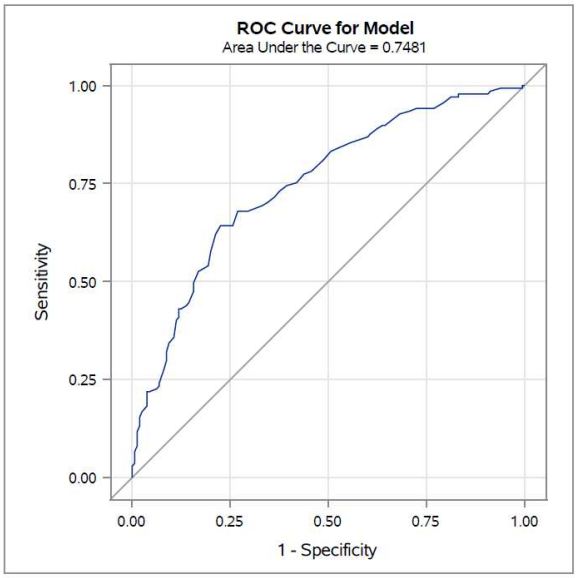
Slika 2.5: ROC krivulja za varijablu **kolesterol** (SAS ispis)



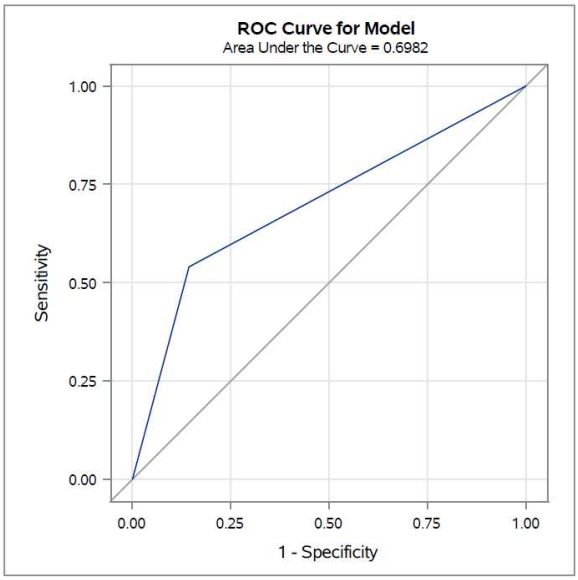
Slika 2.6: ROC krivulja za varijablu **šećer** (SAS ispis)



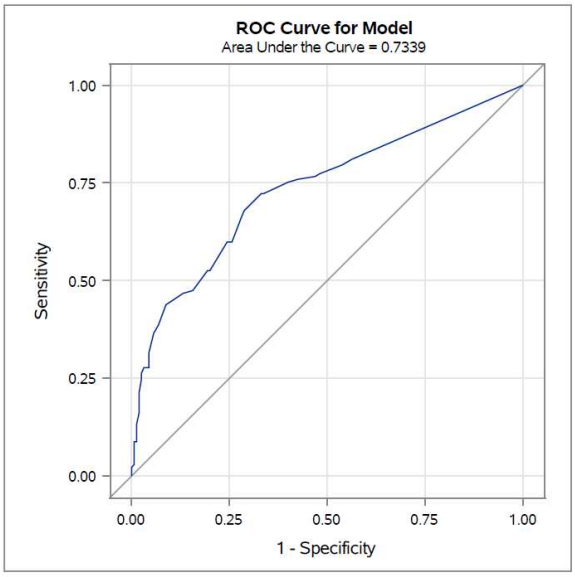
Slika 2.7: ROC krivulja za varijablu **EKG** (SAS ispis)



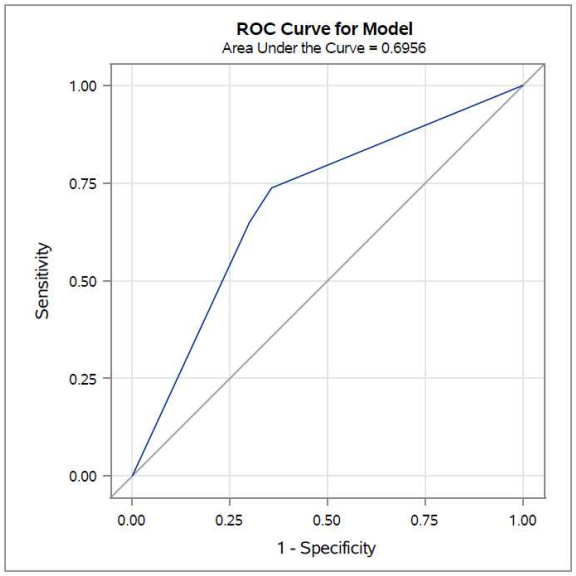
Slika 2.8: ROC krivulja za varijablu **puls** (SAS ispis)



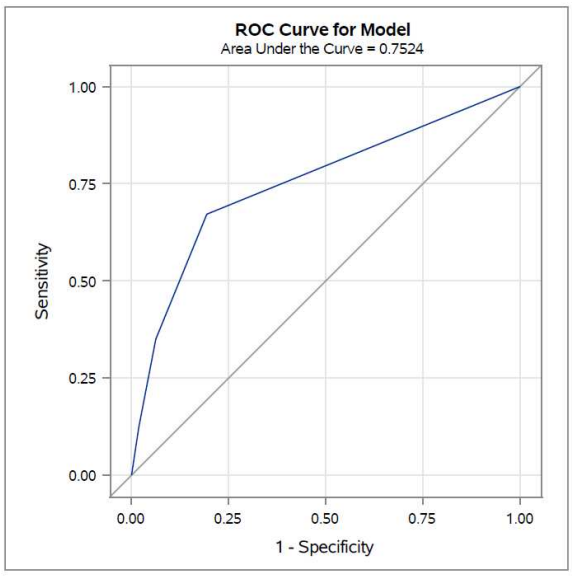
Slika 2.9: ROC krivulja za varijablu **ex_angina** (SAS ispis)



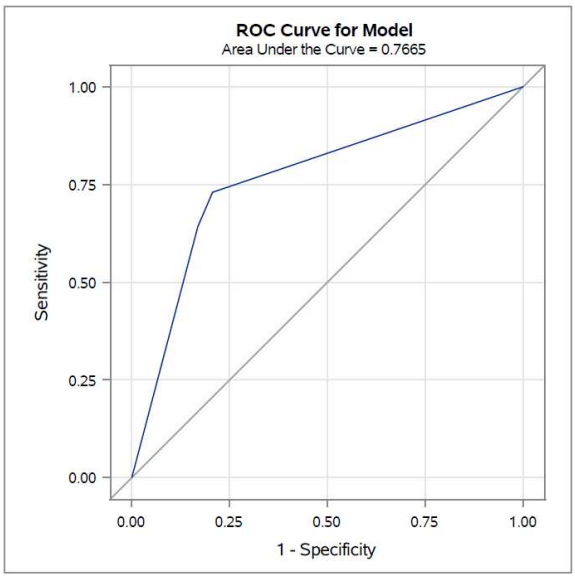
Slika 2.10: ROC krivulja za varijablu **STdepresija** (SAS ispis)



Slika 2.11: ROC krivulja za varijablu **STnagib** (SAS ispis)

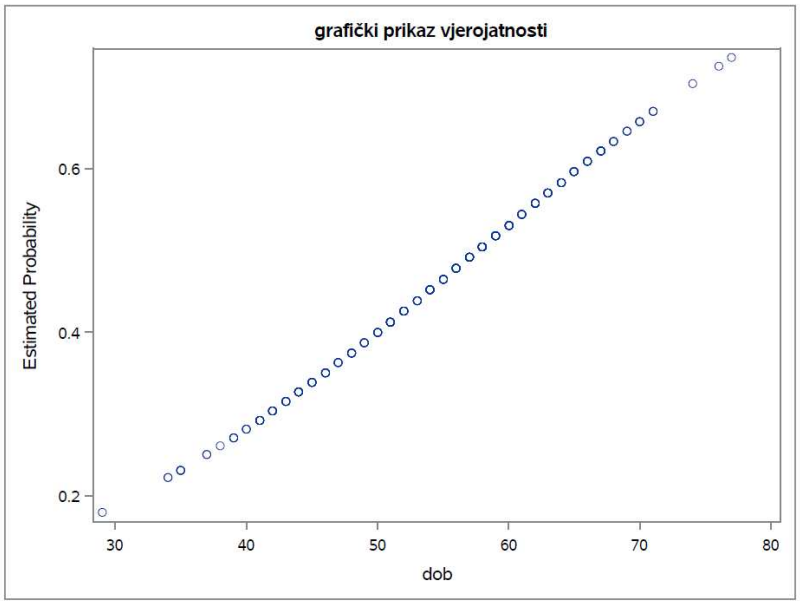


Slika 2.12: ROC krivulja za varijablu **krvne_žile** (SAS ispis)

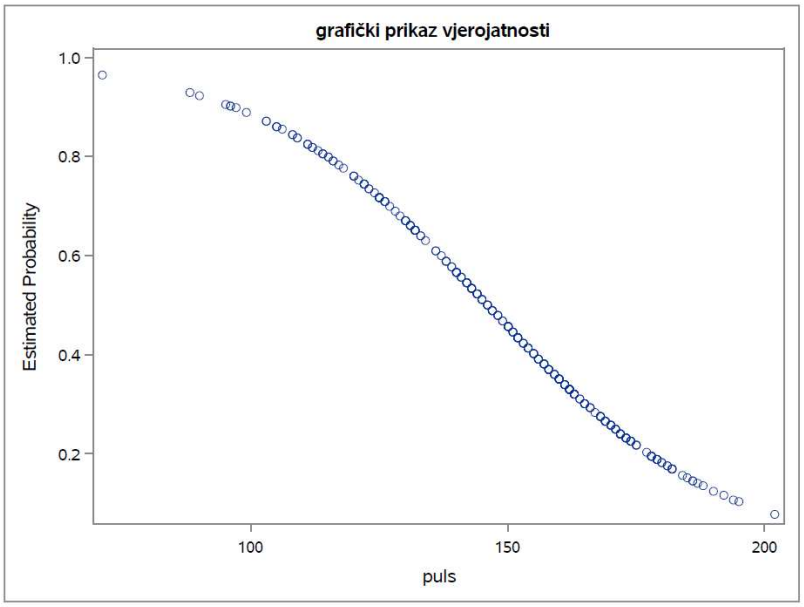


Slika 2.13: ROC krivulja za varijablu **talij_test** (SAS ispis)

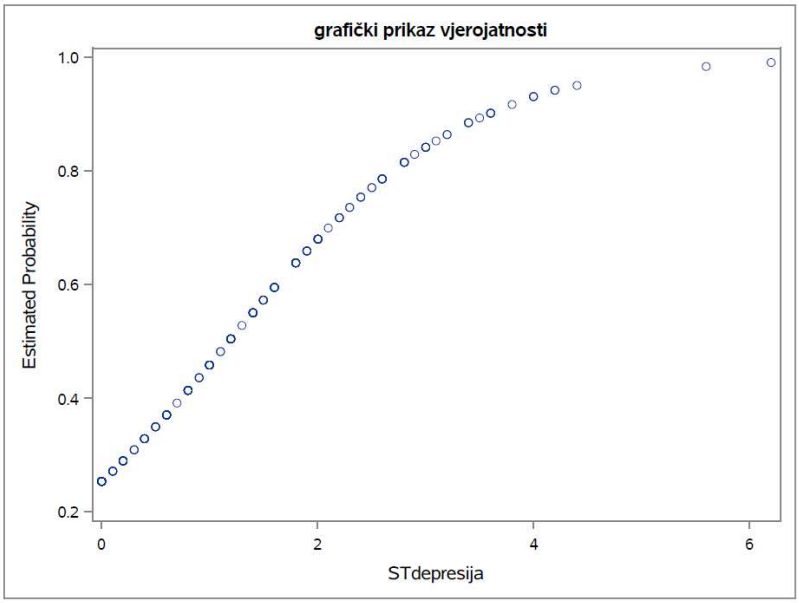
Pogledajmo sada grafički prikaz vjerojatnosti pojave bolesti srca na primjeru nekih nezavisnih varijabli. Što je krivulja vjerojatnosti strmija, to je varijabla značajnija, tj. taj model je bolji klasifikator.
Uočavamo da je krivulja sa slike 2.16 puno strmija od krivulje sa slike 2.14 pa i na ovaj način možemo provjeriti kako je univarijatni logistički model s nezavisnom varijablom ST-depresija bolji klasifikator od univarijatnog logističkog modela s nezavisnom varijablom dob. Sa slike 2.15 iščitavamo da se povećanjem nezavisne varijable puls, smanjuje vjerojatnost dobivanja bolesti srca.



Slika 2.14: Grafički prikaz vjerojatnosti za nezavisnu varijablu **dob** (SAS ispis)



Slika 2.15: Grafički prikaz vjerojatnosti za nezavisnu varijablu **puls** (SAS ispis)



Slika 2.16: Grafički prikaz vjerojatnosti za nezavisnu varijablu **STdepresija** (SAS ispis)

2.4 Multivarijatna logistička regresija

U ovome dijelu ispitat ćemo statističku značajnost nezavisnih varijabli te njihov utjecaj na pojavu bolesti srca na način da u model stavimo sve one varijable koje su se pokazale značajnima u univarijatnoj logističkoj regresiji. Sve prediktorske varijable stavit ćemo u model istovremeno. Dakle, nezavisne varijable koje ulaze u model su: dob, spol, bol, tlak, EKG, puls, ex_angina, STdepresija, STnagib, krvne_zile i talij_test.

Glavni rezultati dobiveni multivarijatnom logističkom regresijom prikazani su u tablicama 2.13 do 2.15.

Tablica 2.13: Rezultati analize multivarijatnog logističkog modela provedene u SAS-u

df	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vr
16	409.946	193.752	216.1943	<.0001

Konvergencija procedure je zadovoljena. Iz tablice 2.13 vidimo da je p-vrijednost <.0001 pa zaključujemo kako je model statistički značajan. Drugim riječima, dane varijable statistički značajno pomažu u klasifikaciji ima li osoba bolest srca ili nema.

Tablica 2.14: Rezultati procjene parametara metodom maksimalne vjerodostojnosti

Varijabla	Procjena parametra	Stand.greška	Wald χ^2	p-vr
intercept	-3.7186	2.5962	2.0516	0.1520
dob	-0.0110	0.0244	0.2034	0.6520
spol	1.3492	0.5031	7.1928	0.0073
bol_tip	-0.2307	0.6534	0.1247	0.7240
bol_atip	1.0950	0.6204	3.1149	0.0776
bol_asimp	1.9579	0.4907	15.9170	<.0001
tlak	0.0229	0.0111	4.2878	0.0384
EKG_abnorm	0.7596	2.2757	0.1114	0.7385
EKG_hiper	0.5361	0.3764	2.0288	0.1543
puls	-0.0165	0.0109	2.2928	0.1300
ex_angina	0.6650	0.4340	2.3485	0.1254
STdepresija	0.4035	0.2290	3.1048	0.0781
ST_uzdignut	-1.1277	0.4704	5.7475	0.0165
ST_spusten	-0.7616	0.8468	0.8090	0.3684
krvne_zile	1.2419	0.2688	21.3430	<.0001
talij_fiks	-0.1357	0.7716	0.0309	0.8604
talij_reverz	1.4471	0.4250	11.5961	0.0007

Prema dobivenim p-vrijednostima, iz tablice 2.14 uočavamo kako su varijable spol, bol_asimp, tlak, ST_uzdignut, krvne_zile i talij_reverz statistički značajni prediktori imanja bolesti srca.
Također, možemo iščitati jednadžbu dobivenog multivarijatnog logističkog modela koja glasi:

$$logit(p) = -3.7186 - 0.0110 \times dob + 1.3492 \times spol - 0.2307 \times bol\_tip + 1.0950 \times bol\_atip + 1.9579 \times bol\_asimp + 0.0229 \times tlak + 0.7596 \times EKG\_abnorm + 0.5361 \times EKG\_hiper - 0.0165 \times puls + 0.6650 \times ex\_angina + 0.4035 \times STdepresija - 1.1277 \times ST\_uzdignut - 0.7616 \times ST\_spusten + 1.2419 \times krvne\_zile - 0.1357 \times talij\_fiks + 1.4471 \times talij\_reverz$$

Pokažimo sada na jednom primjeru kako pomoću gornje jednadžbe možemo izračunati vjerojatnost da osoba s određenim karakteristikama razvije bolest srca.

Primjer: Ženska osoba, 68 godina, s atipičnom anginom, tlak=138 mmHg, uredan nalaz EKG-a, max.puls=121, bez boli u prsima prilikom tjelesne aktivnosti, depresija ST-segmenta=1.2, ST-segment u pokretu je ravan, 2 vidljive krvne žile pri fluoroskopiji i s normalnim rezultatima talij testa.

Pomoću jednadžbi (1.8) i (1.9) dobivamo da vjerojatnost da ta osoba ima bolest srca iznosi:

$$p(68,0,1,138,0,121,0,1.2,1,2,0) = \frac{e^*}{1+e^*} = 0.6814, \tag{2.5}$$

pri čemu je

$* = -3.7186 - 0.0110 \times 68 + 1.3492 \times 0 - 0.2307 \times 0 + 1.0950 \times 1 + 1.9579 \times 0 + 0.0229 \times 138 + 0.7596 \times 0 + 0.5361 \times 0 - 0.0165 \times 121 + 0.6650 \times 0 + 0.4035 \times 1.2 - 1.1277 \times 0 - 0.7616 \times 0 + 1.2419 \times 2 - 0.1357 \times 0 + 1.4471 \times 0 = 0.7601.$

Tablica 2.15: Rezultati procjene omjera šansi multivarijatnog modela

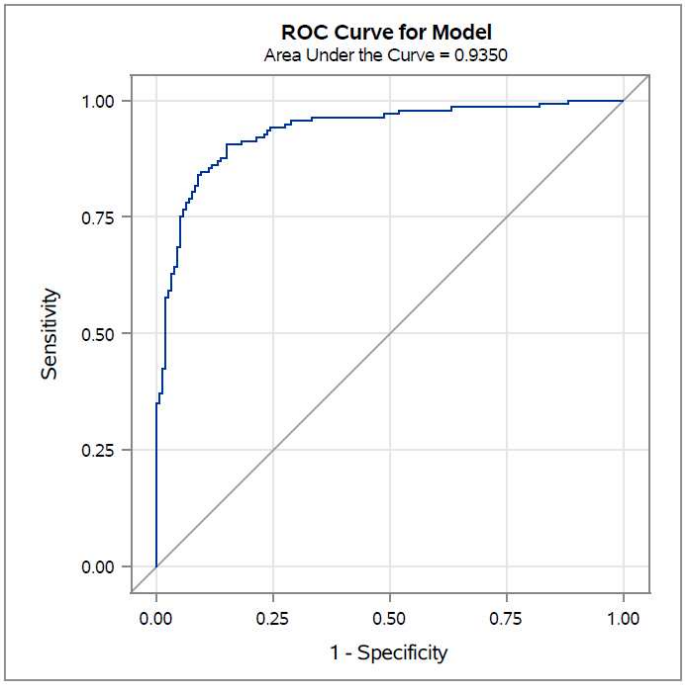
Varijabla	OR procjena	95% pouzdani interval
dob	0.989	0.943 - 1.038
spol	**3.854**	**1.438 - 10.331**
bol_tip	0.794	0.221 - 2.857
bol_atip	2.989	0.886 - 10.085
bol_asimp	**7.084**	**2.708 - 18.536**
tlak	**1.023**	**1.001 - 1.046**
EKG_abnorm	2.137	0.025 - 184.904
EKG_hiper	1.709	0.817 - 3.575
puls	0.984	0.963 - 1.005
ex_angina	1.945	0.831 - 4.552
STdepresija	1.497	0.956 - 2.345
ST_uzdignut	**0.324**	**0.129 - 0.814**
ST_spusten	0.467	0.089 - 2.455
krvne_zile	**3.462**	**2.044 - 5.864**
talij_fiks	0.873	0.192 - 3.961
talij_reverz	**4.251**	**1.848 - 9.777**

Gledajući 95% pouzdane intervale, iz tablice 2.15 ponovno možemo zaključiti da su statistički značajne varijable spol, bol_asimp, tlak, ST_uzdignut, krvne_zile te talij_reverz jer njihovi intervali ne sadrže jedinicu (*Opaska*: one su u tablici 2.15 istaknute podebljanim slovima.)

Na temelju podataka iz stupca *OR procjena*, dobivamo sljedeće statistički značajne rezultate:

- Muškarci, u odnosu na žene, imaju 3.854 puta veći omjer šansi da razviju bolest srca.
- Asimptomatske osobe imaju 7.084 puta veći omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli.
- Povećanje tlaka za 1 jedinicu povećava omjer šansi da osoba razvije bolest srca 1.023 puta.
- Osobe s uzdignutim ST-segmentom u pokretu imaju manji omjer šansi da razviju bolest srca u odnosu na osobe s ravnim ST-segmentom (OR=0.324).
- Povećanje broja vidljivih krvnih žila prilikom fluoroskopije za 1 povećava omjer šansi da osoba razvije bolest srca 3.462 puta.
- Osobe s mogućnosti naknadne jednake raspodjele talija imaju 4.251 puta veći omjer šansi da razviju bolest srca u odnosu na osobe s normalnim rezultatima talij testa.

Na kraju ćemo grafički prikazati ROC-krivulju za dobiveni multivarijatni logistički model. Vrijednost *c* statistike je 0.935, što znači da model ima prilično jaku prediktivnu vrijednost (prediktivna snaga pojave bolesti srca ovim modelom iznosi 93.5%).



Slika 2.17: ROC krivulja za multivarijatni logistički model (SAS ispis)

2.5 Stepwise procedura

Postoji nekoliko metoda selekcije varijabli kojima nastaju različiti multivarijatni logistički modeli. Cilj je da je kombinacija nezavisnih varijabli jako korelirana sa zavisnom, ali također da su one međusobno nekorelirane. Izbor nezavisnih varijabli u model vrlo je osjetljiv jer ubacivanje ili izbacivanje bilo koje varijable može jako promijeniti jednadžbu ravnine. [5] Postoje tri glavne procedure pomoću kojih se odvija selekcija nezavisnih varijabli. U ovome potpoglavlju analizirat ćemo multivarijatni logistički model koji se dobiva tzv. Stepwise procedurom, no prije toga ukratko ćemo opisati svaku od njih.

Selekcija unaprijed (eng. *Forward*) - metoda koja počinje bez ijedne varijable u modelu. U svakom sljedećem koraku dodaje se po jedna varijabla koja ima najveći (i statistički značajan) doprinos adekvatnosti modela.

Selekcija unatrag (eng. *Backward*) - metoda koja počinje punim modelom (sve varijable su uključene). U svakom sljedećem koraku eliminira se iz modela po jedna varijabla, i to ona koja ima najmanji (i statistički ne značajan) doprinos adekvatnosti modela.

Stepwise selekcija - metoda koja predstavlja kombinaciju Backward i Forward procedure. Počinje kao Forward, samo što varijabla koja je ušla u model, ne ostaje nužno

do kraja u njemu. Nakon svakog uključivanja pojedine varijable, testira se može li se uključena varijabla izostaviti iz modela, a da pri tome ne dođe do značajnog smanjenja njegove adekvatnosti. Procedura završava kada više nema varijabli koje se mogu dodati ili ako je trenutni model jednak modelu iz nekog od prethodnih koraka. Stepwise se ujedno i najčešće koristi pri odabiru najadekvatnijeg modela.

Glavni rezultati multivarijatnog logističkog modela dobivenog procedurom Stepwise prikazani su u tablicama 2.16 do 2.20.

Tablica 2.16: Rezultati analize multivarijatnog logističkog modela dobivenog Stepwise procedurom

Varijabla	df	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vr
bol_asimp	1	409.946	330.031	79.9156	<.0001
talij_reverz	2	409.946	283.686	126.2603	<.0001
krvne_zile	3	409.946	242.437	167.5093	<.0001
ST_uzdignut	4	409.946	222.440	187.5063	<.0001
spol	5	409.946	214.325	195.6214	<.0001
tlak	6	409.946	208.654	201.2922	<.0001
ex_angina	7	409.946	204.736	205.2106	<.0001

Konvergencija procedure je zadovoljena. Iz tablice 2.16 vidimo da je p-vrijednost <.0001 za sve varijable pa zaključujemo kako je model statistički značajan.

Tablica 2.17: Rezultati procjene parametara metodom maksimalne vjerodostojnosti

Varijabla	Procjena parametra	Stand.greška	Wald χ^2	p-vr
intercept	-5.6582	1.4381	15.4804	<.0001
spol	1.2828	0.4414	8.4479	0.0037
bol_asimp	1.7845	0.3866	21.3096	<.0001
tlak	0.0213	0.00974	4.7862	0.0287
ex_angina	0.8178	0.4107	3.9653	0.0464
ST_uzdignut	-1.4737	0.3758	15.3767	<.0001
krvne_zile	1.2799	0.2415	28.0823	<.0001
talij_reverz	1.3850	0.3822	13.1332	0.0003

Prema dobivenim p-vrijednostima, iz tablice 2.17 uočavamo kako su varijable spol, bol_asimp, tlak, ex_angina, ST_uzdignut, krvne_zile i talij_reverz statistički značajni prediktori imanja bolesti srca.
Međutim, kod kategorijskih varijabli modeliranih pomoću *dummy* varijabli, ukoliko se jedna kategorija ispostavi značajnom, moramo sve komponente varijable ostaviti u modelu. Iz tog razloga trebamo provesti multivarijatnu logističku regresiju sa svim nezavisnim varijablama koje su se pokazale statistički značajnima provedbom Stepwise-a, ali uključujući i sve pripadne *dummy* varijable.

Tablica 2.18: Rezultati analize multivarijatnog logističkog modela dobivenog Stepwise procedurom (uključene sve pripadne pomoćne varijable)

df	-2LogL (Intercept Only)	-2LogL (Intercept and Covariates)	Likelihood Ratio (χ^2)	p-vr
11	409.946	201.900	208.0467	<.0001

Tablica 2.19: Rezultati procjene parametara metodom maksimalne vjerodostojnosti (uključene sve pripadne pomoćne varijable)

Varijabla	Procjena parametra	Stand.greška	Wald χ^2	p-vr
intercept	-6.1682	1.5199	16.4699	<.0001
spol	1.4351	0.4863	8.7102	0.0032
bol_tip	-0.1475	0.6486	0.0517	0.8201
bol_atip	0.8675	0.5973	2.1093	0.1464
bol_asimp	2.0078	0.4751	17.8601	<.0001
tlak	0.0233	0.0102	5.1993	0.0226
ex_angina	0.8516	0.4190	4.1311	0.0421
ST_uzdignut	-1.6863	0.4195	16.1604	<.0001
ST_spusten	-0.4383	0.7069	0.3845	0.5352
krvne_zile	1.3344	0.2494	28.6281	<.0001
talij_fiks	-0.0121	0.7349	0.0003	0.9869
talij_reverz	1.4309	0.4086	12.2648	0.0005

Sada iz tablice 2.19 možemo iščitati punu jednadžbu multivarijatnog modela dobivenog Stepwise procedurom:

$$logit(p) = -6.1682 + 1.4351 \times spol - 0.1475 \times bol\_tip + 0.8675 \times bol\_atip + 2.0078 \times bol\_asimp + 0.0233 \times tlak + 0.8516 \times ex\_angina - 1.6863 \times ST\_uzdignut - 0.4383 \times$$

$ST\_spusten + 1.3344 \times krvne\_zile - 0.0121 \times talij\_fiks + 1.4309 \times talij\_reverz$

Pokažimo opet na primjeru kako pomoću gornje jednadžbe možemo izračunati vjerojatnost da osoba s određenim karakteristikama dobije bolest srca.

Primjer: Ženska osoba, s asimptomatskom anginom, tlak=140 mmHg, osjeća bol u prsima prilikom tjelesne aktivnosti, ST-segment u pokretu je uzdignut, 3 vidljive krvne žile pri fluoroskopiji, normalni rezultati talij testa.
Vjerojatnost da ta osoba ima bolest srca je:

$$p(0,3,140,1,0,3,0) = \frac{e^*}{1+e^*} = 0.9064, \tag{2.6}$$

pri čemu je

$* = -6.1682 + 1.4351 \times 0 - 0.1475 \times 0 + 0.8675 \times 0 + 2.0078 \times 1 + 0.0233 \times 140 + 0.8516 \times 1 - 1.6863 \times 1 - 0.4383 \times 0 + 1.3344 \times 3 - 0.0121 \times 0 + 1.4309 \times 0 = 2.2701.$

Tablica 2.20: Rezultati procjene omjera šansi Stepwise multivarijatnog modela

Varijabla	OR procjena	95% pouzdani interval
spol	**4.200**	**1.619 - 10.894**
bol_tip	0.863	0.242 - 3.076
bol_atip	2.381	0.738 - 7.677
bol_asimp	**7.447**	**2.935 - 18.896**
tlak	**1.024**	**1.003 - 1.044**
ex_angina	**2.343**	**1.031 - 5.327**
ST_uzdignut	**0.185**	**0.081 - 0.421**
ST_spusten	0.645	0.161 - 2.578
krvne_zile	**3.798**	**2.329 - 6.191**
talij_fiks	0.988	0.234 - 4.172
talij_reverz	**4.183**	**1.878 - 9.316**

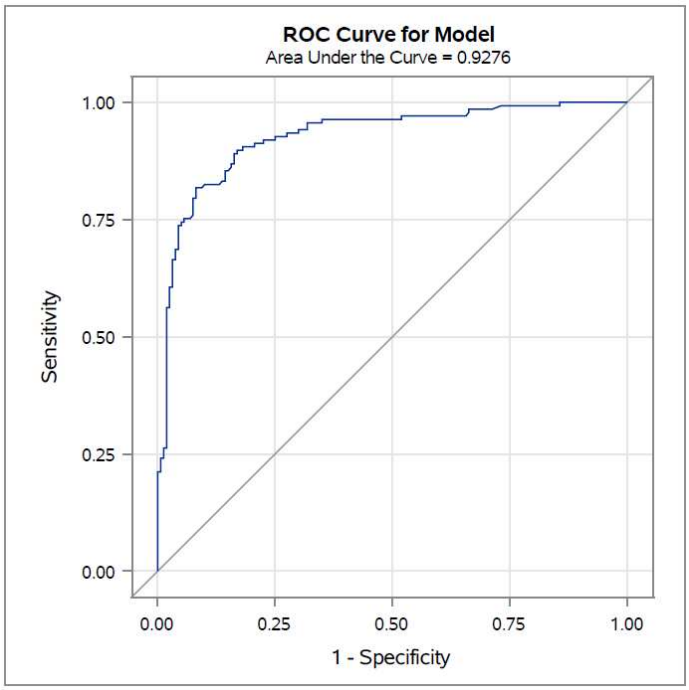
Prema 95% pouzdanim intervalima, iz tablice 2.20 potvrđujemo statističku značajnost varijabli: spol, bol_asimp, tlak, ex_angina, ST_uzdignut, krvne_zile i talij_reverz. (*Opaska*: one su u tablici 2.20 istaknute podebljanim slovima.)

Na temelju vrijednosti iz stupca *OR procjena* možemo interpretirati:

- Muškarci, u odnosu na žene, imaju 4.200 puta veći omjer šansi da razviju bolest srca, i taj rezultat je statistički značajan.

- Osobe s tipičnom anginom imaju manji omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli (OR=0.863).
- Osobe s atipičnom anginom imaju 2.381 puta veći omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli.
- Asimptomatske osobe imaju 7.447 puta veći omjer šansi da razviju bolest srca u odnosu na osobe bez anginalne boli (statistički značajan rezultat).
- Povećanje tlaka za 1 jedinicu povećava omjer šansi da osoba razvije bolest srca 1.024 puta, i to je statistički značajno.
- Osobe koje su osjetile bol u prsima prilikom tjelesne aktivnosti imaju 2.343 puta veći omjer šansi da razviju bolest srca u odnosu na one koje pritom ne osjete bol u prsima (statistički značajan rezultat).
- Osobe s uzdignutim ST-segmentom u pokretu imaju manji omjer šansi da razviju bolest srca u odnosu na osobe s ravnim ST-segmentom (OR=0.185) i to je statistički značajan rezultat.
- Osobe sa spuštenim ST-segmentom u pokretu imaju manji omjer šansi da razviju bolest srca u odnosu na osobe s ravnim ST-segmentom (OR=0.645).
- Povećanje broja vidljivih krvnih žila prilikom fluoroskopije za 1 povećava omjer šansi da osoba razvije bolest srca 3.798 puta, i to je statistički značajno.
- Osobe s nemogućnosti jednake raspodjele talija imaju manji omjer šansi da razviju bolest srca u odnosu na osobe s normalnim rezultatima talij testa (OR=0.988).
- Osobe s mogućnosti naknadne jednake raspodjele talija imaju 4.183 puta veći omjer šansi da razviju bolest srca u odnosu na osobe s normalnim rezultatima talij testa (statistički značajan rezultat).

Prikažimo još ROC-krivulju za dobiveni Stepwise multivarijatni logistički model. Vrijednost *c* statistike je 0.928, što je nešto manje nego u punom multivarijatnom modelu, ali i dalje predstavlja jaku prediktivnu vrijednost (prediktivna snaga pojave bolesti srca ovim modelom iznosi 92.8%).



Slika 2.18: ROC krivulja za multivarijatni logistički model dobiven procedurom Stepwise (SAS ispis)

Gledajući dobivene rezultate procjene omjera šansi za nezavisne varijable, uočavamo kako nema velikih razlika kod univarijatnih i multivarijatnog modela. Najveće razlike vide se kod varijabli talij_test i ex_angina.
Primjerice, ako pogledamo varijablu talij_reverz (koja se svugdje ispostavila statistički značajnom), vidimo da je ona prilično jak prediktor, i to najjači kada je sama u modelu, dakle univarijatno. Tada joj je procjena OR-a jednaka 11.187. Također, vrijednost c statistike je pritom 0.766, što upućuje na vrlo dobru prediktivnu vrijednost. Kada smo joj u model pridodali druge varijable, tada se OR smanjio jer su predikciju preuzele i druge statistički značajne varijable. Međutim, talij_reverz je i dalje ostao jedna od najznačajnijih. Vrijednost OR-a u Stepwise multivarijatnom modelu za varijablu talij_reverz iznosi 4.183, što je vrlo slično kao i za varijablu spol. Stoga bi u nekim daljnjim analizama bilo zanimljivo promatrati ima li razlike u talij_reverzu po svakom spolu zasebno.

2.6 Zaključak

Pri modeliranju vjerojatnosti da osoba razvije bolest srca na temelju danih podataka, korsitili smo se trima različitim pristupima, tj. metodama, a to su univarijatna logistička regresija, multivarijatna logistička regresija i Stepwise procedura.

Univarijatna logistička regresija pokazala je da varijable koje imaju statistički značajan utjecaj na imanje bolesti srca su: dob, spol, tip boli u prsima, krvni tlak, nalaz EKG-a, maksimalni puls u pokretu, angina izazvana tjelesnom aktivnošću, depresija ST-segmenta, vršni nagib ST-segmenta u pokretu, broj vidljivih glavnih krvnih žila tijekom fluoroskopije te rezultati talij testa. Dakle, jedino se varijable kolesterol i šećer u krvi nisu pokazale značajnima na razini značajnosti 5%, iako bismo možda intuitivno mislili da će ispasti značajne.

Puni multivarijatni logistički model obuhvatio je sve varijable koje su ispale značajne u univarijatnoj regresiji. Kao statistički značajne, označio je varijable spol, tip boli u prsima, krvni tlak, vršni nagib ST-segmenta u pokretu, broj vidljivih glavnih krvnih žila tijekom fluoroskopije te rezultate talij testa, dok je u multivarijatnom modelu dobivenom Stepwise procedurom, uz sve navedene, pridodana i varijabla angina izazvana tjelesnom aktivnošću. Dakle, ta dva modela razlikuju se u jednoj varijabli. Razlog tome može se nalaziti u medicinskoj pozadini istraživanja ili u samom uzorku jer i manje promjene u vrijednostima podataka mogu nekada dovesti do većih razlika u modelima. No, oba modela imaju jaku prediktivnu snagu (visoka vrijednost c-statistike). Također, SAS kao statistički softver ima vrlo dobro razvijenu proceduru Stepwise, koja ponekad daje i bolju predikciju nego puni multivarijatni model.

Poglavlje 3

Dodatak

3.1 Korišteni SAS kod

```
data heart_nova;
input dob spol bol tlak kolesterol secer EKG puls ex_angina
STdepresija STnagib krvne_zile talij_test bolest;
cards;
```

Baza podataka korištena pri analizi dostupna je na web stranici https://www.kaggle.com/cherngs/heart-disease-cleveland-uci.

```
/*Deskriptivna statistika za numeričke varijable:*/
title "Deskriptiva za numeričke";
proc means data=heart_nova n mean std min median max fw=8;
var dob tlak kolesterol puls STdepresija krvne_zile;
run;

/*Deskriptivna statistika za kategorijske varijable:*/
title "Deskriptiva za kategorijske";
proc freq data=heart_nova;
table spol bol secer EKG ex_angina STnagib talij_test bolest;
run;

title "Modeliranje varijable bol pomocu dummy varijabli,
baseline= 2-neanginalna bol";
data dummy; set heart_nova;
if bol="0" then do; bol_tip=1;bol_atip=0;bol_asimp=0; end;
```

```
if bol="1" then do; bol_tip=0;bol_atip=1;bol_asimp=0; end;
if bol="2" then do; bol_tip=0;bol_atip=0;bol_asimp=0; end;
if bol="3" then do; bol_tip=0;bol_atip=0;bol_asimp=1; end;
run;

title "Modeliranje varijable EKG pomocu dummy varijabli,
baseline= 0-uredan nalaz";
data dummy; set dummy;
if EKG="0" then do; EKG_abnorm=0;EKG_hiper=0; end;
if EKG="1" then do; EKG_abnorm=1;EKG_hiper=0; end;
if EKG="2" then do; EKG_abnorm=0;EKG_hiper=1; end;
run;

title "Modeliranje varijable STnagib pomocu dummy varijabli,
baseline= 1-ravan";
data dummy; set dummy;
if STnagib="0" then do; ST_uzdignut=1;ST_spusten=0; end;
if STnagib="1" then do; ST_uzdignut=0;ST_spusten=0; end;
if STnagib="2" then do; ST_uzdignut=0;ST_spusten=1; end;
run;

title "Modeliranje varijable talij_test pomocu dummy varijabli,
baseline= 0-jednaka raspodjela";
data dummy; set dummy;
if talij_test="0" then do; talij_fiks=0;talij_reverz=0; end;
if talij_test="1" then do; talij_fiks=1;talij_reverz=0; end;
if talij_test="2" then do; talij_fiks=0;talij_reverz=1; end;
run;

/*Univarijatna logistička regresija:*/
title "Univarijatna logistička regresija - dob";
proc logistic data=heart_nova descending;
model bolest=dob /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - spol";
proc logistic data=heart_nova descending;
model bolest=spol /lackfit rsq outroc=rocgraf;
run;
```

```
title "Univarijatna logistička regresija - bol";
proc logistic data=dummy descending;
model bolest=bol_tip bol_atip bol_asimp /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - tlak";
proc logistic data=heart_nova descending;
model bolest=tlak /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - kolesterol";
proc logistic data=heart_nova descending;
model bolest=kolesterol /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - secer";
proc logistic data=heart_nova descending;
model bolest=secer /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - EKG";
proc logistic data=dummy descending;
model bolest=EKG_abnorm EKG_hiper /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - puls";
proc logistic data=heart_nova descending;
model bolest=puls /lackfit rsq outroc=rocgraf;
run;

title"Univarijatna logistička regresija - ex_angina";
proc logistic data=heart_nova descending;
model bolest=ex_angina /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - STdepresija";
proc logistic data=heart_nova descending;
model bolest=STdepresija /lackfit rsq outroc=rocgraf;
```

```
run;

title "Univarijatna logistička regresija - STnagib";
proc logistic data=dummy descending;
model bolest=ST_uzdignut ST_spusten /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - krvne_zile";
proc logistic data=heart_nova descending;
model bolest=krvne_zile /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - talij_test";
proc logistic data=dummy descending;
model bolest=talij_fiks talij_reverz /lackfit rsq outroc=rocgraf;
run;

title "Crtanje vjerojatnosti p"; /*za dob*/
proc logistic data=heart_nova descending;
model bolest=dob;
output out=crtanje predicted=prob xbeta=logit;
run;
title "grafički prikaz vjerojatnosti";
proc sgplot data=crtanje;
scatter x=dob y=prob;
run;

/*za puls*/
proc logistic data=heart_nova descending;
model bolest=puls; output out=crtanje predicted=prob xbeta=logit;
run;
title "grafički prikaz vjerojatnosti";
proc sgplot data=crtanje;
scatter x=puls y=prob;
run;

/*za STdepresija*/
proc logistic data=heart_nova descending;
model bolest=STdepresija;
```

```
output out=crtanje predicted=prob xbeta=logit;
run;
title "grafički prikaz vjerojatnosti";
proc sgplot data=crtanje;
scatter x=STdepresija y=prob;
run;

/*Multivarijatna logistička regresija*/
title "Multivarijatna logistička regresija";
proc logistic data=dummy descending;
model bolest=dob spol bol_tip bol_atip bol_asimp tlak EKG_abnorm
EKG_hiper puls ex_angina STdepresija ST_uzdignut ST_spusten
krvne_zile talij_fiks talij_reverz /lackfit rsq outroc=rocgraf;
run;

/*Multivarijatna logistička regresija - STEPWISE procedura*/
title "Multivarijatna logistička regresija - STEPWISE";
proc logistic data=dummy descending;
model bolest=dob spol bol_tip bol_atip bol_asimp tlak EKG_abnorm
EKG_hiper puls ex_angina STdepresija ST_uzdignut ST_spusten
krvne_zile talij_fiks talij_reverz /selection=stepwise outroc=outroc;
run;

title "Multivarijatna logistička regresija sa značajnim
varijablama iz Stepwise-a";
proc logistic data=dummy descending;
model bolest=spol bol_tip bol_atip bol_asimp tlak ex_angina
ST_uzdignut ST_spusten krvne_zile talij_fiks talij_reverz
/lackfit rsq outroc=rocgraf;
run;
```

Bibliografija

[1] P. D. Allison, *Logistic Regression Using SAS: Theory and Application*, SAS Institute Inc., Cary, NC, USA, 1999.

[2] L. K. Hermann, S. D. Weingart, Y. M. Yoon, N. G. Genes, B. P. Nelson, P. L. Shearer, W. L. Duvall, M. J. Henzlova, *Comparison of frequency of inducible myocardial ischemia in patients presenting to emergency department with typical versus atypical or nonanginal chest pain*, dostupno na `https://pubmed.ncbi.nlm.nih.gov/20494662/` (prosinac, 2020.)

[3] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression (Second Edition)*, John Wiley and Sons, New York, USA, 2000.

[4] M. Huzak, *Matematička statistika*, PMF-MO, nastavni materijali, 2018.

[5] A. Jazbec, *Odabrane statističke metode u biomedicini*, PMF-MO, nastavni materijali, 2019.

[6] *MSD priručnik dijagnostike i terapije*, dostupno na `http://www.msd-prirucnici.placebo.hr/msd-prirucnik/kardiologija/kardioloske-dijagnosticke-pretrage/elektrokardiografija` (prosinac, 2020.)

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

Sažetak

U ovome radu opisane su glavne značajke logističkog regresijskog modela te njegova usporedba s linearnom regresijom. Predstavljen je način procjene i interpretacije parametara u modelu te odabir najadekvatnijeg modela. Zatim je u drugom dijelu ta teoretska podloga primijenjena na konkretnim podacima. Točnije, baza podataka sastoji se od 297 opservacija koje predstavljaju pacijente i 14 varijabli koje predstavljaju određene podatke o njihovu zdravstvenom stanju na prijemu u bolnicu. Na temelju toga, cilj je bio procijeniti koje varijable i koliko statistički značajno utječu na pojavu bolesti srca kod osobe. Pri izradi najadekvatnijeg modela korištene su univarijatna i multivarijatna logistička regresija te procedura Stepwise. Dobiveno je da na razvoj bolesti srca najviše utječu spol, tip boli u prsima, krvni tlak u mirovanju, vršni nagib ST-segmenta u pokretu, broj vidljivih glavnih krvnih žila tijekom fluoroskopije, rezultati talij testa te angina izazvana tjelesnom aktivnošću.

Summary

This thesis describes the main features of the logistic regression model and its comparison with linear regression. It presents the method of estimation and interpretation of model parameters as well as the selection of the most adequate model. In the second part, this theoretical basis was applied to concrete data. Specifically, the database consists of 297 observations representing patients and 14 variables representing information about their health status at hospital admission. Based on this, the aim was to assess which variables have statistically significant effect on the occurrence of heart disease. Univariate and multivariate logistic regression and Stepwise procedure were used in the development of the most adequate model. The results have shown that heart disease is mostly affected by gender, chest pain type, resting blood pressure, the slope of the peak exercise ST-segment, number of major vessels colored by fluoroscopy, thallium test results and exercise induced angina.

Životopis

Rođena sam 20. siječnja 1995. godine u Zagrebu, gdje sam i pohađala osnovnu školu Vrbani te I. gimnaziju (opću). Nakon srednjoškolskog obrazovanja, akademske godine 2013./2014., upisala sam preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu, koji sam kao prvostupnica završila 2018. godine. Na istome fakultetu upisala sam diplomski sveučilišni studij, smjer Matematička statistika, akademske godine 2018./2019. Tijekom preddiplomskog studija povremeno sam davala instrukcije iz matematike učenicima osnovne i srednje škole. Tijekom ljeta 2019. radila sam u A1 Hrvatska (studentska stručna praksa iz područja rada *Data scientista*).