Matematički fakultet

Univerzitet u Beogradu

Tema

Logistička regresija

Student: Sanja Pašanjski 1107/2013

Mentor: Slobodanka Janković

Sadržaj

1 Uvod

Regresija se koristi za opisivanje i predviđanje zavisne promenljive na osnovu skupa nezavisnih i prvi put se pojavila krajem 19. veka. Ukoliko je zavisna promenljiva binarna, tada logistička regresija daje bolje rezultate od linearne, pa se češće i koristi.

Kako se binarne promenljive pojavljuju u mnogim sferama života, logistička regresija ima široku primenu. Prvi put se pojavila sredinom 20. veka a do danas je našla primene u medicini, biologiji, mašinskom kodiranju, marketingu i raznim drugim istraživanjima. Na osnovu nje se može proučiti efikasnost leka, koji su spoljni uticaji na bolesti, da li se dati tekst i u kojoj meri poklapa sa proizvoljnim drugim tekstom, da li su ispitanici zainteresovani za kupovinu proizvoda i slično. Logistička funkcija je prepoznatljiva po svom S-obliku pomoću koga se najlakše može proceniti stopa rasta neke populacije ili odnos broja starosedelaca u odnosu na broj doseljenika nekog grada.

2 Logistička funkcija

2.1 Maltusov populacioni model

Tomas Robert Maltus (Thomas Robert Malthus) je u svom poznatom delu *An essay on the principle of population as it affects the future improvement of society* objavljenom 1789. godine, izneo svoje populacionističko gledište koje je povezao sa biološko-evolucionističkim i organističkim shvatanjem čoveka i društva i pokušao da skrene pažnju javnosti na problem prenaseljenosti. Maltus je tvrdio da čovečanstvo može opstati samo ako rast populacije bude povremeno prekidan povećanjem smrtnosti – epidemijama, ratovima, oskudicama, ograničenim rađanjem i katastrofama. [20] Prema njemu, sa povećanjem broja stanovnika, povećava se i potrebna količina hrane i drugih resursa neophodnih za preživljavanje, ali to povećanje raste aritmetičkom progresijom jer je zemlja proizvodni resurs sa ograničenim kapacitetom, dok broj stanovnika raste geometrijskom progresijom. Kako se broj stanovnika povećava brže od količine resursa, posle izvesnog vremena zavladaće oskudice. Takvo stanje se naziva *Maltusova (demografska) katastrofa*.

Ako broj stanovnika na planeti u nekom trenutku t_0 označimo sa $P(0)$, a sa r označimo parametar koji opisuje priraštaj stanovništva, tada imamo da će u sledećem trenutku broj stanovništva biti $P(1) = rP(0)$. Pod pretpostavkom da su vremenski intervali jednake dužine u sledećem trenutku dobijamo da broj stanovnika iznosi $P(2) = rP(1) = r^2P(0)$... Posle n vremenskih intervala dobijamo da broj stanovnika na planeti iznosi $P(n) = r^nP(0)$. Kako je parametar priraštaja $r > 1$, populacija se povećava i broj stanovnika, bez obzira na početnu vrednost $P(0)$, teži beskonačnosti.

Ako se sa γ označi konstantna brzina rađanja u jedinici vremena (stopa nataliteta), a sa δ konstantna brzina umiranja (stopa mortaliteta), tada važi da je konstantan priraštaj u jedinici vremena $\lambda = \gamma - \delta$. Na osnovu toga, ako se sa $P(t)$ označi veličina populacije u trenutku t, posle vremena Δt, pri čemu je Δt malo, imamo

$$P(t + \Delta t) = P(t) + \lambda P(t)\Delta t.$$

Prelaskom na limes kad $\Delta t \to 0$, dobijamo

$$P'(t) = \lambda P(t) \tag{1}$$

$$P(0) = P_0$$

odakle dobijamo:

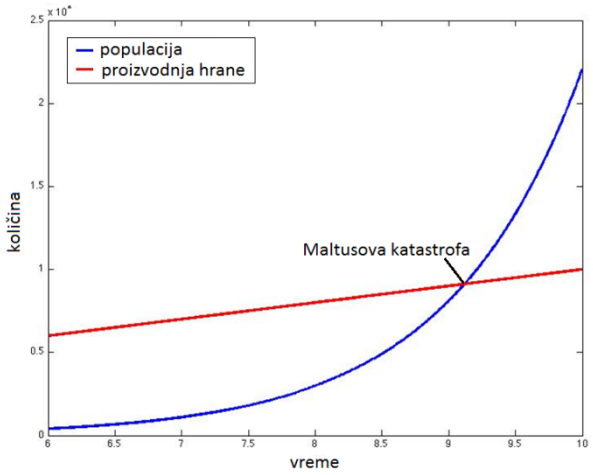
$$\frac{dP(t)}{dt} = \lambda P(t)$$

$$\frac{dP(t)}{P(t)} = \lambda dt$$

$$ln|P(t)| = \lambda t + c$$

$$P(t) = Ce^{\lambda t}.$$

Iz početnog uslova imamo da je $C = P_0$, odakle sledi da je $P(t) = P_0 e^{\lambda t}$, odnosno da veličina populacije eksponencijalno zavisi od vremena. Ovakav model se naziva *Maltusov populacioni model*. [20]



Slika 1. Maltusov populacioni model

Glavni nedostatak ovog modela je taj što se pretpostavlja da je stopa rasta populacije konstantna čime se, bez obzira na to kolika je ta konstanta, za kratko vreme može dostići nerealna veličina populacije.

2.2 Logistički Verhulstov model

Pjer Fransoa Verhulst (Pierre Francois Verhulst) je oko 1830. godine modifikovao Maltusov populacioni model, a skoro vek kasnije taj isti model su pronašli američki naučnici Loel Rid (Lowell Jacob Reed) i Rejmond Perl (Raymond Pearl). [8] Smatrajući da nijedna sredina ne može da održava neograničen rast populacije, Verhulst je napravio model u kome su resursi ograničeni, pa samim

tim je i rast populacije ograničen konstantom. Osnovna razlika u odnosu na Maltusov model je u tome što stope nataliteta i mortaliteta nisu konstantne, već zavise od veličine populacije i date su sa:

$$\gamma(t) = \gamma_0 - \gamma_1 P(t)$$

$$\delta(t) = \delta_0 + \delta_1 P(t),$$

gde su γ_0, γ_1, δ_0 i δ_1 nenegativne konstante. U ovom modelu se smanjuje brzina rađanja i povećava brzina umiranja.

Označimo sa $a = \gamma_0 - \delta_0$ maksimalan priraštaj populacije i neka je $b = \gamma_1 - \delta_1$, $a, b > 0$. Tada imamo da priraštaj λ iznosi

$$\lambda(t) = \gamma_0 - \delta_0 - (\gamma_1 + \delta_1)P(t),$$

odnosno

$$\lambda(t) = a - bP(t).$$

Početni problem (1) sada se transformiše u

$$P'(t) = \big(a - bP(t)\big)P(t)$$

$$P'(t) = aP(t) - bP(t)^2$$

$$P'(t) = a\left(1 - \frac{b}{a}P(t)\right)P(t)$$

$$P'(t) = a\left(1 - \frac{1}{K}P(t)\right)P(t),$$

uz uslov

$$P(0) = P_0.$$

Na osnovu prethodne jednačine može se zaključiti da veličina populacije na početku raste eksponencijalno sa stopom rasta a, odnosno ponaša se u skladu sa Maltusovim modelom, a kasnije se taj rast smanjuje kako se veličina populacije približava svom ograničenju $K = \frac{a}{b}$ jer izraz u zagradi teži nuli. Konstanta K se naziva maksimalni kapacitet populacije. Za ljudsku populaciju je izračunato da je $a = 0.029$, $b = 2.695 \cdot 10^{-12}$, odnosno da je $K \approx 10{,}76 \cdot 10^9$. Osim na ljudsku ovaj model se može primeniti i na životinjsku i biljnu populaciju. Poznat je čuveni eksperiment

nemačkog biologa Gausa (Gause) koji je stavio pet jedinki Paramecium Caudatum-a u epruvetu u odgovarajaću sredinu, zbog čega je došlo do zasićenja posle četiri dana.

Primetimo da je:

$$\left(ln\frac{K-P(t)}{P(t)}\right)' = -\frac{P'(t)}{P(t)\left(1-\frac{P(t)}{K}\right)}.$$

Odatle sledi:

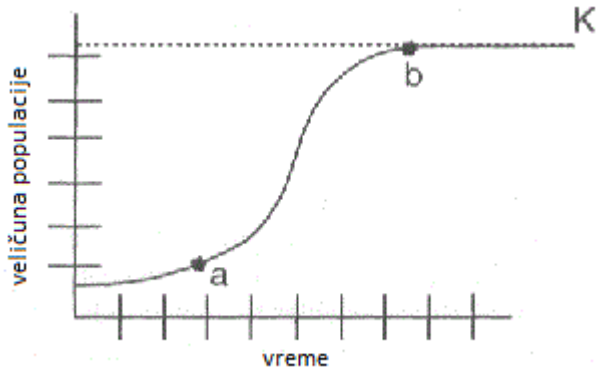
$$ln\left|\frac{K-P(t)}{P(t)}\right| = -at - c$$

$$\frac{K}{P(t)} - 1 = Ce^{-at}.$$

Iz početnog uslova sledi da je $C = \frac{K-P_0}{P_0}$, pa na osnovu toga rešenje sistema je dato sa:

$$P(t) = \frac{\frac{a}{b}}{1+\frac{a-bP_0}{bP_0}e^{-at}} = \frac{K}{1+Ce^{-at}}.$$

Ova kriva je poznata pod nazivom *logistička kriva* a prepoznatljiva je po svom S-obliku zbog čega se još naziva i S-kriva (slika 2).



Slika 2. Logistički Verhulstov model

Ako je $0 < P_0 < K$, tada je stopa rasta pozitivna, pa populacija raste kako raste t. Kada $t \to \infty$, kriva $P(t)$ ima horizontalnu asimptotu $P(t) = K$. Slučaj $P_0 > K$ se u praksi retko dešava a označava smanjenje populacije jer je tada $\lambda(t)$ negativno. Kod ljudske populacije to je posledica

vanrednih situacija kao što su ratovi, ekonomske krize, velike epidemije, loši klimatski uslovi i slično, koji dovode do velikog manjka hrane.

3 Logistička regresija

3.1 Regresija

Prilikom analiziranja podataka neophodno je naći oblik povezanosti između zavisne (kriterijumske) i nezavisnih (prediktorskih) promenljivih. Povezavnost između promenljivih se analizira regresijom. Reč regresija potiče od latinske reči *regressio* koja znači vraćanje, opadanje, odstupanje. U statistiku je dospela 1855. godine kada je Fransis Galton (Francis Galton) objavio publikaciju pod nazivom *Regression towards mediocrity in hereditary stature* u kojoj je analizirao visinu sinova u zavisnosti od visine očeva. Zaključak ove studije bio je da sinovi ekstremno visokih očeva nisu toliko visoki, odnosno da teže ka proseku.

Regresija podrazumeva zavisnost jedne slučajne promenljive od druge ili više njih. [6] Opšti model zavisnosti je

$$Y = f(X) + \varepsilon,$$

gde je ε slučajna promenljiva nezavisna od X, pri čemu X može biti skalarna ili vektorska veličina. [10] Ako je X skalarna veličina, radi se o jednostrukoj regresiji, inače se radi o višestrukoj. Funkcija f opisuje povezanost između X i Y.

Karakteristike povezanosti između promenljivih su

- smer koji može biti pozitivan ili negativan u zavisnosti od toga da li rastom jedne promenljive raste ili opada druga promenljiva
- stepen (jačina) povezanosti koji uzima vrednosti između -1 i 1. Ako je apsolutna vrednost stepena povezanosti bliska jedinici, govorimo o jakoj vezi između promenljivih, a ako je bliska nuli, govorimo o slaboj
- oblik (forma) koji može biti linearan ili nelinearan u zavisnosti od toga da li se veza između promenljivih može opisati linearnom ili nelinearnom funkcijom. Ako je promenljiva X vektorska promenljiva $(X_1, X_2, \dots, X_p)$, tada je linearni model zadat jednačinom $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, gde su β_i $i = 0, \dots, n$ koeficijenti linearne regresije.

Još jedan pokazatelj povezanosti između promenljivih je dijagram rasturanja (povezanosti). Svaka tačka na dijagramu rasturanja predstavlja par podataka o jednoj statističkoj jedinici. Sa dijagrama se može mnogo toga zaključiti o vezi između promenljivih. Na primer, ako su tačke raspršene bez ikakvog pravila, onda se na osnovu dijagrama zaključuje da ne postoji veza između njih. Takođe, ako je prava linija na dijagramu rasturanja najprihvatljivija za date opservacije onda postoji linearan odnos između promenljivih.

Cilj analize podataka je nalaženje funkcije koja ih najbolje aproksimira. Takav postupak se često naziva određivanje regresione linije. Na osnovu analize podataka se može utvrditi koliko je jaka zavisnost između X i Y, kolika je relativna važnost svake nezavisne promenljive u objašnjavanju zavisne, koja je najbolja predviđena vrednost zavisne promenljive za bilo koju kombinaciju nezavisnih promenljivih i koji se obim promene zavisne promenljive može očekivati za svaku jedinicu promene svake nezavisne promenljive. Upotrebom regresionih metoda takođe možemo za poznate vrednosti promenljive X predvideti vrednosti promenljive Y.

3.2 Uopšteni linearni modeli

Uopšteni linearni modeli (*general linear models*) predstavljaju uopštenje linearne regresije tako što dopuštaju obeležju da ima raspodelu iz eksponencijalne familije raspodela: binomne, normalne, Puasonove, gama raspodele i slično. Za slučajnu promenljivu Y se kaže da pripada eksponencijalnoj familiji raspodela ako se njena gustina ili raspodela verovatnoća može napisati na sledeći način:

$$f(y,\theta) = s(y)t(\theta)e^{a(y)b(\theta)} = e^{a(y)b(\theta)+\ln(s(y))+\ln(t(\theta))},$$

gde je θ parametar od koga zavisi raspodela, a a, b, s i t realne funkcije.

Osnovne karakteristike uopštenog linearnog modela su *komponenta slučajnsoti*, *komponenta sistematičnosti* i *funkcija veze*. [14], [11]

- Komponenta slučajnosti definiše uslovnu raspodelu obeležja Y_i (za i-tu od n nezavisnih vrednosti) za date vrednosti nezavisnih promenljivih u modelu.
- Komponenta sistematičnosti (linearno predviđanje, prediktor) predstavlja linearnu funkciju nezavisnih promenljivih i data je sa

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

- Funkcija veze je funkcija koja povezuje komponentu sistematičnosti sa funkcijom $\mu_i = E(Y_i)$ i data je sa

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

Zbog invertibilnosti funkcije veze važi da je [7]

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\big(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\big),$$

pa se uopšteni linearni modeli mogu posmatrati i kao linearni modeli transformacija srednje vrednosti obeležja ili kao nelinearni regresioni modeli obeležja.

3.3 Logistička regresija

Logistička regresija je jedan od modela uopštene linearne regresije u kome zavisna promenljiva uzima samo dve vrednosti (binarna promenljiva), a ređe više od dve, dok nezavisne promenljive mogu biti numeričke, kategorijalne ili njihova kombinacija. Zbog prirode zavisne promenljive, logistički regresioni model se naziva još i *binarni logistički regresioni model* (*Binary Logistic Regression Model*). Zavisna promenljiva se kodira tako što se jednom ishodu dodeljuje 1, a drugom 0, pri čemu je svejedno koji se ishod kodira jedinicom, a koji nulom. Na primer, zavisna promenljiva može biti da li je kupljen proizvod A ili proizvod B, da li je proizvod prošao kontrolu kvaliteta ili ne, da li je pacijent izlečen ili ne i slično. Zbog svoje prirode, ovakav vid regresije je pogodan u mnogim marketinškim, ekonomskim i demografskim istraživanjima.

Na primer, ako je zavisna promenljiva da li je proizvod prošao kontrolu proizvoda, onda se sa 0 može kodirati - proizvod nije prošao kontrolu proizvoda, a sa 1 - proizvod je prošao kontrolu proizvoda. Kao nezavisne promenljive možemo imati temperaturu, kvalitet materijala, vreme pravljenja i slično.

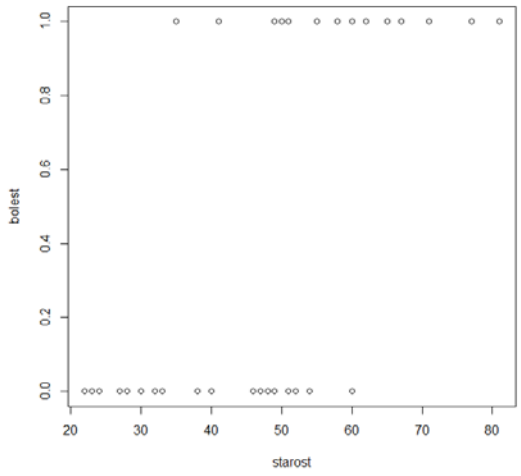
Sledećim primerom ćemo pokazati nedostatak primene linearne regresije kod binarne zavisne promenljive.

Primer 1 (Podaci su uzeti iz literature [23].): U tabeli 1 su dati podaci koji govore o tome da li određena osoba u zavisnosti od starosti boluje od povišenog krvnog pritiska ili ne. Rezultujuća promenljiva je binarna i uzima vrednost 0 ako osoba ne boluje od povišenog krvnog pritiska i vrednost 1 ako osoba boluje od povišenog krvnog pritiska. U ispitivanju su učestvovale 33 osobe.

starost	bolest	starost	bolest	starost	bolest
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Tabela 1.

U slučaju da je rezultujuća promenljiva neprekidna, dobija se sledeći grafik zavisnosti postojanja povišenog krvnog pritiska u odnosu na godine.



Slika 3. Veza između starosti i bolesti

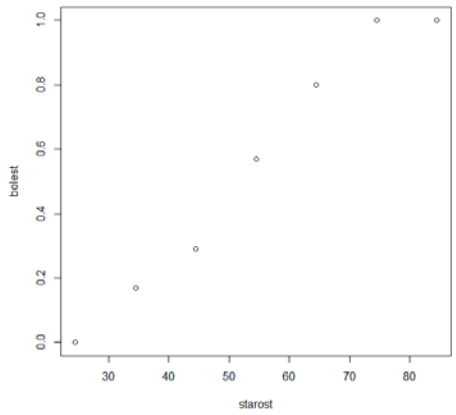
Sa grafika se jasno vidi da sve tačke pripadaju jednoj od dve paralelne prave koje predstavljaju prisustvo (bolest=1) odnosno odsustvo (bolest=0) povišenog krvnog pritiska. Grafik ne daje jasnu sliku o vezi između starosti i pojave povišenog krvnog pritiska, ali se može zaključiti da su starije osobe sklonije tom oboljenju od mlađih. Da bi se dala jasnija slika o ovoj zavisnosti, isptanici se dele u grupe prema starosti, odnosno formiraju se intervali za nezavisnu promenljivu i

računaju se proporcije za rezultujuću promenljivu $\frac{broj\ osoba\ sa\ povišenim\ krvnim\ pritiskom}{ukupan\ broj\ osoba\ u\ datoj\ grupi}$. Na taj način se dobijaju podaci koji su dati u tabeli 2.

starost	veličina grupe	broj obolelih	broj obolelih u procentima
20 – 29	5	0	0
30 – 39	6	1	17
40 – 49	7	2	29
50 – 59	7	4	57
60 – 69	5	4	80
70 – 79	2	2	100
80 – 89	1	1	100

Tabela 2.

Na slici 4 je prikazan grafik sa koga se jasno vidi da se povećanjem godina povećava i rizik od povišenog krvnog pritiska.



Slika 4. Veza između starosti i bolesti

3.4 Logit transformacija

Logistička ili logit transformacija se koristi za predviđanje verovatnoće nastupanja pojave koja je kodirana jedinicom. [6] Na primer, ako je jedinicom kodirana kupovina proizvoda A, a nulom kupovina proizvoda B, traži se verovatnoća da će određeni kupac kupiti proizvod A. Primenom (višestruke) linearne regresije se dobija rešenje u kojem će zavisna promenljiva uzeti vrednost između 0 i 1. Ako je ta vrednost bliža jedinici, tada se može zaključiti da će kupac kupiti priizvod A. Međutim, često se dešava da se primenom linearne regresije dobiju i vrednosti van opsega $[0, 1]$. Pošto se ovakve vrednosti ne mogu tumačiti kao verovatnoća, linearna regresija u ovakvim situacijama nije dobro rešenje.

Da bi se dobio logistički regresioni model, potrebno je izvršiti određenu transformaciju zavisne promenljive, koja se naziva *logit transformacija*. Od velikog značaja za logit transformaciju je očekivana vrednost zavisne promenljive u odnosu na datu vrednost nezavisne promenljive X, $E(Y|\boldsymbol{x})$. U opštem slučaju $E(Y|\boldsymbol{x})$ uzima vrednost između $-\infty$ i $+\infty$, ali kako je promenljiva Y binarna, $E(Y|\boldsymbol{x})$ uzima vrednosti ne manje od 0 i ne veće od 1. Promena u $E(Y|\boldsymbol{x})$ po jedinici promene za X postaje progresivno manja kako uslovna sredina postaje bliža 0 ili 1. [2]

Pošto je zavisna promenljiva Y dihotomna, promenljiva $Y|x$ takođe uzima dve vrednosti. [11] Neka je raspodela promenljive Y data na sledeći način:

$$Y: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix},$$

gde je $0 \leq \pi \leq 1$. Tada je slučajna promenljiva $Y|x$ određena raspodelom:

$$Y|\boldsymbol{x}: \begin{pmatrix} 0 & 1 \\ 1-\pi(\boldsymbol{x}) & \pi(\boldsymbol{x}) \end{pmatrix},$$

a njeno matematičko očekivanje je:

$$E(Y|\boldsymbol{x}) = 0 \cdot \big(1-\pi(\boldsymbol{x})\big) + \pi(\boldsymbol{x}) = \pi(\boldsymbol{x}).$$

U modelu logističke regresije odnos između verovatnoće π i promenljive X se predstavlja na sledeći način:

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0 + \sum \beta_k x_k}}{1 + e^{\beta_0 + \sum \beta_k x_k}}. \qquad (2)$$

Funkcija $\pi(\boldsymbol{x})$ ima pozitivan prvi izvod, pa je rastuća na svom domenu i ima horizontalnu asimptotu $\pi(\boldsymbol{x}) = 1$, odnosno ima S-oblik. Takođe, funkcija je nelinearna po parametrima

$\beta_0,\ \beta_1, \dots, \beta_p,$ ali se određenom logit transformacijom može dovesti do linearnog oblika. Važi sledeće: [3], [4]

$$\pi(\boldsymbol{x}) = \pi\left(x_1, x_2, \dots, x_p\right) = P\left(Y = 1 \middle| X_1 = x_1, X_2 = x_2, \dots, X_p = x_p\right) = \frac{e^{\beta_0 + \sum \beta_k x_k}}{1 + e^{\beta_0 + \sum \beta_k x_k}}$$

$$1 - \pi(\boldsymbol{x}) = 1 - \pi\left(x_1, x_2, \dots, x_p\right) = P\left(Y = 0 \middle| X_1 = x_1, X_2 = x_2, \dots, X_p = x_p\right) = \frac{1}{1 + e^{\beta_0 + \sum \beta_k x_k}}$$

$$\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = e^{\beta_0 + \sum \beta_k x_k}$$

$$g(\boldsymbol{x}) = ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Funkcija $ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ se naziva *logit transformacija* ili *logit funkcija.* Kako vrednost $\pi(\boldsymbol{x})$ pripada intervalu $[0,1]$, vrednost logit funkcije se kreće od $-\infty$ do $+\infty$. [1], [7]

Kod logističke regresije vrednost rezultujuće promenljive Y za dato $\boldsymbol{x}$ se može izraziti kao $Y|\boldsymbol{x} = \pi(\boldsymbol{x}) + \varepsilon$. Najčešći je slučaj da promenljiva ε ima normalnu raspodelu sa matematičkim očekivanjem koje je jednako nuli I konstantnom varijansom. Međutim, pošto je slučajna promenljiva Y binarna, ε ima binarnu raspodelu. Iz činjenice da je $Y|\boldsymbol{x} - \pi(\boldsymbol{x}) = \varepsilon$ sledi da promenljiva ε uzima vrednost $1 - \pi(\boldsymbol{x})$ sa verovatnoćom $\pi(\boldsymbol{x})$ (kada je $y = 1$) i vrednost $-\pi(\boldsymbol{x})$ sa verovatnoćom $1 - \pi(\boldsymbol{x})$ (kada je $y = 0$). Dakle, promenljiva ε ima sledeću raspodelu:

$$\varepsilon: \begin{pmatrix} 1 - \pi(\boldsymbol{x}) & -\pi(\boldsymbol{x}) \\ \pi(\boldsymbol{x}) & 1 - \pi(\boldsymbol{x}) \end{pmatrix}$$

a njeno matemetičko očekivanje i varijansa iznose $E(\varepsilon) = 0$ i $D(\varepsilon) = \pi(\boldsymbol{x})\left(1 - \pi(\boldsymbol{x})\right)$.

3.5 Verovatnoća i šansa događaja

Posmatrajmo binarnu promenljivu Y koja uzima vrednost 1 ako se posmatrani događaj desio i vrednost 0 ako se događaj nije desio. Srednja vrednost promenljive Y predstavlja verovatnoću ostvarenih događaja tj. $\mu = P(Y = 1)$. Na primer, ako je $\mu = 0.83$, tada se od svih posmatranih događaja ostvarilo njih 83%. [4]

Šansa događaja (*odds*) predstavlja količnik verovatnoće da se događaj desio i verovatnoće da se događaj nije desio. [8] Ako sa P označimo verovatnoću da se događaj desio, tada šansa događaja iznosi:

$$odds = \frac{P}{1-P}. \tag{3}$$

Na primer, ako je verovatnoća da se događaj desi 0.83, tada je šansa događaja $odds = 0.83/0.17 = 4.88$, što znači da događaj ima 4.88 puta veće šanse da se desi nego da se ne desi. Iz formule (3) se vidi da šansa događaja rastuća funkcija jer je prvi izvod šanse pozitivna funkcija.

Ako je poznata šansa događaja, verovatnoća događaja se može izračunati kao

$$P = \frac{odds}{1+odds}.$$

U slučaju da je verovatnoća ostvarenja događaja 0.5, tada je šansa događaja 1 i jednaka je šansi komplementarnog događaja. Što je šansa nekog događaja veća, veća je i verovatnoća da se taj događaj desi, pa na taj način možemo videti da li je veća verovatnoća da se desi neki događaj ili njemu komplementarni događaj.

4 Parametri logističke regresije

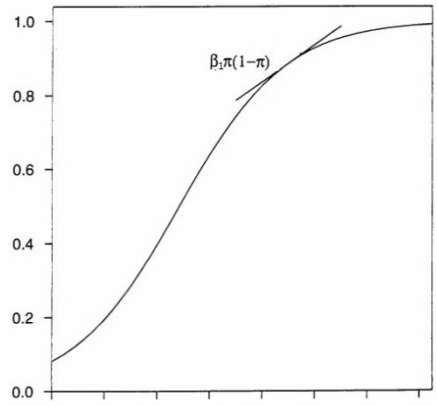
4.1 Značenje parametara

Posmatrajmo logistički model u kome je slučajna promenljiva X jednodimenzionalna, odnosno funkcija $\pi(x)$ je definisana na sledeći način:

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

Koeficijent β_1 se naziva koeficijent nagiba i ukazuje na stopu rasta ili opadanja funkcije $\pi(x)$ u zavisnosti od toga da li je β_1 pozitivna ili negativna konstanta. [8] Ako je $\beta_1 = 0$, tada je $\pi(x)$ konstantna za sve vrednosti x, pa kriva logističke regresije prelazi u horizontalnu pravu. Nagib tangente logističke krive je dat izrazom $\beta_1\pi(x)\big(1 - \pi(x)\big)$. Ako je, na primer, $\pi(x) = 0.5$ tangenta logističke krive ima nagib $0.25\beta_1$, a ako je $\pi(x) = 0.9$ ili $\pi(x) = 0.1$, taj nagib iznosi $0.09\beta_1$. Nagib se približava vrednosti 0, kako se $\pi(x)$ približava 0 ili 1. Najveći nagib tangente se postiže kada je $\pi(x) = 0.5$. [4]

Koeficijent β_1 se može izraziti kao razlika logita u tačkama $x + 1$ i x, pa se može protumačiti kao promena zavisne promenljive koja odgovara jediničnoj logita, odnosno $\beta_1 = g(x + 1) - g(x)$. [7]



Slika 5. S-kriva i njena tangenta

Neka je dat model sa dvodimenzionalnom slučajnom promeljivom X u kome se analizira da li su građani za ili protiv zakonodavstva u zavisnosti od pola i stepena privrženosti jednoj od dve ideologije: konzervatizmu ili liberalizmu. Pol (X_1) je binarna promenljiva koja uzima vrednost 0 ako je ispitanik žensko, odnosno 1 ako je muško. Stepen privrženosti ideologiji (X_2) može da uzme

vrednosti od -3 do 3. Ukoliko je negativan, ispitanik je okrenut više ka konzervatizmu, a ukoliko je pozitivan, okrenut je više ka liberalizmu. Stepen 0 govori da ispitanik nije okrenut nijednoj ideologiji. Zavisna promenljiva uzima vrednost 0, ako ispitanik ne podržava zakonodavstvo, a u suprotnom uzima vrednost 1. Pretpostavimo da je logit zadat formulom:

$$g(\boldsymbol{x}) = ln\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = 1.555 - 1.712x_1 - 0.513x_2.$$

Na primer, predviđena vrednost logita za muškarca kod koga je stepen privrženosti ideologiji 2 iznosi

$$g\big((1,2)\big) = 1.555 - 1.712 \cdot 1 - 0.513 \cdot 2 = -1.183.$$

Eksponencijalnom transformacijom ove vrednosti dobija se $e^{-1.183} = 0.306$ što znači da je predviđena šansa za glasanje u korist zakonodavstva kod ovakvog profila ljudi (muškarac, stepen privrženosti ideologiji iznosi 2) približno $1/3$. Na sličan način se mogu odrediti i ostale šanse.

Slobodan koeficijent koji u ovom modelu iznosi 1.555 predstavlja vrednost logita u tački $(0,0)$. Promenljiva pol uzima vrednost 0, ako je osoba ženskog pola, pa tako kod osoba ženskog pola čiji je stepen privrženosti ideologiji 0, logit uzima vrednost 1.555. Ako je posmatrana osoba muškog pola sa istom stepenom privrženosti ideologiji, njen logit iznosi -0.157. Eksponencijalnom transformacijom ovih vrednosti, dobijaju se predviđene vrednosti šansi za muški pol $e^{-0.157} = 0.855$ i za ženski pol $e^{1.555} = 4.735$. Odnos šansi u ovom slučaju iznosi $\frac{0.855}{4.735} = 0.1805$, što znači da kod osoba koje nisu okrenute nijednoj od ove dve ideologije muškarci imaju 0.1805 puta veće šanse da glasaju za zakonodavstvo od žena.

Koeficijent uz promenljivu koja označava pol je -1.712. Njegovom eksponencijalnom transformacijom se dobija $e^{-1.712} = 0.1805$ što je u stvari prethodno izračunata vrednost. Dakle, za binarnu promenljivu koja uzima vrednosti 0 i 1 važi da se, ako su ostale vrednosti promenljivih fiksirane, eksponencijalnom transformacijom koeficijenta koji stoji uz nju u jednačini logita, dobija odnos šansi kod binarne promenljive.

Ukoliko se fiksira vrednost promenljive koja označava pol (na primer, $X_1 = 0$), možemo izračunati vrednosti $g(\boldsymbol{x})$ i šanse glasanja u korist zakonodavstva $e^{g(x)}$ za različite vrednosti ideologije.

vrednosti promenljive X_2	$g(\boldsymbol{x})$	$e^{g(x)}$
3	0.016	1.017
2	0.529	1.697
1	1.042	2.835
0	1.555	4.735
-1	2.068	7.909
-2	2.581	13.21
-3	3.094	22.065

Tabela 3

U tabeli 3 možemo da uočimo pravilo po kojem se računaju šanse glasanja u korist zakonodavstva. Prilikom povećanja vrednosti promenljive X_2 za jednu jedinicu, vrednost šanse se pomnoži faktorom 0.599. Na primer, ako se vrednost šanse koja je dobijena za vrednost $x_2 = -3$ pomnoži sa 0.599, dobija se $22.065 \cdot 0.599 = 13.21$ što je vrednost šanse koja se dobije za vrednost $x_2 = -2$. Eksponencijalnom transformacijom koeficijenta koji stoji uz promenljivu X_2, dobija se vrednost $e^{-0.513} = 0.599$, što je upravo koeficijent kojim smo množili vrednosti šanse. Kod kvantitativne promenljive, ako su ostale promenljive fiksirane, eksponencijalnom transformacijom koeficijenta koji stoji uz tu promenljivu u jednačini logita, dobija se multiplikativni faktor kojim se, ako se pomnoži vrednost šanse, dobija vrednost šanse u tački te promenljive povećane za jednu jedinicu. [4]

4.2 Ocenjivanje parametara

Za ocenjivanje nepoznatih parametara kod višestruke linearne regresije se koristi metod najmanjih kvadrata koji minimizira sumu kvadrata odstupanja registrovanih vrednosti od predviđenih vrednosti dobijenih na osnovu modela. Na taj način se dobijaju ocene nepoznatih parametara koje poseduju osobine nepristrasnosti, efikasnosti i osobine BLUE (best linear unbiased estimator – najbolja linearna nepristrasna ocena) za mali broj podataka i osobine asimptotske nepristrasnosti, asimptotske efikasnosti i konzistentnosti za veliki broj podataka. Kada bi se metod najmanjih kvadrata primenio na binarni model, ocene ne bi imale ove karakteristike.

Prilikom ocenjivanja parametara logističke regresije koristi se metod maksimalne verodostojnosti. Ovaj metod daje vrednosti parametara $\beta_i, i = 0, \dots, p$ koje maksimiziraju verovatnoću dobijanja registrovanog skupa podataka. Funkcija verodostojnosti $\mathcal{L}(\boldsymbol{\beta}) = \mathcal{L}(\beta_0, \beta_2, \dots, \beta_p)$ koja se javlja u ovom metodu je funkcija nepoznatih parametara koja predstavlja verovatnoću koja kombinuje doprinose svih subjekata u ispitivanju. Da bi se našle ocene nepoznatih parametara pomoću metode maksimalne verodostojnosti, potrebno je naći maksimum funkcije verodostojnosti. Pod uslovom da je diferencijabilna, traže se one vrednosti parametara

koje su rešenje jednačine $\frac{d\mathcal{L}(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0$. Kako je logaritamska funkcija monotono rastuća na svom domenu, rešenja jednačina $\frac{d\mathcal{L}(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0$ se poklapaju sa rešenjima jednačine $\frac{dln\mathcal{L}(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = 0$, pa je nekad jednostavnije posmatrati funkciju $dln\mathcal{L}(\boldsymbol{\beta})$. Neka je dat uzorak od n nezavisnih registrovanih vrednosti parova $(\boldsymbol{x_i}, y_i)$, gde je $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \dots, x_{pi})$, a y_i uzima vrednost 0 ili 1 za svako $i = 1, \dots, n$.

Neka je raspodela za Y data sa

$$Y: \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}.$$

Na osnovu našeg modela imamo da je:

$$\pi(\boldsymbol{x_i}) = P(Y = 1 | X_i = \boldsymbol{x_i}) = \frac{e^{\beta_0 + \sum \beta_k x_{ki}}}{1 + e^{\beta_0 + \sum \beta_k x_{ki}}}$$

$$1 - \pi(\boldsymbol{x_i}) = P(Y = 0 | X_i = \boldsymbol{x_i}) = \frac{1}{1 + e^{\beta_0 + \sum \beta_k x_{ki}}}$$

za svako $i = 1, \dots, n$. Ako je $y_i = 1$, doprinos para $(\boldsymbol{x_i}, y_i)$ funkciji verodostojnosti je $\pi(\boldsymbol{x_i})$, a ako je $y_i = 0$, onda je doprinos jednak $1 - \pi(\boldsymbol{x_i})$, pa je doprinos para $(\boldsymbol{x_i}, y_i)$ funkciji verodostojnosti

$$\pi(\boldsymbol{x_i})^{y_i} \big(1 - \pi(\boldsymbol{x_i})\big)^{1-y_i}.$$

Registrovane vrednosti parova su međusobno nezavisne pa je funkcija verodostojnosti jednaka proizvodu pojedinačnih doprinosa parova $(\boldsymbol{x_i}, y_i)$ za svako $i = 1, \dots, n$, odnosno: [11]

$$\mathcal{L}(\boldsymbol{\beta}) = P\big(y_1, y_2, \dots, y_n\big) = P(y_1) \cdot P(y_2) \cdot \dots \cdot P\big(y_p\big)$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x_i})^{y_i} \big(1 - \pi(\boldsymbol{x_i})\big)^{1-y_i}$$

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right)^{y_i} \big(1 - \pi(\boldsymbol{x_i})\big).$$

Radi jednostavnosti, posmatraćemo logaritam funkcije verodostojnosti.

$$ln\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i ln \left(\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}\right) + ln\big(1 - \pi(\boldsymbol{x_i})\big) \right]$$

$$ln\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \boldsymbol{x_i^T \beta} - ln\left(1 + e^{x_i^T \boldsymbol{\beta}}\right) \right].$$

Izjednačavanjem izvoda funkcije $ln\mathcal{L}(\boldsymbol{\beta})$ sa nulom se dobija

$$\frac{dln\mathcal{L}(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = \sum_{i=1}^{n} \left(y_i - \frac{e^{x_i^T \boldsymbol{\beta}}}{1 + e^{x_i^T \boldsymbol{\beta}}} \right) \boldsymbol{x_i^T} = 0. \tag{4}$$

Kako je sistem jednačina (4) nelinearan po $\boldsymbol{\beta}$, rešenje sistema $\widehat{\boldsymbol{\beta}}$ se može naći nekom iteratvinom metodom. [3] Sistem se može napisati i u ekvivalentnom obliku:

$$\frac{dln\mathcal{L}(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = X^T\big(\boldsymbol{y} - \boldsymbol{E}(\boldsymbol{Y})\big) = X^T(\boldsymbol{y} - \boldsymbol{p}),$$

gde je $\boldsymbol{p}$ vektor generisan sa $p_i = P(Y = 1|\boldsymbol{x_i}) = \pi(\boldsymbol{x_i}) = \pi_i = \frac{e^{x_i^T \boldsymbol{\beta}}}{1+e^{x_i^T \boldsymbol{\beta}}}$ i $X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$.

Ako sa V označimo dijagonalnu matricu $V = \begin{bmatrix} \pi_1(1-\pi_1) & & \\ & \ddots & \\ & & \pi_n(1-\pi_n) \end{bmatrix}$, tada je Hesijanova matrica (kvadratna matrica generisana drugim parcijalnim izvodima) logaritma funkcije verodostojnosti data sa $\frac{d^2 ln\mathcal{L}(\boldsymbol{\beta})}{d\boldsymbol{\beta}^2} = -X^T V X$. Iterativna jednačina za dobijanje ocene $\widehat{\boldsymbol{\beta}}$ nepoznatog parametra $\boldsymbol{\beta}$ je:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \left(X^T V^{(m)} X\right)^{-1} X^T\left(\boldsymbol{y} - \boldsymbol{p}^{(m)}\right), \tag{5}$$

pa tako za početne vrednosti parametra $\beta = \beta^{(0)}$ nalazi se vrednost parametra $\beta = \beta^{(1)}$:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \left(X^T V^{(0)} X\right)^{-1} X^T\left(\boldsymbol{y} - \boldsymbol{p}^{(0)}\right),$$

odnosno:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \left(-\frac{d^2 ln\mathcal{L}\left(\boldsymbol{\beta}^{(0)}\right)}{d\boldsymbol{\beta}^{(0)^2}} \right)^{-1} \frac{dln\mathcal{L}\left(\boldsymbol{\beta}^{(0)}\right)}{d\boldsymbol{\beta}^{(0)}}.$$

Slično, zamenom u jednačinu (5), mogu se naći i vrednosti $\boldsymbol{\beta}^{(2)}$, $\boldsymbol{\beta}^{(3)}$... Ove vrednosti konvergiraju ka oceni nepoznatog parametra $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p\right)$. Posledica jednakosti (5) je da važi $\sum_{i=1}^{n} \boldsymbol{y_i} = \sum_{i=1}^{n} \widehat{\boldsymbol{\pi}}_i$, gde je gde je $\widehat{\boldsymbol{\pi}}_i = \widehat{\boldsymbol{\pi}}(\boldsymbol{x_i})$ ocena dobijena na osnovu metode maksimalne verodostojnosti, odnosno da je suma registrovanih vrednosti za $\boldsymbol{y}$ jednaka sumi predviđenih (očekivanih) vrednosti na osnovu modela. Za dovoljno velik uzorak važi da $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}\left(\boldsymbol{\beta}, \left(-\frac{d^2 ln\mathcal{L}(\widehat{\boldsymbol{\beta}})}{d\widehat{\boldsymbol{\beta}}^2}\right)^{-1}\right)$. [3]

Matrica $I = X^T V X$ se naziva informaciona matrica. Inverz te matrice, tj. $I^{-1} = Var(\widehat{\boldsymbol{\beta}})$ predstavlja kovarijacionu matricu na čijoj se dijagonali nalaze elementi Var$(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})$, a van dijagonale vrednosti kovarijacija $Cov(\hat{\beta}_i, \hat{\beta}_j)$.

4.2.1 GSK metod

GSK metod (Grizzle, Starmer, Koch metod) je zasnovan na neiterativnom metodu najmanjih kvadrata sa težinama. [3] Glavno ograničenje ovog metoda je to što zahteva da ocena verovatnoće π_i za većinu vrednosti x_i ne bude 0 ili 1. Podelom uzorka prema vrednosti zavisne promenljive, dobijaju se dve grupe za $y = 0$ i $y = 1$. Ako nezavisna promenljiva ima asimpotski normalnu raspodelu na svakoj od tih grupa sa različitim matematičkim očekivanjima i jednakim disperzijama, tj.

$$X|Y \sim N(\mu_j, \sigma^2), \qquad j = 0{,}1$$

tada su logistički koeficijenti dati sa:

$$\beta_0 = ln\frac{\theta_1}{\theta_0} - \frac{0.5(\mu_1^2 - \mu_0^2)}{\sigma^2}$$

$$\beta_1 = \frac{\mu_1 - \mu_0}{\sigma^2},$$

gde je $\theta_j = P(Y = j)$, za $j = 0{,}1$. Da bi se izračunale vrednosti β_0 i β_1, potrebno je iz uzorka oceniti parametre θ_0, θ_1, μ_0, μ_1 i σ^2. Parametar μ_j se ocenjuje kao sredina odgovarajuće podgrupe (formirane u zavisnosti od vrednosti y) tj. $\hat{\mu}_j = \overline{x}_j$, za $j = 0{,}1$, $\hat{\theta}_1 = n_1/n$, $\hat{\theta}_0 = 1 - \hat{\theta}_1$ i

$$\hat{\sigma}^2 = \frac{(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2}{n_0 + n_1 - 2},$$

gde je s_j^2 nepristrasna ocena σ^2 izračunata u odgovarajućoj podgrupi za $y = j$, $j = 0{,}1$, a n_0 i n_1 veličine odgovarajućih grupa.

Ovaj metod se može primeniti i kod višestruke logističke regresije. Slično kao u jednodimenzionalnoj logističkoj regresiji uzorak se deli u dve grupe u zavisnosti od ishoda zavisne promenljive. Pretpostavlja se da je uslovna raspodela $\boldsymbol{X}$ u svakoj od tih podgrupa p-dimenzionalna normalna raspodela sa vektorom matematičkih očekivanja koji zavisi od vrednosti y i kovarijacionom matricom koja ne zavisi od vrednosti y, tj:

$$\boldsymbol{X}|Y \sim N(\mu_j, \Sigma), \qquad j = 0,1$$

gde je $\boldsymbol{\mu_j}$ p-dimenzioni vektor matematičkih očekivanja za $j = 0,1$, a Σ kovarijaciona matrica dimenzije $p \times p$. Tada su koeficijenti dati jednačinama:

$$\beta_0 = ln\frac{\theta_1}{\theta_0} - 0.5(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^T\Sigma^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu_0})$$

$$\beta^* = (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^T\Sigma^{-1}$$

gde je $\boldsymbol{\beta}^* = (\beta_1, \beta_2, \dots, \beta_p)$ i $\theta_j = P(Y = j)$, za $j = 0,1$. Za ocenjivanje β_0 i β^* moraju se najpre oceniti parametri $\boldsymbol{\mu_0}$, $\boldsymbol{\mu_1}$, θ_0, θ_1 i Σ. Matematičko očekivanje μ_j se ocenjuje odgovarajućom sredinom podgrupe, tj. $\widehat{\boldsymbol{\mu}}_{\boldsymbol{j}} = \overline{\boldsymbol{x}}_{\boldsymbol{j}}$. Parametri θ_0 i θ_1 se ocenjuju kao i u jednodimenzionalnom slučaju. Ocena kovarijacione matrice je data sa:

$$S = \frac{(n_0 - 1)S_0 + (n_1 - 1)S_1}{n_0 + n_1 - 2}$$

gde su S_j nepristrasne ocene kovarijacionih matrica dimenzije $p \times p$, izračunate u odgovarajućoj podgrupi.

Primer 2: Ako je uzorak isti kao u primeru 1, nađimo ocene nepoznatih parametara β_0 i β_1. Prikazaćemo postupak ocenjivanja parametara u programskom jeziku R.

```
podaci<-read.table("C:/Users/User/Desktop/podaci.txt", header=T)
#ucitavamo podatke
model<-glm(bolest~starost, data=podaci, family="binomial")
#pravimo logisticki model
summary(model)

Call:
glm(formula = bolest ~ starost, family = "binomial", data = podaci)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7025  -0.6497  -0.2377   0.6508   2.1075
```

Coefficients:
Estimate Std. Error z value Pr(>|z|)
*(Intercept) -6.70846 2.35397 -2.850 0.00437 ***
starost 0.13150 0.04634 2.838 0.00454 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

(Dispersion parameter for binomial family taken to be 1)

*Null deviance: 44.987 on 32 degrees of freedom
Residual deviance: 28.672 on 31 degrees of freedom
AIC: 32.672*

Number of Fisher Scoring iterations: 5

Iz priloženog se može zaključiti da je ocena koeficijenta β_1 koji stoji uz promenljivu *starost* $\hat{\beta}_1$=0.13150, dok je $\hat{\beta}_0$=-6.70846.

4.3 Testiranje značajnosti parametara

Nakon ocenjivanja parametara višestruke logističke regresije, potrebno je naći najbolji model, odnosno model koji najbolje opisuje zavisnu promenljivu. Ovo uključuje formulisanje i testiranje statističkih hipoteza za određivanje da li su nezavisne promenljive u modelu značajno povezane sa zavisnom promenljivom, tj. da li značajno utiču na ponašanje zavisne promenljive. Testira se hipoteza H_0: promenljiva nije značajna protiv alternative H_1: promenljiva je značajna. Prilikom takvog testiranja treba porediti registrovane vrednosti rezultujuće promenljive sa vrednostima dobijenih pomoću dva modela, od kojih jedan sadrži a drugi ne sadrži promenljivu čija se značajnost testira. Ako su predviđene vrednosti na osnovu modela koji sadrži tu promenljivu bolje ili tačnije nego vrednosti koje su predviđene na osnovu modela koji ne sadrži tu promenljivu, tada je promenljiva u modelu značajna i model koji sadrži tu promenljivu se smatra boljim od modela koji je ne sadrži.

Postoji mnogo načina za testiranje značajnosti parametara a kao najefikasniji se pokazao test količnika verodostojnosti (*likelihood ratio test*).

4.3.1 Test količnika verodostojnosti

U logističkoj regresiji poređenje registrovane i predviđene vrednosti dobijene iz modela koji sadrži nezavisnu promenljivu i modela koji je ne sadrži, bazirano je na logaritmu funkcije

verodostojnosti. Pri tome se smatra da je registrovana vrednost zavisne promenljive ona predviđena vrednost koja se dobija na osnovu zasićenog modela. Zasićen, potpun ili kompletan model (*saturated model*) je onaj model koji ima onoliko parametara koliko i registrovanih vrednosti, tj. n . On reprodukuje podatke tačno, bez pojednostavljivanja, pa prema tome nije previše pogodan za interpretaciju. Najjednostavniji primer zasićenog modela je model proste linearne regresije koji ima samo dve tačke ($n = 2$).

Za poređenje registrovanih sa predviđenim vrednostima koristi se funkcija verodostojnosti:

$$D = -2ln\frac{\mathcal{L}}{\mathcal{L}_{sat}} = -2\sum_{i=1}^{n}\left[y_i ln\frac{\hat{\pi}_i}{y_i} + (1-y_i)ln\frac{1-\hat{\pi}_i}{1-y_i}\right],$$

gde je $\mathcal{L}_{sat}$ funkcija verodostojnosti zasićenog modela. Statistika $\frac{\mathcal{L}}{\mathcal{L}_{sat}}$ se naziva količnik verodostojnosti. Za statistiku $-2ln\frac{\mathcal{L}}{\mathcal{L}_{sat}}$ se koristi i naziv devijacija, pa se zbog toga ta statistika označava sa D. Dobija se da statistika D iznosi: [11]

$$D = 2ln\frac{\mathcal{L}_{sat}}{\mathcal{L}} = 2ln\mathcal{L}_{sat} - 2ln\mathcal{L}. \qquad (6)$$

Iz definicije zasićnog modela sledi da je $\hat{\pi}_i = y_i$, pa je funkcija verodostojnosti

$$\mathcal{L}_{sat} = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{1-y_i} = 1$$

odakle sledi da je $D = -2ln\mathcal{L}$.

Pri ispitivanju značajnosti nezavisne promenljive posmatraju se modeli sa i bez nezavisne promenljive. Označimo sa G promenu koja nastaje u D prilikom uključivanja nezavisne promenljive, odnosno: [9]

$$G = D(model\ bez\ nezavisne\ promenljive) - D(model\ sa\ nezavisnom\ promenljivom)$$

$$G = -2ln(verodostojnost\ modela\ bez\ nezavisne\ promenljive)$$

$$+2ln(verodostojnost\ modela\ sa\ nezavisnom\ promenljivom)$$

$$G = -2ln\frac{verodostojnost\ modela\ bez\ nezavisne\ promenljive}{verodostojnost\ modela\ sa\ nezavisnom\ promenljivom}.$$

Ako koristimo sledeće oznake: $n_1 = \sum y_i$, a $n_0 = n - \sum y_i$, statistika G je data sa:

$$G = -2ln\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i{}^{y_i}(1-\hat{\pi}_i)^{1-y_i}},$$

odnosno

$$G = 2\left\{\sum_{i=1}^{n}[y_i ln\hat{\pi}_i + (1-y_i)ln\hat{\pi}_i] - [n_1 lnn_1 + n_0 lnn_0 - nlnn]\right\}.$$

Pod hipotezom da je β_1 jednako nuli, statistika G ima hi-kvadrat raspodelu sa jednim stepenom slobode.

4.3.2 Wald-ov test

Dva asimptotski ekvivalentna testa koja se koriste za proveru značajnosti koeficijenata su Wald-ov test i score test. [3] Wald-ov test za jednodimenzionalnu logističku regresiju koristi statistiku koja je jednaka kvadratu količnika ocene maksimalne verodostojnosti parametra β_1, $\hat{\beta}_1$ i njene standardne greške $SE\hat{\beta}_1 = \sqrt{Var(\hat{\beta}_i)}$. Pod pretpostavkom da je $\beta_1 = 0$, statistika $Z = \frac{\hat{\beta}_1}{SE\hat{\beta}_1}$ ima približno normalnu $N(0,1)$ raspodelu. Wald statistika jednodimenzione logističke regresije je [19]:

$$Z^2 = \frac{\hat{\beta}_1{}^2}{SE\hat{\beta}_1{}^2} : \chi^2_{1.}$$

Ako se odbaci nulta hipoteza, tada se može zaključiti da će uklanjanje te promenljive iz modela značajno promeniti model, odnosno da je ta promenljiva statistički značajna.

Isti princip testiranja se primenjuje kod višestruke logističke regresije uz određenu modifikaciju test statistike. Takođe, u slučaju višestruke logističke regresije može se koristiti statistika

$$W = \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\left(Var(\hat{\boldsymbol{\beta}})\right)^{-1}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}X^T V X\hat{\boldsymbol{\beta}},$$

ako želimo da testiramo značajnost svih $p+1$ primenljivih. Statistika W pod pretpostavkom da su svih $p+1$ koeficijenata jednaki 0, ima χ^2_{p+1} raspodelu. Ukoliko ne želimo da testiramo značajnost koeficijenta β_0, tada se statistika W transformiše tako što se iz vektora $\hat{\boldsymbol{\beta}}$ eliminiše $\hat{\beta}_0$ a iz matrica X i V odgovarajuće kolone i vrste. U tom slučaju statistika W ima χ^2_p raspodelu.

Za velike uzorke, Wald test i test količnika maksimalne verodostojnosti daju približno iste rezultate. Međutim, za male uzorke se pokazalo da se rezultati mogu razlikovati i da u tim

situacijama test količnika verodostojnosti daje tačnije rezultate. Takođe, kod višestruke logističke regresije Wald-ov test često ne odbacuje nultu hipotezu iako su koeficijenti značajni.

4.3.3 Score test

Score test je zasnovan na prvom izvodu logaritma funkcije verodostojnosti i prilikom testiranja značajnosti koeficijenata ne zahteva njihovo prethodno ocenjivanje. Test je jednostavan za primenu, ali se, za razliku od Wald-ovog testa, ne nalazi u mnogim softverskim paketima, pa se zbog toga retko koristi. [5]

Radi jednostavnosti, posmatraćemo slučaj jednostruke logističke regresije. Verovatnoće $\pi(x_i)$ i $1-\pi(x_i)$ su zadate na sledeći način:

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

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

Logaritam funkcije maksimalne verodostojnosti je dat sa:

$$ln\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i(\beta_0+\beta_1 x_i) - ln\left(1+e^{\beta_0+\beta_1 x_i}\right)\right].$$

Diferenciranjem logaritma funkcije maksimalne verodostojnosti po parametrima β_0 i β_1 dobijaju se jednačine verodostojnosti:

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

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \qquad (7)$$

Score test koristi vrednosti ovih suma izračunatih za $\beta_0 = ln(n_1/n_0)$ i $\beta_1 = 0$. Koristeći ocene $\hat{\pi} = n_1/n = \overline{y}$ leva strana jednačine (7) postaje $\sum_{i=1}^{n} x_i[y_i - \overline{y}]$. Ocena varijanse je data formulom $\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i-\overline{x})^2$ gde je $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, a statistika koju koristi score test je data sa

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i-\overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

i ima normalnu $N(0,1)$ raspodelu. [3]

4.4 Intervali poverenja za parametre

Interval poverenja (pouzdanosti) nepoznatog parametra je interval koji sa verovatnoćom $(1-\alpha)100\%$ sadrži taj parametar, gde je α nivo značajanosti. Za nivo značajnosti se obično biraju vrednosti bliske 0: 0.1, 0.05, 0.01. Na primer, ako je nivo značajnosti 0.1, tada je pouzdanost da će nepoznati parametar biti sadržan u intervalu 90%. Interval poverenja je određen na sledeći način:

$$(ocenjena\ vrednost\ parametra - greška, ocenjena\ vrednost\ parametra + greška).$$

Kod proste logističke regresije intervali poverenja nepoznatih parametara se baziraju na ocenama koje su dobijene Wald-ovim testom (*Wald-based confidence intervals*). Intervali poverenja za koeficijente β_0 i β_1 sa nivoom značajnosti α su zadati sa:

$$\left(\hat{\beta}_0 - z_{1-\alpha/2} SE(\hat{\beta}_0),\ \ \hat{\beta}_0 + z_{1-\alpha/2} SE(\hat{\beta}_0)\right)$$

$$\left(\hat{\beta}_1 - z_{1-\alpha/2} SE(\hat{\beta}_1),\ \ \hat{\beta}_1 + z_{1-\alpha/2} SE(\hat{\beta}_1)\right)$$

gde je $SE(\cdot)$ ocena standardne greške odgovarajućeg parametra iz modela koji se koristi kao bazni, a $z_{1-\alpha/2}$ tablična vrednost standardne normalne raspodele za koju važi $P\left(\left|\hat{\beta}_i\right| \le z_{1-\alpha/2}\right) = 1-\alpha$, za $i = 0$ ili $i = 1$.

Takođe, na ovaj način se može naći i interval poverenja linearne funkcije koja povezuje promenljive, tj. interval poverenja za logit funkciju. Ocena za logit funkciju je data sa

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

Da bi se našao interval poverenja za logit, potrebno je oceniti disperziju prethodne sume:

$$Var\left(\hat{g}(x)\right) = Var\left(\hat{\beta}_0\right) + x^2 Var\left(\hat{\beta}_1\right) + 2xCov\left(\hat{\beta}_0, \hat{\beta}_1\right)$$

Interval poverenja sa nivoom značajnosti α za logit je dat sa: [3]

$$\left(\hat{g}(x) - z_{1-\alpha/2} SE\left(\hat{g}(x)\right),\ \ \hat{g}(x) + z_{1-\alpha/2} SE\left(\hat{g}(x)\right)\right)$$

gde je $SE\left(\hat{g}(x)\right)$ pozitivan koren od ocene za $Var\left(\hat{g}(x)\right)$.

Na osnovu ocene za logit može se naći ocena logističke verovatnoće π_i i njen interval poverenja. Ocena logističke verovatnoće iznosi:

$$\hat{\pi}_i = \frac{e^{\hat{g}(x_i)}}{1 + e^{\hat{g}(x_i)}},$$

a interval poverenja sa nivoom značajnosti α logističke verovatnoće iznosi:

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

Kod višestruke logističke regresije intervali poverenja za nepoznati parametar se dobijaju na isti način kao i kod jednostruke logističke regresije. Ocena logita se dobija pomoću formule

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p = x^T \hat{\beta}$$

gde je $\hat{\beta}^T = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, a $x^T = (1, x_1, \dots, x_p)$, tj. $x_0 = 1$.

Ocena varijanse logita je:

$$Var(\hat{g}(x)) = \sum_{j=0}^{p} x_j^2 \, Var(\hat{\beta}_j) + \sum_{j=o}^{p} \sum_{k=j+1}^{p} 2 \, x_j x_k Cov(\hat{\beta}_j, \hat{\beta}_k).$$

Koristeći jednakost $Var(\hat{\beta}) = (X^T \hat{V} X)^{-1}$, $Var(\hat{g}(x))$ se može da zapisati u sledećem obliku:

$$Var(\hat{g}(x)) = x^T Var(\hat{\beta}) x = x^T (X^T \hat{V} X)^{-1} x$$

Dalja izračunavanje se vrše analogno jednostrukoj logističkoj regresiji.

5 Tumačenje modela

Pod pretpostavkom da su promenljive u modelu značajne, želimo da izvedemo zaključak o modelu na osnovu koeficijenata. Radi jednostavnosti posmatrajmo model jednodimenzione logističke regresije čiji je logit zadat formulom $g(x) = \beta_0 + \beta_1 x$.

Prilikom tumačenja modela posmatraju se dva problema a to su:

- određivanje funkcionalne veze između zavisne i nezavisne promenljive
- definisanje jedinice promene za nezavisnu promenljivu.

5.1 Binarna nezavisna promenljiva

Slučaj binarne nezavisne promenljive predstavlja teorijsku osnovu za ostale slučajeve – sa polihotomnom i neprekidnom nezavisnom promenljivom. [3] Pretpostavimo da je promenljiva $\boldsymbol{X}$ kodirana nulom i jedinicom. Tada koeficijent nagiba β_1 možemo izraziti kao promenu logita po jedinici promene nezavisne promenljive, odnosno:

$$g(1) - g(0) = \beta_0 + \beta_1 - \beta_0 = \beta_1.$$

Da bismo bolje interpretirali koeficijent nagiba, izračunaćemo šanse u zavisnosti od vrednosti nezavisne promenljive. Uzimajući u obzir da je raspodela za $Y|X$ data sa:

$$Y|x: \begin{pmatrix} 0 & 1 \\ 1-\pi(x) & \pi(x) \end{pmatrix},$$

pa primenom formule (3) važi da je šansa da zavisna promenljiva uzme vrednost 1, kada nezavisna promenljiva uzme vrednost 0 je data sa

$$\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} = \frac{P(Y=1|X=0)}{P(Y=0|X=0)} = \frac{\pi(0)}{1-\pi(0)}.$$

Šansa da zavisna promenljiva uzme vrednost 1, kada nezavisna promenljiva uzme vrednost 1 je data sa

$$\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} = \frac{P(Y=1|X=1)}{P(Y=0|X=1)} = \frac{\pi(1)}{1-\pi(1)}.$$

Odnos šansi (odds ratio) je pozitivna konstanta data sa:

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \tag{8}$$

i predstavlja odnos šansi ostvarenja događaja u grupi u kojoj nezavisna promenljiva uzima vrednost 1 i u grupi u kojoj nezavisna promenljiva uzima vrednost 0.

Koristeći to da je $\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$ i $\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$, daljim izračunavanjem se dobija:

$$OR = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Dakle, ako je nezavisna promenljiva binarna, koeficijent β_1 je jednak prirodnom logaritmu odnosa šansi. [9] Takođe, možemo naći i interval poverenja slučajne promenljive OR. Raspodela promenljive $lnOR = \beta_1$ je normalna, a za velike uzorke i OR ima normalnu raspodelu. Interval poverenja za OR sa nivoom značajnosti α dat je sa:

$$\left(e^{\widehat{\beta}_1 - z_{1-\alpha/2} SE(\widehat{\beta}_1)},\ e^{\widehat{\beta}_1 + z_{1-\alpha/2} SE(\widehat{\beta}_1)}\right)$$

gde je $SE\left(\hat{\beta}_1\right)$ standardno odstupanje slučajne promenljive $lnOR$. Njegova vrednost se može izračunati iz kovarijacione matrice, ali kako je slučajna promenljiva X jednodimenzionalna i uzima samo dve vrednosti, ono se može izračunati i na sledeći način:

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{1}{1-\pi(0)} + \frac{1}{\pi(0)} + \frac{1}{1-\pi(1)} + \frac{1}{\pi(1)}}.$$

Primer 3 (Podaci su uzeti iz literature [20].): U ispitivanju povezanosti konzumiranja alkohola i raka jednjaka je učestvovalo 975 osoba. Nezavisna promenljiva označava konzumiranje alkohola i kodirana je sa 0 ako osoba ne konzumira alkohol, a sa 1 ako konzimira. Zavisna promenljiva je kodirana sa 0 ako osoba ne boluje od raka jednjaka, a sa 1 ako osoba boluje. Na osnovu istraživanja, dobijeni su sledeći rezultati:

prisustvo raka jednjaka	konzumiranje alkohola		ukupno
	da	ne	
da	96	104	200
ne	109	666	775
ukupno	205	770	975

Tabela 4

Pomoću formule za odnos šansi izračunaćemo koliko puta veću šansu da dobiju rak jednjaka imaju alkoholičari.

$$\widehat{OR} = \frac{96/109}{104/666} = 5.64.$$

Dakle, osobe koje konzumiraju alkohol imaju 5.64 puta veće šanse da obole od raka jednjaka nego osobe koje ne konzumiraju alkohol. Koeficijent β_1 i standardno odstupanje promenljive $ln\widehat{OR}$ iznose $\beta_1 = ln\widehat{OR} = 1.7299,\ SE\left(ln\widehat{OR}\right) = 0.1752.$ Interval poverenja sa

nivoom značajnosti $\alpha = 0.05$ za 1 je $e^{1.7299 \mp 1.96 \cdot 0.1752} = (4, 7.95)$, a interval poverenja sa nivoom značajnosti $\alpha = 0.1$ je $e^{1.7299 \mp 1.645 \cdot 0.1752} = (4.23, 7.52)$.

Relativni rizik (*relative risk*) predstavlja odnos verovatnoća uspeha u okviru dve grupe.

$$RR = \frac{P(Y = 1|X = 1)}{P(Y = 1|X = 0)} = \frac{\pi(1)}{\pi(0)} = \frac{1 - \pi(1)}{1 - \pi(0)} OR.$$

U slučaju kada $\frac{1-\pi(1)}{1-\pi(0)} \to 1$, relativni rizik i odnos šansi imaju približno jednake vrednosti. To se dešava u slučaju kada su verovatnoće uspeha u obe grupa približne, odnosno kad je verovatnoća $\pi(x)$ mala bilo da je $x = 0$ ili $x = 1$. U praksi ova situacija se javlja kod ispitivanja retkih bolesti. Ukoliko je vrednost $RR > 1$, veću verovatnoću ostvarenja posmatrani događaj ima u grupi u kojoj nezavisna promenljiva uzima vrednost 1, a ako važi $RR < 1$, tada veću verovatnoću ostvarenja ima u grupi u kojoj nezavisna promenljiva uzima vrednost 0.

Prilikom utvrđivanja intervala poverenja za relativni rizik, koristi se statistka $ln\widehat{RR}$ koja ima normalnu raspodelu. Interval poverenja sa nivoom značajnosti α je dat sa

$$\left(e^{ln\widehat{RR} - z_{1-\alpha/2} SE(ln\widehat{RR})},\ e^{ln\widehat{RR} + z_{1-\alpha/2} SE(ln\widehat{RR})}\right)$$

gde je $SE\left(ln\widehat{RR}\right)$ standardno odstupanje $ln\widehat{RR}$ dato formulom

$$SE\left(ln\widehat{RR}\right) = \sqrt{\frac{1}{N_0 \pi(0)} + \frac{1}{N_1 \pi(1)} + \frac{1}{N_0} + \frac{1}{N_1}},$$

gde je $N_1 = \sum x_i$, a sa $N_0 = N - \sum x_i$.

Primer 4 (Podaci su uzeti iz literature [5]): Ispitujemo povezanost starosti (nezavisna promenljiva) sa postojanjem HIV virusa (zavisna promenljiva). U istraživanju je učestvovalo 1496 osoba koje su podeljene u dve starosne grupe – mlađe (kodirano sa 0) i starije od 30 godina (kodirano sa 1). Podaci su dati u tabeli 6.

Y/X	mlađi od 30 godina	30 godina i stariji	ukupno
HIV negativni	623	816	1439
HIV pozitivni	18	39	57
ukupno	641	855	1496

Tabela 5

Relativnim rizikom upoređujemo verovatnoću oboljenja od HIV-a u ove dve grupe.

$$\widehat{RR} = \frac{39/855}{18/641} = 1.62$$

Na osnovu ovoga može se zaključiti da osobe koje imaju 30 godina i više imaju veću verovatnoću oboljenja od HIV-a. Posle računanja se dobija da je $ln\widehat{RR} = 0.48$, $SE\left(ln\widehat{RR}\right) = \sqrt{\frac{1}{18} + \frac{1}{39} + \frac{1}{641} + \frac{1}{855}} = 0.29$, pa je interval poverenja sa nivoom značajnosti $\alpha = 0.05$ dat sa $e^{0.48 \mp 1.96 \cdot 0.29} = (0.915,\ 2.853)$.

5.2 Politohomna nezavisna promenljiva

U logističkoj regresiji se često javljaju kategorijalne (kvalitativne) promenljive, tj. promenljive koje se mogu meriti samo u smislu pripadanja nekoj grupi (kategoriji), pri čemu su te grupe disjunktne. Najčešći primeri takvih promenljivih su rasa, pol, stepen obrazovanja, opština boravka u republici... Da bi se izučavala kategorijalna promenljiva koja ima više od 2 kategorije, ona se pretvara u niz indikatora. Koeficijent za svaki takav indikator predstavlja efekat te kategorije upoređene sa kategorijom koja nije uključena u taj model (referentna kategorija). Kada se prave indikatori, jedna od najvažnijih stavki je izbor referentne kategorije. Uglavnom se bira ona kategorija koja je opravdana sa biološkog stanovišta (tj. kategorija za koju postoji opravdanje da bude referentni nivo) i ona koja ima prihvatljiv broj posmatranja. [9] . Cilj ovakvog istraživanja je da se uporede šanse određene grupe u odnosu na referentnu grupu.

Primer 5 (Podaci su uzeti iz literature [22].): U sledećoj tabeli su dati podaci koji govore o tome da li je student položio izabrani kurs iz prvog pokušaja ili nije. Student bira jedan od tri kursa koje smo označili sa *kurs_1*, *kurs_2* i *kurs_3*. Ukupno je ispitano 394 studenta.

status	izabrani kurs			
	kurs_1	kurs_2	kurs_3	ukupno
nije položio	27	7	124	236
položio	8	24	204	158
ukupno	35	31	328	394
$\widehat{OR}$	0.1801	2.0840	1	
$ln\widehat{OR}$	-1.7142	0.7343	0	

Tabela 6

Kao referentnu promenljivu izabraćemo *kurs_3*. Dizajn promenljiva je data u tabeli 7. Na primer, šansa za *kurs_1* iznosi $\widehat{OR} = (8/27)/(204/124) = 0.1801$, što znači da studenti koji su izabrali *kurs_1* imaju 0.1801 puta veće šanse da polože iz prve nego studenti koji su izabrali *kurs_3*.

izabrani kurs (kod)	izabrani_kurs_1	izabrani_kurs_2
kurs_1	1	0
kurs_2	0	1
kurs_3	0	0

Tabela 7

U programskom jeziku R ćemo oceniti koeficijente sledećeg logističkog modela:

status=$\beta_1 \cdot$ *izabrani_kurs_1*+$\beta_2 \cdot$ *izabrani_kurs_2*+β_0

gde promenljiva *status* može da uzme vrednost 0 (ako kurs nije položen iz prve) i vrednost 1 (ako je kurs položen iz prve).

Kod programa je sledeći:

```
podaci1<-read.table("C:/Users/User/Desktop/podaci1.txt", header=T)
model<-glm(status~izabrani_kurs_1 + izabrani_kurs_2, data=podaci1, family="binomial")
summary(model)

Call:
glm(formula = status ~ izabrani_kurs_1 + izabrani_kurs_2, family = "binomial",
    data = podaci1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7251  -1.3948   0.9746   0.9746   1.7181

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.4978     0.1139   4.372 1.23e-05 ***
izabrani_kurs_1  -1.7142     0.4183  -4.098 4.17e-05 ***
izabrani_kurs_2   0.7343     0.4444   1.652   0.0985 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

Null deviance: 530.66 on 393 degrees of freedom
Residual deviance: 505.74 on 391 degrees of freedom
AIC: 511.74

Number of Fisher Scoring iterations: 4

Ocene koeficijenata i njihovo standardno odstupanje su dati sa:

promenljiva	ocena koeficijenta	standardno odstupanje
izabrani_kurs_1	-1.7142	*0.4183*
izabrani_kurs_2	0.7343	*0.4444*
slobodan koeficijent	*0.4978*	*0.1139*

Tabela 8.

Ocene koeficijenata smo mogli da nađemo i na sledeći način:

$$ln\left(\widehat{OR}(izabrani\_kurs\_1, kurs\_3)\right) = \hat{\beta}_1 = -1.7142$$

$$ln\left(\widehat{OR}(izabrani\_kurs\_2, kurs\_3)\right) = \hat{\beta}_2 = 0.7343.$$

Sada ćemo pokazati da ovo važi zbog načina na koji smo kreirali dizajn promenljivu. Za primer ćemo uzeti $izabrani\_kurs\_1$ i $kurs\_3$.

$$ln\left(\widehat{OR}(izabrani\_kurs\_1, kurs\_3)\right) = \hat{g}(izabrani\_kurs\_1) - \hat{g}(kurs_3) =$$

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot (izabrani\_kurs\_1 = 1) + \hat{\beta}_2 \cdot (izabrani\_kurs\_2 = 0) -$$

$$\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot (izabrani\_kurs\_1 = 0) + \hat{\beta}_2 \cdot (izabrani\_kurs\_2 = 0)\right) = \hat{\beta}_1$$

Na osnovu standardnog odstupanja i vrednosti koeficijenata, mogu se naći intervali poverenja što je prikazano u prethodnom primeru.

5.3 Neprekidna nezavisna promenljiva

Opšti postupak ocene odnosa šansi za jednu jedinicu priraštaja promenljive $\boldsymbol{X}$ je data je u sledećim koracima:

1. iz jednačine za logit $g(x) = \beta_0 + \beta_1(x)$ dobija se $g(x+1) = \beta_0 + \beta_1(x+1)$
2. ocena koeficijenta β_1 na osnovu logita je $\hat{\beta}_1 = \hat{g}(x+1) - \hat{g}(x)$
3. ocena odnosa šansi je data sa $\widehat{OR} = e^{\hat{\beta}_1}$.

Međutim, u slučaju neprekidne promenljive promena za jednu jedinicu nije previše interesantna. Na primer, povećanje telesne težine za jedan kilogram kod odraslih osoba je od manjeg značaja nego povećanje težine za 10 kilograma. Ili, ako je opseg slučajne promenljive interval $[0,1]$, promena za jednu jedinicu je suviše velika i realnije je posmatrati promenu za 0.01 ili 0.05.

Da bi se obezbedila pravilna interpretacija, metod procene odnosa šansi se prilagođava tačkama intervala na kome je definisana promenljiva X. Taj interval se deli na određeni broj jedinica. Ako se desila promena od c jedinica, razlika ocene logita u tačkama $x + c$ i x iznosi:

$$\hat{g}(x+c) - \hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1(x+c) - \hat{\beta}_0 - \hat{\beta}_1 x = c\hat{\beta}_1.$$

Na osnovu toga, ocena odnosa šansi je data sa $\widehat{OR} = \widehat{OR}(x+c, x) = e^{c\hat{\beta}_1}$. Za ocenu standardnog odstupanja $c\hat{\beta}_1$ važi $SE(c\hat{\beta}_1) = |c|SE(\hat{\beta}_1)$. Interval poverenja sa nivoom značajnosti α za $\widehat{OR}(x+c, x)$ je dat sa

$$\left(e^{c\hat{\beta}_1 - z_{1-\alpha/2}|c|SE(\hat{\beta}_1)},\ e^{c\hat{\beta}_1 + z_{1-\alpha/2}|c|SE(\hat{\beta}_1)}\right).$$

Primer 6 (Podaci su uzeti iz literature [3]): Neka zavisna promenljiva predstavlja prisustvo ili odsustvo srčanog oboljenja, a nezavisna promenljiva X starost pacijenta i neka je logit ocenjen sa $\hat{g}(x) = -1.44 + 0.038x$. Ocenjeni odnos šansi ima sledeći oblik: $\widehat{OR}(x+c, x) = e^{0.038c}$ Ako bismo hteli da vidimo kako povećanje starosti od 10 godina utiče na pojavu srčane bolesti u prethodnom izrazu c ćemo zameniti sa 10 i dobiti $\widehat{OR}(10) = e^{0.038 \cdot 10} = 1.46$. Dakle, sa svakim povećanjem starosti od 10 godina rizik za pojavu srčanog oboljenja se povećava 1.46 puta.

5.4 Interakcija između promenljivih

Ukoliko je uticaj nezavisne promenljive na zavisnu posredan, odnosno zavisi od vrednosti neke treće promenljive, tada kažemo da su takve promenljive u međusobnoj interakciji, a promenljivu koja utiče na promenu odnosa između promenljivih nazivamo uticajna promenljiva (*moderator variable*) i označavamo sa $\boldsymbol{Z}$. Na primer, ako psihoterapija smanjuje depresiju kod muškaraca više nego kod žena, kažemo da pol utiče na vezu između psihoterapije i depresije. [13]

Pretpostavimo da je dat model sa dve nezavisne promenljive, od kojih je jedna neprekidna (X), a druga binarna (Z), čiji je logit zadat sledećom jednačinom:

$$g(x,z) = \beta_0 + \alpha x + \beta_2 z. \qquad (9)$$

Ako se uticaj promenljive X na zavisnu promenljivu razlikuje u zavisnosti od vrednosti promenljive Z, onda kažemo da se javlja efekat interakcije između promenljivih. [4] Koeficijent koji odražava efekat delovanja promenljive X na zavisnu promeljivu je linearna funkcija od Z i zadat je jednačinom:

$$\alpha = \beta_1 + \beta_3 z.$$

Prema ovoj formulaciji, pri promeni promenljive Z za jednu jedinicu, vrednost koeficijenta β_1 se promeni za β_3 jedinica. Zamenom u izraz (9) dobija se

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

$$= \beta_0 + \beta_1 x + \beta_3 xz + \beta_2 z$$

$$= \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$$

gde je sa xz označena vrednost interakcije između promenljivih X i Z.

Da bismo ocenili odnos šansi, posmatraćemo logite za oba vrednosti binarne promenljive. Eksponencijalnom transformacijom njihove razlike, dobija se ocena odnosa šansi.

Logit ovog modela dat je jednačinom

$$g(x,z) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz.$$

Ako sa z_0 i z_1 označimo vrednosti binarne promenljive, tada se zamenom u prethodnu jednačinu dobija:

$$g(x,z_0) = \beta_0 + \beta_1 x + \beta_2 z_0 + \beta_3 xz_0$$

$$g(x, z_1) = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 x z_1.$$

Računanjem razlike ova dva izraza, dobija se logaritam odnosa šansi.

$$g(x, z_1) - g(x, z_0) = \beta_0 + \beta_1 x + \beta_2 z_0 + \beta_3 x z_0 - (\beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 x z_1) =$$

$$\beta_2(z_1 - z_0) + \beta_3 x(z_1 - z_0)$$

$$OR = e^{\beta_2(z_1 - z_0) + \beta_3 x(z_1 - z_0)}.$$

Da bismo našli interval poverenja, potrebno je naći varijansu ocene $ln\widehat{OR}$

$$Var\left(ln\widehat{OR}\right) = (z_1 - z_0)^2 Var\left(\hat{\beta}_2\right) + x^2(z_1 - z_0)^2 Var\left(\hat{\beta}_3\right) + 2x(z_1 - z_0)^2 Cov\left(\hat{\beta}_2, \hat{\beta}_3\right).$$

Interval poverenja za $\widehat{OR}$ sa nivoom značajnosti α je dat sa

$$\left(e^{\hat{\beta}_2(z_1 - z_0) + \hat{\beta}_3 x(z_1 - z_0) - z_{1-\alpha/2} SE(ln\widehat{OR})},\ e^{\hat{\beta}_2(z_1 - z_0) + \hat{\beta}_3 x(z_1 - z_0) + z_{1-\alpha/2} SE(ln\widehat{OR})}\right).$$

gde je $SE\left(ln\widehat{OR}\right)$ pozitivan koren veličine $Var\left(ln\widehat{OR}\right)$.

5.5 Identifikacija važnih opservacija

5.5.1 Autlajeri

Autlajeri predstavljaju ekstremne vrednosti u uzorku, tj. vrednosti koje vidljivo odstupaju od drugih. Statističari na različite načine definišu autlajere, u zavisnosti od strukture podataka. Hokins (Hawkins, 1980) definiše autlajer kao opažanje koje mnogo odstupa od drugih opažanja pod sumnjom da su generisani različitim mehanizmom. Barnet i Luis (Barnett i Lewis, 1994) ukazuju da udaljeno opažanje, ili autlajer, je opažanje koje deluje da odstupa upadljivo od drugih članova posmatranog uzorka. Džonson (Johnson, 1992) definiše autlajer kao opažanje u skupu podataka koje deluje da je nekonzistentno sa ostatkom skupa podataka. [5]

Autlajeri se često smatraju greškom merenja, mada to ne moraju biti. To mogu biti podaci dobijeni za neke ekstremne vrednosti nezavisne promenljive. U logističkoj regresiji najčešće nastaju kao pogrešna vrednost zavisne promenljive. Neotkriveni autlajeri mogu voditi ka pogrešnoj specifikaciji modela, pristrasnoj oceni parametara, smanjenju značajnosti promenljivih i netačnim rezultatima. Stoga je važno identifikovati autlajere pre analize podataka. Autlajeri nekad mogu nositi važnu informaciju o modelu, pa bi njihovo uklanjanje povlačilo gubitak te informacije. Pre uklanjanja autlajera potrebno je ispitati njihov uticaj na model, kao i detaljnu proveru modela sa i bez autlajera (broj promenljivih, značajnost koeficijenata i standardne greške) i odabrati najbolji model.

Ako model sadrži jednu ili dve nezavisne promenljive, autlajer se može identifikovati grafičkom metodom. Nakon kreiranja modela koji opisuje podatke i njegovim grafičkim predstavljanjem, potrebno je pronaći tačke koje su izdvojene od drugih. Takve podatke smatramo mogućim autlajerima. U slučaju da model sadrži više od dve nezavisne promenljive, ova metoda nije primenljiva.

Jedan od najpopularnijih načina utvrđivanja da li je nešto autlajer ili ne je Kukovo rastojanje (Cook's distance). [21] Kukovo rastojanje se računa po formuli

$$CD_i = \frac{\left(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}}\right)^T (X^T V X)\left(\widehat{\boldsymbol{\beta}}^{(-i)} - \widehat{\boldsymbol{\beta}}\right)}{(p+1)\hat{\sigma}^2} \qquad i = 1, \dots, n$$

gde je $\widehat{\boldsymbol{\beta}}^{(-i)}$ ocena vektora $\boldsymbol{\beta}$ bez i-te opservacija, a $\hat{\sigma}$ pozitivan koren srednje kvadratne greške ocenjivanja, tj. $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. Ukoliko je $CD_i \geq \frac{4}{n-(p+1)}$, tada se i-ta opservacija može smatrati autlajerom.

U mnogim statističkim softverima postoji ugrađena funkcija koja prema Kukovom rastojanju izdvaja autlajere.

5.5.2 Delta-beta statistika

Statistika delta-beta (deletion displacement) meri promenu izazvanu brisanjem svih zapažanja koja se poklapaju sa odabranim elementom iz skupa vrednosti nezavisnih promenljivih i koristi se za otkrivanje zapažanja koje imaju jak uticaj na procenu modela. [16] Na ovaj način se može utvrditi da li uklanjanje nekog od elemenata skupa vrednosti nezavisnih promenljivih ima značajnog uticaja na ocene odnosa šansi i koeficijenata u modelu.

Pretpostavimo da model sadrži p različitih nezavisnih promenljivih $\boldsymbol{x^T} = (x_1, x_2, \dots, x_p)$ i neka se uzorak sastoji od n elemenata. Neka je m_j, $j = 1, \dots, J$ broj elemenata koji imaju iste vrednosti nezavisnih promenljivih kao određen element skupa vrednosti nezavisne promenljive, a y_j broj elemenata za koje zavisna promenljiva uzima vrednost 1 među m_j elemenata i neka je $\sum y_j = n_1$. [1]

Da bismo ovo objasnili, posmatrajmo primer sa nezavisnim promenljivama koje označavaju pol, rasu i visinu. Veličina uzorka je $n = 8$, dok je broj različitih vrednosti nezavisnih promenljivih $J = 6$. Podelom u grupe prema različitim vrednostima imamo:

- jednu belkinju visine 167, $m_1 = 1$
- jednog belca visine 202, $m_2 = 1$
- dve crnkinje visine 156, $m_3 = 2$
- jednog crnca visine 189, $m_4 = 1$
- dva belca visine 139, $m_5 = 2$

- jednu crnkinju visine 142, $m_6 = 1$.

Promena koja se dešava prilikom izbacivanja svih opservacija koje se poklapaju sa j –im elementom skupa vrednosti nezavisne promenljive se izračunava kao [15]:

$$\Delta\beta_j = \frac{r(y_j, \hat{\pi}_j)^2 h_{jj}}{1 - h_{jj}},$$

gde je h_{jj} j-ta dijagonalna vrednost matrice $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$, a sa W dijagonalna matrica koja je generisana elementom $w_{jj} = m_j\hat{\pi}_j(1 - \hat{\pi}_j)$, a sa $r(y_j, \hat{\pi}_j)$ *Pirsonov rezidual* (Pearson's residual) koji se definiše kao [15]:

$$r(y_j, \hat{\pi}_j)\frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}}, \quad j = 1, \dots, J.$$

Matrica $W^{1/2}$ je generisana elementima ${w_{jj}}^{1/2}$. Umesto delta-beta vrednosti, može se koristiti i standardizovana delta-beta vrednost (*standardized deletion displacement*) koja se izračunava na sledeći način:

$$\Delta\beta s_j = \frac{rs(y_j, \hat{\pi}_j)^2 h_{jj}}{1 - h_{jj}},$$

gde je sa $rs(y_j, \hat{\pi}_j), j = 1, \dots, J$ označen *standardizovan Pirsonov rezidual* koji se definiše na sledeći način:

$$rs(y_j, \hat{\pi}_j)\frac{r(y_j, \hat{\pi}_j)}{\sqrt{1 - h_{jj}}}.$$

Veće vrednosti ovih veličina ukazuju na značajnost posmatranog elementa skupa vrednosti nezavisnih promenljivih za model.

6 Kreiranje modela i procena slaganja modela sa podacima

6.1 Izbor promenljivih u model

Proces izgradnje logističkog regresionog modela je sličan procesu izgradnje modela linearne regresije. Prilikom ovakve analize potrebno je međusobno upoređivanje modela i izbor modela koji najbolje objašnjava date podatke. Takav model bi trebalo da ima što manje promenljivih jer povećanje broja promenljivih povlači sa sobom povećanje standardnih grešaka ocena, pa model postaje numerički nestabilan. Izbor promenljivih koje najbolje opisuju podatke ćemo prikazati kroz nekoliko koraka. [12]

Korak 1. Deskriptivna analiza

Tokom deskriptivne analize potrebno je utvrditi raspodelu svake od promenjivih, na primer konstrukcijom štapićastih dijagrama ili kreiranjem histograma ili box-plot-ova. Takođe, neophodno je proveriti disperziju i matematičko očekivanje svake promenljive u zavisnosti od ishoda zavisne promenljive. Ukoliko promenljiva ima malu disperziju ili veliki broj nedostajućih podataka, najbolje bi bilo izostaviti je iz modela.

Korak 2. Jednodimenziona analiza svake promenljive

Jednodimenzionom analizom se testira odnos jedne promenljive sa zavisnom promenljivom (ishodom), bez obzira na ostale promenljive koje se javljaju u modelu. Cilj ovakve analize je priprema modela za višedimenzionu analizu i smanjenje broja promenljivih i ovaj korak je posebno važan ako model sadrži veliki broj promenljivih. Hi-kvadrat testom količnika verodostojnosti ili nekim drugim testom neophodno je oceniti koeficijente a zatim testirati njihovu značajnost (na primer Wald-testom) i oceniti standardne greške i intervale poverenja. Ukoliko promenljiva ne pokazuje veliki stepen povezanosti sa zavisnom promenljivom, tj. nije značajna za model, treba je izbaciti iz modela. Za promenljivu se može reći da je značajna ukoliko je p-vrednost testa manja od 0.25. Posebnu pažnju treba obratiti na autlajere i testirati model sa i bez observacija koje se smatraju autlajerima. Takođe, potrebno je oceniti pojedinačni odnos šansi promenljivih sa zavisnom promenljivom. Ukoliko je odnos šansi 1, može se zaključiti da ne postoji povezanost. Ako je odnos šansi manji ili veći od 1, onda postoji negativna, odnosno pozitivna povezanost sa ishodom.

Korak 3. Višedimenziona analiza

Jednodimenziona analiza ignoriše mogućnost da skup promenljivih, od kojih je svaka slabo povezana sa zavisnom promenljivom, može postati važan prediktor ukoliko ih uzmemo zajedno u razmatranje. Višedimenzionom analizom se bira podskup promenljivih koje najbolje opisuju date podatke. Postoje dva načina na koji se može obaviti multivarijantna analiza. Prvi način za izbor

promenljivih počinje od modela sa jednom promenljivom, njenom analizom, dodavanjem još jedne promenljive i analizom takvog modela, itd. Drugi način podrazumeva analiziranje modela koji uključuje sve promenljive i postepenim uklanjanjem promenljivih iz modela. Ova tehnika se zove još i *tehnika najboljeg podskupa*.

Korak 4. Verifikacija značajnosti promenljivih uključenih u model

Nakon višedimenzione analize potrebno je opet ispitati značajnost svake promenljive u modelu. Promenljive koje nisu značajne za model treba izbaciti, a zatim testom količnika maksimalne verodostojnosti uporediti nov model sa modelom koji sadrži tu promenljivu. Ovaj postupak treba primenjivati sve dok se ne pokaže da promenljive koje su izbačene iz modela statistički nisu značajne. Za neprekidne promenljive je takođe potrebno proveriti pretpostavku o linearnosti logita. Ukoliko logit nije linearan, potrebno je izvršiti odgovarajuću transformaciju promenljive.

6.2 Procena slaganja modela sa podacima

Često se dešava da prilikom analiziranja podataka i traženja najboljeg modela imamo više od jednog modela za koje smatramo da dobro opisuju podatke. Da bismo odabrali najbolji od njih, proveravamo koliko se svaki od njih uklapa u rezultujuću promenljivu i upoređujemo rezultate. Pretpostavimo da smo iz modela uklonili sve promenljive koje nisu značajno uticale na rezultujuću i da su promenljive koje su ostale u modelu unete u korektnom funkcionalnom obliku. Provera uklapanja modela sa rezultujućom promenljivom se naziva *test slaganja* (*goodness of fit test*).

Pre provere korektnosti i uklapanja modela moramo navesti osnovne kriterijume koji ukazuju na to šta se podrazumeva pod dobrim modelom. Pretpostavimo da smo ishode zavisne promenljive i njene ocene predstavili u vektorskom obliku, tj. da smo sa $\boldsymbol{y^T} = (y_1, y_2, \dots, y_n)$ označili vrednosti zavisne promenljive a sa $\hat{\boldsymbol{y}}^{\boldsymbol{T}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ procenjene (fitovane) vrednosti vektora $\boldsymbol{y^T}$. Model se smatra prilagođenim podacima ako zadovoljava dva uslova:

- sveukupna mera rastojanja između $\boldsymbol{y}$ i $\hat{\boldsymbol{y}}$ je zanemarljivo mala
- doprinos svakog para $(y_i, \hat{y}_i), i = 1, \dots, n$ je mali u odnosu na grešku modela.

Prilikom procene modela neophodno je merenje rastojanja između vektora $\boldsymbol{y}$ i $\hat{\boldsymbol{y}}$, kao i merenje rastojanja između pojedinačnih komponenti. Osim računskim putem, mnogi zaključci o slaganju modela sa podacima se mogu dobiti i primenom grafičke metode.

Testom slaganja se testira nulta hipoteza H_0: *model je korektan* protiv alternative H_1: *model nije korektan*. [12] Kako je idealan model koji ne sadrži greške merenja i koji se uklapa u sve opservacije praktično nemoguće naći, ovim testiranjem u zavisnosti od praga značajnosti

odbacujemo ili prihvatamo nultu hipotezu i tražimo onaj model za koji za što manji nivo značajnosti ne odbacuje nultu hipotezu. Ovaj test se bazira na grupisanju vrednosti određenih pomoću vrednosti nezavisnih promenljivih u modelu.

Ako se broj vrednosti koje uzimaju nezavisne promenljive povećava sa povećanjem veličine uzoraka, tada svaka od vrednosti m_j ima tendenciju da bude mala. Za takve raspodele koje su dobijene pod pretpostavkom da n uzima dovoljno velike vrednosti, kažemo da su n-asimptotske raspodele. U prethodnom primeru ako imamo veliki uzorak, kako je visina neprekidna promenljiva, velike su šanse da ćemo imati i veliki broj različitih vrednosti. Ako se fiksira broj grupa, tada se sa povećanjem obima uzorka povećava i broj elemenata u svakoj od tih grupa. Odnosno, ako je $J < n$ fiksirana vrednost, tada povećanjem veličine n, povećavaju se i veličine m_j. Takve raspodele se nazivaju m-asimptotske. Primer ovakvih raspodela je model koji sadrži dve nezavisne promenljive – pol i rasu.

Posmatrajući neki model, prilikom testiranja testom slaganja pretpostavljamo da je $\boldsymbol{J \approx n}$. Ovakvi slučajevi su česti i javljaju se kad u modelu postoji bar jedna neprekidna promenljiva. Razmotrićemo nekoliko načina kojima se može ustanoviti slaganje modela sa podacima.

6.2.1 Pirsonov χ^2 test i odstupanje

Da bi model bio dobar, potrebno je da reziduali (razlika između stvarnih i ocenjenih vrednosti zavisne promenljive $\boldsymbol{y - \hat{y}}$) budu što manji. U logističkoj regresiji ocenjene vrednosti zavisne promenljive se mogu predstaviti kao

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}}, \quad j = 1, \dots, J$$

gde je $\hat{g}(x_j) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ ocena logita za određenu vrednost iz skupa vrednosti nezavisne promenljive. Funkciju maksimalne verodostojnosti i njen logaritam možemo napisati na sledeći način:

$$\mathcal{L}(\beta) = \prod_{j=1}^{J} \binom{m_j}{y_j} \left(\pi(x_j)\right)^{y_j} \left(1 - \pi(x_j)\right)^{m_j - y_j}$$

$$ln\mathcal{L}(\beta) = \sum_{j=1}^{J} \left[ln \binom{m_j}{y_j} + y_j x_j^T \beta - m_j ln\left(1 + e^{x_j^T \beta}\right) \right].$$

Devijaciju D datu formulom (6) možemo predstaviti kao

$$D = 2\sum_{j=1}^{J}\left[y_j ln\frac{y_j/m_j}{\hat{y}_j/m_j} + \left(m_j - y_j\right) ln\frac{1-y_j/m_j}{1-\hat{y}_j/m_j}\right] =$$

$$2\sum_{j=1}^{J}\left[y_j ln\frac{y_j}{\hat{y}_j} + \left(m_j - y_j\right) ln\frac{m_j - y_j}{m_j - \hat{y}_j}\right].$$

Devijaciju takođe možemo izraziti pomoću *reziduala odstupanja* (deviance residual) definisanih na sledeći način: [5]

$$d\left(y_j, \hat{\pi}_j\right) = \begin{cases} -\sqrt{2m_j\left|ln\left(1-\hat{\pi}_j\right)\right|} & \text{za } y_j = 0 \\ \sqrt{2\left[y_j ln\frac{y_j}{m_j\hat{\pi}_j} + \left(m_j - y_j\right) ln\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}\right]} & \text{za } 0 < y_j < m_j, y_j/m_j > \hat{\pi}_j \\ -\sqrt{2\left[y_j ln\frac{y_j}{m_j\hat{\pi}_j} + \left(m_j - y_j\right) ln\frac{m_j - y_j}{m_j\left(1-\hat{\pi}_j\right)}\right]} & \text{za } 0 < y_j < m_j, y_j/m_j < \hat{\pi}_j \\ \sqrt{2m_j\left|ln\hat{\pi}_j\right|} & \text{za } y_j = m_j \end{cases}$$

Statistika testa saglasnosti se definiše kao

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

Još jedna statistika za procenu slaganja modela sa podacima je Pirsonova χ^2 statistika testa saglasnosti koja je bazirana na Pirsonovim rezidualima je data sa [12]

$$\chi^2 = \sum_{j=1}^{J}\left(r\left(y_j, \hat{\pi}_j\right)\right)^2.$$

Pod pretpostavkom da je fitovan model korektan, statistike D i χ^2 imaju $\chi^2_{J-(p+1)}$ raspodelu.

7 Zaključak

U ovom radu je obrađen pojam logističke regresije koja služi za opisivanje zavisne binarne promenljive i nalaženje najbolje veze između zavisne i skupa nezavisnih promenljivih.

Rad se sastoji iz pet poglavlja. U prvom poglavlju smo se upoznali sa populacionim modelima, poreklom logističke krive i njenim oblikom. U drugom poglavlju smo objasnili pojam zavisnosti promenljivih i karakterisike te zavisnosti i detaljnije se upoznali sa logističkom regresijom.

U sledećem poglavlju upoznali smo se sa parametrima logističke regresije i njihovim značenjem, metodama ocenjivanja parametara i testovima kojim se može utvrditi da li neka promenljiva utiče na promenu nezavisne promenljive i u kojoj meri. Takođe smo se upoznali i sa intervalima poverenja (pouzdanosti) za parametre.

Poslednja dva poglavlja govore o kreiranju modela. U njima smo se detaljnije upoznali sa nezavisnom promenljivom i opservacijama koje mogu da dovedu do pogrešnih rezultata prilikom testiranja. Videli smo, takođe, na koji način možemo da izaberemo promenljive u model i koliko se izabrani model slaže sa podacima.

8 Literatura

[1] Hallet D., Goodness of fit tests in logistic regression University of Toronto, 1999
[2] Harrell F., Regression Modeling Strategies, Springer, 2001
[3] Hosmer D, Lemeshow S, Studirvant R, Applied logistic regression 3rd ed., Wiley, 2013
[4] Jaccard J., Interaction Effects in Logistic Regression, Sage publications, 2001
[5] Kleinbaum D., Klein M., Logistic Regression, Springer, 2010
[6] Liu W, Simultaneous Inference In Regression, CRC Press, 2011
[7] Loader C, Local Regression and Likelihood, Springer, 1999
[8] Malthus T., An Essay on the Principle of Population, Paul's church-yard, 1798
[9] Pampel C. Fred, Logistic regression: A Primer, Sage publications, 2000
[10] Stivenson M., An Introduction to Logistic Regression, Massey University, 2012
[11] Seber G., Wild C, Nonlinear Regression, Wiley, 1989
[12] Ying L., On goodness of fit of logistic regression model, Kansas, 2007
[13] http://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logistic2.pdf
[14] http://sydney.edu.au/vetscience/biostat/macros/logistic_tut_model1.shtml
[15] http://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logfit.pdf
[16] http://davidakenny.net/cm/moderation.htm
[17] http://www.sagepub.com/upm-data/21121_Chapter_15.pdf
[18] http://www.statsdirect.com/help/default.htm#regression_and_correlation/logistic.htm
[19] http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/logistic-regression/what-are-delta-statistics/
[20] http://www.artnit.net/društvo/item/638-tomas-robert-maltus-porast-stanovništva.html
[21] http://www.model.u-szeged.hu/cd/content/Interreg_2008/Development%20in%20teaching%20science.pdf
[22] http://stats.stackexchange.com/questions/59085/how-to-test-for-simultaneous-equality-of-choosen-coefficients-in-logit-or-probit
[23] http://www.iarc.fr/en/publications/pdfs-online/stat/sp32/SP32_vol1-6.pdf