Modeliranje dijabetesa pomoću logističke regresije s nominalnom zavisnom varijablom

Dodik, Anto

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Anto Dodik

MODELIRANJE DIJABETESA POMOĆU LOGISTIČKE REGRESIJE S NOMINALNOM ZAVISNOM VARIJABLOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, srpanj 2015.

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. _____

Veliko hvala mojim roditeljima na strpljenju i ljubavi.
Hvala svima koji su vjerovali u mene.

Sadržaj

Uvod

Metoda ispitivanja i analize ovisnosti jedne varijable (zavisne) o jednoj ili više drugih (nezavisnih) varijabli naziva se regresijska analiza. Linearna i logistička regresija najpopularnije su metode u regresijskoj analizi. Glavna razlika između te dvije metode je u klasifikaciji zavisne varijable, koja je kategorijska kod logističke regresije, dok je kod linearne regresije ona kontinuirana.

Kako je čest slučaj da je ishod zavisne varijable diskretan, u posljednjem desetljeću logistička regresija postala je neizostavna metoda u biomedicini, biomatematici, kemiji, ekonomiji i općenito u statistici. Korijeni logističke regresije sežu još od 19. stoljeća, kada je *Pierre-Francois Verhulst* proučavao logističku funkciju za modeliranje rasta populacije. Logistička funkcija ponovno je otkrivena 1920. godine od strane *Raymonda Pearla* te *Lowella Reeda*, neovisno o radu *Verhulsta*. Nakon pronalaska probit funkcije, inverzne funkcije od funkcije distribucije standardne normalne razdiobe, uslijedila je pojava logit funkcije. Logit je inverzna funkcija od logističke funkcije. Pokazalo se da logit funkcija ima jasnu prednost u praktičnom aspektu računanja nad probit funkcijom. *Daniel McFadden* je 1973. godine prvi povezao logističku funkciju sa teorijom diskretnog izbora (eng. Theory of discrete choice). To je osiguralo teorijsku osnovu logističkog regresijskog modela.

U ovom radu upravo ćemo proučavati logističku regresiju i pripadni logistički model te njihovu konkretnu primjenu.

U prvom poglavlju upoznat ćemo se sa pojmom linearne regresije, gdje ćemo obraditi linearni regresijski model te navesti mjere adekvatnosti takvog modela. Nakon osnovnih pojmova o logističkoj regresiji, u drugom poglavlju, objasnit ćemo metodu maksimalne vjerodostojnosti, omjer šanse te izvesti pripadajući logistički regresijski model. Treće poglavlje sadrži konkretnu primjenu na modeliranju dijabetesa logističkom regresijom s nominalnom zavisnom varijablom.

Poglavlje 1

Linearna regresija

1.1 Pojam linearne regresije

Jedan od rezultata svake regresijske analize je regresijski model. Regresijski model je matematička jednadžba koja definira, tj. kvantificira povezanost između zavisne varijable s nezavisnim. Ako je povezanost između zavisne i nezavisnih varijabli linearna, govorimo o linearnoj regresiji i pripadajućem linearnom modelu. Linearna regresija je bila prvi tip regresijske analize koja je detaljno proučavana i koja se ekstenzivno koristila u praktičnim primjenama. Razlog za ovo je taj, što se modeli kod kojih imamo linearnu ovisnost između zavisne i nezavisnih varijabli lakše modeliraju i interpretiraju nego modeli sa nelinearnom ovisnošću. Također, statistička svojstva rezultirajućih procjenitelja se lakše određuju.

Ideja regresije je da na temelju izmjerenih ili prikupljenih podataka napravimo matematički model kojem je cilj predviđanje ili prognoza budućih vrijednosti. Promotrimo sljedeću situaciju. Imamo jednu ili više kontroliranih varijabli $x_1, x_2 \ldots$ i slučajnu veličinu Y mjerenu u ovisnosti o $X = (x_1, x_2, \ldots)$. Y zovemo zavisna varijabla – varijabla koju želimo opisati ili procijeniti. X je nezavisna varijabla – varijabla pomoću koje želimo opisati zavisnu varijablu.

Ovdje se linearna regresija koristi za podešavanje prediktivnog modela prema promatranom skupu podataka vrijednosti X i Y. Nakon razvoja ovakvog modela, ako je dana vrijednost za X bez pripadajuće vrijednosti Y, podešeni model će se koristiti za predviđanje vrijednosti Y.

1.2 Linearni regresijski model

Linearni model ovisnosti veličine Y o $X = (x_1, x_2, \ldots, x_k)$ zadan je sa:

$$Y = \theta_0 + \theta_1 x_1 + \ldots + \theta_k x_k + \xi, \tag{1.1}$$

gdje je ξ slučajna varijabla koju zovemo pogreška, a $\theta = (\theta_0, \theta_1, \ldots, \theta_k)^\tau$ je parametar modela.

Neka je sproveden niz od n nezavisnih mjerenja veličine Y za zadane (ne nužno sve različite) vrijednosti od X te neka je dobiven niz realizacija (x_{ij}, Y_i), za $i = 1, \ldots, n$ te $j = 1, \ldots, k$. Tada imamo model:

$$Y_i = \theta_0 + \sum_{j=1}^{k} \theta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.2}$$

gdje su $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ slučajne pogreške za koje vrijede Gauss-Markovljevi uvjeti.
Kraće to zapisujemo u matričnom obliku:

$$\mathbf{Y} = \mathbf{X}\theta + \varepsilon,$$

gdje je $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\tau$, $\theta = (\theta_0, \theta_1, \ldots, \theta_k)^\tau$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\tau$ i

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}.$$

Cilj je minimizirati funkciju

$$L(\theta) = \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 x_{i1} - \ldots - \theta_k x_{ik})^2, \tag{1.3}$$

da bi mogli procijeniti parametre modela $\theta_0, \theta_1, \ldots, \theta_k$.

Iz relacije (1.3) vidimo da zapravo minimiziramo sumu kvadriranih reziduala, tj. minimiziramo

$$L(\theta) = \sum_{i=1}^{n} \varepsilon_i^2. \tag{1.4}$$

Zapišimo relaciju (1.3) u obliku norme[1] :

$$L(\theta) = \|\mathbf{Y} - \mathbf{X}\theta\|^2 = \|\varepsilon\|^2$$
$$L(\theta) \longrightarrow \min .$$

Uz uvjet da je $\mathbf{X}^\tau\mathbf{X}$ regularna, dobivamo da je najbolja ocjena za θ :

$$\widehat{\theta} = (\mathbf{X}^\tau\mathbf{X})^{-1}\mathbf{X}^\tau\mathbf{Y}. \tag{1.5}$$

$\widehat{\theta}$ je procjenitelj za $\theta = (\theta_0, \theta_1, \ldots, \theta_k)^\tau$.

Ovaj postupak procjene parametara modela naziva se metoda najmanjih kvadrata (LS[2]).

Čak štoviše, lako se pokaže (jer vrijede Gauss-Markovljevi uvjeti) da je $\widehat{\theta}$ nepristran procjenitelj za $\theta = (\theta_0, \theta_1, \ldots, \theta_k)^\tau$ (vidi definiciju 2.3.5.).

Jedan od osnovnih rezultata linearne regresije, osim prediktorske uloge, je da procijenjeni regresijski parametri (lako vidimo iz samog modela) ukazuju koliko će se promijeniti vrijednost zavisne varijable ako se nezavisna varijabla poveća za jednu jedinicu. Ta promjena će biti upravo vrijednost procijenjenog parametra, u smislu povećanja, odnosno smanjenja zavisne varijable za tu vrijednost. To je jako korisna činjenica u različitim studijama i analizama, gdje nam je od značaja takozvani monitoring zavisne varijable, a posebno kod primjene logističke regresije, gdje je princip zaključivanja ipak malo drugačiji. No, o tome ćemo nešto više reći u *poglavlju 2*.

[1] Euklidska norma
[2] eng. Least Squares

1.3 Osnovni pojmovi u odabiru adekvatnog modela

Vidjeli smo da se rezidualna odstupanja koriste za mjerenje prilagođenosti regresije opažanjima iz uzorka. Naime, nakon što se odredi procijenjeni regresijski pravac postavlja se pitanje je li izračunata regresija dobra. Odgovor na to pitanje nam daje analiza varijance. Općenito se smatra da je regresija dobro prilagođena opažanjima iz uzorka ako je velik dio proporcije varijance Y protumačen modelom. Ako je odstupanje protumačeno modelom (odstupanje odgovarajuće procijenjene ili regresijske vrijednosti od prosjeka) dosta veće od neprotumačenog (rezidualno odstupanje), tada je regresijski model položen „dosta blizu" opažanjima iz uzorka. Još jedna mjera adekvatnosti modela je koeficijent determinacije, u oznaci R^2. On nam govori koliko je od ukupne varijabilnosti zavisne varijable objašnjeno samim modelom. Isti poprima vrijednosti na intervalu $[0, 1]$, a promatrani je model to reprezentativniji što je koeficijent determinacije bliži jedinici. Također valja odrediti statističku značajnost nezavisnih varijabli u modelu. To radimo pomoću univarijatne regresijske analize (jedna nezavisna varijabla u modelu) te pomoću multivarijatne regresijske analize (barem dvije nezavisne varijable u modelu). No, nisu sve nezavisne varijable uvijek pogodne u odabiru najboljeg modela. Pitanje je kako odrediti koje nezavisne varijable najbolje opisuju zavisnu varijablu.

Cilj nam je da je kombinacija nezavisnih varijabli jako korelirana sa zavisnom, ali isto tako da su one međusobno nekorelirane. Ako su nezavisne varijable međusobno korelirane, može doći do određenih problema. Naime, ako varijabla koja prva uđe u model objasni svu varijabilnost zavisne varijable, tada se utjecaj njoj korelirane varijable gubi. Dakle, izbor nezavisnih varijabli u modelu je jako osjetljiv. Ubacivanje ili izbacivanje bilo koje varijable može jako promijeniti jednadžbu ravnine. Imamo tri procedure koje služe sa selekciju nezavisnih varijabli. To su *Forward*, *Backward* i *Stepwise* procedura.

Ugrubo, ovako izgleda opis njihovih algoritama:

Forward – Takozvana selekcija unaprijed. Ova metoda počinje bez ijedne varijable u modelu. U svakom sljedećem koraku dodaje se po jedna varijabla koja maksimizira adekvatnost modela.

Backward – Takozvana eliminacija odostraga. Metoda počinje cijelim modelom i eliminira iz modela varijablu po varijablu koja ima najmanji doprinos adekvatnosti modela.

Stepwise – Kombinacija forwarda i backwarda. Počinje kao forward samo što varijabla koja je ušla u model, ne mora tamo i ostati.

Stepwise procedura se pokazala kao najpouzdanija te se ona većinom i upotrebljava. Nju ćemo koristiti i u ovom radu kod traženja najadekvatnijeg modela logističkom regresijom u *poglavlju 3*.

Ukoliko je nezavisna varijabla ordinalna ili nominalna, vrlo često je korisno da se za procjenu parametara modela koriste tzv. dummy varijable. Kod ordinalnih varijabli gdje postoji uređaj kodiranja brojem, regresija će procijeniti parametar tako da prelazak nezavisne varijable iz niže u višu kategoriju mijenja vrijednost zavisne varijable za vrijednost parametra. Međutim, ukoliko ovisnost nezavisne i zavisne varijable nije linearna ili pak razmak između kategorija nije ekvidistantan, kao i kod svih nominalnih varijabli gdje ne postoji uređaj između kategorija, treba koristiti dummy varijable kako bi se dobro procijenio odgovarajući regresijski parametar. U *poglavlju 3* vidjet ćemo kodiranje dummy varijabli pomoću statističkog programa **SAS**.

Poglavlje 2

Logistička regresija

2.1 Pojam logističke regresije

Regresijsku analizu u kojoj je zavisna varijabla kategorijska zovemo logistička regresija. Zavisna kategorijska varijabla u logističkoj regresiji je najčešće dihotomna, tj. ima samo dvije kategorije. Takva logistička regresija zove se i binarna logistička regresija. Primarna zadaća takve regresije je procijeniti omjer šanse prelaska zavisne varijable iz jedne kategorije u drugu, uz određen pomak nezavisne varijable. Općenitiji slučaj je kada je zavisna varijabla dana sa više kategorija: ordinalna ili nominalna.

Logistička regresija se najčešće koristi kada relacija između zavisne i nezavisne varijable nije linearna. Ista je pogodna za rješavanje problema kada su u pitanju demografske varijable, jer su one uglavnom kategorijske (bračni status, zanimanje, lokacija, itd.).

Problemi ove vrste se mogu riješiti i preko višestruke linearne regresije, tako što bi dvije vrijednosti varijable obilježili sa dva cijela broja, obično sa 0 i 1. Dobili bismo regresijski model koji bi mogao predvidjeti vrijednost zavisne varijable, zajedno sa regresijskim koeficijentima, koji bi pokazivali relativni utjecaj svake nezavisne varijable. Ipak, logistička odnosno logit regresija je adekvatnije rješenje, jer bismo vrlo lako mogli dobiti jako loš model.

Potrebno je izvršiti određenu vrstu matematičke transformacije da bi se dobio logistički regresijski model. Nešto više o tome ćemo opisati u *podpoglavlju 2.4*, no prvo navedimo razlike i sličnosti sa linearnom regresijom.

2.2 Razlike i sličnosti između linearne i logističke regresije

Vidjeli smo u samom opisu logističke regresije da se ona razlikuje od linearne u tome što zavisna varijabla ne može biti kontinuirana. Također, vidjet ćemo da povećanje nezavisne varijable za jednu jedinicu kazuje koliki je omjer šanse za prelazak zavisne varijable iz jedne kategorije u drugu.

Ostali aspekti regresijske analize identični su kao i kod obične regresije: ocjenjivanje statističke značajnosti nezavisnih varijabli, koreliranost, stepwise regresija, rješavanje problema nezavisnih varijabli koje su kategorijske sa više od dvije kategorije, rješavanje problema vrijednosti koje nedostaju, itd.

Napomenimo još da metoda najmanjih kvadrata (*LS*) kod logističke regresije ne funkcionira te zbog toga koristimo metodu maksimalne vjerodostojnosti (*ML*[1]) za procjenu parametara modela. Upravo tu metodu obradit ćemo u idućem podpoglavlju.

2.3 Metoda maksimalne vjerodostojnosti

Ovdje ćemo obraditi osnovnu teorijsku podlogu metode maksimalne vjerodostojnosti u matematičkoj statistici.

Definicija 2.3.1. *Neka je* $(\Omega, \mathcal{F})$ *izmjeriv prostor i* $\mathcal{P}$ *familija vjerojatnosnih mjera na* $(\Omega, \mathcal{F})$*. Tada je uređena trojka* $(\Omega, \mathcal{F}, \mathcal{P})$ ***statistička struktura****.*

Napomena. *Ako je* $\mathcal{P}$ *jednočlana familija, tada je statistička struktura vjerojatnosni prostor.*

Definicija 2.3.2. ***n-dimenzionalni slučajni uzorak*** *na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je niz* $X_1, X_2, \ldots, X_n$ *slučajnih varijabli (vektora) na* $(\Omega, \mathcal{F})$ *takvih da su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost* $\mathbb{P} \in \mathcal{P}$*.*

Definicija 2.3.3. ***Statistika*** *na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *je svaka slučajna varijabla (vektor)* $T : \Omega \to \mathbb{R}^d$ *takva da postoji* $n \in \mathbb{N}$ *i n-dimenzionalni slučajni uzorak* $(X_1, \ldots, X_n)$ *na* $(\Omega, \mathcal{F}, \mathcal{P})$ *te izmjerivo preslikavanje* $t : \mathbb{R}^n \to \mathbb{R}^d$ *takvo da je* $T = t(X_1, \ldots, X_n)$*.*

[1] eng. Maximum Likelihood

Iz definicije 2.3.3. zapravo vidimo da je ***statistika*** funkcija slučajnog uzorka.

Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ slučajni uzorak iz modela $\mathcal{P} = \{f(\cdot;\theta)\colon \theta \in \Theta\}$. Na osnovu zadanog uzorka želimo procijeniti vrijednost parametra θ, ili općenito, neke njegove funkcije $\tau(\theta) \in \tau(\Theta) \subseteq \mathbb{R}^k$.

Definicija 2.3.4. ***Procjenitelj*** *od $\tau(\theta)$ je statistika $T = t(\mathbb{X}) = t(X_1, \ldots, X_n)$ u $\mathbb{R}^k$.*

Definicija 2.3.5. *Procjenitelj $T = t(\mathbb{X})$ za $\tau(\theta) \in \mathbb{R}$ je* ***nepristran*** *za $\tau(\theta)$ ako vrijedi*

$$\mathbb{E}_\theta[T] = \tau(\theta),\ \theta \in \Theta.$$

Ako je $\mathbf{x} = (x_1, \ldots, x_n)$ realizacija slučajnog uzorka $\mathbb{X}$, tada je **vjerodostojnost** funkcija

$$L\colon \Theta \to \mathbb{R}, \quad L(\theta \mid \mathbf{x}) = L(\theta) := f_{\mathbb{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Definicija 2.3.6. *Statistika $\widehat{\theta} = \widehat{\theta}(\mathbb{X})$ je procjenitelj* ***maksimalne vjerodostojnosti*** *za θ (**MLE**[2]) ako vrijedi*

$$L(\widehat{\theta} \mid \mathbb{X}) = \max_{\theta\in\Theta} L(\theta \mid \mathbb{X}). \tag{2.1}$$

Primjer. *Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ slučajni uzorak iz modela $U(0,\theta)$, $\theta > 0$ i $\mathbf{x}$ jedna realizacija od $\mathbb{X}$. Tada je vjerodostojnost*

$$L(\theta) := \prod_{i=1}^{n} \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbb{1}_{[x_{(n)},\infty\rangle}(\theta).$$

Zato vidimo da je $\widehat{\theta}(\mathbf{x}) = x_{(n)}$ procjena maksimalne vjerodostojnosti pa je $\widehat{\theta}(\mathbb{X}) = X_{(n)}$ MLE za θ.

Primjer. *Neka je $\mathbb{X} = (X_1, \ldots, X_n)$ slučajni uzorak iz modela $N(\mu,\sigma^2)$, $\theta = (\mu,\sigma^2) \in \mathbb{R} \times \langle 0,\infty\rangle =: \Theta$, te neka je $\mathbf{x}$ realizacija od $\mathbb{X}$. Nađimo MLE za θ. Vjerodostojnost je dana s*

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{(\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Budući da je prirodni logaritam strogo rastuća diferencijabilna funkcija, dovoljno je naći globalni ekstrem funkcije $l(\theta) = \log L(\theta)\colon \Theta \to \mathbb{R}$,

$$l(\theta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi).$$

[2] eng. Maximum Likelihood Estimator

Nađimo njene stacionarne točke. Imamo

$$\frac{\partial l}{\partial \mu}(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \Rightarrow \widehat{\mu} = \overline{x},$$

$$\frac{\partial l}{\partial \sigma^2}(\mu, \sigma^2) = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0 \Rightarrow \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n}s^2.$$

$$\Rightarrow \widehat{\theta}(\mathbf{x}) = \left(\overline{x}, \frac{n-1}{n}s^2\right).$$

Hesseova matrica funkcije l je

$$\text{Hess } l(\mu, \sigma^2) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu) & -\frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{n}{2(\sigma^2)^2} \end{bmatrix},$$

a kako je

$$\text{Hess } l(\widehat{\mu}, \widehat{\sigma^2}) = \begin{bmatrix} -\frac{n}{\widehat{\sigma^2}} & 0 \\ 0 & -\frac{n}{2(\widehat{\sigma^2})^2} \end{bmatrix}$$

negativno definitna matrica, $\widehat{\theta}$ je lokalni maksimum funkcije l. No, $\widehat{\theta}$ je i globalni maksimum funkcije l (npr. za svaki fiksirani σ^2 je $l(\mu, \sigma^2)$ konkavna funkcija definirana na konveksnom skupu sa stacionarnom točkom $\widehat{\mu}$ pa je $\widehat{\mu}$ globalni maksimum te funkcije) pa zaključujemo da je

$$\widehat{\theta}(\mathbb{X}) = \left(\overline{X}_n, \frac{n-1}{n}S_n^2\right)$$

MLE za θ.

Iz zadnjeg primjera vidimo algoritam prema kojem prvo iz empirijskih podataka i pripadajućih funkcija gustoće nalazimo produkt istih, koji zatim logaritmiramo te parcijalno deriviramo. Na taj način nalazimo procjenitelj maksimalne vjerodostojnosti za parametar modela.

2.4 Logistički regresijski model

2.4.1 Omjer šansi

Pojmove šansa (*eng. odds*) i omjer šansi (*eng. odds ratio*) objasnit ćemo na primjeru. Promotrimo sljedeću tablicu.

Tablica 2.1: Ocjena rizika

		Bolestan	Zdrav	Ukupno
Izloženost faktoru	Pušač	a	b	a + b
	Nepušač	c	d	c + d
Ukupno		a + c	b + d	a + b + c + d

Neka je p vjerojatnost nekog događaja, u našem slučaju vjerojatnost imanja neke bolesti. Šansa (*odds*) da razvijemo tu bolest je omjer između vjerojatnosti da ju imamo i vjerojatnosti da ju nemamo, tj.

$$odds = \frac{p}{1-p}. \tag{2.2}$$

Na primjer, ako je vjerojatnost da imamo bolest 0.7, onda je šansa da razvijemo bolest $2.\dot{3}$, tj. $odds = 0.7/0.3 = 2.\dot{3}$.

Zašto nam treba *odds*? Zbog razumnijeg višestrukog uspoređivanja. Primjerice, ako je vjerojatnost da osoba A ima bolest 0.4, a osoba B 0.8, razumno je zaključiti da je vjerojatnost od osobe A duplo veća od vjerojatnosti osobe B. Međutim, ako je vjerojatnost da osoba A ima bolest 0.8, onda bi vjerojatnost osobe B trebala biti 1.6, ali to nije moguće, jer najveća vjerojatnost može biti 1. Zato je za uspoređivanje puno razumnije koristiti *odds*.

Omjer šansi (*odds ratio*), u našem slučaju, je *odds* izloženih kroz *odds* neizloženih rizičnom faktoru. *Odds* izloženih je omjer vjerojatnosti izloženih rizičnom faktoru da imaju bolest i vjerojatnosti da ju nemaju. Analogno vrijedi za *odds* neizloženih rizičnom faktoru. Preciznije, imamo sljedeću situaciju:

$$OR = \frac{\frac{p_{izlozeni}}{1-p_{izlozeni}}}{\frac{p_{neizlozeni}}{1-p_{neizlozeni}}}.$$

Iz tablice (2.1) slijedi

$$OR = \frac{a/(a+b)}{b/(a+b)} \Big/ \frac{c/(c+d)}{d/(c+d)}.$$

$$\Rightarrow OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}. \tag{2.3}$$

Primjer. *(*Odds *i* odds ratio*)*

Tablica 2.2: Primjer

		Bolestan	Zdrav	Ukupno
Izloženost faktoru	Pušač	16	7	23
	Nepušač	10	27	37
Ukupno		26	34	60

odds(pušači) = $16/7 = 2.286$. Pušači imaju 2.286 puta veću šansu da dobiju bolest nego da ju ne dobiju.

odds(nepušači) = $10/27 = 0.370$. Nepušači imaju 0.370 puta veću šansu da dobiju bolest nego da ju ne dobiju. Dakle, imaju manju šansu da dobiju bolest.

$OR = \frac{16 \cdot 27}{7 \cdot 10} = 6.171$. Pušači imaju 6.171 puta veći omjer šanse da dobiju bolest.

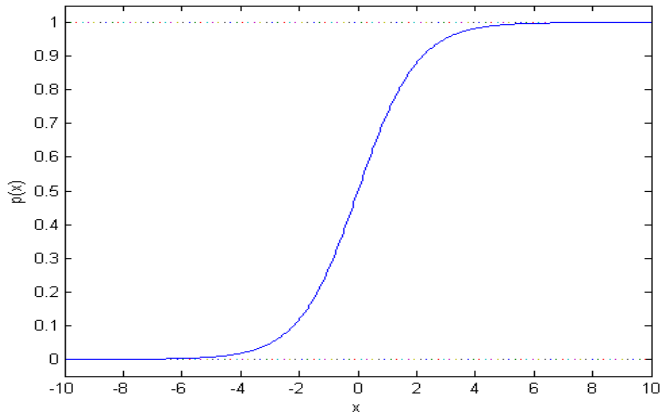
2.4.2 Logistički regresijski model

Osnovna ideja logističke regresije je koristiti mehanizam već razvijen za linearnu regresiju modeliranjem vjerojatnosti p. Promatrat ćemo regresijski model sa zavisnom dihotomnom varijablom. Relacija između zavisne i nezavisne varijable je stoga nelinearna. Cilj je linearizirati takav model.

Neka je

$$p\colon \mathbb{R} \to (0,1), \quad p(x) := \frac{1}{1+e^{-x}}.$$

Tu funkciju zovemo logistička funkcija. Graf je S − krivulja (Slika 2.1). Kako poprima samo vrijednosti između 0 i 1, pogodna je za modeliranje raznih problema u biologiji, biomatematici, kemiji, ekonomiji, statistici, itd. Logistička funkcija se često primjenjuje za modeliranje rasta populacije, za što ju je izvorno koristio i *Pierre-Francois Verhulst* 1838. godine, pri čemu je stopa razmnožavanja razmjerna postojećoj populaciji i količini raspoloživih resursa.



Slika 2.1: Krivulja logističke funkcije

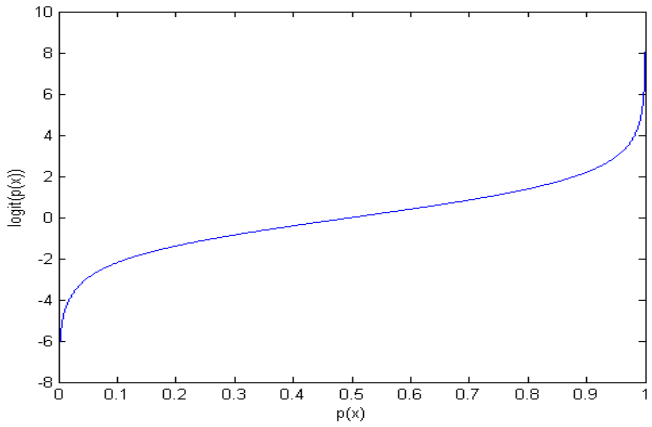
Dakle, $p(x)$ gledamo kao vjerojatnost nekog događaja. Neka je kao u relaciji 2.2

$$odds\,(x) = \frac{p(x)}{1-p(x)}. \tag{2.4}$$

Logit je inverzna funkcija od logističke funkcije:

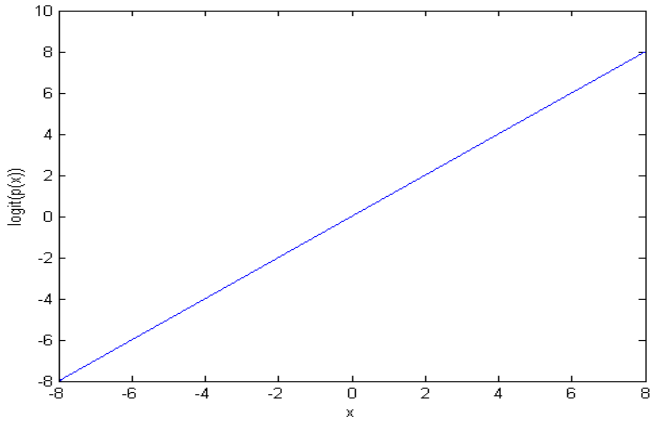
$$logit(p)\colon (0,1) \to \mathbb{R}, \quad logit(p(x)) := \ln\left(\frac{p(x)}{1-p(x)}\right), \quad p(x) \in (0,1). \tag{2.5}$$

Ta funkcija raspoređuje vjerojatnosti preko cijelog skupa realnih brojeva $\mathbb{R}$, a njezin graf vidimo na slici 2.2.



Slika 2.2: Krivulja logit funkcije

Transformacijom $\frac{p(x)}{1-p(x)}$, odnosno *odds(x)*, mičemo gornju granicu, a logaritmiranjem donju. Pripadna transformacija zove se *logit transformacija* te smo iz nelinearne relacije (Slika 2.1) dobili linearnu.



Slika 2.3: Linearna relacija

Dakle, naš logistički regresijski model izgleda:

$$logit(p(x)) = \ln(odds(x)) = \beta_0 + \beta_1 x. \tag{2.6}$$

Model je sličan linearnom regresijskom modelu, ali raspodjela je binomna, a ne normalna. Koeficijenti β_0 i β_1 se ne određuju pomoću metode najmanjih kvadrata, već pomoću metode maksimalne vjerodostojnosti.
Iz relacije 2.6 slijedi

$$odds(x) = \frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}. \tag{2.7}$$

Nadalje, iz relacije 2.7 slijedi

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}. \tag{2.8}$$

Model sa više prediktora (nezavisnih varijabli) $x_1, x_2, \ldots, x_k$:

$$logit(p(x)) = \ln(odds(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k. \tag{2.9}$$

Parametre modela procjenjujemo algoritmom opisanom u drugom primjeru *podpoglavlja 2.3*. Dakle, prvo iz empirijskih podataka i pripadajućih funkcija gustoće nalazimo produkt istih, koji zatim logaritmiramo te parcijalno deriviramo.

Pretpostavimo da imamo n nezavisnih observacija $(x_i, y_i),\ y_i \in \{0, 1\},\ i = 1, \ldots, n.$

Pripadna funkcija vjerodostojnosti (*eng. Likelihood*) dana je sa

$$L(\beta) = \prod_{i=1}^{n} \left[p(x_i)^{y_i} \cdot (1 - p(x_i))^{1 - y_i} \right], \tag{2.10}$$

gdje je $\beta = (\beta_0, \beta_1)$ parametar modela.

Funkcija $x \longrightarrow lnx$ je strogo rastuća pa je dovoljno maksimizirati funkciju $ln(L(\beta))$. Ta se funkcija u logističkoj regresiji zove *Log-likelihood* (*LL*). Dakle, dovoljno je maksimizirati funkciju

$$ln(L(\beta)) = LL(\beta) = \sum_{i=1}^{n} \left[y_i \, ln(p(x_i)) + (1 - y_i) \, ln(1 - p(x_i)) \right]. \tag{2.11}$$

Kada parcijalno deriviramo po β_0 i β_1 te izjednačimo sa nulom, dobijemo

$$\sum_{i=1}^{n} \left[y_i - p(x_i) \right] = 0 \quad i \quad \sum_{i=1}^{n} x_i \left[y_i - p(x_i) \right] = 0,$$

te nakon toga nalazimo ***MLE*** za $\beta = (\beta_0, \beta_1)$.

Adekvatnost modela (*eng. Goodness of fit*) ne testiramo analizom varijance kao kod linearne regresije. U logističkoj regresijskoj analizi koristi se *devijanca*, u oznaci [3]D. Ako je saturirani model dostupan (model sa teoretski savršenim „fitom"), odstupanje se izračunava usporedbom procijenjenog modela sa saturiranim modelom na sljedeći način:

$$D = -2ln\left[\frac{Likelihood\ proci jen jenog\ modela}{Likelihood\ saturiranog\ modela}\right], tj.$$

$$D = -2\sum_{i=1}^{n}\left[y_i\, ln\left(\frac{\hat{p}(x_i)}{y_i}\right) + (1-y_i)\, ln\left(\frac{1-\hat{p}(x_i)}{1-y_i}\right)\right] \approx \chi^2. \tag{2.12}$$

Sa aspekta logističke regresije u **SAS**-u, traženju najboljeg modela pristupamo numerički, tako da tražimo najmanje moguće odstupanje između opaženih i prediktivnih vrijednosti, koristeći iterativne računalne metode.

2.4.3 Interpretacija parametara modela

Parametar β_0 je neophodan za model, ali nema značaja za interpretaciju. On predstavlja vrijednost ln($odds$) kada je prediktor jednak 0. Promotrimo:

$g(x) := logit(p(x)) = \ln(odds(x)) = \beta_0 + \beta_1 x.$

$g(x+1) = \beta_0 + \beta_1(x+1).$

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

$logit(p(x+1)) - logit(p(x)) = \beta_1.$

$ln(odds\,(x+1)) - ln(odds\,(x)) = \beta_1.$

$$ln\,\frac{odds\,(x+1)}{odds\,(x)} = \beta_1.$$
$$\Rightarrow\ \frac{odds\,(x+1)}{odds\,(x)} = e^{\beta_1}.$$

[3] eng. Deviance

Izraz $\dfrac{odds\,(x+1)}{odds\,(x)}$ zapravo predstavlja *odds ratio* definiran u *podpoglavlju 2.4.1.*

Ako uzmemo da je i nezavisna varijabla dihotomna te joj pridružimo vrijednosti 0 i 1, imamo sljedeću situaciju:

$$za\ x = 1 \Rightarrow\ odds\,(1) = \frac{p(1)}{1-p(1)}.$$

$$za\ x = 0 \Rightarrow\ odds\,(0) = \frac{p(0)}{1-p(0)}.$$

$$g(1) - g(0) = ln\,\frac{odds\,(1)}{odds\,(0)} = ln\,(odds\ ratio\,(1,0)) = \beta_1.$$

$$\Rightarrow\ odds\ ratio\,(1,0) = e^{\beta_1}.$$

Upravo izračunati odnos za *odds ratio* i parametar β_1 možemo analizirati promatrajući primjer izloženosti rizičnom faktoru iz *podpoglavlja 2.4.1.* Stanja „Bolestan" i „Zdrav" možemo shvatiti kao vrijednosti zavisne varijable y (1 i 0) te izloženost faktoru „Pušač" i „Nepušač" možemo shvatiti kao vrijednosti nezavisne varijable x (1 i 0). *Odds ratio* u datom primjeru iznosi 6.171 te označava da pušači imaju 6.171 puta veći omjer šanse da dobiju bolest nego nepušači.

$$e^{\beta_1} = 6.171.$$

Sa aspekta logističke regresije to možemo protumačiti tako da prelaskom nezavisne (dihotomne) varijable iz niže u višu kategoriju se povećava omjer šanse za prelazak zavisne varijable iz niže u višu kategoriju za 6.171. Analogan zaključak vrijedi i ako je nezavisna varijabla kontinuirana, samo tada gledamo povećanje nezavisne varijable za jednu jedinicu. No, vrlo često je kod kontinuiranih varijabli pomak za jednu jedinicu biološki neznačajan. Na primjer, pomak za 1 mmHg sistoličkog tlaka je premali da bi biološki nešto predstavljao, ali pomak za 10 mmHg je već interesantniji.

Ako imamo povećanje kontinuirane nezavisne varijable za proizvoljan $c \in \mathbb{R}$, onda imamo sljedeći odnos između *odds ratio* i parametra β_1 :

$$odds\ ratio\,(x+c, x) = e^{c\beta_1}.$$

To znači da povećanje nezavisne varijable za c, povećava omjer šanse za prelazak zavisne varijable u višu kategoriju za $e^{c\beta_1}$.

Poglavlje 3

Modeliranje dijabetesa logističkom regresijom s nominalnom zavisnom varijablom pomoću statističkog programa SAS

3.1 Opis problema

U ovom poglavlju ćemo logističkom regresijom obraditi podatke koji su prikupljeni na 80 osoba. Ti podaci nalaze se u bazi koja se koristi na predavanjima iz kolegija *Odabrane statističke metode u biomedicini* na *PMF-MO*. Za 80 osoba, izmjerene su vrijednosti sistoličkog i dijastoličkog tlaka, hemoglobina, glukoze, leukocita, indeksa tjelesne mase (BMI[1]), itd. (18 varijabli). Također, za svaku osobu je evidentirano da li boluje od dijabetesa (tip 1 ili tip 2) ili ne boluje. To je spremljeno u varijablu *TIP*. Varijabla *TIP* je kategorijska i poprima tri vrijednosti: 0, 1 i 2. Zbog lakšeg razumijevanja nazvat ćemo te kategorije: *TIP 0*, *TIP 1* i *TIP 2*. *TIP 0* označuje da osoba nema dijabetes. *TIP 1* označuje da osoba boluje od dijabetesa tip 1, dok *TIP 2* označuje da osoba boluje od dijabetesa tip 2. Kako ta dva tipa dijabetesa nisu u nikakvoj međusobnoj vezi, tj. ne može se iz jednog stupnja dijabetesa preći u drugi, to znači da je kategorijska varijabla *TIP* nominalna. U ovoj analizi želimo pronaći model koji najbolje predviđa pojavnost dijabetesa tip 1, odnosno tip 2, u ovisnosti o ostalim izmjerenim vrijednostima (nezavisne varijable).

[1] eng. Body Mass Index

Navedimo prvo neke osnovne činjenice o oba tipa dijabetesa. Dijabetes tip 1 je rjeđi od tipa 2 i predstavlja svega 7 do 10% svih slučajeva dijabetesa. Tip 1 se može pojaviti u bilo kojoj dobi, ali obično se javlja u razdoblju između djetinjstva i 30. godine života. To je bolest koja nastaje disfunkcijom u organizmu, točnije, potpunim nedostatkom inzulina i traje cijeli život. Nedostatak inzulina uzrokovan je oštećenjem beta stanica gušterače te gubitkom njihove funkcije. Pojavljivanje dijebetesa tip 1 povezuje se također i sa godišnjim dobom, virusnim epidemijama, stresnim događajima te još nekim čimbenicima rizika, za sada znanstveno nedokazanog djelovanja. Dijabetes tip 2 najčešći je oblik dijabetesa, a ujedno i jedno od najčešćih oboljenja suvremenog društva te rapidno rastući javnozdravstveni problem. Za pandemijske razmjere obolijevanja od dijabetesa tip 2 krive su životne navike – pretežno sjedilački način života, manjak tjelesne aktivnosti, prevelik svakodnevni energetski unos i prehrana osiromašena funkcionalnim sastojcima, porast pretilosti, produženje životnog vijeka i sve veće količine nezaobilaznog stresa.

Iz navedenog je sasvim jasno da ne postoji uređaj između ova 2 tipa bolesti.

U prethodnim poglavljima upoznali smo se sa regresijskom analizom, gdje smo obradili pojmove linearne i logističke regresije. U teorijskoj obradi logističke regresije prvenstveno smo dali naglasak na logistički regresijski model sa dihotomnom zavisnom varijablom. Kako je varijabla *TIP* nominalna (tri kategorije) te ne možemo gledati prelazak iz jedne kategorije u drugu , a zanima nas šansa obolijevanja od dijabetesa tip 1, odnosno tip 2, dolazimo na ideju da promatramo dva slučaja logističke regresije sa dihotomnom zavisnom varijablom. Prvi slučaj, gdje ćemo promatrati zavisnu varijablu *TIP* sa vrijednostima *TIP 0* i *TIP 1* te drugi slučaj sa vrijednostima *TIP 0* i *TIP 2*. Na taj način ćemo za prikupljene podatke pokušati ustanoviti koji čimbenici utječu na vjerojatnost dobivanja dijabetesa tip 1 kod zdrave osobe, odnosno dobivanja tipa 2.

Za obradu prikupljenih podataka koristit ćemo statistički program **SAS**, u kojem ćemo primijeniti univarijatnu, odnosno multivarijatnu logističku regresiju, kao i proceduru *Stepwise* kako bi našli što bolje modele.

3.2 Univarijatna logistička regresija

Cilj ovog podpoglavlja je, za svaku od nezavisnih varijabli iz baze, provesti univarijatnu logističku regresiju. Na taj način ćemo odrediti statističku značajnost svake varijable zasebno, tj. statističku značajnost pripadnih (procijenjenih) parametara. Univarijatni logistički model, za svaku od nezavisnih varijabli, dan je relacijom 2.6 te izgleda:

$$logit(p(x)) = \ln(odds(x)) = \beta_0 + \beta_1 x.$$

Za parametar β_1 kažemo da je statistički značajan ako se on statistički značajno razlikuje od 0. Primjerice, ako je β_1 „približno" jednak 0, tada nezavisna varijabla nije statistički značajna za model, jer je njezin utjecaj na zavisnu varijablu zanemariv. Vidimo da preko parametra β_1 ispitujemo da li postoji statistički značajna veza između zavisne i nezavisne varijable. Preciznije, testiramo sljedeće hipoteze:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

Prvo trebamo zadati razinu značajnosti koja nam daje eliminatorni prag za određivanje statističke značajnosti parametra β_1, tj. pripadne nezavisne varijable. Najčešća razina značajnosti je 5%. Nakon toga koristimo p-vrijednost, kao mjeru odbacivanja, odnosno zadržavanja nulte hipoteze H_0. Ako je p-vrijednost manja od zadane razine značajnosti (5%), tada odbacujemo H_0 u korist alternative H_1 te kažemo da je nezavisna varijabla statistički značajna, na razini značajnosti od 5%.

Ovim pristupom moći ćemo u daljnjoj analizi suziti izbor nezavisnih varijabli koje najbolje opisuju modele, za slučaj 1, odnosno slučaj 2.

Slučaj 1 - TIP 0 i TIP 1

Ovdje analiziramo univarijatnu logističku regresiju sa zavisnom varijablom *TIP* koja prima dvije vrijednosti, *TIP 0* i *TIP 1*. Prvo unesimo cijelu bazu u **SAS**:

```
data baza;
input DOB TIP TRAJANJE CRP FIB HCY ADN
BMI BMI3$ BMI3kat SPOL RRS RRD HbA1c GUKb
LEUK Ac_UR AER;
cards;
 ----------------------------------------
proc print;
run;
```

Zbog povjerljivosti podataka nismo unijeli pripadne vrijednosti u ovaj rad.

Sada ćemo odvojiti podatke koji u varijabli *TIP* imaju vrijednosti *TIP 0*, odnosno *TIP 1* te ćemo ih spremiti u dataset tip1 s kojim ćemo obrađivati prvi slučaj.

```
data tip1;
set baza;
if tip=2 then delete;
run;
```

U datasetu tip1 imamo 33 opservacije, od čega 19 sa vrijednostima *TIP 0*, a 14 sa vrijednostima *TIP 1*. To jest, od 33 osobe, 19 ih je zdravih, dok preostalih 14 boluje od dijabetesa tip 1.
Modelirajmo kategorijsku varijablu *BMI3kat* pomoću dummy varijabli. Pripadni kod je:

```
data tip1;
set tip1;
if BMI3kat=1 then do; BMI25=1;BMI30=0;end;
if BMI3kat=2 then do; BMI25=0;BMI30=0;end;
if BMI3kat=3 then do; BMI25=0;BMI30=1;end;
run;
```

Kada je nezavisna varijabla kontinuirana, tada je veza između zavisne i nezavisne varijable linearna, dok kod nezavisnih kategorijskih varijabli sa više od dvije kategorije to ne znamo, te je stoga najbolje koristiti dummy (pomoćne) varijable. Dummy varijable su binarne varijable te mogu primiti samo dvije vrijednosti, 0 i 1.

Ako nezavisna kategorijska varijabla ima $k \geq 3$ kategorija, tada se koristi $k-1$ dummy varijabli, na način da se jedna kategorija fiksira te zatim konstruira $k-1$ dummy varijabli za preostalih $k-1$ kategorija. Fiksirana kategorija će uvijek biti referentna u odnosu na ostale. Varijabla *BMI3kat* u našem slučaju ima tri kategorije te su stoga konstruirane dvije dummy varijable: *BMI25* i *BMI30*.

Provedimo (za svaku od nezavisnih varijabli) univarijatnu logističku regresiju u **SAS**-u: (*Opaska*: Zamijetimo da nezavisnu varijablu *TRAJANJE* nema smisla gledati.)

```
title " Univarijatna logistička regresija - DOB ";
proc logistic data=tip1 descending;
model TIP=DOB /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - CRP ";
proc logistic data=tip1 descending;
model TIP=CRP /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - FIB ";
proc logistic data=tip1 descending;
model TIP=FIB /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - HCY ";
proc logistic data=tip1 descending;
model TIP=HCY /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - ADN ";
proc logistic data=tip1 descending;
model TIP=ADN /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - BMI ";
proc logistic data=tip1 descending;
model TIP=BMI /lackfit rsq outroc=rocgraf;
run;
```

```
title " Univarijatna logistička regresija - SPOL ";
proc logistic data=tip1 descending;
model TIP=SPOL /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - RRS ";
proc logistic data=tip1 descending;
model TIP=RRS /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - RRD ";
proc logistic data=tip1 descending;
model TIP=RRD /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - HbA1c ";
proc logistic data=tip1 descending;
model TIP=HbA1c /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - GUKb ";
proc logistic data=tip1 descending;
model TIP=GUKb /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - LEUK ";
proc logistic data=tip1 descending;
model TIP=LEUK /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - Ac_UR ";
proc logistic data=tip1 descending;
model TIP=Ac_UR /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - AER ";
proc logistic data=tip1 descending;
model TIP=AER /lackfit rsq outroc=rocgraf;
run;
```

Nezavisna varijabla *BMI3kat* predstavlja indeks tjelesne mase podijeljen u 3 kategorije: 1−pothranjenost, 2−idealna težina, 3−pretilost. Model sa kategorijskom nezavisnom varijablom *BMI3kat* nije iskonvergirao, jer u datasetu tip1 nema niti jedna pretila osoba koja je oboljela od dijabetesa tip 1. To i nije toliko čudno ako uzmemo u obzir da imamo samo 33 podatka. Slično, model sa nezavisnom varijablom *AER* također nije iskonvergirao, jer 18 od 19 vrijednosti za *TIP 0* nedostaju („missing values").

Glavni rezultati prikazani su u tablici 3.1.

Tablica 3.1: Slučaj 1 - TIP 0 i TIP 1

Univarijatna logistička regresija								
Var	DF	Estimate	Standard Error	Wald χ^2	p value	OR	95% C.L.	c
DOB	1	-0.0372	0.0275	1.8192	0.1774	0.964	0.913 – 1.017	0.639
CRP	1	-0.1880	0.1356	1.9233	0.1655	0.829	0.635 – 1.081	0.722
FIB	1	-0.3965	0.4436	0.7988	0.3715	0.673	0.282 – 1.605	0.557
HCY	1	-0.2292	0.1327	2.9822	0.0842	0.795	0.613 – 1.031	0.700
ADN	1	0.1089	0.0671	2.6325	0.1047	1.115	0.978 – 1.272	0.707
BMI25	1	1.4917	0.8515	3.0690	0.0798	4.444	0.838 – 23.58	0.774
BMI30	1	-11.726	185.4	0.0040	0.9496	0.001	0.001 – 999.9	0.774
SPOL	1	-0.9061	0.7260	1.5580	0.2120	0.404	0.097 – 1.677	0.611
RRS	1	-0.0171	0.0198	0.7503	0.3864	0.983	0.946 – 1.022	0.547
RRD	1	-0.1218	0.0515	5.5930	0.0180	0.885	0.800 – 0.979	0.767
HbA1c	1	2.2755	0.9035	6.3437	0.0118	9.733	1.657 – 57.19	0.866
GUKb	1	0.2530	0.2191	1.3327	0.2483	1.288	0.838 – 1.979	0.562
LEUK	1	0.0842	0.1712	0.2419	0.6228	1.088	0.778 – 1.522	0.562
Ac_UR	1	-0.0200	0.00820	5.9357	0.0148	0.980	0.965 – 0.996	0.760
AER	1	0.2628	0.4577	0.3297	0.5659	1.301	0.530 – 3.189	0.786

Iz tablice 3.1 i izračunatih p−vrijednosti vidimo da su statistički značajne varijable *HbA1c*, *RRD* te *Ac_UR*, na razini značajnosti od 5%. Drugi kriterij kojim također možemo provjeriti statističku značajnost svake pojedine varijable je 95% pouzdani interval. Ako 95% pouzdani interval ne sadrži jedinicu, onda je varijabla statistički značajna. Opet vidimo da su jedino varijable *HbA1c*, *RRD* te *Ac_UR* statistički značajne, na razini značajnosti od 5%.

Zadnji stupac u tablici, u oznaci ^2c, predstavlja prediktivnu snagu modela. Isti poprima vrijednosti na intervalu $[0, 1]$, a što je bliži jedinici to je snaga modela za predikciju (u našem slučaju dijabetesa tip 1) bolja.

Definirajmo da se događaj dogodio s 1, a da se nije dogodio s 0. Za par opservacija s različitim odgovorima, kažemo da je *concordant* ako opservacija koja ima više rangirani odgovor (npr. 2 „događaj se ne dogodi"), ima nižu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom (npr. 1 „događaj se dogodi"). Za par opservacija s različitim odgovorima, kažemo da je *discordant* ako opservacija koja ima više rangirani odgovor, ima višu prediktivnu vjerojatnost da se događaj dogodi od opservacije s niže rangiranim odgovorom. Ako par opservacija nije ni *concordant* ni *discordant*, kažemo da je *tie* (jednak odgovor).

$$c = \frac{nc + 0.5\,(t - nc - nd)}{t}, \tag{3.1}$$

pri čemu je t − broj parova s različitim vrijednostima odgovora, nc − broj *concordant* parova, a nd − broj *discordant* parova.

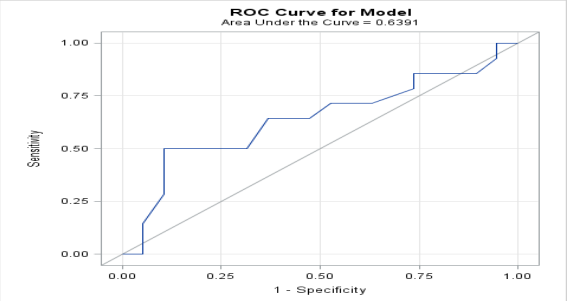
c zapravo predstavlja površinu ispod *ROC* krivulje.

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. Analiza osjetljivosti i specifičnosti testa ovisno o postavljanju granice koja odvaja „test-pozitivne" od „test-negativnih", naziva se *ROC* analiza. Jedan od efikasnijih načina da se prikaže veza između osjetljivosti i specifičnosti testa je takozvana *ROC*[3] krivulja. Glavna ideja takve krivulje je prikaz odnosa proporcija lažno pozitivnih (1-specifičnost) i stvarno pozitivnih (osjetljivost).

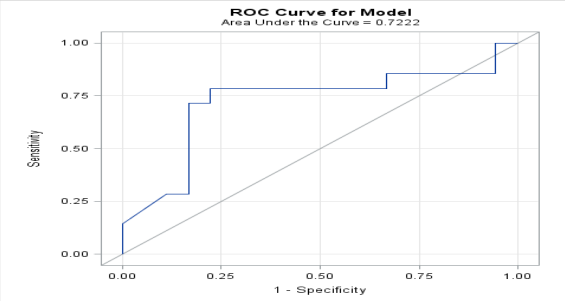
ROC krivulje u slučaju 1 su prikazane narednim grafovima.

[2] eng. Concordance Index

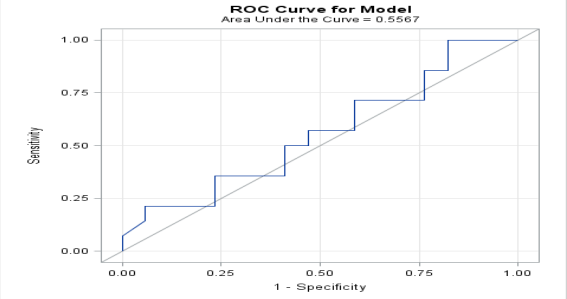
[3] eng. Receiver Operating Characteristic



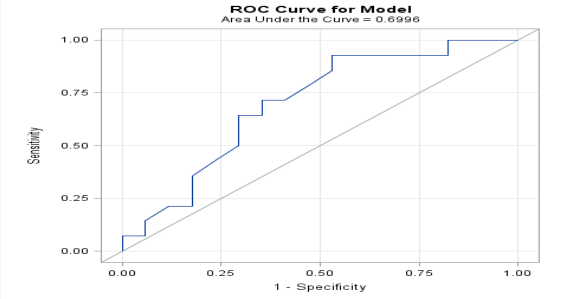
Slika 3.1: DOB



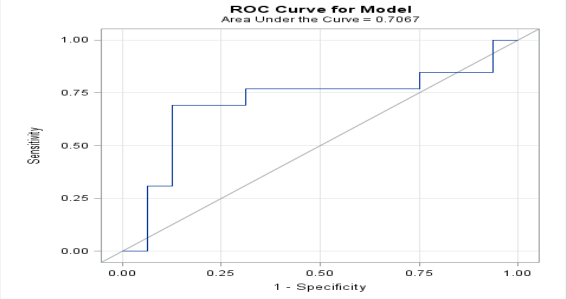
Slika 3.2: CRP



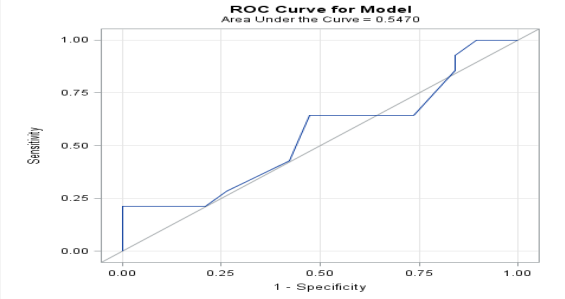
Slika 3.3: FIB



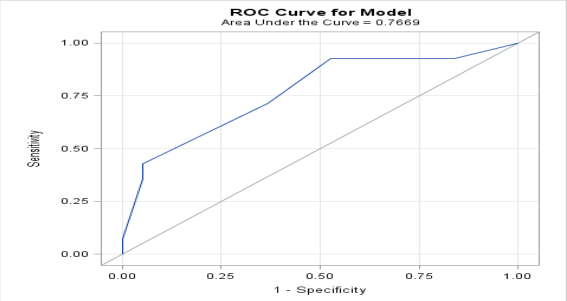
Slika 3.4: HCY



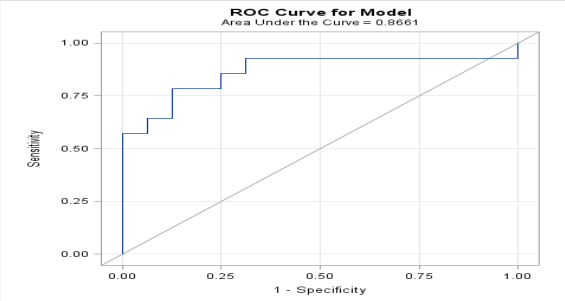
Slika 3.5: ADN



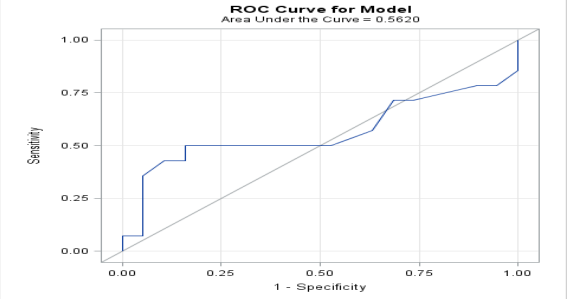
Slika 3.6: RRS



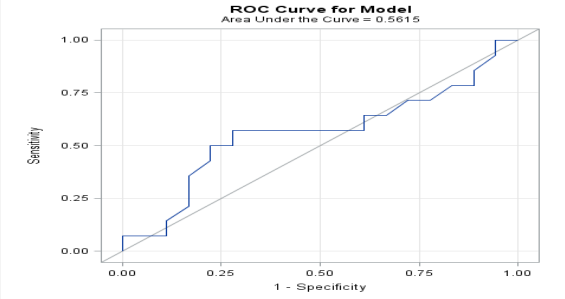
Slika 3.7: RRD



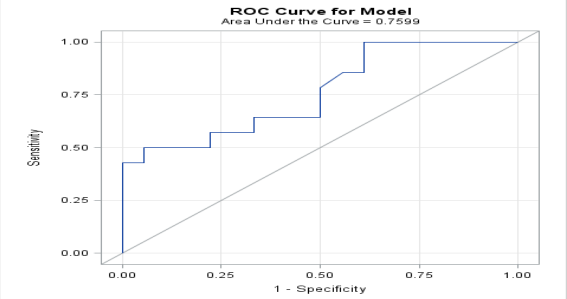
Slika 3.8: HbA1c



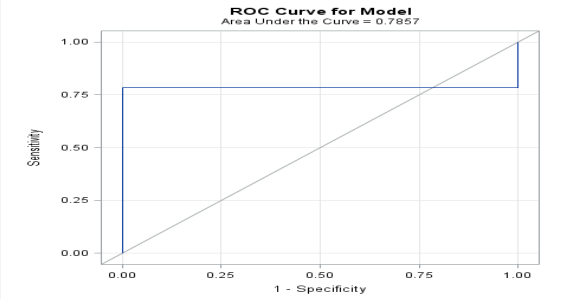
Slika 3.9: GUKb



Slika 3.10: LEUK



Slika 3.11: Ac_UR



Slika 3.12: AER

Slučaj 2 - TIP 0 i TIP 2

Ovdje analiziramo univarijatnu logističku regresiju sa zavisnom varijablom *TIP* koja prima dvije vrijednosti, *TIP 0* i *TIP 2*. Analogno kao u slučaju 1, odvojimo podatke koji u varijabli *TIP* imaju vrijednosti *TIP 0*, odnosno *TIP 2* te ih spremimo u dataset tip2.

```
data tip2;
set baza;
if TIP=1 then delete;
run;
```

U datasetu tip2 imamo 66 opservacija, od čega 19 sa vrijednostima *TIP 0*, a 47 sa vrijednostima *TIP 2*. To jest, od 66 osoba, 19 ih je zdravih, dok preostalih 47 boluje od dijabetesa tip 2.

Također, modelirajmo kategorijsku varijablu *BMI3kat* pomoću dummy varijabli kao u slučaju 1:

```
data tip2;
set tip2;
if BMI3kat=1 then do; BMI25=1;BMI30=0;end;
if BMI3kat=2 then do; BMI25=0;BMI30=0;end;
if BMI3kat=3 then do; BMI25=0;BMI30=1;end;
run;
```

Provedimo (za svaku od nezavisnih varijabli) univarijatnu logističku regresiju u **SAS**-u: (*Opaska*: Zamijetimo da nezavisnu varijablu *TRAJANJE* nema smisla gledati.)

```
title "Univarijatna logistička regresija - DOB";
proc logistic data=tip2 descending;
model TIP=DOB /lackfit rsq outroc=rocgraf;
run;

title "Univarijatna logistička regresija - CRP";
proc logistic data=tip2 descending;
model TIP=CRP /lackfit rsq outroc=rocgraf;
run;
```

```
title " Univarijatna logistička regresija - FIB ";
proc logistic data=tip2 descending;
model TIP=FIB /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - HCY ";
proc logistic data=tip2 descending;
model TIP=HCY /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - ADN ";
proc logistic data=tip2 descending;
model TIP=ADN /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - BMI ";
proc logistic data=tip2 descending;
model TIP=BMI /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - SPOL ";
proc logistic data=tip2 descending;
model TIP=SPOL /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - RRS ";
proc logistic data=tip2 descending;
model TIP=RRS /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - RRD ";
proc logistic data=tip2 descending;
model TIP=RRD /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - HbA1c ";
proc logistic data=tip2 descending;
model TIP=HbA1c /lackfit rsq outroc=rocgraf;
run;
```

```
title " Univarijatna logistička regresija - GUKb ";
proc logistic data=tip2 descending;
model TIP=GUKb /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - LEUK ";
proc logistic data=tip2 descending;
model TIP=LEUK /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - Ac_UR ";
proc logistic data=tip2 descending;
model TIP=Ac_UR /lackfit rsq outroc=rocgraf;
run;

title " Univarijatna logistička regresija - AER ";
proc logistic data=tip2 descending;
model TIP=AER /lackfit rsq outroc=rocgraf;
run;
```

Svi modeli su uspješno iskonvergirali.

Glavni rezultati prikazani su u tablici 3.2.

Tablica 3.2: Slučaj 2 - TIP 0 i TIP 2

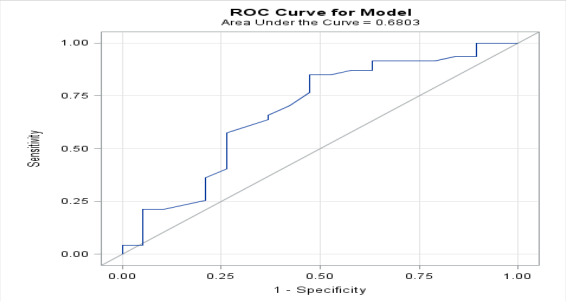
Univarijatna logistička regresija								
Var	DF	Estimate	Standard Error	Wald χ^2	p value	OR	95% C.L.	c
DOB	1	0.0598	0.0252	5.6192	0.0178	1.062	1.010 – 1.115	0.680
CRP	1	-0.0558	0.0877	0.4050	0.5245	0.946	0.796 – 1.123	0.531
FIB	1	0.0525	0.2988	0.0309	0.8605	1.054	0.587 – 1.893	0.516
HCY	1	-0.0455	0.0656	0.4818	0.4876	0.956	0.840 – 1.087	0.559
ADN	1	-0.0280	0.0471	0.3530	0.5524	0.972	0.887 – 1.067	0.533
BMI25	1	0.2389	0.7784	0.0942	0.7589	1.270	0.276 – 5.839	0.543
BMI30	1	0.3567	0.6081	0.3440	0.5576	1.429	0.434 – 4.705	0.543
SPOL	1	0.0693	0.5516	0.0158	0.9000	1.072	0.364 – 3.160	0.508
RRS	1	0.00993	0.0136	0.5346	0.4647	1.010	0.983 – 1.037	0.579
RRD	1	0.00160	0.0259	0.0038	0.9508	1.002	0.952 – 1.054	0.510
HbA1c	1	3.0592	0.9074	11.3672	0.0007	21.31	3.600 – 126.2	0.939
GUKb	1	1.3921	0.4168	11.1548	0.0008	4.023	1.777 – 9.107	0.913
LEUK	1	0.2047	0.1690	1.4670	0.2258	1.227	0.881 – 1.709	0.621
Ac_UR	1	-0.0024	0.00344	0.4667	0.4945	0.998	0.991 – 1.004	0.579
AER	1	0.4756	0.5117	0.8639	0.3527	1.609	0.590 – 4.386	0.936

Iz tablice 3.2 i izračunatih p−vrijednosti vidimo da su statistički značajne varijable *HbA1c*, *DOB* te *GUKb*, na razini značajnosti od 5%. Drugi kriterij kojim također možemo provjeriti statističku značajnost svake pojedine varijable je 95% pouzdani interval. Ako 95% pouzdani interval ne sadrži jedinicu, onda je varijabla statistički značajna. Opet vidimo da su jedino varijable *HbA1c*, *DOB* te *GUKb* statistički značajne, na razini značajnosti od 5%. Primijetimo jako visoke *c* vrijednosti kod varijabli *HbA1c* (0.939) te *GUKb* (0.913).

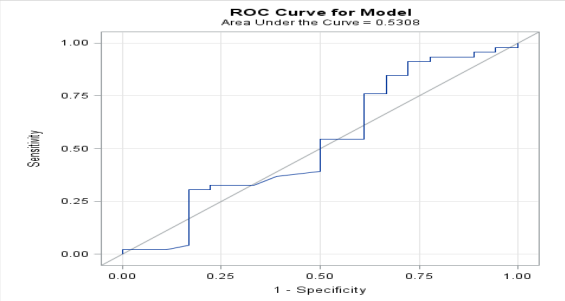
c, tj. površina ispod *ROC* krivulje mjera je točnosti testa:

0.90 − 1 = izvrstan test
0.80 − 0.90 = dobar test
0.70 − 0.80 = osrednji test
0.60 − 0.70 = slabiji test
0.50 − 0.60 = test bez uspjeha

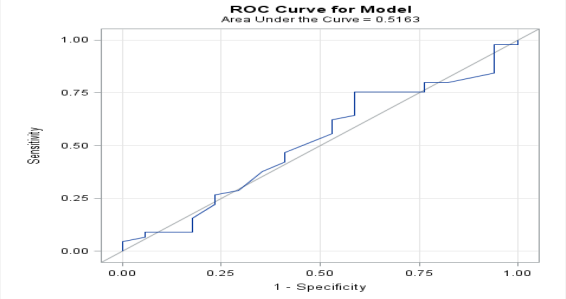
ROC krivulje u slučaju 2 su prikazane narednim grafovima.



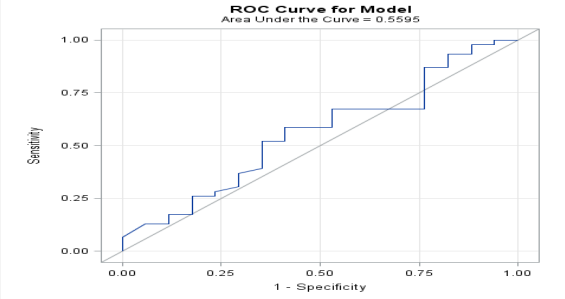
Slika 3.13: DOB



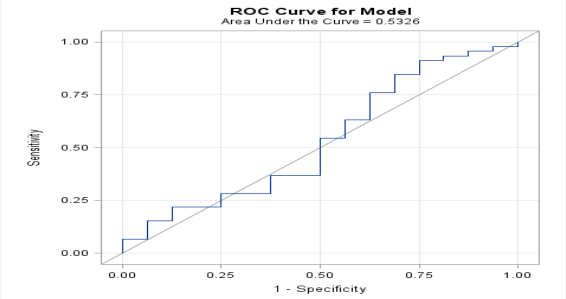
Slika 3.14: CRP



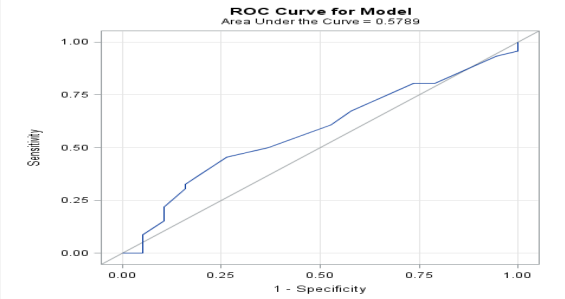
Slika 3.15: FIB



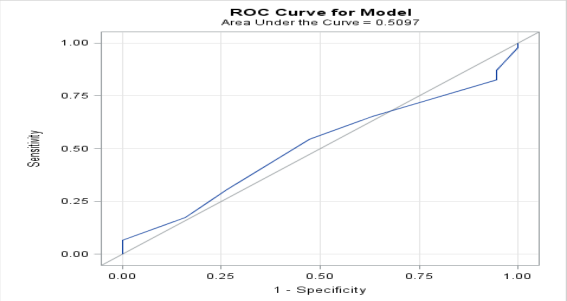
Slika 3.16: HCY



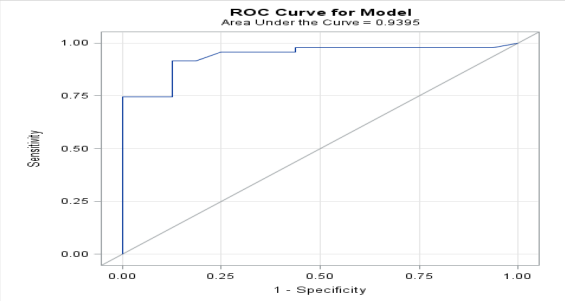
Slika 3.17: ADN



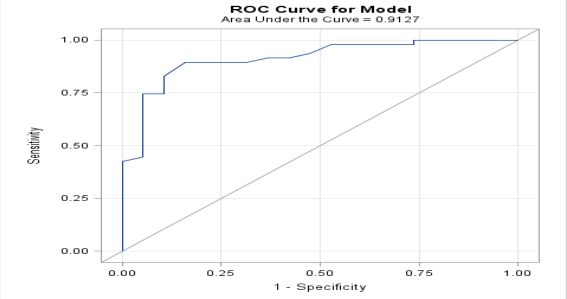
Slika 3.18: RRS



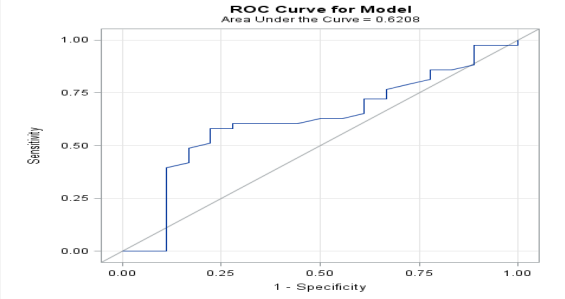
Slika 3.19: RRD



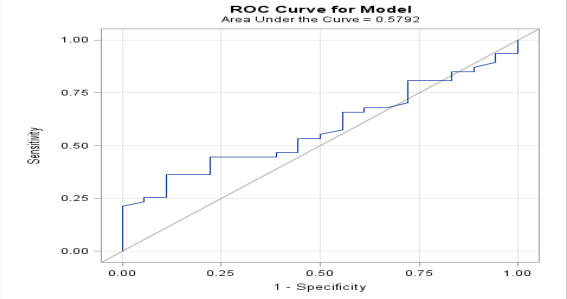
Slika 3.20: HbA1c



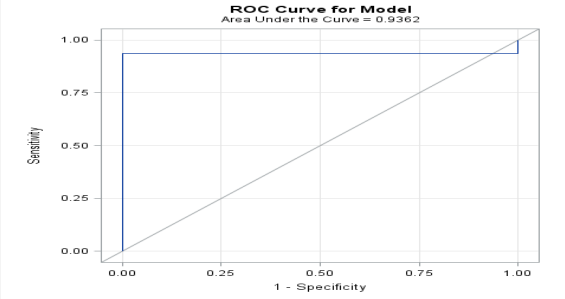
Slika 3.21: GUKb



Slika 3.22: LEUK



Slika 3.23: Ac_UR



Slika 3.24: AER

3.3 Stepwise procedura

Ovdje ćemo testirati razne modele *Stepwise* procedurom. Princip rada *Stepwise* procedure spomenut je u *podpoglavlju 1.3*.

Slučaj 1 - TIP 0 i TIP 1

Varijanta a) – puni model (sve nezavisne varijable u modelu)

```
title "Stepwise metoda";
proc logistic data=tip1 descending;
model TIP = DOB CRP FIB HCY ADN BMI25 BMI30 SPOL RRS
RRD HbA1c GUKb LEUK Ac_UR AER /selection=stepwise;
run;
```

Nedostaju vrijednosti za 19 podataka od 33. *BMI30* ušao u model te je na kraju ipak uklonjen iz modela. Dakle, niti jedna varijabla nije statistički značajna za puni model. Uočavamo da je problem u 19 vrijednosti koje nedostaju. Izbacimo varijablu *AER* iz modela, jer znamo od prije da varijabli *AER* nedostaje 18 od 19 vrijednosti za *TIP 0*.

Varijanta b) – krnji puni model (sve nezavisne varijable u modelu, osim *AER*)

```
title "Stepwise metoda";
proc logistic data=tip1 descending;
model TIP = DOB CRP FIB HCY ADN BMI25 BMI30 SPOL
RRS RRD HbA1c GUKb LEUK Ac_UR/selection=stepwise;
run;
```

Nedostaju vrijednosti za 10 podataka. *HbA1c* ušao prvi u model, nakon njega i *BMI30*, ali je ipak na kraju uklonjen. Dakle, jedina statistički značajna varijabla je *HbA1c*. c vrijednost iznosi 0.846.

Varijanta c) – *HbA1c*, *RRD* i *Ac_UR* (statistički značajne varijable iz univarijatne regresije)

```
title "Stepwise metoda";
proc logistic data=tip1 descending;
model TIP = HbA1c RRD Ac_UR/selection=stepwise;
run;
```

Nedostaju vrijednosti za samo 4 podatka, što je već bolje, s obzirom da ukupno imaju samo 33 podatka. *HbA1c* ušao prvi u model, nakon njega i *RRD*, ali je ipak na kraju uklonjen. Opet je samo *HbA1c* ostao u modelu, ali c je nešto bolji, $c = 0.871$.

Varijanta d) – *HbA1c* i *RRD*

```
title "Stepwise metoda";
proc logistic data=tip1 descending;
model TIP = HbA1c RRD/selection=stepwise;
run;
```

Nedostaju vrijednosti za samo 3 podatka. U model ušli i *HbA1c* i *RRD*. c dosad daleko najbolji, $c = 0.951$. Ovaj model je zasad obećavajući.

Varijanta e) – *HbA1c* i *Ac_UR*

```
title "Stepwise metoda";
proc logistic data=tip1 descending;
model TIP = HbA1c Ac_UR/selection=stepwise;
run;
```

Nedostaju vrijednosti za 4 podatka. U model ušao samo *HbA1c*, $c = 0.871$.

Varijanta f) – *RRD* i *Ac_UR*

```
title "Stepwise metoda";
proc logistic data=tip1 descending;
model TIP = RRD Ac_UR/selection=stepwise;
run;
```

Nedostaje vrijednost za samo 1 podatak. U model ušli i *RRD* i *Ac_UR*, $c = 0.851$.

Testirane su još razne druge varijante, ali ovdje su navedene samo one od konkretnog značaja.

Na temelju varijanti $a) - f)$, u *podpoglavlju 3.4* ćemo odabrati najadekvatniji model koji predviđa pojavnost dijabetesa tip 1 kod zdravih osoba te provesti multivarijatnu logističku regresiju. Nakon toga ćemo interpretirati dobivene rezultate.

Slučaj 2 - TIP 0 i TIP 2

Varijanta a) – puni model (sve nezavisne varijable u modelu)

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
model TIP = DOB CRP FIB HCY ADN BMI25 BMI30 SPOL RRS
RRD HbA1c GUKb LEUK Ac_UR AER /selection=stepwise;
run;
```

Nedostaju vrijednosti za 26 podataka od 66. Niti jedna varijabla nije ušla u model. Vjerojatno je opet problem u varijabli *AER* pa ćemo ju izbaciti za iduću varijantu.

Varijanta b) – krnji puni model (sve nezavisne varijable u modelu, osim *AER*)

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
model TIP = DOB CRP FIB HCY ADN BMI25 BMI30 SPOL
RRS RRD HbA1c GUKb LEUK Ac_UR/selection=stepwise;
run;
```

Nedostaju vrijednosti za 17 podataka. *HbA1c* ušao prvi u model, nakon njega i *GUKb*, ali je ipak na kraju uklonjen. Dakle, jedina statistički značajna varijabla je *HbA1c*. c vrijednost iznosi 0.922.

Varijanta c) – *HbA1c*, *GUKb* i *DOB* (statistički značajne varijable iz univarijatne regresije)

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
model TIP = HbA1c Gukb DOB/selection=stepwise;
run;
```

Nedostaju vrijednosti za samo 3 podatka, što je već puno bolje u usporedbi sa prethodna dva modela. *HbA1c* ušao prvi u model, nakon njega i *GUKb*, ali je ipak na kraju uklonjen. Opet je samo *HbA1c* ostao u modelu, ali c je nešto bolji, $c = 0.939$.

Varijanta d) – *HbA1c* i *GUkb*

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
```

```
model TIP = HbA1c GUKb/selection=stepwise;
run;
```

Nedostaju vrijednosti za 3 podatka. *HbA1c* ušao prvi u model, nakon njega i *GUKb*, ali je ipak na kraju uklonjen. Opet je samo *HbA1c* ostao u modelu, a *c* je isti kao u prethodnoj varijanti.

Varijanta e) – *HbA1c* i *DOB*

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
model TIP = HbA1c DOB/selection=stepwise;
run;
```

Nedostaju vrijednosti za 3 podatka. U model ušao samo *HbA1c*, a *c* opet iznosi 0.939.

Varijanta f) – *GUKb* i *DOB*

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
model TIP = GUKb DOB/selection=stepwise;
run;
```

Svi podaci u modelu. U model ušao samo *GUKb*. *c* je nešto lošiji i iznosi 0.913.

Varijanta g) – *HbA1c*, *BMI25* i *BMI30*

```
title "Stepwise metoda";
proc logistic data=tip2 descending;
model TIP = HbA1c BMI25 BMI30/selection=stepwise;
run;
```

Nedostaju vrijednosti za 3 podatka. U model ušao samo *HbA1c*, a *c* opet iznosi 0.939. Razlog zbog kojeg smo testirali ovu varijantu je spoznaja o mogućem utjecaju indeksa tjelesne mase na šansu obolijevanja od dijabetesa tip 2. Kao i u univarijatnoj logističkoj regresiji, tako i u *Stepwise* proceduri vidimo da kategorijska varijabla *BMI3kat* nije relevantna za predikciju dijabetesa tip 2, barem za prikupljene podakte iz dataseta tip2.

Na temelju ovih testiranja, tj. varijanti *a*) – *g*), u *podpoglavlju 3.4* ćemo odabrati najadekvatniji model koji predviđa pojavnost dijabetesa tip 2 kod zdravih osoba te provesti multivarijatnu logističku regresiju. Nakon toga ćemo interpretirati dobivene rezultate.

3.4 Multivarijatna logistička regresija

3.4.1 Odabir modela

Slučaj 1 - TIP 0 i TIP 1

Na temelju *Stepwise* procedure te univarijatne logističke regresije, odlučujemo se za model pod varijantom d), tj. model sa nezavisnim varijablama *HbA1c* i *RRD*. Naime, vidjeli smo kroz *Stepwise* proceduru da varijabla *HbA1c*, kao statistički „najjača", uvijek ulazi u model u svim varijantama (osim varijante a)). U kombinaciji sa varijablom *RRD* iznimno se pospješuje prediktivna snaga modela *c* i iznosi 0.951 te stoga biramo upravo taj model.

Provedimo multivarijatnu logističku regresiju sa nezavisnim varijablama *HbA1c* i *RRD*.

```
title "Multivarijatna logistička regresija";
proc logistic data=tip1 descending;
model TIP = HbA1c RRD;
run;
```

Nedostaju vrijednosti za 3 podatka te je model uredno iskonvergirao. Rezultati se nalaze u sljedećim tablicama.

Tablica 3.3: Statistička značajnost modela

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.8956	2	<.0001
Score	17.5014	2	0.0002
Wald	7.3888	2	0.0249

Tablica 3.4: Procjena i statistička značajnost parametara modela

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald χ^2	Pr > ChiSq
Intercept	1	-0.7108	7.9383	0.0080	0.9287
HbA1c	1	2.4088	1.0743	5.0275	0.0249
RRD	1	-0.1624	0.0820	3.9159	0.0478

Broj stupnjeva slobode (DF[4]) u tablici 3.3 iznosi 2, jer se u modelu nalaze dvije nezavisne varijable. Iz tablice 3.3 pod *Likelihood Ratio* i *Wald* iščitavamo p-vrijednosti manje od 5% te zaključujemo da je model statistički značajan, na razini značajnosti od 5%. Također, iz tablice 3.4 i pripadnih p-vrijednosti zaključujemo da su obe nezavisne varijable statistički značajne. Isti zaključak možemo donijeti ako promatramo 95% pouzdani interval iz tablice 3.8, jer jedinica u oba slučaja nije u intervalu. Tablica 3.8 se nalazi u idućem podpoglavlju, gdje ćemo interpretirati rezultate. Napomenimo još da je $c = 0.951$.

Slučaj 2 - TIP 0 i TIP 2

Na temelju *Stepwise* procedure te univarijatne logističke regresije, dalo bi se naslutiti da bi najbolji odabir modela trebao biti univarijatni logistički model sa nezavisnom varijablom *HbA1c*. No, ipak ćemo se odlučiti za multivarijatni logistički model sa nezavisnim varijablama *HbA1c* i *GUKb*. Analizirajući takav multivarijatni model u nastavku, nastojat ćemo opravdati ovo razmišljanje.

```
title "Multivarijatna logistička regresija";
proc logistic data=tip2 descending;
model TIP = HbA1c GUKb;
run;
```

Nedostaju vrijednosti za 3 podatka te je model uredno iskonvergirao. Rezultati se nalaze u sljedećim tablicama.

Tablica 3.5: Statistička značajnost modela

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	45.3404	2	<.0001
Score	23.4718	2	<.0001
Wald	10.3313	2	0.0057

[4]eng. Degrees Of Freedom

Tablica 3.6: Procjena i statistička značajnost parametara modela

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald χ^2	Pr > ChiSq
Intercept	1	-21.8143	7.0377	9.6078	0.0019
HbA1c	1	3.2537	1.1662	7.7835	0.0053
GUKb	1	0.7131	0.3750	3.6150	0.0573

Tablica 3.7: Prediktivna snaga modela

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.5	Somers' D	0.931
Percent Discordant	3.5	Gamma	0.931
Percent Tied	0.0	Tau-a	0.358
Pairs	752	c	$0.965*$

Iz tablice 3.5 pod *Likelihood Ratio* i *Wald* iščitavamo p-vrijednosti manje od 5% te zaključujemo da je model statistički značajan, na razini značajnosti od 5%. Iz tablice 3.6 vidimo da je nezavisna varijabla *HbA1c* statistički značajna. Također, mogli bismo reći i da je varijabla *GUKb* statistički „tijesno" značajna, jer joj je p-vrijednost jednaka 0.0573, što je približno jednako 5%.

Vidjeli smo kroz *Stepwise* proceduru da varijabla *HbA1c* jedina ulazi u modele u svim varijantama (osim varijante a)) u kojima je razmatrana. U svim varijantama je pripadna c vrijednost jednaka 0.939 ili manja. Također, u univarijatnoj logističkoj regresiji, kada je varijabla *HbA1c* sama u modelu, c vrijednost je opet jednaka 0.939 (vidi tablicu 3.2).

Ako promotrimo tablicu 3.7, vidimo da je c u modelu sa nezavisnim varijablama *HbA1c* i *GUKb* jednak 0.965 ($*$), što nije neznatan pomak u prediktivnoj snazi.

Dakle, ako uzmemo u obzir da je nezavisna varijabla *GUKb* „tijesno" značajna te se povedemo statističkom intuicijom, mogli bismo reći da je promatrani model pogodan za predikciju dijabetesa tip 2.

3.4.2 Interpretacija rezultata

Slučaj 1 - TIP 0 i TIP 1

Najadekvatniji model za predikciju pojavnosti dijabetesa tip 1, baziran na logističkoj regresijskoj analizi iz prethodnih poglavlja, je model sa nezavisnim varijablama *HbA1c* i *RRD*. *HbA1c* predstavlja hemoglobin, dok *RRD* predstavlja dijastolički krvni tlak, tzv. „donji" tlak. Kroz tablicu 3.8 objasnit ćemo povezanost između vrijednosti hemoglobina, odnosno dijastoličkog tlaka i šanse obolijevanja od dijabetesa tip 1. Tablica 3.8 dobivena je multivarijatnom logističkom regresijom u prethodnom podpoglavlju.

Tablica 3.8: Omjer šansi

Odds Ratio Estimates		
Effect	Point Estimate	95% C.L.
HbA1c	11.120	1.354 – 91.315
RRD	0.850	0.724 – 0.998

Vrijednosti u tablici pod *Point Estimate* predstavljaju omjer šanse obolijevanja od dijabetesa tip 1, uz pomak vrijednosti hemoglobina, odnosno dijastoličkog tlaka za jednu jedinicu. Te vrijednosti dobivene su već poznatom relacijom

$$odds\ ratio\,(x+1, x) = e^{\beta},$$

pri čemu je β parametar modela.

Na primjer, vrijednost 11.120 je dobivena kao $e^{2.4088}$, gdje je 2.4088 procijenjeni parametar pripadne nezavisne varijable *HbA1c*, dobiven multivarijatnom logističkom regresijom u prethodnom podpoglavlju (vidi tablicu 3.4).

To znači da ukoliko imamo povećanje vrijednosti hemoglobina za jednu jedinicu, povećava se omjer šanse za obolijevanje od dijabetesa tip 1 za približno 11 puta, tj. za približno 1000%. Kako 1000% nije nikako zanemariv rezultat, možemo reći da je hemoglobin ozbiljan prediktor u pojavnosti dijabetesa tip 1.
Međutim, valja i napomenuti da u ovoj analizi koristimo samo 33 podatka te da treba biti oprezan u globalnoj ocjeni.

S druge strane, vrijednost 0.850, dobivena kao $e^{-0.1624}$ (vidi tablicu 3.4), kazuje kako povećanje vrijednosti dijastoličkog tlaka za jednu jedinicu (1 mmHg), povećava omjer šanse za obolijevanje od dijabetesa tip 1 za 0.850 puta, tj. smanjuje za 15%.

Ovaj rezultat se kosi sa intuicijom, no takvi podaci su prikupljeni i ne moraju nužno prezentirati čitavu populaciju.

Slučaj 2 - TIP 0 i TIP 2

Najadekvatniji model za predikciju pojavnosti dijabetesa tip 2, baziran na logističkoj regresijskoj analizi iz prethodnih poglavlja, je model sa nezavisnim varijablama *HbA1c* i *GUKb*. *HbA1c* predstavlja hemoglobin, dok *GUKb* predstavlja glukozu. Kao u slučaju 1, objasnit ćemo vezu između vrijednosti hemoglobina, odnosno glukoze i šanse obolijevanja od dijabetesa tip 2, kroz sljedeću tablicu.

Tablica 3.9: Omjer šansi

Odds Ratio Estimates		
Effect	Point Estimate	95% C.L.
HbA1c	25.885	2.632 – 254.542
GUKb	2.040	0.978 – 4.255

Iz tablice 3.9 iščitavamo da povećanje hemoglobina za jednu jedinicu, povećava omjer šanse za obolijevanje od dijabetesa tip 2 za 25.885 puta. Vidimo da je hemoglobin i u slučaju dijabetesa tip 2 iznimno relevantan prediktor. No, to se već i dalo naslutiti kroz obradu *Stepwise* procedurom kao i univarijatnom logističkom regresijom. Nadalje, povećanje glukoze za jednu jedinicu, povećava omjer šanse za obolijevanje od dijabetesa tip 2 za 2.040 puta.

Napomenimo još da smo u obradi slučaja 2 očekivali da će indeks tjelesne mase imati utjecaj na predikciju pojavnosti dijabetesa tip 2. To jest, očekivali smo rezultat u kojem će pretile osobe u odnosu na osobe idealne težine imati veći omjer šanse obolijevanja od dijabetesa tip 2. No, to se nije dogodilo. Takav ishod možemo pripisati obujmu, ali i prirodi prikupljenih podataka. Naime, analizirajući deskriptivu od 47 oboljelih osoba od dijabetesa tip 2, vidimo da ih je 8 pothranjenih, 21 idealne težine te 18 pretilih. Sličan omjer imamo i u deskriptivi od 19 zdravih osoba. Da bi dobili rezultate koji su u skladu sa očekivanjem, broj pretilih osoba oboljelih od dijabetesa tip 2, intuitivno, morao bi biti kudikamo veći od broja oboljelih idealne težine.

Bibliografija

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

[2] M. E. Stokes, C. S. Davis, G. G. Koch, *Categorical Data Analysis Using SAS System, Second Edition*, Cary, NC: SAS Institute Inc., USA, 2000.

[3] A. Jazbec, *Odabrane statističke metode u biomedicini*, PMF-MO slideovi, 2014.

[4] M. Huzak, *Matematička statistika*, PMF-MO skripta, 2015.

[5] M. Huzak, *Statistički praktikum*, PMF-MO slideovi, 2014.

[6] J. S. Cramer, *Logit Models from Economics and Other Fields*, Cambridge University Press, http://papers.tinbergen.nl/02119.pdf , 2003.

Sažetak

Logistička regresija, kao moćan statistički alat, koristi se u istraživačkim problemima koji uključuju kategorijsku zavisnu varijablu. Ista je jako robusna regresijska metoda, jer zahtjeva minimalan skup temeljnih pretpostavki. Rezultat je širok raspon primjene, kako u znanstvenim krugovima tako i u kontekstu prakse.

U ovom radu objašnjeni su glavni koncepti logističke regresijske analize. Između ostalog, obrađeni su pojmovi omjera šanse, metode maksimalne vjerodostojnosti te logističkog regresijskog modela. Logistički regresijski model koristi se za predviđanje vjerojatnosti događaja putem prilagođavanja podataka logističkoj krivulji.

Teorijska podloga logističke regresije sa dihotomnom zavisnom varijablom implementirana je u modeliranju dijabetesa tip 1, odnosno tip 2. Analizirana je baza podataka sa predavanja iz kolegija *Odabrane statističke metode u biomedicini*. Na temelju univarijatne, odnosno multivarijatne logističke regresije, kao i *Stepwise* procedure u **SAS**-u, pronađeni su najadekvatniji modeli koji predviđaju pojavnost dijabetesa tip 1, odnosno tip 2. Ispostavilo se da su hemoglobin i dijastolički krvni tlak glavni prediktori za pojavnost dijabetesa tip 1, dok su hemoglobin i glukoza prediktori za pojavnost dijabetesa tip 2.

Summary

Logistic regression, a powerful statistic tool, is used in researches which include categorical dependent variable. Logistic regression is robust thanks to the fact that it requires a minimum set of assumptions. As a result, it can be used widely, in scientific as well as in practical context.

This thesis describes the basic concepts of logistic regression. Among others, the terms of odds ratio, maximum likelihood estimation and logistic regression model have been analysed. Logistic regression model is used to predict the probability of the event occuring by adjusting the data to the logistic curve.

Theoretical background of logistic regression with dichotomous dependent variable is implemented in modeling of type 1 and type 2 diabetes. The database from the course *Statistical methods in Biomedical Research* has been examined. Based on univariate and multivariate logistic regression, as well as on *Stepwise* procedure in **SAS**, the most adequate model for the prediction of type 1 and type 2 diabetes has been found. The results have shown that hemoglobin and diastolic blood pressure are the main predictors of type 1 diabetes, while the hemoglobin and glucose are predictors for type 2 diabetes.

Životopis

- rođen sam 23. studenoga 1985. godine u Metkoviću
- od 1992. do 2000. godine pohađam OŠ Stjepana Radića u Metkoviću
- od 2000. do 2004. godine pohađam Matematičku gimnaziju u Metkoviću
- od 2006. do 2013. godine pohađam preddiplomski sveučilišni studij Matematika na PMF-MO u Zagrebu
- 2013. godine upisujem diplomski sveučilišni studij Matematička statistika na PMF-MO u Zagrebu