UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA
MATEMATIKU I INFORMATIKU

Gorica Gvozdić

Primenjena logistička regresija

-master rad-

Novi Sad, *2011*.

Sadržaj

Predgovor 3
1 Poreklo logističke funkcije 5
1.1 Populacioni modeli 5
1.2 Sigmoid funkcija 9
1.3 Uopšteni linearni modeli (*GLM*) 10
2 Logistički regresioni model 12
2.1 Logistički i linearni regresioni model 12
2.2 Logit model 15
3 Slaganje logističkog regresionog modela sa podacima 17
3.1 Metod maksimalne verodostojnosti (ML) 17
3.2 Testiranje značajnosti koeficijenata 21
3.2.1 Test količnika verodostojnosti 21
3.2.2 Wald test 23
3.3 Interval poverenja za ocenu 24
4 Interpretacija fitovanog logističkog modela 28
4.1 Dihotomna nezavisna promenljiva 28
4.2 Polihotomna nezavisna promenljiva 33
4.3 Neprekidna nezavisna promenljiva 34
4.4 Interakcija i ometanje 35
4.5 Ocena OR pri interakciji 38
5 Metode i postupci za građenje modela u logističkoj regresiji 41
5.1 Izbor promenljivih 41
5.2 Logistička regresiija "korak po korak" 44
6 Procena slaganja modela sa podacima 50
6.1 Osnovne mere za goodness-of-fit (*GOF*) 50
6.1.1 Pirsonova hi-kvadrat statistika i odstupanje 52
6.1.2 Hosmer-Lemeshow test(*HL* test) 54
6.1.3 Tabele klasifikacije 56
6.1.4 ROC kriva 58
7 Konstrukcija logističkog regresionog modela sa zavisnom promenljivom gojaznost 63
8 Zaključak 69
Literatura 70

PREDGOVOR

Ovaj rad opisuje primenu logističke regresije koja se pokazala veoma značajnom u modeliranju širokog opsega pojava, pre svega u ekonomskim, populacionim, biološkim, marketinškim, medicinskim i mnogim drugim istraživanjima.

Rad obuhvata dva dela. Prvi deo rada,koji sadrži šest poglavalja, predstavlja teorijski pristup logističkoj regresiji sa mnoštvom ilustrativnih primera, dok drugi deo rada predstavlja primenu logističke regresije na model koji utvrđuje povezanost gojaznosti sa potencijalnim faktorima rizika.

U prvom poglavlju je predstavljeno poreklo logističke funkcije i dat je osvrt na uopštene linearne modele kao klasu kojoj pripadaju logistički regresioni modeli. Drugo poglavlje ima za cilj da nas putem linearne regresije, kao metoda kod kog je zavisna promenljiva neprekidna, uvede u logističke regresione modele kod kojih je zavisna promenljiva diskretna. U ovom poglavlju vršimo poređenje ove dve metode i takođe se upoznajemo sa logit funkcijom.

Kad smo se upoznali sa logističkim regresionim modelom proveravamo koliko se dati logistički model slaže sa podacima, ovo je opisano u trećem poglavlju. Metoda za ocenjivanje koeficijenata koju smo ovde predstavili je metoda maksimalne verodostojnosti. Nakon ocenjivanja koeficijenata predstavićemo i ispitivanje značajnosti promenljivih u modelu pomoću test količnika verodostojnosti ili wald test-a kao i testiranje intervala poverenja za parametre koji nas interesuju. Interpretacija fitovanog modela, koja podrazumeva izvođenje zaključaka na osnovu ocenjenih koeficijenata u modelu, je predstavljena u četvrtom poglavlju. Interpretaciju dajemo u tri slučaja u zavisnosti da li je nezavisna promenljiva dihotomna, polihotomna ili neprekidna.

Peto poglavlje predstavlja metode i postupke za građenje modela, gde se upoznajemo sa široko rasprostranjenom metodom za izbor promenljivih, *korak po korak*, koja se zasniva na test statistici količnika verodostojnosti. Nakon građenja modela se, u šestom poglavlju, upoznajemo sa *goodness-of-fit* koje predstavljaju opšti pokazatelj koliko dobro se model slaže sa podacima. Statistike sa kojima ćemo se ovde upoznati su: Pirsonova Hi-kvadrat statistika i odstupanje, Hosmer-Lemeshow test, tabele klasifikacije i *ROC* kriva(*Receiver Operating Characteristic Curve*).

Na kraju rada, u sedmom poglavlju, prezentujemo primenu prethodno opisanih metoda i tehnika na konstrukciju modela sa zavisnom promenljivom gojaznost i faktorima

rizika kao što su: starost, obim kukova, konzumiranje određenih namirnica, bavljenje rekreacijom i slično.

***

Izuzetnu zahvalnost dugujem svom mentoru, prof. dr Zagorki Lozanov-Crvenković, na sugestijama, savetima, strpljenju i pomoći prilikom izrade ovog master rada. Želela bih da se zahvalim i članovima komisije dr Ivani Štajner-Papuga i dr Ljiljani Gajić na saradnji. Takođe, veliko hvala mojoj porodici i prijateljima, a posebno ocu Milošu, na bezuslovnoj podršci i razumevanju tokom studiranja.

Novi Sad 8.12.2011. Gorica Gvozdić

1 POREKLO LOGISTIČKE FUNKCIJE

Tomas Maltus (1776-1834), ekonomista iz Engleske je u svom radu ''*An essay on the principle of population as it affects the future improvement of society''* iz 1789 godine izložio svoje gledište da se sa povećanjem broja stanovnika povećava i količina proizvedenih resursa, hrane i slično, ali ovo povećanje raste aritmetičkom regresijom, dok rast broja stanovnika prati geometrijsku regresiju. Posle određenog broja godina, resursa će biti manje, a stanovnika koji će ih koristiti više, pa će tako zavladati oskudice. Ovo stanje će se vremenom pogoršavati i dobilo je naziv-*demografska (Maltusova) katastrofa*. Ovakvim rezonovanjem došlo se do zakjlučka da je jedini način da se izbegne izbegne ili odloži katastrofa smanjenje priraštaja, što se može postići povećanjem smrtnosti - namerno izazvanim ratovima, bolestima, oskudicama, ili ograničenim rađanjem.

1.1 POPULACIONI MODELI

Osnovni Maltusov model

Prebrojavanjem dolazimo do podatka da u nekom trenutku t_0 na Zemlji živi $p(0)$ stanovnika. Populacija u sledećem trenutku je srazmerna populaciji u prethodnom jer rast stanovništva prati geometrijsku regresiju, odnosno $p(1) = rp(0)$, gde je r parametar koji opisuje neto priraštaj stanovništva i može se dobiti iz postojećih podataka.

Ako sa γ označimo konstantnu brzinu rađanja u jedinici vremena po jedinki (stopa nataliteta), a sa δ konstantnu brzina umiranja u jedinici vremena po jedinki (stopa mortaliteta), tada važi da je konstantan priraštaj $\lambda = \gamma - \delta$.

Ako je sa $p(t)$ označen broj jedinki u trenutku t, onda je on posle nekog vremenskog intervala Δt jednak

$$p(t + \Delta t) = p(t) + \lambda p(t)\Delta t$$

Vidimo da je rast srazmeran postojećoj populaciji i vremenu.

Odnosno imamo problem:

$$\begin{aligned} p'(t) &= \lambda p(t) \\ p(0) &= p_0 \end{aligned} \qquad (1)$$

Rešavanjem ove diferencijalne jednačinu dobijamo:

$$\frac{dp(t)}{p(t)} = \lambda dt$$
$$\ln|p(t)| = \lambda t + c$$

$$p(t) = e^{\lambda t}e^{c}$$
$$p(t) = Ce^{\lambda t}$$

Važi da je

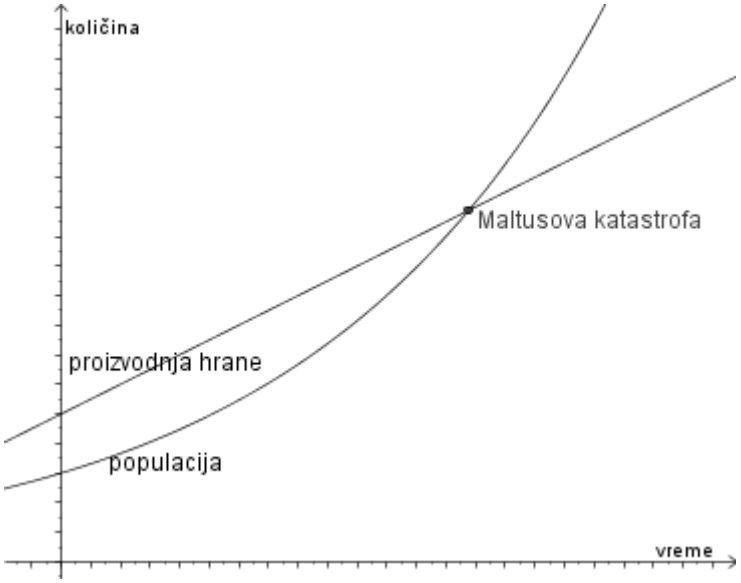
$$p(0) = p_0 = Ce^0 = C,$$

pa je rešenje jednačine (1)

$$p(t) = p_0 e^{\lambda t} \qquad (2)$$

Vidimo da populacija raste eksponencijalno po vremenu (Slika 1).

Ovaj model se naziva *osnovni (Maltusov) populacioni model* [9].



Slika 1.

Modifikacija Maltusovog modela

Maltusov model je imao bitna nedostatke. Pjer Fransoa Verhlust (1804-1849) je izvršio modifikaciju modela. On smatra da nijedna sredina ne može na sebi da održava neograničen broj jedinki, odnosno da rast populacije treba ograničiti do neke maksimalne fiksne vrednosti karakteristične za sistem koji se posmatra, odnosno do nekog maksimalnog nosivog kapaciteta sredine K. Ograničeni resursi usporavaju rast populacije i populacija teži ka graničnom zasićenu. Takođe linearne brzine rađanja i umiranja nisu konstante, već su date sa:

$$\gamma(t) = \gamma_0 - \gamma_1 p(t)$$
$$\delta(t) = \delta_0 + \delta_1 p(t)$$
$$\gamma_0 > \delta_0 > 0, \qquad \gamma_1, \delta_1 > 0$$

i smanjuju brzinu rađanja, a uvećavaju brzinu umiranja sa rastom populacije.

Verhlustov logistički model

Maksimalni priraštaj označićemon sa a, gde je

$$a = \gamma_0 - \delta_0$$

Sada važi da je prirodni priraštaj

$$\lambda(t) = (\gamma_0 - \delta_0) - (\gamma_1 + \delta_1)p(t) = a - bp(t)$$

gde smo sa b označili $b = \gamma_1 + \delta_1$

Maltusova jednačina sada ima oblik

$$p'(t) = \lambda(t)p(t)$$

$$p'(t) = ap(t) - bp^2(t)$$

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

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

$$a > b > 0, \qquad p(0) = p_0 \tag{3}$$

Populacija P u početku raste eksponencijalno sa stopom rasta a, ali se taj rast smanjuje kako se populacija približava maksimalnom (nosivom) kapacitetu sistema $\frac{a}{b} = K$. Matematički takvo ponašanje možemo modelirati *logističkom jednačinom*:

$$\frac{dp(t)}{dt} = ap(t)(1 - \frac{p(t)}{K})$$

$$p(0) = p_0$$

Odnosno važi da kada ja populacija P mala u odnosu na kapacitet K, populacija se ponaša prema Maltusovom populacionom modelu. Kada se populacija približi maksimalnom kapacitetu, tada izraz u zagradi teži nula što koči rast populacije. Rešimo jednačinu:

$$\frac{dp(t)}{dt} = ap(t)(1 - \frac{p(t)}{K})$$

$$\int \frac{1}{p(1 - \frac{p}{K})}\,dp = at + c$$

$$\ln\left|\frac{K - p}{p}\right| = -at - c$$

$$\left|\frac{K-p}{p}\right| = e^{-at}e^{-c}$$

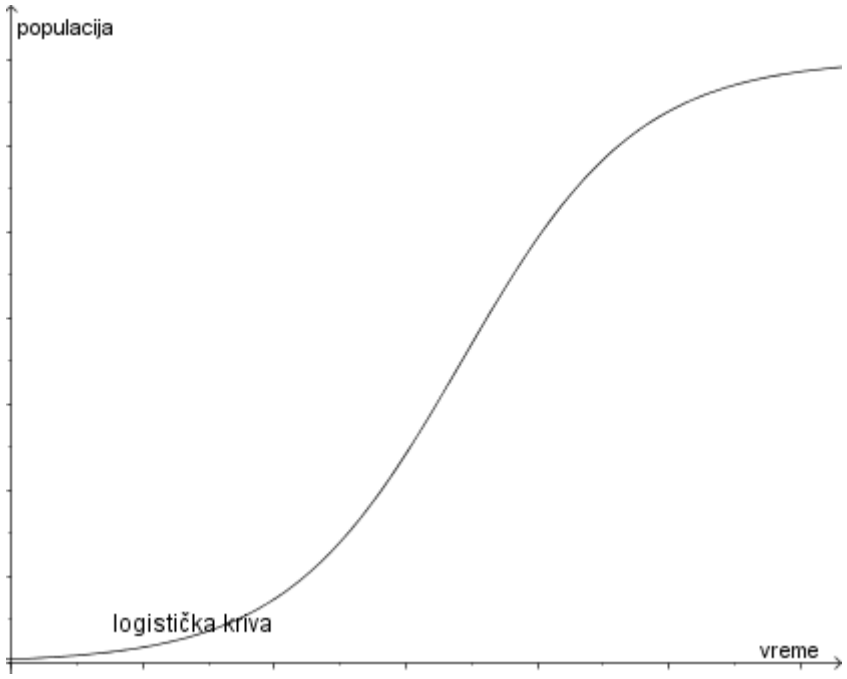
$$\frac{K}{p} - 1 = Ce^{-at}$$

$$p(t) = \frac{K}{1+Ce^{-at}}$$

Vidimo da kada $t \to \infty$ funkcija $p(t) \to K$. Opšte rešenje ove jednačine je *logistička funkcija* Konstantu C dobijamo kad uvrstimo početni uslov, odnosno:

$$p(0) = p_0 = \frac{K}{1+C} \to C = \frac{K-p_0}{p_0}$$

Kriva $p(t)$ ima S-oblik i naziva se *logistička kriva* (Slika 2). Ovaj model je bolji nego Maltusov model, ali ima nedostatke jer nisu uzeti u obzir i mnogi spoljašnji uticaji [17].



Slika 2.

1.2 SIGMOID FUNKCIJA

Postoje različiti oblici logističke funkcije a jedan od specijalnih slučajeva je *sigmoid funkcija* ili *sigmoid kriva* koja je još poznata i pod nazivom *standardna logistička funkcija* ili *osnovna logistička funkcija* i data je sa:

$$P(t) = \frac{1}{1 + e^{-t}} \tag{4}$$

Standardna sigmoid funkcija se dobija kao rešenje diferencijalne jednačine prvog reda:

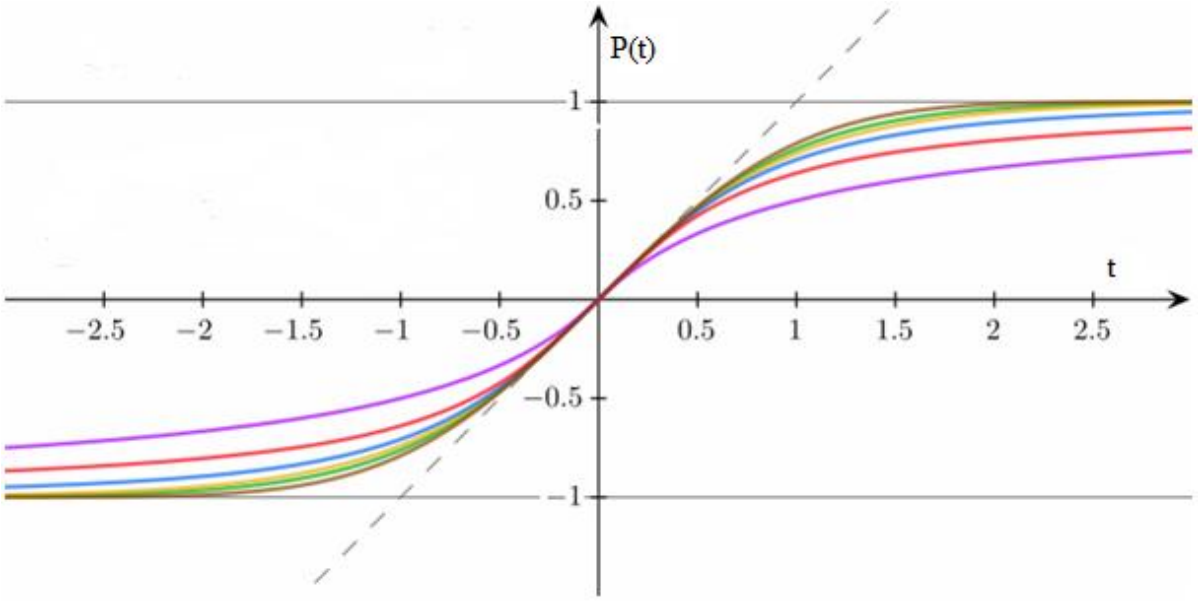
$$\frac{dP}{dt} = P(1 - P)$$

$$P(0) = \frac{1}{2}$$

Ona je strogo rastuća funkcija koja se može prikazati i u sledećem obliku:

$$P(t) = \frac{1}{1 + e^{-at}}$$

gde je a parametar nagiba sigmoidne funkcije. Menjajući vrednost parametra a, dobijaju se različiti oblici, što je prikazano na Slici 3**.**



Slika 3

Posmatrajmo izraz (4). P-predstavlja verovatnoću da se neki događaj desi, pod uticajem nekih nezavisnih rizičnih faktora, promenljiva t se definiše kao:

$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, gde su $\beta_i, i = 1 \ldots k$ regresioni koeficijenti koji opisuju veličinu doprinosa odgovarajućeg rizičnog faktora x_i. Kada su regresioni koeficijenti pozitivni tada nezavisne promenljive x_i povećavaju verovatnoću pozitivnog ishoda, a kada su negativni, onda smanjuju tu verovatnoću.

Primer1

Ispitujemo verovatnoću da osoba u narednih 10 godina umre od bolesti srca, posamtrajući rizične faktore: x_1= godine preko 50, x_2 – pol (muško-0, žensko-1),

x_3 - nivo holesterola preko 5 $mmol/l$. Neka su nam regresioni koeficijenti dati sa:

$$\beta_0 = -5, \beta_1 = 2, \beta_2 = -1, \beta_3 = 1.2$$

Posmatrajmo: Muškarca koji ima 50 godina i 7 mmol/l holesterola u krvi.

Verovatnoća da on umre u narednih 10 godina je tada data sa:

$$P(t) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}}$$

$$P(t) = \frac{1}{1 + e^{-(-5 + 2(50-50) - 1 \cdot 0 + 1.2(7-5))}} = 0.07$$

Odnosno verovatnoća da ova osoba umre u narednih 10 godina je 7%.

1.3 UOPŠTENI LINEARNI MODELI (*GLM*)

Uopšteni linearni modeli se koriste za regresiono modeliranje zavisne promenljive koja ne mora da ima normalnu raspodelu, odnosno za modeliranje promenljive iz eksponencijalne familije raspodela kako što su Poasonova, Binomna, Multinomna i slično.

GLM modeli sastoje se iz tri komponente i to:

1. Komponenta slučajnosti
2. Komponenta sistematičnosti
3. Funkcija veze (Link funkcija)

Komponenta slučajnosti- se indentifikuje sa zavisnom promenljivom i prihvata njenu funkciju verovatnoće. Neka su $Y_1, Y_2, \ldots Y_n$ realizacije zavisne promenljive i one su po pretpostavkama *GLM* medjusobno nezavisne. Realizacije od *Y* mogu biti dihotomne (binarne, sa samo dva moguće ishoda (uspeh ili neuspeh)) tada zavisna promenljiva *Y* ima binomnu raspodelu ili se realizacije mogu dobiti prebrojavanjem tada zavisna promenljiva ima Poasonovu raspodelu[2].

Komponenta sistematičnosti: - predstavlja lineranu funkciju nezavisnih promenljivih koje opisuju zavisnu promenljivu:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Ovakva linerana kombinacija se naziva linearno predviđanje, i promenljive x_i ne moraju biti linerano nezavisne.

Funkcija veze(Link funkcija) – Funkcija veze povezuje linerno predviđanje sa funkcijom $\mu = E(Y)$, odnosno povezuje komponentu slučajnosti i komponentu sistematičnosti. Ona predstavlja neku funkciju $g(\cdot)$ tako da važi:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Funkcija $g(\cdot)$ je monotona i ne mora da bude linearno preslikavanje.

Logistički regresioni model je jedan od *GLM* modela kod kojeg je zavisna promenljiva binomna.

2 LOGISTIČKI REGRESIONI MODEL

Regresione metode su sastavni deo svake analize podataka koja se bavi opisivanjem veze između zavisnih i nezavisnih promenljivih. Cilj analize koja koristi ovaj metod je naći model koji je najbolje prilagođen podacima, odnosno najekonomičniji, ali ipak prihvatljiv model koji opisuje vezu između zavisne promenljive i skupa nezavisnih promenljivih koji nju opisuju. Logistička regresija se koristi za:

- predviđanje zavisne promenljive na osnovu vrednosti nezavisnih promenljivih
- rangiranje nezavisnih promenljivih po važnosti
- procenu efekta interakcije.

Zavisna promenljiva u logističkom regresionom modelu je diskretna, obično binarna, a u ređim slučajevima može da ima više od dve kategorije. Na primer, zavisna promenljiva može biti da li je pacijent izlečen ili ne; da li je neki proizvod prošao kontrolu kvaliteta ili ne; da li je životinja na kojoj se vršio neki eksperiment preživela isti ili ne i slično. Tada u zavisnosti od merne skale zavisne varijable, govorimo o *Nominalnim*, odnosno *Ordinalnim logističkim regresionim modelima*. Nezavisne promenljive mogu biti kategorijalne ili kombinacija kategorijalnih i neprekidnih, pri čemu u logističkoj regresiji ne postoje pretpostavke o raspodeli za ove promenljive. Zavisnu promenljivu označavaćemo sa Y, dok nezavisne označavamo sa x [6].

Na primer, ukoliko je pacijent izlečen, ishod je „uspeh“, a ako nije ishod je „neuspeh“; ako proizvod prođe kontrolu kvaliteta ishod je „uspeh“, u suprotnom „neuspeh“. Ukoliko zavisna promenljiva označava to da li je osoba zdrava ili ne, onda bismo npr. sa 0 kodirali - osoba nije zdrava, a sa 1 – osoba je zdrava.

Često je potrebno izvršiti grupisanje podataka, tako da se u okviru jedne grupe nalaze svi subjekti koji imaju iste vrednosti nezavisnih promenljivih. Kada su podaci grupisani, lakše je zabeležiti broj „uspeha“, odnosno broj „neuspeha“, jer ih beležimo za svaku grupu posebno, dok bismo u slučaju negrupisanih podataka dobijali dugačke nizove 0 i 1.

2.1 LOGISTIČKI I LINEARNI REGRESIONI MODEL

Ilustrovaćemo razlike između linearne i logističke regresije primerom.

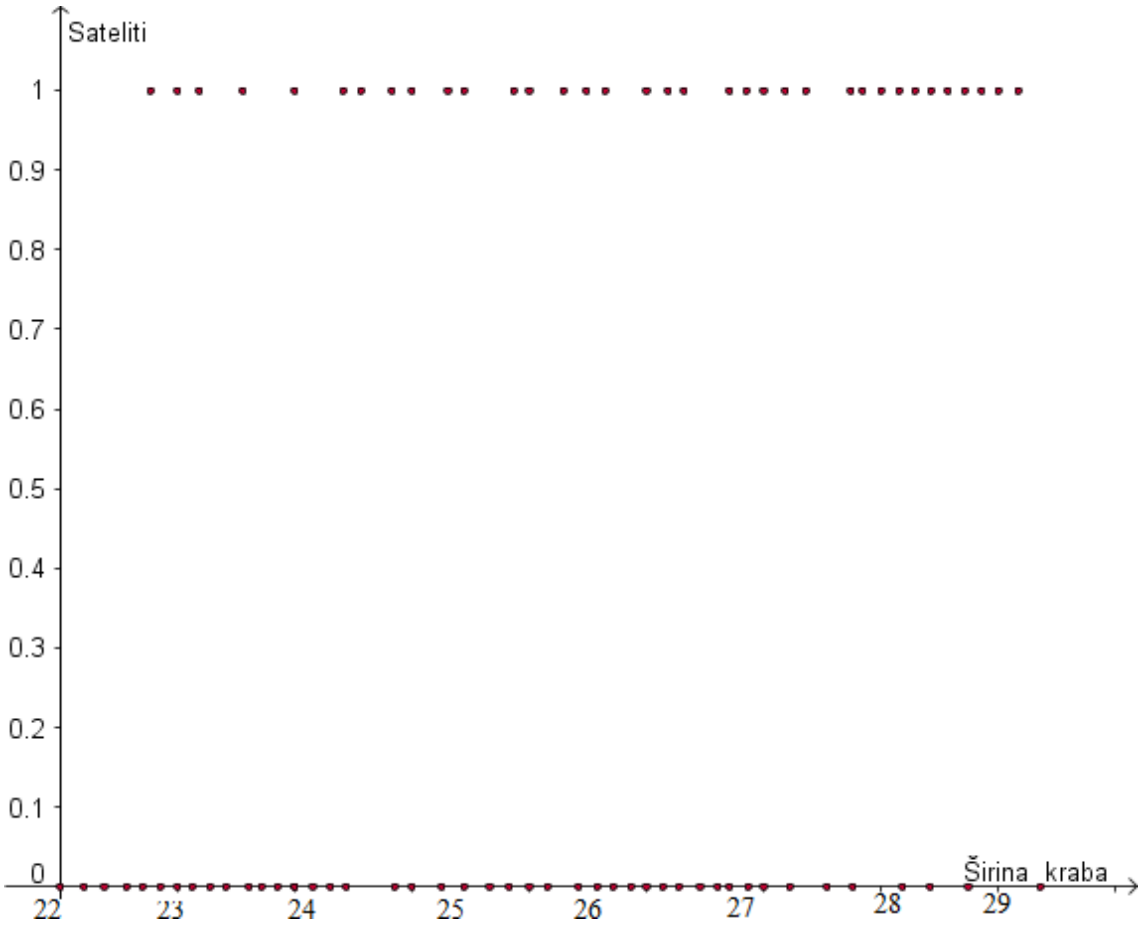
Određujemo da li ženske krabe u odnosu na njihovu širinu izraženu u centimetrima imaju mužjake-satelite (*SAT*) ili ne. Izabrano je 173 krabe da učestvuju u istraživanju koje su grupisane u kategorije (*ŠIR-KAT*) u odnosu na širinu (Tabela 1 na cd-u u prilogu).

Rezultujuća promenljiva je *SAT*, koja je binarna i prima vrednost nula ($Y = 0$) ako nema mužjaka, odnosno vrednost jedan ($Y = 1$) ako ima mužjaka.

Širina kraba	n	SAT		Proporcija
		da	ne	
<23.25	14	3	11	0.21
23.25-24.25	14	4	10	0.29
24.25-25.25	28	13	15	0.46
25.25-26.25	39	21	18	0.54
26.25-27.25	22	15	7	0.58
27.25-28.25	24	20	4	0.83
28.25-29.25	18	15	3	0.83
>29.25	14	14	0	1.00

Tabela 2

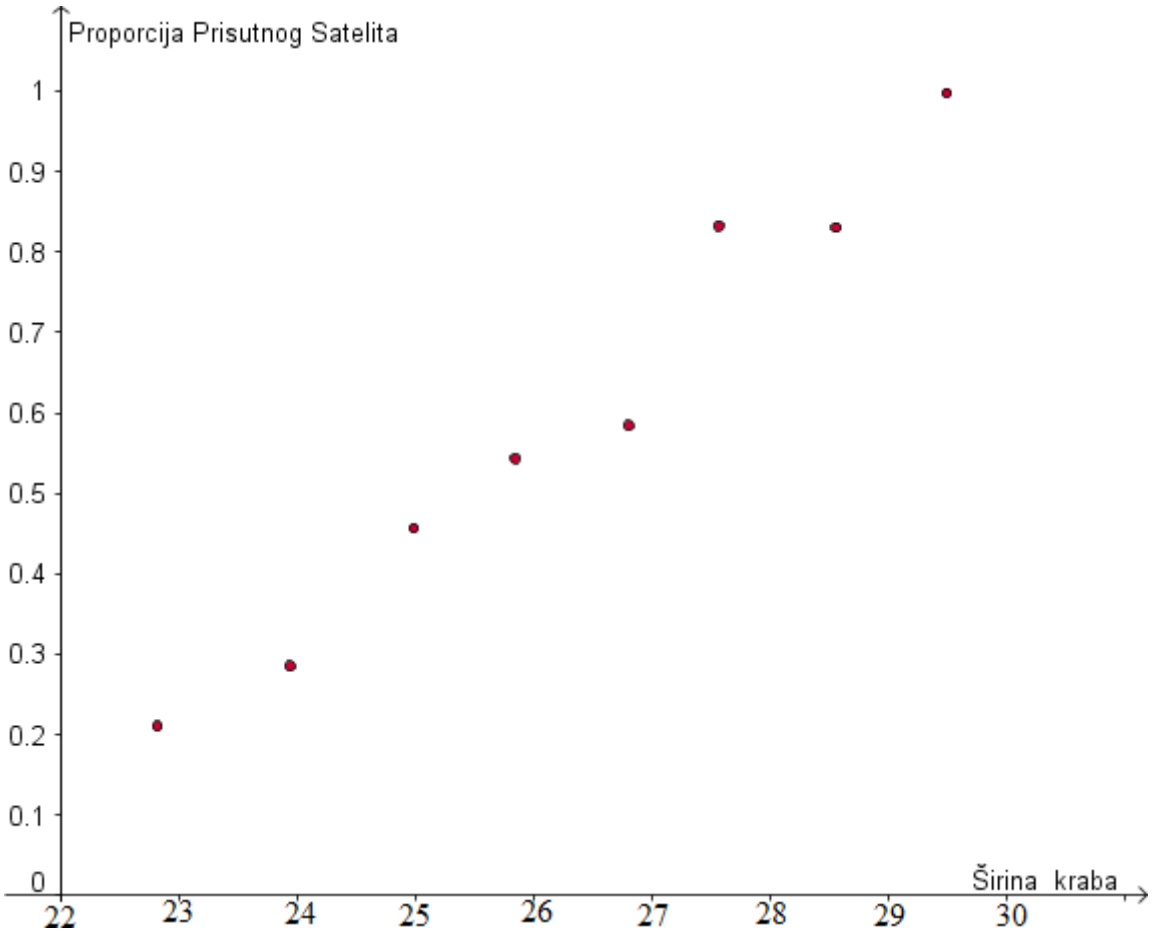
Ako bi rezultujuća promenljiva bila neprekidna, tada bismo koristili dijagram rasipanja rezultata u odnosu na nezavisnu promenljivu(Slika 4) [18].



Slika 4

Sa ovog grafika vidimo da sve tačke pripadaju jednoj od dve paralelne prave koje predstavljaju prisustvo ($Y = 1$), odnosno odsustvo satelita ($Y = 0$). Može se uočiti tendencija

da se kod širih kraba češće javlja prisustvo satelita, ali je ipak na ovaj način teško opisati vezu između širine kraba i prisustva satelita. Problem je što je varijabilnost za promenljivu *SAT* za sve širine velika, pa je samim tim teško opisati funkcionalnu vezu izmedju širine i *SAT*. Uobičajni metod eliminisanja nekih promenljivih sa ciljem održavanja strukture veze između rezultata i nezavisne promenljive je formiranja intervala za nezavisnu promenljivu i računanje proporcije ($\frac{broj\ onih\ koji\ imaju\ satelite}{ukupan\ broj\ kraba\ u\ datoj\ grupi}$) za rezultujuću promenljivu (Tabela 2). Na ovaj način dobijamo novi grafik (Slika 5) sa kojeg se jasno vidi da povećanjem širina kraba povećava proporcija onih koje imaju satelite [12].



Slika 5

U bilo kom regresionom modelu ključno je odrediti očekivanu vrednost zavisne promenljive za datu vrednost nezavisne promenljive, u oznaci $E(Y|x)$ [12]. Kako je zavisna

promenljiva dihotomna, za uslovnu sredinu važi $0 \leq E(Y|x) \leq 1$. Promena u $E(Y|x)$ po jedinici promene za *x* postaje progresivno manja kako uslovna sredina postaje bliža 0 ili 1. Kako je zavisna promenljiva dihotomna i uzima vrednosti 0 i 1, uzećemo da uzima vrednost 1 sa verovatnoćom π, a vrednost 0 sa verovatnoćom $1-\pi$, tj. $Y = \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$.

Slučajna promenljiva $Y|x$ će takođe uzimati vrednosti 0 i 1, sa verovatnoćama $1-\pi(x),\ \pi(x)$ redom, tj. $Y|x = \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$.

Kako nas interesuje očekivana vrednost od Y za dato x, izračunaćemo je:

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

Zbog ovoga, ubuduće ćemo koristiti oznaku $\pi(x)$ za prikazivanje uslovne sredine od Y za dato x kada se koristi logistička raspodela [6].

Poseban oblik regresionog modela koji koristimo je

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

Kod logističke regresije, vrednost rezultujuće promenljive za dato x možemo izraziti kao $Y|x = \pi(x) + \varepsilon$, gde je ε greška koja ima binomnu raspodelu.
Promenljiva ε može uzeti vrednost $-\pi(x)$ i $1-\pi(x)$ i to vrednost $-\pi(x)$ uzima kada promenljiva $Y|x$ uzme vrednost 0, a vrednost $1-\pi(x)$ uzima kada $Y|x$ uzme vrednost 1.
Kako slučajna promenljiva $Y|x$ uzima vrednost 0 sa verovatnoćom $1-\pi(x)$, a vrednost 1 sa verovatnoćom $\pi(x)$, sledi da će i ε uzeti odgovarajuće vrednosti sa tim verovatnoćama, tj.

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

Dakle, ε zaista ima binomnu raspodelu sa sredinom $E(\varepsilon) = 0$ nula i varijansom

$$Var(\varepsilon) = \pi(x)\big(1-\pi(x)\big)$$

2.2 LOGIT MODEL

Odnos između verovatnoće π i nezavisne promenljive X se može predstaviti preko logističkog regresionog modela, koji se predstavlja preko S-krive date na Slici 5. Vidimo da verovatnoća π polako raste sa porastom X, kasnije se rast ubrzava, dok se na kraju ne stabilizuje i ne ide preko vrednosti 1. Verovatnoća π se može predstaviti formulom:

$$\pi = P(Y = 1|X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Model može biti uopšten za slučaj kada imamo više nezavisnih promenljivih i onda izgleda ovako:

$$\pi = P\big(Y = 1\big|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p\big) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1 + e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$

Ova jednakost se naziva logistička regresiona funkcija [8]. Nije linearna po parametrima $\beta_i\,, i = 0 \dots p$, ali se može linearizovati odgovarajućom logit transformacijom. Tada važi:

$$1 - \pi = P\big(Y = 0\big|X_1 = x_1, X_2 = x_2, \dots, X_p = x_p\big) = \frac{1}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$$

Dalje imamo da je:

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}$$

Ako logoritmujemo sa prirodnim logaritmom obe strane gornje jednakosti dobijamo:

$$\ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$$

Ova jednakost se naziva logit i ona je linearna po komponentama $\beta_i\,, i = 1 \dots p$. Primetimo još da vrednost od π pripada intervalu [0,1], dok se vrednost logita kreće od $(-\infty, +\infty)$, pa je logit funkcija najprikladniji izbor za link funkciju[1].

3 SLAGANJE LOGISTIČKOG REGRESIONOG MODELA SA PODACIMA

3.1 METOD MAKSIMALNE VERODOSTOJNOSTI (ML)

U linearnoj regresiji najčešći metod za ocenjivanje regresionih parametara je metod najmanjih kvadrata. U tom metodu, biramo one vrednosti β_0 i β_1, koje minimiziraju sumu kvadrata odstupanja registrovane vrednosti za *Y* od predviđene vrednosti dobijene na osnovu modela. Pod uobičajenim pretpostavkama za linearnu regresiju, metod najmanjih kvadrata daje ocene sa mnoštvom poželjnih statističkih svojstava. Međutim, kada se metod najmanjih kvadrata primeni na model sa dihotomnim ishodom, ocene više nemaju te iste osobine.

Kada je u pitanju logistička regresija za ocenjivanje regresionih koeficijenata koristimo metod maksimalne verodostojnosti. Ovaj metod daje vrednosti za β_i , $i = 0 \dots p$, koje maksimizraju verovatnoću dobijanja registrovanog skupa podataka. Odnosno utvrđujemo verodostojnost (verovatnoće) registrovanih podataka za različite kombinacije vrednosti regresionih koeficijenata, za razliku od metode najmanjih kvadrata. Ovaj metod zahteva dosta iterativnog izračunavanja.

Da bismo opisali ML model, potrebno je da se upoznamo sa funkcijom verodostojnosti, to je funkcija nepoznatih parametara, u našem slučaju regresionih koeficijenata u oznaci $L(\boldsymbol{\beta})$, gde je $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ i predstavlja verovatnoću koja kombinuje doprinose svih subjekata u istraživanju.

Ako je zavisna promenljiva Y: $\begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$, tada izraz

$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}}$ za proizvoljnu vrednost $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$, daje uslovnu verovatnoću $P\{Y = 1|x_i\} = \pi(x_i)$ i $P\{Y = 0|x_i\} = 1 - \pi(x_i)$, gde je $x_i = (1, x_{1i}, x_{2i} \dots x_{pi})$, $i = 1 \dots n$

Za one parove (x_i, y_i) gde je $y_i = 1$ doprinos funkciji verodostojnosti je $\pi(x_i)$, a za one parove (x_i, y_i) gde je $y_i = 0$ doprinos funkciji verodostojnosti je $1 - \pi(x_i)$.

Dakle, za par (x_i, y_i) doprinos funkciji verodostojnosti je dat sledećim izrazom:

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

S obzirom da radimo pod pretpostavkom da su registrovane vrednosti nezavisne, funkcija verodostojnosti je dobijena kao proizvod gornjeg izraza, odnosno:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{p} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i} \tag{5}$$

Verodostojnost se može predstaviti i kao:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{p} \left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i} \left(1-\pi(x_i)\right)$$

gde se izraz $\frac{\pi(x_i)}{1-\pi(x_i)}$ naziva šansa za $P\{Y=1|x_i\}$ i jednak je

$$\frac{\pi(x_i)}{1-\pi(x_i)} = e^{e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_p x_{pi}}} = e^{\boldsymbol{x_i'\beta}}$$

odnosno verodostojnost predstavlja funkciju registrovanih vrednosti zavisne i nezavisnih promenljivih i nepoznatih parametara [6], [13].

Radi jednostavnosti koristićemo logaritam ove funkcije, tj. logaritam verodostojnosti:

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta}) = \sum_{i=1}^{p} [y_i \ln \pi(x_i) + (1-y_i)\ln \pi(x_i)] \tag{6}$$

odnosno:

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

Gde x_i predstavlja kovarijatu registrovanih vrednosti za posmatranje i .

Ocene parametara tražimo tako da maksimiziraju funkciju verodostojnosti. Da bismo našli $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ koji maksimizira funkciju $L(\boldsymbol{\beta})$ diferenciraćemo $L(\boldsymbol{\beta})$ u odnosu na $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ i dobijene jednačine ćemo izjednačiti sa nulom, odnosno važi:

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{p} \left(y_i - \frac{e^{\boldsymbol{x_i'\beta}}}{1+e^{\boldsymbol{x_i'\beta}}}\right)\boldsymbol{x_i'} \tag{7}$$

Ove jednačine su nelinearne po $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$, pa se rešavaju nekim od iterativnih postupaka.

Jedan od najčešće korišćenih iterativnih postupaka za rešavanje jednačine (7) je Njutn-Rapšanov postupak. Radi lakšeg rada sistem (7) ćemo napisati u ekvivalentnom matričnom zapisu, odnosno važi:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = X'(\boldsymbol{y} - \boldsymbol{p})$$

gde je $\boldsymbol{p} = P\{Y = 1|x_i\} = \pi(x_i) = \pi_i = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ i $X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & x_{np} \end{bmatrix}$

Neka je $W = diag(p_i(1 - p_i))$, odnosno:

$$W = \begin{bmatrix} \pi_1(1-\pi_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \pi_n(1-\pi_n) \end{bmatrix}$$

odakle sledi da je

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta^2}} = -X'WX$$

Neka je $\beta^{(0)}$ vektor početnih aproksimacija za svako $\beta^{(k)}$, tada je prva iteracija Njutn-Rapšanovog postupka:

$$\beta^{(1)} = \beta^{(0)} + \left(-\frac{\partial^2 L(\beta^{(0)})}{\partial \beta^{(0)^2}}\right)^{-1} \frac{\partial L(\beta^{(0)})}{\partial \beta^{(0)}}$$

Odnosno:

$$\beta^{(1)} = \beta^{(0)} + \left(X'W^{(0)}X\right)^{-1} X'(\boldsymbol{y} - \boldsymbol{p^{(0)}})$$

Svaku $l + 1$ iteraciju dobijamo:

$$\beta^{(l+1)} = \beta^{(l)} + \left(X'W^{(l)}X\right)^{-1} X'(\boldsymbol{y} - \boldsymbol{p^{(l)}})$$

Vrednost koja se dobija kao rešenje ovih iteracija naziva se ocena maksimalne verodostojnosti i označava se sa $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ [6].

Primer 2

Posmatrajmo podatke iz Tabele 3.

Pr.Test	Pret.Isk.	Završen	Pr.Test	Pret.Isk.	Završen
5	6	0	3	3	0
1	15	0	1	24	1
1	12	0	2	8	0
4	6	0	1	9	0
1	15	1	4	18	0
1	6	0	4	22	1
4	16	1	5	3	1
1	10	1	4	12	0
3	12	0	4	24	1
4	26	1	2	18	1
5	2	1	2	6	0
1	12	0	1	8	0
3	18	0	5	12	0

Tabela 3

U tabeli se nalaze podaci vezani za odgovarajući kurs, odnosno: *Pret.Isk* - predstavlja trajanje prethodnog iskustava u mesecima bitnog za dati kurs , *Pr.Test* - predstavlja ocenu dobijenu na prijemnom ispitu za dati kurs, *Završen* – je promenljiva koja je kodirana sa 1 ako je kurs završen, odnosno sa 0 ako kurs nije završen.

Promenljiva	β	Standardna greška
Pr.Test	0.28758	0.33918
Pret.Isk.	0.14688	0.04334
Konstanta	-3.16567	0.02627
-2L(β)=28.879		

Tabela 4

Korišćenjem logističke regresije sa neprekidnom nezavisnim promenljivim *Pret.Isk* i *Pr.Test* i zavisnom promenljivom *Završen* dobijamo Tabelu 4 sa ocenama maksimalne verodostojnosti parametara $\beta_0, \beta_1 \; i \; \beta_2$, kao i ocenama njihovih standardnih greški i logaritma verodostojnosti.

Takođe dobijamo i ocenu logita, odnosno važi:

$$\hat{g}(x) = -3.16567 + 0.14688Pret.Isk + 0.28758Pr.Test$$

Kao i fitovane vrednosti koje dobijamo iz jednakosti:

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

3.2 TESTIRANJE ZNAČAJNOSTI KOEFICIJENATA

Nakon ocenjivanja koeficijenata, dalje razmatranje fitovanog modela se uopšteno odnosi na ocenjivanje značajnosti promenljivih u modelu. Ovo obično uključuje formulisanje i testiranje statističkih hipoteza za određivanje da li su nezavisne promenljive u modelu *značajno* povezane sa rezultujućom promenljivom.

Pitanje koje ovde postavljamo je sledeće: *Da li nam model koji sadrži promenljivu, govori više o rezultujućoj promenljivoj nego model koji ne sadrži tu promenljivu?*

Odgovor na ovo pitanje je dobijen upoređivanjem registrovane vrednosti rezultujuće promenljive sa predviđenom vrednosti pomoću svakog od dva modela; prvim sa, i drugi bez te promenljive. Ako su predviđene vrednosti na osnovu modela koji sadrži tu promenljivu bolje, ili tačnije u nekom smislu, nego vrednosti koje su predviđene na osnovu modela koji ne sadrža tu promenljivu, tada kažemo da je promenljiva u modelu značajna.

3.2.1 TEST KOLIČNIKA VERODOSTOJNOSTI

Poređenje registrovane i predviđene vrednosti dobijene iz modela koji sadrži nezavisnu promenljivu i modela koji je ne sadrži, je bazirano na logaritmu funkcije verodostojnosti. Pri tome se smatra da je registrovana vrednost zavisne promenljive ona predviđena vrednost koja se dobija iz zasićenog modela. Zasićen model je onaj model koji sadrži toliko mnogo parametara koliko ima podataka[20].

Za poređenje registrovanih sa predviđenim vrednostima na osnovu modela koristimo funkcije verodostojnosti:

$$D = -2\ln\frac{l_f}{l_z} \qquad (8)$$

Gde je l_f-verodostojnost fitovanog modela, dok je l_z-verodostojnost zasićenog modela, dok se izraz $\frac{l_f}{l_z}$ naziva *količnik verodostojnosti*.

Koristili smo -2ln da bismo dobili veličinu čija nam je raspodela poznata, tako da ovu statistiku možemo koristiti za testiranje hipoteza.

Korišćenjem izraza (5) izraz (8) postaje:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi_i}}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi_i}}{1-y_i}\right)\right] \quad (9)$$

gde je $\widehat{\pi_i} = \widehat{\pi_i}(x_i)$

Statistika *D*, u jednakosti (9), se naziva *odstupanje*[6].

U cilju procenjivanja značajnosti nezavisne promenljive, upoređujemo vrednost *D* za model koji sadrži nezavisnu promenljivu i model koji je ne sadrži. Promena u *D* koja nastaje zbog uključivanja nezavisne promenljive u model je data sa:

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

Kako obe vrednosti *D* imaju isti imenilac (verodostojnost zasićenog modela), *G* se može se izraziti kao:

$$G = -2\ln\left(\frac{verodostojnost\ modela\ bez\ nezavisne\ promenljive}{verodostojnost\ modela\ sa\ nezavisnom\ promenljivom}\right)$$

Kada je u pitanju univarijabilni slučaj lako se pokazuje da kada promenljiva nije u modelu maksimalna verodostojnost od β_0 je $\ln\frac{n_1}{n_0}$, gde je $n_1 = \sum y_i$, a $n_0 = \sum 1 - y_i$, dok je predviđenan vrednost konstantna i iznosi $\frac{n_1}{n}$. U tom slučaju vrednost G je :

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi(x_i)}^{y_i}\left(1-\widehat{\pi(x_i)}\right)^{1-y_i}}\right]$$

odnosno

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i \ln\widehat{\pi(x_i)} + (1-y_i)\ln\widehat{\pi(x_i)}\right] - \left[n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)\right]\right\}$$

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

Primer 3

Posmatrajmo Tabelu 5 (na cd-u u prilogu) u njoj se nalaze podaci vezani za zavisnu promenljivu manjinsko stanovništvo koja je kodiran sa 0 ako osoba ne pripada manjinama, odnosno sa 1 ako osoba pripada manjinskom stanovništvu. Ostale promenljive su: *god_obrazovanja*-koja predstavlja duzinu skolovanja ispitanika, *vrsta_zaposlenja* koja je kodirana sa 1 - ako je osoba službenik, 2 - obezbeđenje, odnosno sa 3 - ako osoba pripada menadžmentu kompanije, *prethodno_iskustvo*-predstavlja radno iskustvo u mesecima. Ispitivaćemo koliko *prethodno_iskustvo* utiče na činjenicu da je osoba manjinskog porekla. Za neka izračunavanja koristićemo statistički paket *SPSS 17* [4].

Iz tabele vidimo da je $n_1 = 370$, dok je $n_0 = 104$, takođe važi da je na osnovu formule (6) logoritam verodostojnosti jednak -242.574, pa važi da je:

$$G = 2\left(-242.574 - \left(370ln(370) + 104ln(104) - 474ln(474)\right)\right) = 13.651$$

Kako je p vrednost ove statistike jednaka : $P\{\chi_1^2 > 13{,}651\} = 0.0002 < 0.002$ pokazali smo da je promenljiva *prethodno_iskustvo* značajna u određivanju da li je osoba manjinskog porekla.

Kada je u pitanju multivarijabilni logistički regresioni model test količnika verodostojnosti za ukupnu značajnost p koeficijenata za nezavisne promenljive u modelu je izveden na isti način kao i u univarijabilnom slučaju. Jedina razlika je da su fitovane vrednosti za model, $\hat{\pi}$, bazirane na vektoru $\widehat{\boldsymbol{\beta}}$ koji sadrži $p + 1$ parametar. Tada G ima hi-kvadrat raspodelu sa p stepeni slobode pod nultom hipotezom da je svih p koeficijenata nagiba za kovarijate u modelu jednako 0.

3.2.2 WALD TEST

Još jedan od pristupa ispitivanju značajnosti koeficijenata jeste da koristimo test koji povezuje koeficijente sa njihovim standardnim greškama. *Wald test* predstavlja količnik ocene maksimalne verodostojnosti koeficijenta $\hat{\beta}$ sa njegovom standardnom greškom $S_{\widehat{\beta}}$ i statistički ima približno standardnu normalnu raspodelu $N(0,1)$ pod hipotezom da je $\beta = 0$. Kvadrat ove Z statistike za univarijabilni slučaj ima približno χ^2 raspodelu sa jednim stepenom slobode. Odnosno Wald statistika za univarijantni slučaj je [3]:

$$Z = \frac{\hat{\beta}}{S_{\widehat{\beta}}} : N(0,1)$$

$$Z^2 : \chi_1^2$$

Test statistika količnika verodostojnosti i Wald statistika daju približno iste vrednosti kad su u pitanju veliki uzorci, pa ako je neka studija dovoljno obimna nije bitno koju statistiku koristimo, međutim ako su uzorci malog obima statisitke mogu značajno da se razlikuju i pokazano je da je test statistika količnika verodostojnosti u ovakvim situacijama tačnija.

Primer 4

Posmatrajmo Tabelu 5 na cd-u u prilogu.

Korišćenjem SPSS 17 dobijamo sledeće:

Promenljiva	β	Stan.gr.	Wald(Z^2)	p
Prethodno_iskustvo	0.003	0.001	9.557	0.002
Konstanta	-1.586	0.157	101.406	0.000

Tabela 6

Kako je p vrednost Waldove test statistike jednaka $P\{\chi_1^2 > 9{,}557\} = 0.001 < 0.002$ sledi da odbacujemo hipotezu da je $\beta = 0$, odnosno dobijamo i ovim testom da je promenljiva *prethodno_iskustvo* značajna u određivanju da li je osoba manjinskog porekla.

Waldova test statistika za multivarijabilni slučaj je:

$$W = \widehat{\boldsymbol{\beta}}'\left[\widehat{Var}(\widehat{\boldsymbol{\beta}})\right]^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})\widehat{\boldsymbol{\beta}}$$

I ima χ^2 raspodelu sa $p+1$ stepenom slobode pod početnom hipotezom da je svaki od $p+1$ koeficijenata jednak nuli. Statistiku za samo p koeficijenata nagiba dobijamo kad eliminišemo $\widehat{\beta_0}$ iz vektora $\widehat{\boldsymbol{\beta}}$ kao i odgovarajuće redove, odnosno kolone iz $'\boldsymbol{V}\boldsymbol{X}$.

Wald-ov test često ima nedostatak da se ne odbacuje nulta hipoteza iako su koeficijenti značajni tako da je preporučljivije koristiti test količnika verodostojnosti

3.3 INTERVAL POVERENJA ZA OCENU

Nakon testiranja značajnosi koeficijenata interpretiraćemo testiranje intervala poverenja za parametre koji nas interesuju. Prvo razmatramo intervale poverenja za univarijabilni slučaj.

Baza za konstrukciju intervala ocene je ista statistička teorija koju smo koristili za formulisanje testa za značajnost modela. Intervali poverenja ocene za nagib i odsečak su bazirani na njihovim odgovarajućim Wald testovima. Krajnje tačke za 100(1 – α)% interval poverenja za ocenjeni koeficijent nagiba su[14]:

$$\widehat{\beta_1} \mp z_{1-\alpha/2}S_{\widehat{\beta_1}} \qquad (10)$$

a za odsečak:

$$\widehat{\beta_0} \mp z_{1-\alpha/2} S_{\widehat{\beta_0}} \quad (11)$$

gde je sa $S_{\widehat{\beta_\circ}}$ označena ocena standardne greške (zasnovane na modelu) za odgovarajuću ocenu parametra, a $z_{1-\alpha/2}$ je gornja tačka standardne normalne raspodele.

Kao primer, posmatrajmo slaganje modela sa podacima (Tabela 5 na cd-u u prilogu) koji povezuje promenljivu *prethodno_iskustvo* sa *manjina*. Rezultati su prikazani u Tabeli 6. Krajnje tačke 95%-tnog intervala poverenja za koeficijent nagiba su $0.003 \mp 1.96 * 0.001$, što daje interval (0.00104, 0.00496). Odnosno rezultat pokazuje da ukoliko se *prethodno_iskustvo* poveća za jedan mesec promena logoritma šanse za zavisnu promenljivu *manjina* iznosi 0.003, te sa 95- procentnim poverenjem zaključujemo da može može biti u intervalu od (0.00104, 0.00496)

Kao u slučaju bilo kog regresionog modela, konstanta daje ocenu ishoda za *x* = 0, izuzev ako je nezavisna promenljiva centrirana u nekoj vrednosti koja ima smisla u kliničkoj praksi. Za naš primer, konstanta daje ocenu logaritma odnosa šansi za *manjina* kada je *prethodno_iskustvo* jednako 0. Kao rezultat, konstantanta sama po sebi nema značajnu interpretaciju. Krajnje tačke za 95% interval poverenja za konstantu su $-1.58 \mp 1.96 * 0.157$, što daje interval (-1.893,-1.278). Konstanta je važna kada se razmatra tačkasta i intervalna ocena za logit.

Logit je linearan deo logističkog regresionog modela, i kao takav podseća na fitovanu pravu u linearnoj regresiji. Ocena za logit je

$$\widehat{g(x)} = \widehat{\beta_0} + \widehat{\beta_1} x$$

Ocena varijanse za ocenjeni logit zahteva korišćenje varijansu od sume. U tom slučaju je

$$\widehat{Var}\left(\widehat{g(x)}\right) = \widehat{Var}\left(\widehat{\beta_0}\right) + x^2 \widehat{Var}\left(\widehat{\beta_1}\right) + 2x\widehat{Cov}\left(\widehat{\beta_0}, \widehat{\beta_1}\right) \quad (12)$$

Krajnje tačke za 100(1 – α)% interval poverenja za logit (na osnovu Wald statistike) su:

$$\widehat{g(x)} \mp z_{1-\alpha/2} S_{\widehat{g(x)}}$$

gde je $S_{\widehat{g(x)}}$ pozitivan kvadratni koren varijanse ocenjene u (12)[6].

Ocenjen logit za osobu sa radnim iskustvom od 300 meseci je

$$\widehat{g(300)} = -1.586 + 0.003 * 300 = -0.686$$

Ocenjena kovarijansna matrica za ocenjene koeficijenata iz Tabele 6:

	prethodno_iskustvo	Konstanta
prethodno_iskustvo	0.000001	-0.00011
Konstanta	-0.00011	0.024649

Tabela 7

Ocenjena varijansa je :

$$\widehat{Var}\left(\widehat{g(300)}\right) = 0{,}024649 + 300^2 * 0.000001 + 2 * 300 * (-0.00011) = 0.048$$

dok je ocenjena standardna greška :

$$S_{\widehat{g(x)}} = \sqrt{0.048} = 0.220$$

Tako da su krajnje tačke 95% intervala poverenja za logit za osobu sa *prethodnim_iskustvom* od 300 meseci

$$-0{,}686 \mp 1.96 * 0{,}220 = (-1.117, -0.254)$$

Ocena za logit i njegov interval poverenja su osnova za ocenu fitovanih vrednosti, u ovom slučaju logističke verovatnoće i njenog intervala poverenja. Konkretno, korišćenjem

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

Za osobu sa radnim iskustvom od 300 meseci ocenjena logistička verovatnoća je:

$$\hat{\pi}(65) = \frac{e^{\hat{g}(300)}}{1 + e^{\hat{g}(300)}} = \frac{e^{-0.686}}{1 + e^{-0.686}} = 0.334 \qquad (13)$$

I krajnje tačke 95% intervala poverenja su dobijene iz odgovarajućeh krajnjih tačaka intervala poverenja za logit.Tako da su krajnje tačke za $100(1-\alpha)\%$ interval poverenja (baziran na Wald-ovom testu) za fitovanu vrednost:

$$\frac{e^{\widehat{g(x)} \mp z_{1-\alpha/2}\hat{S}_{\widehat{g(x)}}}}{1 + e^{\widehat{g(x)} \mp z_{1-\alpha/2}\hat{S}_{\widehat{g(x)}}}}$$

Tako da je za osobu sa radnim iskustvom od 300 meseci donja granica intervala:

$$\frac{e^{-1.117}}{1 + e^{-1.117}} = 0.246$$

odnosno gornja granica:

$$\frac{e^{-0.254}}{1 + e^{-0.254}} = 0.436$$

Fitovana vrednost izračunata u (13) je analogna odgovarajućoj tački na pravoj dobijenoj linearnom regresijom, U linearnoj regresiji svaka tačka na fitovanoj pravoj predstavlja ocenu proporcije zavisne promenljive u populaciji sa kovarijatom x. Tako da je vrednost 0.358 u stvari ocena proporcije subjekata sa 300 meseci prethodnog radnog iskustva u uzorku populacije koji pripadaju manjinskom stanovništvu. Intervala poverenja nam govori da ova proporcija može da se kreće između 0.246 i 0.436 sa 95% poverenjem.

Kad je u pitanju multivarijabilni slučaj, odnosno višestruka regresija, računanje intervala poverenja za koeficijente se izvodi na isti način kao i za univarijabilni slučaj korišćenjem (10) i (11). Kod računanja intervala poverenja za logit koristimo istu osnovnu ideju kao i kod univarijabilnog slučaja sa tom razlikom da sada imamo više izraza uključenih u sumiranje. Opšti izraz za ocenu logita koji sadrži *p* kovarijati je:

$$\widehat{g(x)} = \widehat{\beta_0} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \widehat{\beta_p} x_p$$

odnosno:

$$\hat{g}(\boldsymbol{x}) = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$$

gde je $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, a $\widehat{\boldsymbol{x}} = (x_0, x_1, \dots, x_p)$ je konstantni vektor gde je $x_0 = 1$

Tada važi da je:

$$\widehat{Var}\left(\widehat{g(x)}\right) = \sum_{j=0}^{p} x_j^2 \widehat{Var}\left(\widehat{\beta_j}\right) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \widehat{Cov}\left(\widehat{\beta_0}, \widehat{\beta_1}\right)$$

Odnosno ako izrazimo u matričnom obliku dobijamo:

$$\widehat{Var}\left(\widehat{g(x)}\right) = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}$$

Dalje izračunavanje se vrši analogno univarijabilnom slučaju.

4 INTERPRETACIJA FITOVANOG LOGISTIČKOG MODELA

Pretpostavićemo da je logistički regresioni model prilagođen podacima, odnosno da je fitovan i da su promenljive u modelu značajne, tj. da su odgovarajući regresioni koeficijenti različiti od nule.

Interpretacija fitovanog modela porazumeva izvođenje zaključaka na osnovu ocenjenih koeficijenata u modelu. Ključno pitanje koje se tu javlja je šta nam, zapravo, ocenjeni koeficijenti govore o pitanjima zbog kojih je i započeto istraživanje. Prilikom interpretacije modela posmatraju se dva problema a to su: određivanje funkcionalne veze između zavisne i nezavisne promenljive i definisanje odgovarajuće jedinice promene za nezavisnu promenljivu.

Funkcionalnu vezu između zavisne i nezavisne promenljive u logističkom regresionom modelu daje logit funkcija, tj.

$$g(x) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Na dalje ćemo, zbog jednostavnosti, raditi samo sa jednom nezavisnom promenljivom.

U logističkom regresionom modelu koeficijent nagiba β_1 predstavlja promenu u logitu po jedinici promene nezavisne promenljive, tj. : $\beta_1 = g(x+1) - g(x)$

Interpretaciju fitovanog logističkog regresionog modela ćemo dati u tri slučaja u zavisnosti od toga da li je nezavisna promenljiva dihotomna, polihotomna ili neprekidna.

4.1 DIHOTOMNA NEZAVISNA PROMENLJIVA

Slučaj kada je nezavisna promenljiva u logističkom regresionom modelu dihotomna predstavlja osnovu za druge slučajeve i podrazumeva da nezavisna promenljiva može uzeti dve vrednosti. U našem slučaju neka je nezavisna promenljiva kodirana sa 0 i 1.

Kako koeficijent β_1 predstavlja stopu promene zavisne promenljive po jedinici promene nezavisne promenljive, važi da je:

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

Da bismo mogli interpretirati dobijeni rezultat uvešćemo pojam *odnos šansi* (unakrsni odnos šansi, odds ratio), koji daje meru povezanosti nezavisne promenljive sa ishodom od interesa.

Šansa je odnos verovatnoća da se događaj desi prema verovatnoći da se događaj ne desi.

Šansa da je zavisna promenljiva uzela vrednost 1, kada nezavisna promenljiva uzme vrednost 1 je:

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

Kada nezavisna promenljiva uzme vrednost 0, šansa da je zavisna promenljiva uzela vrednost 1 je:

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

Odnos šansi (unakrsni odnos šansi), u oznaci OR, je definisan kao odnos ove dve šanse, tj.

$$OR=\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} \tag{14}$$

Moguće vrednosti logističke verovatnoće se mogu predstaviti tablicom $2x2$ na sledeći način:

Rezultujuća promenljiva (Y)	Nezavisna promenljiva (X)	
	$x=1$	$x=0$
$y=1$	$\pi(1)=\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0)=\frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y=0$	$1-\pi(1)=\frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0)=\frac{1}{1+e^{\beta_0}}$
Ukupno	1	1

Tabela 8

Ova tabela opravdava to što se odnos šansi OR još naziva i unakrsni odnos šansi, jer vidimo da se OR dobija kao odnos unakrsnog proizvoda elemenata na glavnoj dijagonali date tabele i elemenata na sporednoj dijagonali [25].

Zamenom izraza iz tabele u (14) dobijamo:

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

Ova jednostavna veza između koeficijenta i odnosa šansi je osnovni razlog zašto se logistička regresija pokazala kao moćan analitički alat.

Primer 5

	Pušači(1)	Nepušači(0)	Ukupno
KSB(1)	16	10	26
KSB(0)	7	27	34
Ukupno	23	37	60

Tabela 9

Posmatrajmo Tabelu 9. Preko OR ćemo ispitati povezanost nezavisna promenljive koja je kodirana sa 1 ako osoba puši, odnosno sa 0 ako je osoba nepušač, sa zavisnom promenljivom *KBS*(koronarno srčano oboljenje), odnosno važi:

$$\widehat{OR} = \frac{16/7}{10/27} = 6.17$$

Odnosno važi da pušači imaju 6.17 puta veću šansu da obole od *KBS* od nepušača.

Uvešćemo još jedan pojam, a to je *relativni rizik*, u oznaci RR. Relativni rizik predstavlja odnos verovatnoća uspeha u okviru dve grupe.

U našem slučaju:

$$RR = \left(\frac{P(Y=1|X=1)}{P(Y=1|X=0)}\right) = \frac{\pi(1)}{\pi(0)}$$

Izraz za odnos šansi se sada može zapisati na sledeći način:

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

Vidimo da odnos šansi aproksimira relativni rizik kada $\frac{1-\pi(0)}{1-\pi(1)} \to 1$, odnosno kada su verovatnoće neuspeha u obe grupe približno jednake. U praksi se srećemo sa ovom situacijom kod ispitivanja relativno retkih bolesti, koje kao takve imaju malu verovatnoću pojave.

Na sledećem primeru ćemo videti razliku između OR i RR.

Primer 6

Dati su podaci u Tabeli 10, koji se odnose na broj preživelih i poginulih putnika na Titanku, gde je bilo ukupno 1313 putnika, od toga 462 žene i 851 muškarac.

	Žene	Muškarci	Ukupno
Preživeli	308	142	450
Poginuli	154	709	863
Ukupno	462	851	1313

Tabela 10

Iz same tabele se vidi da je verovatnije da muškarac umre nego žena, pa ćemo smrt muškaraca uzeti kao referentan ishod, jer ćemo na taj način dobiti vrednost odnosa šansi veću od jedan. Dakle, odnos šansi će porediti odnose šansi za smrt u okviru svake grupe, tj. među muškarcima i ženama.

Šansa za smrt kod žena je:

$$\frac{\frac{154}{462}}{\frac{308}{462}} = 0.5$$

Šansa za smrt kod muškaraca je

$$\frac{\frac{709}{851}}{\frac{142}{851}} = 4.993$$

Dakle, odnos šansi je:

$$\widehat{OR} = \frac{4.993}{0.5} = 9.986$$

što znači da su skoro deset puta veće šanse za smrt muškarca u odnosu na smrt žene.

Relativni rizik poredi verovatnoće za smrt u okviru svake grupe, tj.

$$RR = \frac{verovatnoća\ smrti\ kod\ muškaaca}{verovatnoća\ smrti\ kod\ žena}$$

Verovatnoća smrti kod muškaraca je $\frac{709}{851} = 0.833$, dok je verovatnoća smrt kod žena $\frac{154}{462} = 0.333$. Zamenom ovih vrednosti dobijamo da je relativni rizik $\widehat{RR} = 2.5$, odnosno postoji 2.5 puta veća verovatnoća za smrt muškarca nego za smrt žena.

Vidimo da i OR i RR pokazuju da će muškaraci najverovatnije umreti, ali kod OR je mnogo veća šansa za to nego kod RR. Pitanje je koji od ova dva testa daje bolju procenu?

Ovde postoje tri problema:

1. Relativni rizik meri događaje na način koji je kompatibilan sa ljudskom procenom. Uzmimo npr. dve grupe, prva grupa ima 25% šanse da umre, dok druga ima 75% šanse, vidimo da je relativni rizik ovde 3, dok je odnos šansi čak 9. U ovakvim situacijama relativni rizik ima prednost nad OR.
2. Neke istraživačke metode onemogućavaju upotrebu relativnog rizika. Podaci koji potiču iz retrospektivnih studija se zasnivaju na uzorcima ispitanika sa bolešću (slučajevi) i onih bez nje (kontrole) pa se retrospektivno određuje razlika među njima i samim tim onemogućava se upotreba RR.
3. Relativni rizik može ponekad da dovede do zbunjujćih i dvosmislenih rezultata, ovo proizilazi iz činjenice da su relativne mere često kontra intuitivne[19].

Odnos šansi je parametar od interesa u logističkoj regresiji između ostalog i zbog jednostavne interpretacije. Međutim njegova ocena $\widehat{OR}$ teži da ima raspodelu koja je iskrivljena s obzirom na činjenicu da $\widehat{OR} \in (0, \infty)$ sa nultom vrednošću jednakom 1. Teoretski za dovoljno velike uzorke $\widehat{OR}$ ima normalnu raspodelu. Prema tome zaključci se obično baziraju na uzoračkoj raspodeli za $\ln \widehat{OR} = \widehat{\beta_1}$ koja teži normalnoj za mnogo manje veličine uzoraka. Tako da ćemo $100(1-\alpha)\%$ interval poverenja za $\widehat{OR}$ računati tako što ćemo prvo izračunati krajnje tačke intervala poverenja za koeficijent $\widehat{\beta_1}$ i zatim eksponovati ove vrednosti[15]. Tačnije interval poverenja je dat izrazom:

$$e^{\widehat{\beta_1} \mp z_{1-\alpha/2} S_{\widehat{\beta_1}}}$$

Primer 7.

Posmatrajmo Tabelu 9. Već smo izračunali da je $\widehat{OR} = 6.17$ i važi da je $\beta_1 = 1.820$ i $S_{\widehat{\beta_1}} = 0.585$, tada je 95% interval poverenja za odnos šansi:

$$e^{\widehat{\beta_1} \mp z_{1-\alpha/2} S_{\widehat{\beta_1}}} = e^{1.820 \mp 0.585 * 1.96} = (1.96, 19.425)$$

Odnosno zaključijemo da je populacioni odnos šansi za pojavu *KBS* kod pušača u odnosu na nepušače između 1.96 i 19.425.

4.2 POLIHOTOMNA NEZAVISNA PROMENLJIVA

Pretpostavimo da nezavisna promenljiva može uzeti više od dve vrednosti. Na primer, možemo imati promenljivu koja predstavlja rasu, boju kose, boju očiju i sl. i svaka od ovih promenljivih ima fiksni broj diskretnih vrednosti. Da bismo mogli da manipulišemo ovakvim podacima, neophodno je da napravimo odgovarajuće dizajne promenljive koje odgovaraju svim statističkim paketima[27].

Posmatrajmo Tabelu 11. *Škola* je zavisna promenljiva koja je binarna, dok je *Rasa* nezavisna promenljiva koja ima četiri nivoa.

	RASA				
ŠKOLA	latino-američka	azijska	afričko-američka	bela	ukupno
privatna škola	2	1	2	27	168
javna škola	22	10	18	118	32
ukupno	24	11	20	145	200
OR	0.40	0.44	0.49	1.00	
ln(OR)	-0.923	-0.828	-0.722	0	

Tabela 11

Da bismo napravili odgovarajuću dizajn promenljivu uzećemo *belu* rasu kao referentu. Odgovarajuća dizajn promenljiva za naš problem je data u Tabeli 12. Tako da je npr. OR za latino-američku rasu jednak: $(2 * 118)/(22 * 27) = 0.4$ odnosno latino-amerikanci imaju 0.4 puta veću šansu da će ići u privatnu školu nego belci. U Tabeli 11 su takođe date vrednosti i za ln $\widehat{OR}$.

	Dizajn promenljive		
RASA(kod)	rasa_2	rasa_3	rasa_4
bela(1)	0	0	0
latino-američka(2)	1	0	0
azijska(3)	0	1	0
afričko-američka(4)	0	0	1

Tabela 12

Ako uporedimo ocenjene koeficijente u Tabeli 13 sa ln $\widehat{OR}$ vidimo da važi:

$$\ln\left[\widehat{OR}(rasa\_2, bela)\right] = 0.923 = \widehat{\beta_1}$$

$$\ln\left[\widehat{OR}(rasa\_3, bela)\right] = 0.828 = \widehat{\beta_2}$$

$$\ln\left[\widehat{OR}(rasa\_4, bela)\right] = 0.722 = \widehat{\beta_3}$$

Promenljiva	β_{i-1}	Stan.gr
rasa_2	-0.923	0.769
rasa_3	-0.828	1.07
rasa_4	-0.722	0.775
Konstanta	-1.475	0.213

Tabela 13

Pokazaćemo da ovo važi zbog načina na koji smo kodirali dizajn promenljivu. Uzmimo za primer *rasu_2 i belu rasu*. Odnosno imamo:

$$\ln\left[\widehat{OR}(rasa\_2, bela)\right] = \hat{g}(rasa_2) - \hat{g}(bela)$$
$$= \left(\widehat{\beta_0} + \widehat{\beta_1} * (rasa_2 = 1) + \widehat{\beta_2} * (rasa_3 = 0) + \widehat{\beta_3} * (rasa_4 = 0)\right)$$
$$- \left(\widehat{\beta_0} + \widehat{\beta_1} * (rasa_2 = 0) + \widehat{\beta_2} * (rasa_3 = 0) + \widehat{\beta_3} * (rasa_4 = 0)\right) = \widehat{\beta_1}$$

Pošto posmatramo univarijabilni slučaj binarne logističke regresije, ocenu standardne greške možemo izračunati pomoću odgovarajućih frenkvencija iz Tabele 11, odnosno za koeficijent koji odgovara *rasi_2* važi:

$$\hat{S}_{\widehat{\beta_1}} = \sqrt{\left(\frac{1}{22} + \frac{1}{2} + \frac{1}{118} + \frac{1}{27}\right)} = 0.769$$

Intervali poverenja za $\widehat{OR}$ se računa na isti način kao i kad je u pitanju dihotomna nezavisna promenljiva, odnosno:

$$e^{\widehat{\beta_J} \mp z_{1-\alpha/2} S_{\widehat{\beta_J}}}$$

Tako je interval poverenja za $\widehat{OR}$ koji odgovara *rasi_2* data sa:

$$e^{-0.923 \mp 1.96 * 0.769} = (0.088, 1.793)$$

Odnosno zaključujemo da je odnos šansi da će latino-amerikanci upisati privatnu školu u odnosu na belce u intervalu $(0.088, 1.793)$.

4.3 NEPREKIDNA NEZAVISNA PROMENLJIVA

Sada ćemo posmatrati logistički regresioni model koji sadrži neprekidnu nezavisnu promenljivu i pretpostavićemo da je logit linearan po toj promenljivoj odnosno:

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

U ovom slučaju za razliku od slučaja kada je nezavisna promenljiva diskretna, promena od jedne jedinice nezavisne promenljive najčešće nije interesantna. Na primer, rast sistolnog krvnog pritiska za 1 $mmHg$ može biti suviše mali da bismo ga smatrali važnim dok bi recimo rast od 10 jedinica predstavljao značajniji podatak. Sa druge strane, ako se

vrednosti koje nezavisna promenljiva može uzeti kreću u inervalu od 0 do 1, tada bi promena od jedne jednice bila suviše velika, dok bi promena od 0.01 jedinice bila realnija.

Dakle, da bismo obezbedili pravilnu interpretaciju modela smatraćemo da se desila promena od c jedinica. Tada je promena u logitu sledeća:

$$g(x+c) - g(x) = \beta_0 + \beta_1(x+c) - \beta_0 + \beta_1 x = \beta_1 c$$

Sada je odnos šansi dat izrazom:

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

tj. njegova ocena je:

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

Važi da c može uzeti bilo koju vrednost, pri čemu se mora voditi računa o tome da se na jasan način ukaže kako se menja rizik da je ishod prisutan sa promenom nezavisne promenljive [6].

Ocena standardne greške koja je otrebna za izračunavanje intervala poverenja se dobija množenjem c sa ocenjenom standardnom greškom za $\widehat{\beta_1}$. Krajnje tačke za $100(1-\alpha)\%$ interval poverenja za $\widehat{OR}$ su:

$$e^{\widehat{\beta_1} \mp c z_{1-\alpha/2} S_{\widehat{\beta_1}}}$$

Pošto tačkasta ocena parametara i krajnje tačke intervala poverenja direktno zavise od c neophodno je jasno definisati određenu vrednost c u svim tabelama i izračunavanjima.

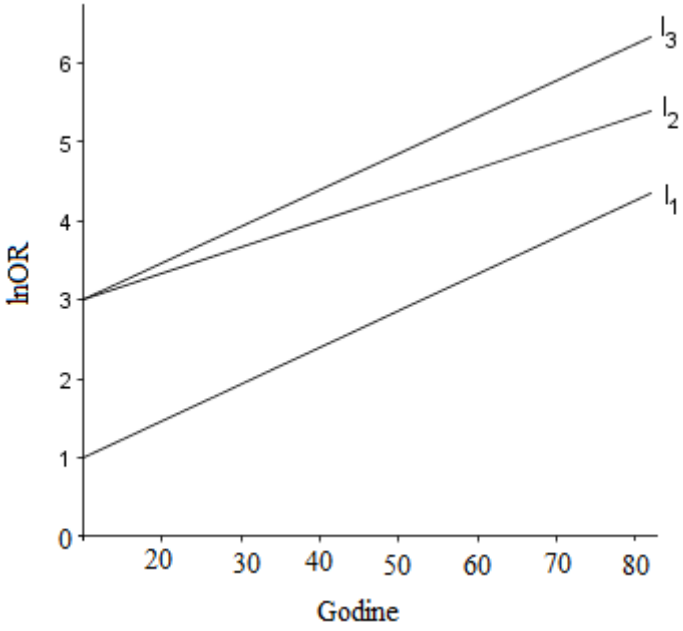
4.4 INTERAKCIJA I OMETANJE

U ovom odeljku ćemo predstaviti koncept interakcije i ometanja i pokazati kako možemo kontrolisati ova dva efekta.

Pretpostavimo, radi lakše interpretacije, da posmatramo model sa neprekidnom kovarijatom i dihotomnim rizičnim faktorom. Ako je povezanost između kovarijate i zavisne promenljive ista za svaki nivo rizičnog faktora tada ne postoji interakcije između kovarijate i rizičnog faktora. Grafički, nedostatak interakcije se predstavlja kao model sa dve paralelne linije. Kada je interakcija prisutna povezanost između rizičnog faktora i zavisne promenljive se razlikuje ili zavisi na neki način od nivoa kovarijate. Odnosno kovarijata modifikuje efekte rizičnog faktora. Epidemiolozi koriste termin *modifikator efekta* da opišu promenljivu koja je u interakciji sa rizičnim faktorom [21].

Najjednostavnija i najčešće korišćen model koji uključuje i interakciju je onaj kod koga je logit ometača linearan, ali pod drugim je uglom u odnosu na rizični faktor.

Figura na Slici 6 predstavlja tri različita logita. Posmatrajmo primer (Tabela 14 na cd-u u prilogu), gde je zavisna promenljiva *Kupljena*-koja označava da li je dodatna garancija za neki proizvod kupljena ili ne, rizični faktor je *Pol*, a kovarijata je *Starost*. Pretpostavimo radi lakše interpretacije da linija označena sa l_1 predstavlja logit za žene u funkciji godina, l_3 predstavlja logit za muškarce. Ove dve linije su paralelne, što znači da je odnos između toga da li je garancija kupljena ili ne i godina ista i za muškarce i za žene. U ovoj situaciji nema interakcije. ln OR za pol, u zavisnosti od godina je predstavljen kao razlika između l_1 i l_3 . Ova razlika je jednaka za vertikalnoj razdaljini između ovih linija, i ona je ista za sve godine.



Slika 6

Pretpostavimo sad da je umesto linije l_3 , dobijena linija l_2 . Ova linija nije paralelna sa l_1 što znači da postoji interakcija između pola i godina. ln OR za pol, u odnosu na godine je i dalje predstavljen kao verikalna razdaljina između l_1 i l_2, ali ova razlika sad zavisi od starosti osobe. Odnosno, ne možemo oceniti OR za pol, a da ne znamo na koje godište se taj OR odnosi. Jednom rečju, godine su u ovom slučaju modifikator efekta. U praksi smatramo da je kovarijata modifikator efekta ako interakcija dodata u model bude i klinički i statistički značajna. Kada je kovarijata modifikator efekta, njen status kao ometača je od manjeg značaja pošto je ocena efekta rizičnog faktora zavisi od specifične vrednosti kovarijate.

Za kovarijatu kažemo da je *ometač* (*confounder*) ako je povezana i sa zavisnom i sa nezavisnom promenljivom od interesa (rizičnim faktorom). Kada postoji ova povezanosti, tada se odnos između zavisne prmenljive i rizičnog faktora zove *ometajući (confounded)*.U

praksi, način određivanja da li je kovarijata ometač ili ne je da uporedimo ocenjene koeficijente riizičnog faktora iz modela koji sadrži i onog koji ne sadrži kovarijatu. Svaka klinički značajna promena u ocenjenim koeficijentima rizičnog faktora ukazuje na to da je kovarijata ometač (ali ne uvek) i trebala bi da bude uključena u model, bez obzira na statističku značajnost njenog ocenjenog koeficijenta [21].

Primer 8

Tabele 15 i 16 predstavljaju rezultate fitivovanja logističkog regresionog modela, za dva različita skupa podataka, jednog iz tabele 14, a drugog hipotetičkog. Interakcija je dodata u model kreiranjem promenljive koja je jednaka proizvodu promenljivih *Pol* i *Starost*. Rezultate koji su nam od interesa smo predstavili u Tabelama 15 i 16 [23], za prvi odnosno drugi slučaj.

Model	Konstanta	Pol	Starost	Pol x Starost	Odstupanje	G
1	0.001	1.421			54.876	
2	-2.911	1.099	0.064		47.621	7.255
3	-2.207	0.72	0.049	0.022	47.462	0.159

Tabela 15

Iz Tabele 15 vidimo da se koeficijent promenljive *Pol* promenio od 1.421 u 1.099, odnosno smanjio se za 30% dodavanjem promenljive *Starost* u model 2. Odavde vidimo da postoji efekat ometanja, odnosno da je promenljiva *Starost* ometač. Kada je interakcija dodata u model 3, imamo promenu odstupanja u iznosu od $G = 0.159$, koji kada uporedimo sa χ_1^2 imamo p-vrednost u iznosu od 0.69, što očigledno nije značajno. Iz ovoga zaključujemo da je promenljiva *Starost* ometač, ali ne i modifikator efekta. Za ove podatke važi da je model 2 bolji od modela 3.

Model	Konstanta	Pol	Starost	Pol x Starost	Odstupanje	G
1	0.201	3.386			53.876	
2	-6.672	1.724	0.074		35.621	18.255
3	-4.825	-3.259	0.029	0.201	27.462	8.159

Tabela 16

Ako posmatramo podatke iz Tabele 16 vidimo da se da se koeficijent promenljive *Pol* promenio od 3.386 u 1.724, odnosno smanjio se za 96% dodavanjem promenljive *Starost* u model 2. Odavde vidimo da postoji efekat ometanja, odnosno da je promenljiva *Starost* ometač. Kada je interakcija dodata u model 3, imamo promenu odstupanja u iznosu od $G = 8.159$, koji kada uporedimo sa χ_1^2 imamo p-vrednost u iznosu od 0.004, što je očigledno

značajno. Iz ovoga zaključujemo da je promenljiva *Starost* ometač ali i modifikator efekta. Ovaj rezultat takođe podrazumeva da se OR za *Pol* mora računati u odnosu na određenu *Starost*, o čemu će biti više reči u narednom odeljku. Za ove podatke važi da je model 3 bolji od modela 2.

Koncepti interakcije i ometanje mogu da se prošire na bilo koju drugu situacije koja uključuje bilo koji broj promenljivih ii bilo koji mernu skalu. Mi smo se ovde bazirali na binarnim i neprekidnim promenljivim jer je rezultat lakše predstaviti, što nije slučaj sa komplikovanijim modelima.

4.5 OCENA OR PRI INTERAKCIJI

Kada postoji interakcija između rizičnog faktora i neke druge nezavisne promenljive ocena OR za rizični faktor zavisi od promenljive sa kojom je on u interakciji. U ovakvom slučaju ocena OR se ne može dobiti prostim eksponovanjem ocenjenog koeficijenta, već se vrši u tri koraka i to:

1. Ispišemo odgovarajuče logite za oba nivo rizičnog faktor
2. Izračunamo razliku ova dva logita
3. Eksponujemo vrednost izraza dobijenog pod 2 [7].

Pretpostavimo da je rizični faktor označen sa R a odgovarujuća kovarijata koja je sa njim u interakciji neka je označena sa K i njihova interakcija neka je $R \times K$. Neka su vrednosti promenljivih $R = r$ i $K = k$, tada je logit za ovakav model:

$$g(r,k) = \beta_0 + \beta_1 r + \beta_2 k + \beta_3 r \times k$$

Neka su vrednosti dva nivo rizičnog faktora data sa r_0 i r_1. Sledeći proceduru, prvi korak je ispisivanje odgovarajućih logita odnosno:

$$g(r_1,k) = \beta_0 + \beta_1 r_1 + \beta_2 k + \beta_3 r_1 \times k$$
$$g(r_0,k) = \beta_0 + \beta_1 r_0 + \beta_2 k + \beta_3 r_0 \times k$$

Drugi korak predstavlja računanje odgovarajuće razlike, odnosno:

$$\begin{aligned} \ln(OR(R = r_1, R = r_0, K = k)) &= g(r_1,k) - g(r_0,k) \\ &= \beta_0 + \beta_1 r_1 + \beta_2 k + \beta_3 r_1 - (\beta_0 + \beta_1 r_0 + \beta_2 k + \beta_3 r_0 \times k) \\ &= \beta_1(r_1 - r_0) + \beta_3 k(r_1 - r_0) \end{aligned} \qquad (15)$$

Treće korak predstavlja eksponovanje date vrednosti i dobijanje traženog OR:

$$OR = e^{\beta_1(r_1 - r_0) + \beta_3 k(r_1 - r_0)}$$

Ocenu OR dobijamo zamenom odgovarajućih vrednost u gornji izraz [22].

Krajne tačke intervala poverenja dobijamo na isti način kao i za modele u kojima nema interakcije. Odnosno računamo krajnje tačke intervala poverenja za $\ln OR$ i onda ih

eksponujemo. Da bi našli krajnje tačke intervala poverenja potreban nam je prvo ocenjivač varijanse od ocenjivča ln OR datog u (15). Odnosno:

$$\widehat{Var}\left(\ln\left(\widehat{OR}(R=r_1,R=r_0,K=k)\right)\right)$$
$$=(r_1-r_0)^2\times\widehat{Var}(\widehat{\beta_1})+\left(k(r_1-r_0)\right)^2\times\widehat{Var}(\widehat{\beta_3})+$$
$$+2k(r_1-r_0)^2\times\widehat{Cov}(\widehat{\beta_1},\widehat{\beta_3}).$$

Dalje imamo da su krajnje tačke $100\times(1-\alpha)\%$ intervala poverenja jednake:

$$\left(\widehat{\beta_1}(r_1-r_0)+\widehat{\beta_3}k(r_1-r_0)\right)\pm z_{1-\alpha/2}S_{\ln\left(\widehat{OR}(R=r_1,R=r_0,K=k)\right)} \qquad (16)$$

gde standardna greška predstavlja pozitvan kvadratni koren iz ocenjene varijanse. Krajnje tačke intervala poverenja za ocenjeni $\widehat{OR}$ dobijamo eksponovanjem vrednosti iz (16) [7], [6].

Kada je rizični faktor dihotoman (neka je $r_1=1, r_0=0$) tada su ocenjivač za ln OR i njegova varijansa dati sa [22]:

$$\ln\left(\widehat{OR}(R=1,R=0,K=k)\right)=\widehat{\beta_1}+\widehat{\beta_3}k$$
$$\widehat{Var}\left(\ln\left(\widehat{OR}(R=1,R=0,K=k)\right)\right)=\widehat{Var}(\widehat{\beta_1})+k^2\widehat{Var}(\widehat{\beta_3})+2k\widehat{Cov}(\widehat{\beta_1},\widehat{\beta_3}).$$

dok su krajnje tačke intervala poverenja date sa:

$$\left(\widehat{\beta_1}+\widehat{\beta_3}k\right)\pm z_{1-\alpha/2}S_{\ln\left(\widehat{OR}(R=1,R=0,K=k)\right)}$$

Primer 9

Posmatrajmo logistički regresioni model (Tabela 17 na cd-u u prilogu) koji ispituje ishoda presude (0-nije kriv, 1-kriv) u zavisnosti od rizičnog faktora *Pol* (0-žena, 1-muškarac) i kovarijate *Atraktivna* (1-osoba je atraktivna,0-osoba nije atraktivna).

Rezultati fitovanja ovog modela dati su u Tabeli 18

Model	Konstanta	Pol	Atraktivna	Pol x Atraktivna	Odstupanje	G	p
0	1.312						
1	1.988	-1.164			161.919		
2	1.974	-1.165	0.03		161.913	0.006	0.94
3	1.504	-0.373	1.414	-1.995	156.906	5.007	0.03

Tabela 18

Kao što vidimo iz Tabele 18, promenljiva *Atraktivna* nije ometač i jer je promena u koeficijentu za *Pol* u modelu dva samo 0.0008%. Promenljiva *Atraktivna* je u interakciji sa promenljivom *Pol*, što sledi iz p-vrednosti od 0.03.

Ako pretpostavimo da u modelu nema interakcije, tada bi ocenjeni OR iznosio:

$$OR=e^{\beta_1}=e^{-1.164}=0.312$$

odnosno važi da muškarci imaju 0.312 puta veću šansu da budu osuđeni nego žene.

Kako u modelu postoji interakcija važi da ocenjeni $\ln OR$ za promenljivu *Presuda*, za osobe čija je atraktivnost ocenjena sa a je:

$$\ln\left(\widehat{OR}(Pol = 1, Pol = 0, Atraktivna = a)\right) = -0.373 - 1.995a$$

Da bismo dobili ocenjenu varijansu, potrebna nam je kovarijansna matrica data u Tabeli 19

Konstanta	0.153			
Pol	-0.153	0.285		
Atraktivna	-0.153	0.153	0.680	
Pol*Atraktivna	-0.209	-0.285	-0.680	0.917
	Konstanta	Pol	Atraktivna	Pol*Atraktivna

Tabela 19

Ocenjena varijansa za $\ln OR$ je data sa:

$$\widehat{Var}\left(\ln\left(\widehat{OR}(Pol = 1, Pol = 0, Atraktivna = a)\right)\right) = 0.285 + a^2 0.917 + 2a * (-1.179)$$

Vrednosti za ocenjeni OR i 95% interval poverenja su dati u Tabeli 20.

Atraktivna	0	1
OR	0.093	0.686
95% I.P.	(0.699,5.669)	(0.231,5.216)

Tabela 20

Odnosno važi da će muškarci koji nisu atraktivni imati 0.093 puta veću šansu da budu osuđeni u odnosu na žene koje nisu atraktivne, takođe važi da će atraktivni muškarci imati 0.686 puta veću šansu da budu osuđeni u odnosu na atraktivne žene.

5 METODE I POSTUPCI ZA GRAĐENJE MODELA U LOGISTIČKOJ REGRESIJI

U prethodnim poglavljima smo se bavili ocenom, testiranjem i interpretacijom koeficijenata logističkog regresiong modela. Primeri koje smo koristili su uglavnom imali samo nekoliko nezavisnih promenljivih, međutim u praksi se često javljaju situacije kada imamo više desetina nezavisnih promenljivih koje mogu biti uključene u model. Zato treba da odredimo odgovarajuće metode i postupke kako bismo rešili ovakve probleme.

Cilj bilo koje metode je izbor onih promenljivih koje daju 'bolji' model za naučni kontekst problema. Radi postizanja ovog cilja moramo imati:

- Plan za izbor promenljivih u model
- Metode za dobijanje modela koji je odgovarajući kako za pojedinačne promenljive, tako i za grupu promenljivih.

5.1 IZBOR PROMENLJIVIH

Kriterijum za uključivanje promenljivih u model može varirati od jednog problema do drugog i od jedne naučne discipline do druge. Građenju statističkog modela uključuje težnju ka modelu sa što manjim brojem promenljivih koji ipak objašnjava podatke. Obrazloženje za minimiziranje broja promenljivih u modelu je da će rezultujući model najverovatnije biti numerički stabilniji, i da će se lakše generalizovati. Ukoliko je više promenljivih uključeno u model, ocene standardne greške postaju veće, i model postaje više zavisan od registrovanih podataka. Postoji nekoliko koraka koje možemo pratiti kao pomoć pri izboru promenljivih za logistički regresioni model. Postupak za izbor modela je prilično sličan onom korišćenom u linearnom regresiji [29].

1. Postupak za izbor promenljivih trebao bi početi univarijabilnom analizom svake promenljive.
   a) Za kategorijalne i neprekidne promenljive sa nekoliko celobrojnih vrednosti trebalo bi analizirati tabelu kontigencije koja sadrži ishod ($y = 1, y = 0$) i k nivoa nezavisne promenljive. Hi-kvadrat test količnika verodostojnosti sa $k - 1$ stepeni slobode je jednak vrednosti testa količnika verodostojnosti za značajnost koeficijenata $k - 1$ dizajn promenljivih u univarijabilnom logističkom regresionom modelu koji sadrži tu jednu nezavisnu promenljivu.

Takođe je potrebno oceniti pojedinačni odnos šansi (zajedno sa granicama poverenja) za promenljive koje pokazuju umeren nivo povezanosti koristeći jedan od nivoa kao referentnu kategoriju.

b) Za neprekidne promenljive analiza bi trebala da obuhvati: fitovanje univarijablinog modela radi dobijanja ocenjenog koeficijenta, standardne greške, testa količnika verodostojnosti za značajnost koeficijent kao i univarijabilne Wald statistike.

2. Nakon univarijabilne analize, biramo promenljive za multivarijabilnu analizu. Bilo koja promenljiva iz univarijabilnog testa koja ima p-vrednost manju od npr. 0.25 je kandidat za multivarijabilni model zajedno sa svim promenljivama za koje se zna da su klinički zanačajne.

Problem sa univarijabilnim pristupom je što on ignoriše mogućnost da skup promenljivih, od kojih je svaka slabo povezana sa rezultatom, može postati važan prediktor rezultata kada ih uzimamo zajedno. Ukoliko je to mogućnost, tada bi mogli izabrati nivo značajnosti dovoljno velik da dopusti i takvim promenljivim da postanu kandidati za uključivanje u multivarijabilni model. Tehnika *'izbora najboljeg podskupa'* je jedna strategija za građenje efikasnog modela za identifikaciju skupova promenljivih koje imaju ovaj tip povezanosti sa rezultujućom promenljivom.

Još jedan pristup izboru promenljivih je korišćenje metode '*korak po korak*' u kojoj se promenljive koje su izabrane uključuju odnosno isključuju iz model u nizu koraka zasnovanih na odgovarajućim statističkim kriterijumima. O ovoj metodi će biti reči u narednom odeljku.

3. Nakon formiranja multivarijabilnog modela, potrebno je verifikovati značajnost svake promenljive uključene u model, a to treba da uključi:

a) Ispitivanje Wald statistike za svaku promenljivu.

b) Upoređivanje svakog ocenjenog koeficijenta sa koeficijentom iz modela koji sadrži samo tu promenljivu.

c) Promenljive koje ne doprinose modelu, a koje su bazirane na ovim kriterijumima treba eliminisati. Novi model zatim treba uporediti sa starijim, većim modelom koristeći test količnika verodostojnosti. Takođe, ocenjeni koeficijenti za promenljive koje ostaju bi trebalo uporediti sa onima iz potpunog modela. Zapravo, trebalo bi da posmatramo one promenljive kod

kojih se veličina koeficijenta značajno menja. Ovo ukazuje da je jedna ili više isključenih promenljivih bila važna u smislu obezbeđivanja potrebnih podešavanja efekta promenljive koja je ostala u modelu. Ovaj postupak brisanja, popravke, i verifikacije se nastavlja, dok se ne pokaže da sve promenljive koje su uključene u model i one koje su isključene su ili klinički, ili statistički nevažne.

d) Na kraju treba bilo koja promenljivu koja nije izabrana za orginalni multivarijabilni model dodati ponovo u model. Ovaj korak može biti koristan u identifikaciji promenljivih koje same za sebe, nisu značajno povezane sa rezultatom, ali daje važan doprinos u prisustvu ostalih promenljivih.

Model na kraju koraka nazivamo *Preliminarni model glavnih efekata*

4. Kada je dobijen model koji sadrži sve promenljive koje su od suštinskog značaja, trebalo bi pažljivije posmatrati promenljive u modelu. Za neprekidne promenljive potrebno je proveriti pretpostavku o linearnosti za logit. Ako logit nije linearan, potrebno je izvršiti odgovarajuću transformaciju promenljive, tako da logit bude linearniji u novoj promenljivoj. Neke od transformacija su npr: metod dizajn promeljivih i metod frakcionih polinoma, ali se ovde nećemo zadržavati na njima. Model sada nazivamo *Model glavnih efekata*

5. Kada imamo *Model glavnih efekata* kontrolišemo interakciju između promenljivih u modelu.

a) Kreiramo listu mogućih parova promenljivih u *Model glavnih efekata* koje imaju naučnu osnovu da budu u interakciji jedna sa drugom. Lista ne mora da sadrži sve promenljive u modelu.

b) Dodajemo interakcione promenljive, jednu po jednu u model koji sadrži sve glavne efekte, i ocenjujemo njihovu značajnost koristeći test količnika verodostojnosti.

c) Dodajemo značajne interakcije *Modelu glavnih efekata* i verifikujemo njihovu značajnost putem Wald testa i testa količnika verodostojnosti za značajnost koeficijenata. Na kraju iz modela izbacujemo neodgovarajuće interakcije.

Model sada nazivamo *Prvi konačni model*.

5.2 LOGISTIČKA REGRESIIJA "KORAK PO KORAK"

Izbor promenljivih korak po korak (engl. *Stepwise Logistic Regression*) je široko rasprostranjen u linearnoj regresiji. Upotreba ovog postupka izbora može da obezbedi brz i efektivan način za proveru, kontrolu velikog broja promenljivih i fitovanja više logističkih regresionih jednačina istovremeno.

Bilo koji postupak korak po korak za izbor ili eliminisanje promenljivih iz modela je bazirana na statističkom algoritmu koji proverava značajnost promenljivih, te ih uključuje ili isključuje iz modela na osnovu utvrđenog pravila odlučivanja. Značajnost promenljive je definisana pomoću statističke značajnosti njenog koeficijenta. Statistika koja je korišćena zavisi od pretpostavki modela. U linearnoj regresiji korak po korak je korišćen F test, jer je pretpostavka da greške imaju normalnu raspodelu. U logističkoj regresiji pretpostavka je da greške imaju binomnu raspodelu, a značajnost je ocenjena putem hi- kvadrat testa količnika verodostojnosti. Dakle, u bilo kom koraku procedure, najvažnija promenljiva, u statističkim terminima je ona koja prouzrokuje najveću promenu u logaritmu verodostojnosti za model koji sadrži promenljivu u odnosu na onaj koji ne sadrži promenljivu (to jest onaj koji bi trebalo rezultirati najvećom statistikom količnika verodostojnosti, *G*).

Opisaćemo i ilustrovati algoritam za izbor "unapred" sa testom za eliminaciju "unazad" u logističkom postupku korak po korak. Svaka ostala varijanta ovog algoritma je predstavlja samo različitu modifikaciju ove procedure [28],[27].

Korak (0):

Pretpostavimo da imamo na raspolaganju ukupno p mogućih nezavisnih promenljivih, gde je za svaku utvrđeno da li je od moguće "kliničke" značajnosti u analiziranju izlazne promenljive.

- Počinjemo fitovanjem modela koji sadrži samo odsečak, odnosno β_0, i izračunavanjem njegovog logaritma verodostojnosti L_0.
- Zatim se fituje svaki od p mogućih univarijabilnih logističkih regresionih modela i upoređuju se njihovi odgovarajući logaritmi verodostojnosti.
- Biramo promenljivu sa najmanjom p-vrednosti testa količnika verodostojnosti, x_1
- Ako je p-vrednost za G_1 manja od α prelazimo na korak (1), inače stajemo.

Odlučujući aspekt u korišćenju logističke regresije korak po korak je izbor "alfa" nivoa za procenu važnosti promenljive. On određuje koliko je promenljivih konačno uključeno u model.. Rezultati dosadašnjih istraživanja su pokazali da je izbor $\alpha = 0.05$ previše strog, jer često isključuje važne promenljive iz modela, dok se izbor vrednosti α u rangu od 0.15 do

0.20 češće preporučuje. Ponekad cilj analize može biti širi, i traže se modeli koji sadrže više promenljivih, da obezbede kompletniju sliku mogućih modela. U tom slučaju, korišćenje $\alpha = 0.25$ ili čak višeg nivoa, može biri opravdan izbor.

Korak (1)

Počinjemo sa fitovanim logističkim regresionim modelom koji sadrži promenljivu iz koraka (0), x_1, koja ima najmanju p vrednost i određujemo da li je bilo koja od $p-1$ preostalih promenljivih značajna dok je promenljiva x_1 u modelu

- Fitujemo $p-1$ logističkih regresionih modela koji sadrže x_1 i novu nezavisnu promenljivu i upoređuju se njihovi odgovarajući logaritmi verodostojnosti.
- Biramo promenljivu sa najmanjom p-vrednosti testa količnika verodostojnosti, x_2
- Ako je p-vrednost za test količnika verodostojnosti manja od α prelazimo na korak (2), inače stajemo.

Korak (2)

Ovaj korak počinje fitovanjem modela koji sadržii i x_1 i x_2 . Može se desiti da kada u model uđe promenljiva x_2, promenljiva x_1 više nije od značaja za model. Dakle, korak (2) uključuje proveru za "*eliminaciju unazad*".

- Računamo logaritam verodostojnosti kada je promenljiva x_1, odnosno x_2 izbačena iz modela
- Računamo količnik verodostojnosti ovakvog i punog modela i p-vrednost ove statistike
- Da bismo odlučili koju promenljivu i da li je treba izbaciti, biramo onu promenljivu koja kad je eliminisana iz modela, model ima veću p-vrednost statistike testa količnika verodostojnosti.
- Odlučujemo da li ovako izabranu promenljivu treba eliminisati tako što upoređujemo odgovarajuću p-vrednost sa unapred izabranim nivoom α_1. Ako ne želimo da izbacimo previše promenljivih koristimo veću α_1, npr. 0.9, važi i obrnuto.

Kad završimo '*eliminaciju unazad*', prelazimo na izbor *'unapred'*.

- Fitujemo $p-2$ logističkih regresionih modela koji sadrže x_1, x_2 i novu nezavisnu promenljivu i upoređuju se njihovi odgovarajući logaritmi verodostojnosti.
- Biramo promenljivu sa najmanjom p-vrednosti testa količnika verodostojnosti, x_2
- Ako je p-vrednost za test količnika verodostojnosti manja od α prelazimo na korak (3), inače stajemo.

Korak (3)

Procedura za korak (3) je identična onoj u koraku (2). Fituje se model koji uključuje promenljivu izabranu u prethodnom koraku, izvodi se kontrola za eliminaciju '*unazad*' i izbor '*unapred'*. Proces se nastavlja na ovaj način, do poslednjeg koraka, koraka (4).

Korak (4):

Do ovog koraka dolazimo kada je svih p promenljivih uneto u model, ili sve promenljive u modelu imaju p-vrednosti za eliminisanje koje su manje od α i promenljive koje nisu uključene u model imaju p-vrednost za unos koja je veća od α_1. Model u ovom koraku sadrži one promenljive koje su značajne u odnosu na kriterijum za α i α_1. One mogu ili ne moraju biti promenljive koje su prikazane u konačnom modelu. Na primer, ako se izabrane vrednosti za α i α_1 slažu sa našom verovanjem u statističku značajnost, tada model na kraju koraka (4) može sadržati značajne promenljive. Međutim, ako smo koristili vrednosti za α_1 i α koje su manje stroge, tada bi trebali birati promenljive za konačni model iz tabele koja prikazuje rezultate *korak po korak* procedure.

Postoje dve metode koje se mogu koristiti za izbor promenljivih iz tabele; one su uporedive sa metodama koje se uobičajeno koriste u linearnoj regresiji korak po korak. Prvi metod je baziran na p-vrednosti za unos u svakom koraku, dok je drugi metod baziran na testu količnika verodostojnosti za model u tekućem koraku u odnosu na model u poslednjem koraku.

Neka "Y" označava jedan proizvoljan korak u proceduri.

- U prvom metodu upoređujemo p-vrednost u koraku Y-1 sa unapred izabranim nivoom značajnosti, kao recimo $\alpha = 0.05$. Ako je p-vrednost manja od α, tada prelazimo na korak Y, a inače se zaustavljamo. Razmatramo model u prethodnom koraku za dalju analizu. U ovom metodu, kriterijum za unos je baziran na testu značajnosti koeficijenta za x_j koji zavisi od $x_1, x_2, \cdots x_{j-1}$ koji su u modelu. Broj stepeni slobode za test je 1 ili $k - 1$ što zavisi od toga da li je x_j neprekidna ili polihotomna promenljiva sa k kategorija.
- U drugom metodu, upoređujemo model u tekućem koraku, koraku Y, ne sa modelom iz prethodnog koraka, koraka Y – 1, već sa modelom u poslednjem koraku, koraku (4). Izračunavamo p-vrednost za test količnika verodostojnosti za ova dva modela i nastavljamo na ovaj način sve dok je p-vrednost $\geq \alpha$. Ovde se testira da li su svi koeficijenti za promenljive koje su dodate u model iz koraka Y do koraka (4) jednaki nuli. U bilo kom datom koraku on ima više stepeni slobode nego test koji je

upotrebljen u prvoj metodi. Iz tog razloga drugi metod može da izabere veći broj promenljivih nego prvi metod.

Dobro je poznato da p-vrednosti izračunate u proceduri izbora korak po korak nisu p -vrednosti u kontekstu tradicionalnog testiranja hipoteza. Umesto toga, one imaju značenje indikatora relativnog značaja među promenljivima.

Primer 10

Posmatrajmo Tabelu 21 na cd-u u prilogu. U tabeli su dati podaci pomoću kojih treba da predvidimo da li će seksualno uznemiravanje žena biti prijavljeno ili ne. Zavisna promenljiva je *prijavljeno* koja je kodirana sa 0-nije prijavljeno, odnosno sa 1 ako jeste. Nezavisne promenljive su: *godine*, *bračni status* kodiran sa 1-osoba je udata, 2-osoba nije udata, *učestalost* kodirana od 0-4, *nivo* seksualnog uznemiravanja kodiran od 0-7, *feministkinja*, što je veći broj, žena ima više izražene feminističke stavove. U ovom primeru ćemo pokazati kako izgleda metoda '*korak po korak'*.

Izabraćemo da radimo sa $\alpha = 0.7$ i $\alpha_1 = 0.9$. Počinjemo od koraka (0).

Prvo računamo logaritam verodostojnosti za model koji sadrži samo odsečak i dobijamo da je on jednak:

$$L_0 = (n_1 \ln(n_1) + n_0 \ln(n_0) - n\ln(n)) =$$
$$= (174 \ln 174 + 169 \ln 169 - 343 \ln 343) = -237.713$$

Zatim računamo logiratam verodostojnosti kada je npr. promenljiva *nivo* uneta u model i dobijamo da je tada logoritam verodostojnosti jednak -221.046.

Sledeći korak je izračunavanje G statistike i odgovarajuće p-vrednosti.

$$G = 2\big(-221.046 - (-237.713)\big) = 33.33$$

Važi da je $P\{\chi_7^2 > 33.33\} = 0.00002 < \alpha = 0.7$. Iz ovoga sledi da promenljivu *nivo* smatramo statistički značajnom.

Isto radimo i sa preostale tri promenljive i dobijamo njihove p-vrednosti, date u Tabeli 22.

promenljive	p-vrednost
godine	0.27961
bračni_status	0.79123
učestalost	0.98814
nivo	0.00002
feministkinja	0.03738

Tabela 22

Kao što vidimo iz tabele promenljiva *bračni_status* i *učestalost* ne igraju veliku ulogu u modelu, pošto je njihova p-vrednost veća od unapred zadatog α i zato ćemo ih eliminisati iz modela.

Kako vidimo najmanju p-vrednost ima promenljiva *nivo* tako da korak(1) počinjemo sa fitovanim logističkim regresionim modelom koji sadrži ovu promenljivu.

Računamo logoritam verodostojnosti modela koji sadrži promenljivu *nivo* i neku od preostalih promenljivih, npr. *godine*. Logaritam ovog novog modela je: -220.018. Računamo G statistiku ova dva modela i odgovarajuću p-vrednost i dobijamo:

$$G = 2(-220.018 - (-221.046)) = 1.732$$

A odgovarajuća p-vrednost je: $P\{\chi_1^2 > 1.732\} = 0.18815 < \alpha = 0.7$

Isti postupak ponavljamo i za promenljivu *feministkinja* i dobijamo da je G statistika u ovom slučaju jednaka 0.395, dok je p-vrednost 0.53 što je takođe manje od nivoa α.

Korak(2) počinjemo sa fitovanim logističkim modelom koji ima manju p-vrednost, u našem slučaju je to model koji sadrži promenljive *nivo* i *godine.* Prvo računamo logaritam verodostojnosti kada je promenljiva $nivo$ odnosno $godine$ izbačena iz modela, pa zatim i količnik verodostojnosti ovakvog i punog modela i p-vrednost ove statistike. Logoritam verodostojnosti kada je samo promenljiva *nivo* u modelu je -221.046, odnosno kada je promenljiva *godine* u modelu je: -237.128. Računamo odgovarajuće G statistike i p-vrednosti i dobijamo:

$$G = 2(-220.018 - (-221.046)) = 1.732$$

koja ima p-vrednost 0.18815

Odnosno za promenljivu *godine* dobijamo:

$$G = 2(-220.018 - (-237.128)) = 34.221$$

koja ima p −vrednost 0.00000.

Kako je p-vrednost G statistike veća kada promenljiva *nivo* nije u modelu, tu promenljivu bismo mogli ipak eliminisati, međutim njena p-vrednost je manja od nivoa $\alpha_1 = 0.9$ iz čega sledi da promenljivu *nivo* ostavljamo u modelu. Dalje nastavljamo kao i u koraku (1), ubacujemo promenljivu feministkinja u ovaj novi model, i računamo p-vrednost odgovarajuće G statistike, odnosno:

$$G = 2\big(-220.087 - (-220.180)\big) = 0.186$$

A njena p-vrednost je 0.66626 što je manje od α pa prelazimo na korak(3) gde proveravamo da li treba eliminisati neku promenljivu, na identičan način kao u koraku(2). Dobijamo da je p-vrednosti kada promenljiva *nivo* nije u modelu jednak 0.00008, odnosno 0.2171 kada

promenljiva godine nisu u modelu i na kraju 0.66626 kada promenljiva *feministkinja* nije u modelu. Odavde sledi da bi iz modela mogli eliminisati promenljivu *feministkinja*, ali to ne radimo jer je p-vrednost manja od zadatog nivoa $\alpha_1 = 0.9$. Na kraju dobijamo model koji sadrži samo značajne promenljive i to: *godine*, *feministkinja* i *nivo.*

6 PROCENA SLAGANJA MODELA SA PODACIMA

Počinjemo razmatranje metoda za procenjivanje slaganja ocenjenog logističkog regresionog modela sa podacima, pretpostavkom da smo zadovoljni našim pokušajima na nivou građenja modela. Odnosno, podrazumevamo da model sadrži one promenljive koje treba da su u modelu, tj. koje su značajne i da su promenljive unete u korektnom funkcionalnom obliku. Sada nas interesuje koliko efikasno naš model opisuje rezultujuću (ishodnu) promenljivu (tzv. *goodness-of-fit*).

Neka su registrovane uzoračke vrednosti rezultujuće promenljive prikazane u vektorskom obliku sa $\boldsymbol{y}' = (y_1, y_2 \cdots y_n)$.. Označićemo fitovane vrednosti, sa $\widehat{\boldsymbol{y}}$, gde je $\widehat{\boldsymbol{y}}' = (\hat{y}_1, \hat{y}_2 \cdots \hat{y}_n)$.

Model je prilagođen podacima ako su:

(1) mere rastojanja između $\boldsymbol{y}$ i $\widehat{\boldsymbol{y}}$ male.

(2) doprinos svakog para,$(y_i, \hat{y}_i)$ $i = 1, 2, 3, \ldots, n$ ovim merama je nesistematski, i mali u odnosu na grešku modela.

Kompletno procenjivanje fitovanog modela obuhvata kako izračunavanje mera rastojanja između $\boldsymbol{y}$ i $\widehat{\boldsymbol{y}}$, tako i ispitivanje pojedinačnih komponenti tih mera.

6.1 OSNOVNE MERE ZA GOODNESS-OF-FIT (*GOF*)

Osnovne mere za *goodness-of-fit* predstavljaju opšti pokazatelj koliko dobro se model slaže sa podacima ali ne govori o tome da li je dati model bolji od nekog drugog modela. U mnogim epidemiološkim analizama cilj je da se nađe najbolji mogući model koji opisuje odnos između izloženosti određenim faktorima i bolesti. Tako da te anlize češće koriste strategije koje porede više različitih modela a ne *GOF* [5].

Statistike *GOF* ne moraju da daju informaciju o pojedinim komponentama modela. Mala vrednost neke od tih statistika ne isključuje mogućnost nekih bitnih pa samim tim i interesantnih odstupanja od vrednosti dobijenih na osnovu fitovanog modela za nekoliko subjekata. Sa druge strane, velika vrednost neke od tih statistika jasno ukazuje na stvarne probleme modela.

Pre razmatranja specifične *GOF* statistike, moramo prvo razmotriti efekat koji fitovan model ima na stepene slobode koji su dostupni za procenu učinka modela. Koristićemo izraz *kovarijatni obrazac* za opisivanje odabranog skupa vrednosti za kovarijate u modelu.

Kovarijatni obrazac predstavlja opservacije sa istim vrednostima za sve nezavisne promenljive[5]. Na primer, ako imamo dve nezavisne promenljive koje označavaju pol i rasu, obe sa 2 kategorije, tada imamo četiri kovarijatna obrasca. Ako imamo npr. još jednu promenljivu koja predstavlja težinu, i posmatramo n-subjekata, tada možemo imati najviše n kovarijatnih obrazaca (tačno n obrazaca imamo ako je težina kod svakog subjekta različita) pošto je težina neprekidna nezavisna promenljiva. Tokom razvijanja modela nije neophodno baviti se brojem kovarijatnih obrazaca. Stepeni slobode za testove su bazirani na razlici u broju parametara za modele koji se upoređuju, a ne na broju kovarijatnih obrazaca. Međutim, kada je procenjeno koliko se model slaže sa podacima, tada sporno pitanje može biti broj kovarijatnih obrazaca.

GOF je procenjen preko grupisanja fitovanih vrednosti određenih pomoću kovarijati u modelu, a ne ukupnog skupa kovarijati. Na primer, pretpostavimo da naš fitovan model sadrži p nezavisnih promenljivih, $\boldsymbol{x}' = (x_1, x_2, \cdots x_p)$ i neka J označava broj različitih vrednosti za registrovano $\boldsymbol{x}$. Ako neki subjekti imaju istu vrednost za $\boldsymbol{x}$, tada je $J < n$[6].

Označimo broj subjekata za koje je $\boldsymbol{x} = \boldsymbol{x_j}$, sa m_j , za $j = 1, 2, 3, \ldots, J,$ odnosno važi da je $\sum m_j = n$. Neka je sa y_j označen broj pozitivnih odgovora, $y = 1$, među m_j subjekata za koje važi $\boldsymbol{x} = \boldsymbol{x_j}$ i neka važi da je $\sum y_j = n_1$.

Raspodela za statistiku *GOF* se dobija, ako pustimo da n bude dovoljno veliko. Ako se broj kovarijatnih obrazaca takođe povećava sa n, tada svaka vrednost m_j teži da bude mala. Za raspodele dobijene pod pretpostavkom da samo n postaje veliko kažemo da su *n-asimptotski*. Ako fiksiramo broj grupa, J, i povećavamo obim uzorka onda će se povećavati broj elemenata u svakoj grupi tj. ako fiksiramo $J < n$ i pustimo n da je dovoljno veliko, tada svaka vrednost m_j takođe teži da postane velika. Za raspodele gde svako m_j postaje veliko, kažemo da su m - *asimptotske*. Slučaj koji se najčešće javlja u praksi je $J \approx n$, kao što i očekujemo kad god postoji bar jedna neprekidna kovarijata u modelu i predstavlja najveći izazov u razvijanju raspodela *GOF* statistike[6].

6.1.1 PIRSONOVA HI-KVADRAT STATISTIKA I ODSTUPANJE

U linearnoj regresiji osnovne mere za procenu slaganja modela sa podacima su funkcije reziduala i definisane su kao razllika između observirane i fitovane vrednosti. Međutim u logističkoj regresiji postoji nekoliko mogućih načina za procenu razlike između ove dve vrednosti. Za isticanje činjenice da su fitovane vrednosti u logističkoj regresiji izračunate za svaki kovarijatni obrazac i da zavise od ocenjene verovatnoće za taj kovarijatni obrazac, označavamo vrednost za j-ti kovarijatni obrazac sa $\hat{y}_j$, i važi da je:

$$\hat{y}_j = m_j\hat{\pi}_j = \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}}$$

gde je $\hat{g}(x_j)$ ocenjen logit.

Počinjemo razmatranjem dve mere rastojanja između registrovane i predviđene vrednosti na osnovu modela, a to su: *Pirsonov rezidual* i *rezidual odstupanja*. Za određen kovarijatni obrazac, Pirsonov rezidual je definisan na sledeći način:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

Statistika koja je bazirana na ovim rezidualima je *Pirsonova hi-kvadrat statistika*:

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

Rezidual odstupanja je definisan kao:

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

Važi:

$$d(y_j, \hat{\pi}_j) = \begin{cases} -\sqrt{2m_j|\ln 1-\hat{\pi}_j|} & , y_j = 0 \\ \sqrt{2m_j|\ln \hat{\pi}_j|} & , y_j = m_j \end{cases}$$

Statistika koja je bazirana na rezidualima odstupanja se naziva *odstupanje* (*Deviance*)

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

Pod pretpostavkom da je fitovani model korektan za sve aspekte, statistike χ^2 i D imaju hi-kvadrat raspodelu sa $J - (p + 1)$ stepeni slobode. Kada je u pitanje *odstupanje* ova izjava sledi iz činjenice da je D test statistika količnika verodostojnosti zasićenog modela sa

J parametara u odnosu na fitovani model sa $p + 1$ parametara. Slična teorija daje nultu raspodelu za χ^2. Problem nastaje kada je $J \approx n$, jer je raspodela n-asimptotska, pa se broj parametara povećava u istom odnosu kao veličina uzorka. Dakle, p-vrednosti, izračunate za ove dve statistike kada je $J \approx n$, a korišćenjem χ^2_{J-p-1} raspodele su nekorektne[8].

Jedan način da se izbegnu navedene smetnje sa raspodelama za χ^2 i D, kada je $J \approx n$ je grupisanje podataka na takav način da se koristi m-asimptotska raspodela. Da bi se razumelo obrazloženje za različite postupke grupisanja, korisno je smatrati χ^2 Pirsonovom i D kao logaritam verodostojnosti hi-kvadrat statistike koja se dobija iz tabele $2 \times J$. Redovi tabele odgovaraju vrednostima rezultujuće promenljive, $y = 1, 0$, a J kolona odgovara J mogućim vrednostima kovarijatnog obrazca. Ocena očekivanih vrednosti pod pretpostavkom da je logistički model u stvari korektan model za ćelije koje odgovaraju $y = 1$ redu i j-toj koloni je $m_j\hat{\pi}_j$. Sledi da je ocena očekivanih vrednosti za ćeliju koja odgovara $y = 0$ i j-toj koloni $m_j(1 - \hat{\pi}_j)$ [6]

Kada su hi-kvadrat testovi izračunati iz tabele kontigencije, p-vrednosti su korektne pod nultom hipotezom da su ocenjene vrednosti suviše "velike" u svakoj ćeliji. Iako ovo pojednostavljuje situaciju, ipak je korektno. U gore opisanoj tabeli $2 \times J$, očekivane vrednosti su uvek prilično male jer se broj kolona povećava kako se n povećava. Da bi se izbegao ovaj problem, možemo smanjiti kolone u fiksiran broj grupa, g, i tada računati registrovane i očekivane frekvencije. Fiksiranjem broja kolona, ocenjene očekivane vrednosti postaju veće, sa povećanjem n[8].

Prednosti ovih statistika su između ostalog što se nalaze u skoro svim softverskim statističkim paketima. Još jedna dobra karakteristika je i ta da se statistike relativno lako računaju korišćenjem elementarnih kalkulacija kao i njihova odgovarajuća p-vrednost. Često se dešava da ove dve statistike imaju različite vrednosti, ako su te razlike jako velike smatramo da χ^2 aproksimacija ovih raspodela nije odgovarajuća [6].

6.1.2 HOSMER-LEMESHOW TEST(*HL* TEST)

Da bi izbegli problematičnu upotrebu odstupanja i da bi obezbedili značajne testove za pristup *GOF*, Hosmer i Lemeshow (1980) i Lemeshow i Hosmer (1982) su predložili grupisanje bazirano na vrednostima ocenjenih verovatnoća. *HL* statistika je široko rasprostranjena nezavisno od toga da li je broj kovrijatnih obrazaca blizak broju opservacija. Mada, statistika zahteva da model ima najmanje tri kovarijatna obrazca, uprkos tome što imamo i slabe rezultate značajnosti ako model sadrži šest obrazaca, a najbolje rezultate pokazuje kada je broj kovarijatnih obrazaca blizu n [10].

Pretpostavimo, u cilju razmatranja, da je $J = n$. U tom slučaju imamo n kolona koje odgovaraju vrednostima ocenjenih verovatnoća, sa prvom kolonom kojoj odgovara najmanja vrednost, i n-tom kolonom sa najvećom vrednosti. Predložena su dva postupka grupisanja i to formiranjem tabele zasnovane na:

- percentilima ocenjenih verovatnoća
- fiksiranim vrednostima ocenjenih verovatnoća

U prvoj metodi, koristi se $g = 10$ grupa, pa tako prva grupa sadrži $n'_1 = n/10$ subjekata koji imaju najmanje ocenjene verovatnoće, dok poslednja grupa sadrži $n'_{10} = n/10$ subjekata sa najvećim ocenjenim verovatnoćama. U drugoj metodi, koristi se $g = 10$ grupa koje sadrže sve subjekte sa ocenjenim verovatnoćama između susednih *nivoa odlučivanja* (definisani su kao vrednosti $k/10, k = 1,2,\ldots,9$). Na primer, prva grupa sadrži sve subjekte čije su ocenjene vrednosti manje ili jednake 0.1, dok deseta grupa sadrži one subjekte čije su ocenjene vrednosti veće od 0.9. Za $y = 1$, ocene očekivanih vrednosti su dobijene sumiranjem ocenjenih verovatnoća za sve subjekte u grupi. Za $y = 0$, ocenjene očekivane vrednosti su dobijene sumiranjem za sve subjekte u grupi jedan minus ocenjene verovatnoće.

Bez obzira koji postupak grupisanja je u pitanju, *HL GOF* statistika $\hat{C}$ je dobijena računanjem Pirsonove hi-kvadrat statistike iz tabele $g \times 2$ sa observiranim i ocenjenim očekivanim frekvencijama.

Statistika $\hat{C}$ je definisana na sledeći način:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

gde je n'_k ukupan broj subjekata u k-toj grupi. Neka c_k označava broj kovarijatnih obrazaca u k - tom decilu, tada važi da je:

$$o_k = \sum_{j=1}^{c_k} y_j$$

broj jedinica među c_k kovarijatnih obrazaca. Još važi da je:

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k}$$

prosečna ocenjena verovatnoća[13].

Hosmer i Lemeshow su pokazali da je za $J = n$ (kao i za $J \approx n$) i kada je fitovani logistički regresioni model korektan model, raspodela $\hat{C}$ statistike dobro aproksimirana sa χ^2_{g-2} raspodelom.

Dodatna istraživanja koja su vršili Hosmer, Lemeshow, i Klar (1988) su pokazala da metod grupisanja baziran na percentilima ocenjenih verovatnoća ima prednost nad onima koji su bazirani na fiksiranim nivoima odlučivanja u smislu boljeg slaganja sa χ^2_{g-2} raspodelom, naročito kada je mnogo ocenjenih vrednosti male vrednosti (to jest, manje od 0.2). Ukoliko nije posebno naglašeno, podrazumeva se da je $\hat{C}$ bazirano na percentilnom tipu grupisanja, obično sa $g = 10$ grupa. Često se za ove grupe koristi termin "*decili rizika*", koji potiče iz zdravstvenih naučnih istraživanja gde rezultat $y = 1$ često predstavlja prisustvo nekog oboljenja [6], [11].

Kako raspodela statistike $\hat{C}$ zavisi od m-asimptotske raspodele, prikladnost p-vrednosti zavisi od validnosti pretpostavki da su ocenjene očekivane frekvencije velike. Smatramo da se model dobro slaže sa podacimo ako je $p-$vrednost odgovarajuće χ^2_{g-2} statistike veća od 0.05.

Primer 11

Posmatrajmo Tabelu 23 na cd-u u prilogu. Pretpostavimo da je model fitovan. Model ispituje koji faktori utiču na premor predavača. Model sadrži zavisnu promenljivu *Premor* koja je kodirana sa 0-osoba nije premorena, 1-osoba je premorena. I nezavisne promenljive: *Os.Kon*-predstavlja koliko predavač ima osećaj kontrole nad studentima (veće vrednosti - manji osećaj kontrole), *Stres* - označava kako se osoba nosi sa stresom uopšteno (veće vrednosti-osoba se slabije nosi sa stresom), *Predavanje* - označava koliko je predavaču stersno predavanje (veće vrednosti - osobi je predavanje stresnije), *Istraživanje* - označava koliko je predavaču stersno istarživanje (veće vrednosti - osobi je istraživanje sve stresnije), *Mentorstvo* - označava koliko je predavaču stersno mentorstvo (veće vrednosti - osobi je

mentorstvo stresnije). Rezultati Hosmer-Lemeshow-og testa su dobijeni korišćenjemi statističkog paketa SPSS 17 i prikazani su u Tabeli 24 i 25.

Hi-kvadart	Stepeni slobode	p
12.399	8	0.134

Tabela 24

	Tabela kontigencije za Hosmer-Lemeshow test				
	Premor = nije premoren		Premor = premoren		Ukupno u grupi
	posmatrano	očekivano	posmatrano	očekivano	
1	47	46.346	0	0.654	47
2	46	45.752	1	1.248	47
3	47	45.016	0	1.984	47
4	45	43.869	2	3.131	47
5	43	42.547	4	4.453	47
6	40	40.210	7	6.790	47
7	29	35.511	18	11.489	47
8	25	28.332	22	18.668	47
9	23	17.089	24	29.911	47
10	3	3.326	41	40.674	44

Tabela 25

Vidimo da p vrednost χ_8^2 statistike iznosi 0.134, tako da zaključujemo da se model dobro slaže sa podacima, odnosno da se posmatrane i očekivane frenkvencije ne razlikuju značajno što vidimo i iz Tabele 25.

6.1.3 TABELE KLASIFIKACIJE

Jedan od načina za sažimanje rezultata fitovanog logističkog regresionog modela je pomoću tabele klasifikacije, koja je rezulat ukrštanja rezultujuće promenljive sa dihotomnom promenljivom čije su vrednosti izvedene iz ocenjenih logističkih verovatnoća.

Da bismo kreirali tabelu klasifikacije 2×2 predviđenih vrednosti iz našeg modela, za ishodnu promenljivu nasuprot tačnoj vrednosti ishodne promenljive, moramo prvo definisati nivo odlučivanja c sa kojim ćemo porediti svaku ocenjenu verovatnoću. Odnosno uzećemo da važi da je $\hat{y} = 1$ ukoliko je $\hat{\pi}_\iota > c$ tj. $\hat{y} = 0$ ukoliko je $\hat{\pi}_\iota \leq c$. Najčešće korišćena vrednost je $c = 0.5$. Još dva bitna pojma za tabele klasifikacije su[19]:

- *Senzitivnost* testa predstavlja verovatnoću da je predviđena vrednost zavisne promenljive jedan, ukoliko je zaista zavisna promenljiva primila vrednost jedan tj. $P(\hat{y} = 1|y = 1)$.

- *Specifičnost* testa je verovatnoća da je predviđena vrednost zavisne promenljive nula, ako je njena stvarna vrednost nula tj. $P(\hat{y} = 0|y = 0)$.

U ovom pristupu, ocenjene verovatnoće se koriste za predviđanje grupe članova. Moguće je da ukoliko model predviđa tačno grupu članova prema nekom kriterijumu, da se klasifikacijom želi dokazati da je model fitovan. Nažalost, ovo može ali i ne mora biti slučaj, jer postoje situacije gde je logistički regresioni model u stvari korektan model, i dakle fitovan, ali da je klasifikacija loša.

Primer 12

Posmatrajmo Tabelu 23 na cd-u u Prilogu. U statističkom paket SPSS smo izračunali tabelu klasifikacije za dati primer i ona je pokazana u Tabeli 26.

	registrovano		
klasifikovano	premor = 0	premor = 1	ukupno
premor = 0	324	24	348
premor = 1	55	64	119
ukupno	379	88	467

Tabela 26

Iz tabele vidimo da je ukupno posmatrano 467 osoba, od kojih 88 osećaju premor. Od njih 88 mi smo dobro klasifikovali njih 64, dok je 24 osobe pogrešno klasifikovano. Od 379 osobe koje ne osecaju premor njih 324 smo dobro klasifikovali, dok je 55 osoba pogrešno klasifikovano.

Senzitivnost testa je:

$$P\{\widehat{premor} = 1|premor = 1\} = \frac{64}{88} = 72.7\%$$

Specifičnost testa je :

$$P\{\widehat{premor} = 0|premor = 0\} = \frac{324}{379} = 85.4\%$$

Dakle, tačno smo klasifikovali 72.7% osoba koje osećaju premor i 85.4% osoba koje ne osećaju premor pa je ukupna stopa tačne klasifikacije:

$$\frac{64 + 324}{467} = 83.1\%$$

Dok je pogrešno klasifikovano:

$$\frac{55 + 24}{467} = 16.9\%$$

posmatranih osoba.

Klasifikacija je osetljiva na relativnu veličinu dve komponente grupe i uvek favorizuje klasifikaciju u veće grupe, što takođe ne zavisi od prilagođenosti modela podacima.

Važan razlog zašto mere izvedene iz tabele klasifikacije 2×2 (kao što su senzitivnost i specifičnost) ne bismo trebali koristiti za procenu koliko je model dobar, je taj da one dosta zavise od raspodele verovatnoća u uzorku.

Zbog razmatranja koje sledi treba da razumemo smisao verovatnoće, a to je da se od n subjekata koji imaju istu verovatnoću ishoda koji nas interesuje, $\hat{\pi}$, očekuje se da će broj onih koji će imati ishod od interesa biti $n\hat{\pi}$, a broj onih za za koje se očekuje da neće imati ishod od interesa je $n(1-\hat{\pi})$. Pretpostavimo da je korišćen nivo odlučivanja $c = 0.50$ u cilju klasifikacije i pretpostavimo da je 100 subjekata imalo verovatnoću $\hat{\pi} = 0.51$. Za sve ove subjekte je predviđeno da će imati rezultat koji se posmatra, ali pretpostavljajući da je model dobro podešen, 51 subjekat bi trebalo da zaista ima ishod od interesa, dok se za njih 49 treba očekivati da neće imati ishod od interesa. Dakle, 49 od 100 pacijenata je pogrešno klasifikovano.

Ne mogu se upoređivati modeli na bazi mera izvedenih iz tabele klasifikacije 2×2, jer ove mere ne možemo posmatrati nezavisno od raspodela verovatnoća u uzorcima na kojima su bazirani. Isti model procenjen u dve populacije, korišćenjem mera senzitivnosti ili specifičnosti bi mogao da da vrlo različite utiske o njegovom učinku.

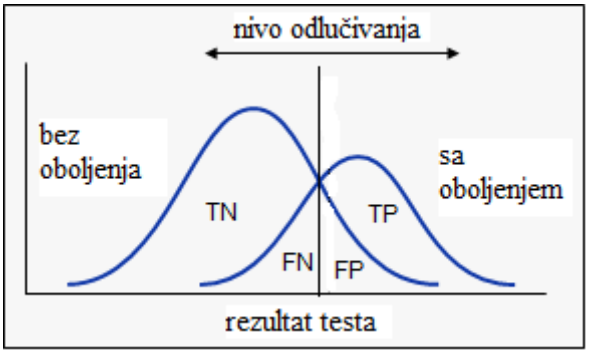
Ukratko, tabela klasifikacije je najprikladnija kada je klasifikacija postavljena kao cilj analize, inače bi trebala da bude samo dopuna mnogo strožijim metodama procene slaganja modela sa podacima.

6.1.4 ROC KRIVA

Receiver Operating Characteristic Curve-ROC kriva je grafička tehnika koja je više od 30 godina veoma popularna posebno u labaratorijskoj medicini. Primena ove tehnike ja započela tokom Drugog svetskog rata za evaluaciju lažno pozitivnih i stvarno pozitivnih signala na ekranu radara. Kasnije je adaptirana od strane radiologa i labaratorijskih naučnika za evaluaciju osetljivosti i specifičnosti medicinskih određivanja pri različitim nivoima odlučivanja.

Kada se senzitivnost i specifičnost testa izračunaju za čitav niz nivoa verovatnoće, nivoa odlučivanja, moguće je konstruisati ROC krivu koja povezuje senzitivnost (verovatnoću tačnog detektovanja prisustva osobine) i 1–specifičnost, (verovatnoću netačnog

detektovanja prisustva osobine). Svaka tačka ROC krive predstavlja uređeni par (senzitivnost, 1-specifičnost) koji odgovara pojedinačnom nivou odlučivanja. Kada razmatramo rezultate određenog testa u dve populacije, npr. jednu populaciju sa oboljenjem, i drugu bez oboljenja, retko ćemo dobiti perfektno razdvajanje između ove dve grupe. Umesto toga raspodela rezultata testa će se preklapati, kao što je prikazano na Slici 7. ROC kriva koje se odlikuje kompletnim razdvajanjem (nema preklapanja raspodele rezultata dve grupe) prolazi kroz gornji levi ugao gde stvarno pozitivni udeo iznosi 1,0 odnosno osetljivost 100%, a lažno pozitivni udeo 0, odnosno 1-specifičnost 100%. Teoretska kriva za test kod koga nema razdvajanja (identična raspodela rezultata dve grupe) je dijagonalna linija od donjeg levog ugla do gornjeg desnog ugla Većina ROC krivih se nalazi između ove dve krajnosti i kvalitativno gledano ona koja je bliža gornjem levom uglu ukazuje na test sa većom tačnošću. Ukoliko je više ROC krivih prikazano na jednom dijagramu ona koja se nalazi iznad i na levo u odnosu na ROC krivu sa kojom se poredi ukazuje na test sa većom posmatranom tačnošću. Relativni položaj dve ili više krivih omogućava kvalitativno poređenje više testova [24].



Slika 7

Za svaku moguću kritičnu vrednost koju smo izabrali da razdvaja dve populacije, postojaće neki slučajevi sa oboljenjem koji su korektno klasifikovani kao pozitivni, (TP = *true positive fraction*), ali će neki slučajevi sa oboljenjem biti klasifikovani kao negativni, to jest *lažno negativni* (FN = *false negative fraction*). Sa druge strane, neki slučajevi bez oboljenja će biti korektno klasifikovani kao negativni (TN = *true negative fraction*), dok će neki slučajevi bez oboljenja biti klasifikovani kao pozitivni, tj. *lažno pozitivni* (FP = *false positive fraction*), što je prikazano u Tabeli 27 [6],[24].

	oboljenje	
test	prisutno	odsutno
pozitivan	tačno pozitivni(TP)	lažno pozitivni(FP)
negativan	lažno negativni(FN)	tačno negativni(TN)

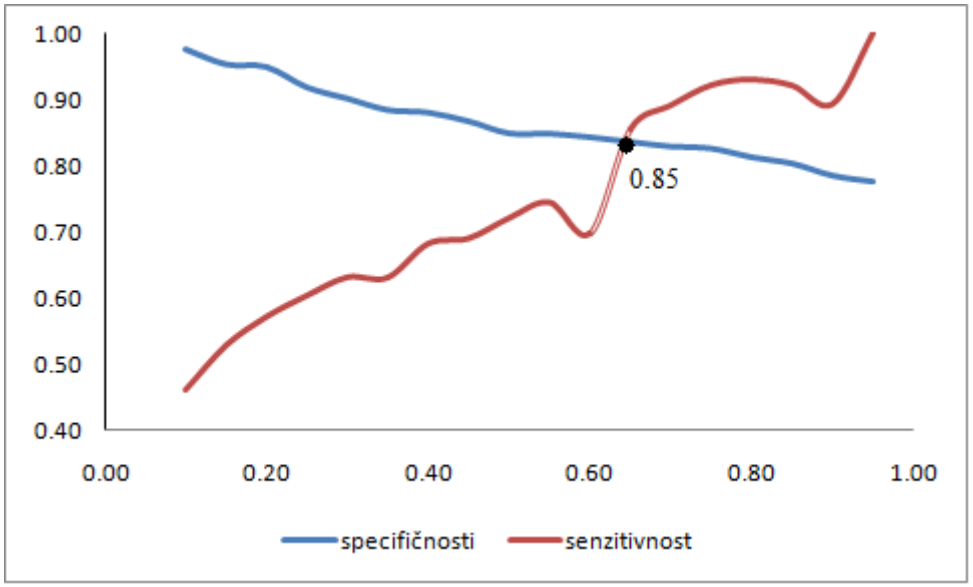
Tabela 27

Prepostavimo da imao model za ocenjivanje verovatnoće iz primera 12. Pravilo koje je prikazano u Tabeli 26, predviđa da će osoba osećati premor, ako je $P(y = 1) \geq 0.50$, odnosno neće osećati premor ako $P(y = 0) < 0.50$. Postoje neke statistički dobre osobine povezane sa korišćenjem $c = 0.5$, ali bi trebali razmatrati i šta se dešava kada koristimo druge vrednosti za *cutpoints* (Tabela 28).

nivo odlučivanja	specifičnost	senzitivnost	1-specifičnost
0.10	0.98	0.46	0.02
0.15	0.95	0.53	0.05
0.20	0.95	0.57	0.05
0.25	0.92	0.60	0.08
0.30	0.90	0.63	0.10
0.35	0.89	0.63	0.11
0.40	0.88	0.68	0.12
0.45	0.87	0.69	0.13
0.50	0.85	0.72	0.15
0.55	0.85	0.74	0.15
0.60	0.84	0.70	0.16
0.65	0.84	0.85	0.16
0.70	0.83	0.89	0.17
0.75	0.83	0.92	0.17
0.80	0.81	0.93	0.19
0.85	0.80	0.92	0.20
0.90	0.79	0.89	0.21
0.95	0.78	1.00	0.22

Tabela 28

Ako je naš cilj izbor optimalnog cutpoint, a u cilju klasifikacije, mogli bismo izabrati onaj za koji je maksimalna i senzitivnost i specifičnost. Na Slici 8 prikazan je primer optimalanog izbora, za nivo odlučivanja, gde se krive senzitivnosti i specifičnosti seku i iznosi $c = 0.85$.



Slika 8

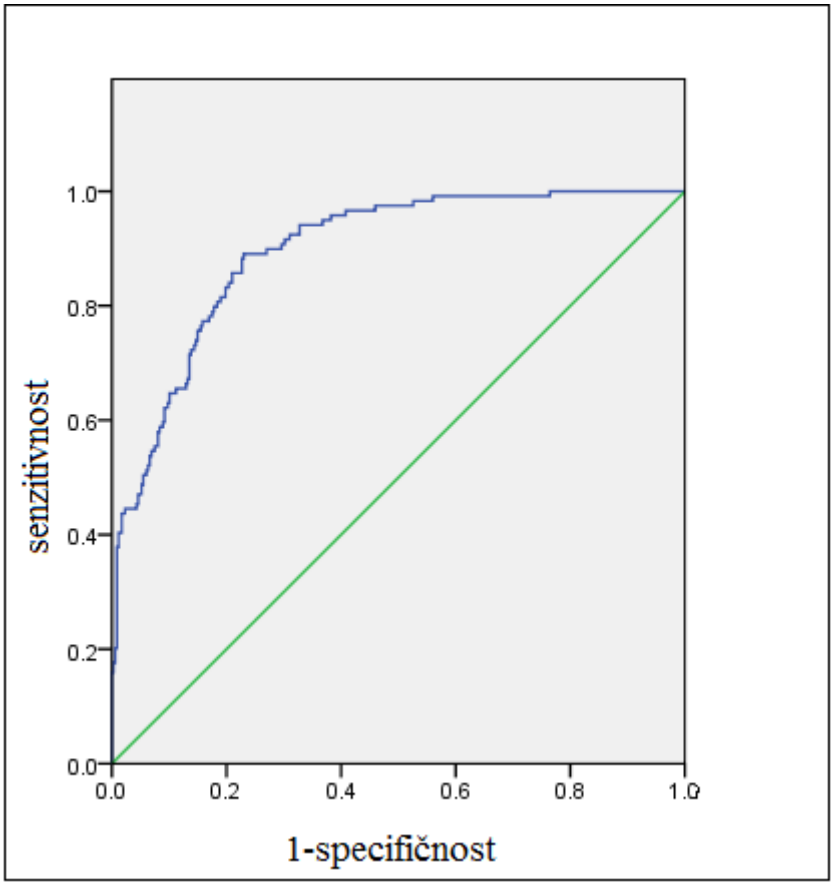
Rezultati korišćenja $c = 0.85$ su prikazani u Tabeli 29, ali ovo se može uraditi za bilo koji mogući izbor cutpoint.

	registrovano		
klasifikovano	premor = 0	premor = 1	ukupno
premor = 0	345	3	348
premor = 1	84	35	119
ukupno	429	38	467

Tabela 29

Ovde je senzitivnost jednaka 0.92, dok je specifičnost 0.80.

Grafikon sezitivnosti u odnosu na 1 – specifičnost za sve moguće nivoe odlučivanja daje, kako što smo već rekli, ROC krivu . ROC krive za naš primer je prikazana na Slici 9. Površina ispod ove krive daje meru razdvajanja koja je, u našem slučaju verovatnoća da će osobe koje osećaju premor imati veću ocenjenu verovatnoću nego oni koji ne osećaju premor.



Slika 9

Površina ispod ROC krive, koja se kreće od nule do jedan, je mera sposobnosti modela u razdvajanju subjekata koji su iskusili događaj koji se posmatra u odnosu na one koji nisu. Površina ispod ROC krive, u oznaci *AUC* (*The Area Under the Curve*), je prihvaćena tradicionalna izvedena mera za ROC krivu[6].

Kao opšte pravilo, koristimo sledeće:

- $AUC = 0.5$ – *nema razdvajanja*
- $0.5 \leq AUC < 0.7$ - *loše razdvajanje*
- $0.7 \leq AUC < 0.8$ - *prihvatljivo razdvajanje*
- $0.8 \leq AUC < 0.9$ - *odlično razdvajanje*
- $AUC \geq 0.9$ - *izvanredno razdvajanje.*

U našem primeru imali smo razdvajanje slučajeva od 0.897 što se smatra *odličnim razdvajanjem.*

7 KONSTRUKCIJA LOGISTIČKOG REGRESIONOG MODELA SA ZAVISNOM PROMENLJIVOM GOJAZNOST

Masovne nezarazne bolesti su vodeći uzrok smrti i nesposobnosti širom sveta. Uzroci visoke učestalosti masovnih nezaraznih bolesti u poslednjim decenijama dvadesetog veka su značajne i brze promene u načinu života savremenih ljudi. Najviše izražene promene su u načinu ishrane, nivou fizičke aktivnosti, povećanoj upotrebi alkohola i duvana. Masovne nezarazne bolesti mogu se sprečiti i kontrolisati njihovim ranim otkrivanjem.

Korišćenje logističke regresije u tom smislu pruža kompletnu sliku o povezanosti zavisne promenljive i nezavisnih promenljivih, pri čemu je ova nezavisna promenljiva kontrolisana za sve druge faktore.

Logistički regresioni model je primenjen na istraživanje povezanosti pojave gojaznosti sa potencijalnim faktorima rizika na populaciji devojaka starosti od 15 do 19 godina. Model će biti testiran da bi se utvrdilo koliko se on dobro slaže sa podacima. Očekuje se da model dobro opisuje podatke i da se na osnovu njega može izvršiti predikcija gojaznosti (Tabela 30 na cd-u u prilogu).

U Tabeli 31 su prikazane promenljive koje smo koristili u istraživanju povezanosti gojaznosti sa potencijalnim faktorima rizika.

Promenljiva	Opis	Kod/Vrednost	Naziv
1	Stanje ishranjenosti	0 = Nije gojazna 1 = Gojaznost	stanje_i
2	Starost	Godine	starost
3	Stanje struka	1=Normalan 2=Povišen 3= Izrazito povišen	stanje_s
4	Obim kukova	Obim u cm	ok
5	Zbir DKN(Debljina Kožnog Nabora)	Debljina u mm	zbir_dkn
6	Mesto boravka	0 = Selo 1 = Grad	mes_bor
7	Doseljenik	0 = Nije doseljenik 1 = Doseljenik je	doselj
8	Uspeh u školi	1 = Nedovoljan 2 = Dovoljan 3 = Dobar 4 = Vrlo dobar 5 = Odličan	uspeh
9	Osoba živi	1 = Sa roditeljima 2 = Sa starateljima 3 = Sa praroditeljima 4 = Sa ocem 5 = Sa majkom	zivi
10	Sprema majke	1 = Osnovna 2 = Srednja 3 = Viša 4 = Visoka	spre_m
11	Sprema oca	1 = Osnovna 2 = Srednja 3 = Viša 4 = Visoka	spre_o
12	Da li osoba doruckuje	0 = Ne 1 = Da	dorucak
13	Broj obroka dnevno	Broj obroka	br_obrok
14	Koliko puta nedeljno osoba jede beli hleb	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	beli_h
15	Koliko puta nedeljno osoba jede crni hleb	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	crni_h
16	Koliko puta nedeljno osoba jede belo pecivo	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	belo_p
17	Koliko puta nedeljno osoba jede variva i salate	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	var_sal
18	Koliko puta nedeljno osoba jede proizvode od voca	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	pro_voc

19	Koliko puta nedeljno osoba jede proizvode od mesa	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	pro_meso
20	Koliko puta nedeljno osoba jede ribu	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	riba
21	Koliko puta nedeljno osoba jede mlecne proizvode	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	mlec_pro
22	Koliko puta nedeljno osoba jede puter i majonez	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	put_maj
23	Koliko puta nedeljno osoba jede grickalice	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	grickal
24	Koliko puta nedeljno osoba pije gazirana pica	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	gazirana
25	Koliko puta nedeljno osoba pije alkohol	0 = Ne jede 2 = Do dva puta 4 = Do četiri puta 7 = Svaki dan	alkohol
26	Broj cigareta na dan	Broj cigareta	br_cig
27	Broj rekreacija nedeljno	0 = Nijednom 1 = do dva puta 2 =Od tri do pet puta 3 = Više od pet puta	br_rek
28	Duzina rekreacije u minutima	0 = Nula minuta 1 = Do sat vremena 2 = Više od sat	duz_re

Tabela 31

Da bismo konstruisali odgovarajući model potrebno je da izaberemo promenljive koje su od značajnosti za model. Prvo fitujemo univarijabilne logističke regresione modele. Rezultati ovog fitovanja su dati u Tabeli 32 na cd-u u prilogu, gde je kao referetna vrednost kategorijalnih promenljivih uzimana prva vrednost iz Tabele 31 izuzev kada su u pitranje promenljive *spre_m* i *riba*, kada je kao refrentna vrednost uzeta poslednja vrednost iz date tabele. Smatramo da je promenljiva značajna ako je odgovarajuća p vrednost manja od 0.25, jer smatramo da bi bilo koja manja p vrednost eliminisala klinički značajne promenljive iz modela. Kako fitovanje univarijabilnih model retko daje adekvatnu analizu podataka u istraživanju, dalje razmatramo multivarijabilnu analizu. Dalje prelazimo na logističku regresiju korak po korak, koristićemo izbor "unapred" sa testom za eliminaciju "unazad".

U Tabeli 33 su predstavljene značajne promenljive dobijene korišćenjem statističkog paketa *SPSS 17* metodom korak po korak, kao i odgovarajući OR i njihova značajnost dobijena korišćenjem G testa[16]

Promenljiva	*Koeficijent*	*Značajnost*	*OR*
starost	-0.678	0.000	0.507
stanje_s(normalan**)**		0.000	1
stanje_s(povišen**)**	1.584	0.000	4.875
stanje_s(izrazito povišen**)**	20.507	0.997	805816269
ok	0.212	0.000	1.236
zbir_dkn	0.066	0.000	1.068
spre_m(visoka**)**		0.003	1
spre_m(viša**)**	2.202	0.007	9.040
spre_m(srednja**)**	2.280	0.004	9.781
spre_m(osnovna**)**	1.220	0.156	3.388
beli_h(ni jedan dan**)**		0.102	1
beli_h(do 2 puta nedeljno**)**	0.668	0.252	1.951
beli_h(do 4 puta nedeljno**)**	0.609	0.301	1.839
beli_h(svaki dan**)**	-0.033	0.952	0.968
pro_meso(svaki dan**)**		0.160	1
pro_meso(do 4 puta nedeljno**)**	1.103	0.031	3.015
pro_meso(do 2 puta nedeljno**)**	0.452	0.192	1.572
pro_meso(ni jedan dan**)**	0.655	0.106	1.926
riba(svaki dan**)**		0.022	1
riba(do 4 puta nedeljno**)**	0.230	0.882	1.259
riba(do 2 puta nedeljno**)**	0.602	0.694	1.825
riba(ni jedan dan**)**	1.462	0.346	4.313
gazirana(ni jedan dan**)**		0.076	1
gazirana(do 2 puta nedeljno**)**	1.082	0.014	2.952
gazirana(do 4 puta nedeljno**)**	0.605	0.214	1.831
gazirana(svaki dan**)**	0.953	0.038	2.593
duz_re(nula minuta**)**		0.040	1
duz_re(do sat vremena**)**	-0.213	0.462	0.808
duz_re(preko sat vremena**)**	0.932	0.030	2.540

Tabela 33

Multivarijabilnom logističkom regresijom uočeno je da osobe čija majka ima nižu stručnu spremu imaju do 10 puta veću šansu za gojaznost nego kod osoba čija majka ima visoku stručnu spremu. Količina pojedinih konzumiranih namirnica nedeljno je takođe značajan prediktor gojaznosti, tako npr. osobe koje ne jedu uopšte ribu imaju četiri puta veću šansu za gojaznost u odnosu na osobe koje jedu ribu svaki dan, ili osobe koje dva puta nedeljno piju gazirana pića imaju četiri puta veću šansu za gojaznost od osoba koje ne piju gazirana pića. Međutim uočili smo i par nepravilnosti koje se ne slažu sa našim intuitivnim shvatanjem, odnosno važi da osobe koje vežbaju duže od sat vremena imaju dva i po puta

veću šansu da ostanu gojazne od osoba koje uopšte ne vežbaju, takođe važi i da osobe koje ne jedu beli hleb i one koje ga jedu svaki dan imaju približno istu šansu za gojaznost, dok osobe koje jedu beli hleb dva puta nedeljno imaju oko dva puta veću šansu za gojaznost od njih . Iz tog razloga smo za ove promenljive proverili da li su u interakciji sa nekim promenljivim i dobili smo da važi da je promenljiva *duz_re* u interakciji sa promenljivom *gazirana*, i važi i da je promenljiva *beli_h* u interakciji sa promenljivom *riba*. Tako da smo ove dve interakcije ubacili u model. Takođe smo ispitali i da li među promenljivim koje nisu ušle u model postoji ometača kao i da li postoji drugih interakcija sem ove i nismo došli do pozitivnih zaključaka.

Koliko se model slaže sa podacim smo prvo testirali upotrebom Hosmer-Lemeshevog testa (Tabela 34) i vidimo da se u modelu observirane i očekivane frekvencije ne razlikuju značajno, te da je model na osnovu ovog testa fitovan.

	Tabela kontigencije za Hosmer-Lemeshow test				
	Premor = nije premoren		Premor = premoren		Ukupno u grupi
	posmatrano	očekivano	posmatrano	očekivano	
1	216	215.977	0	0.023	216
2	216	215.905	0	0.095	216
3	216	215.777	0	0.223	216
4	216	215.551	0	0.449	216
5	215	215.066	1	0.934	216
6	216	214.178	0	1.822	216
7	211	211.976	5	4.024	216
8	206	206.367	10	9.633	216
9	180	181.534	36	34.466	216
10	43	42.668	176	176.332	219

Tabela 34

Hi-kvadart	Stepeni slobode	p
2.974	8	0.936

Tabela 35

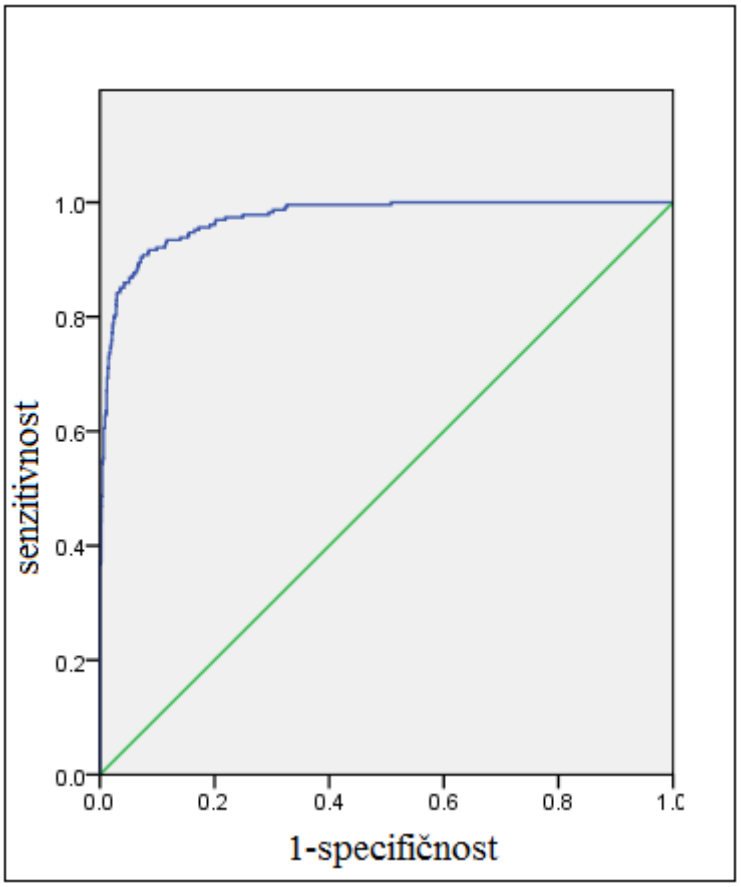
Fitovanje model možemo oceniti i pomoću tabele klasifikacije (Tabela 36), koja je značajna jer možemo da ocenimo koliko model dobro predviđa one osobe koje su zaista gojazne – senzitivnost kao i one koje nisu gojazne – specifičnost, kao i kolika je ukupna predikcija modela.

	registrovano		
klasifikovano	gojaznost = 0	gojaznost = 1	ukupno
gojaznost = 0	1906	29	1935
gojaznost = 1	62	166	228
ukupno	1968	195	2163

Tabela 36

Važi da je senzitivnost testa 85.1%, dok je specifičnost testa 96.8%, dok je ukupna stopa tačne klasifikacije modela 95.8%. Odnosno i pomoću tabela klasifikacije smo videli da je model dobro fitovan.

Površina ispod ROC krive, mera koja predstavlja verovatnoću da će osobe koje su gojazne imati veću ocenjenu verovatnoću nego oni koji nisu gojazni iznosi 0.973 (Slika 10) što predstavlja *izvanredno razdvajanje*.



Slika 10

8 Zaključak

U ovom radu smo se upoznali sa logističkim regresionim modelom čija je zavisna promenljiva dihotomna, dok su faktori rizika kategorijalne ili neprekidne promenljive.

Kroz mnoštvo primera smo prikazali sve nivoe građenja logističkog regresionog modela počev od ocenjivanja koeficijenata, do procene slaganja modela sa podacima. Takođe smo se upoznali sa bitnim statističkim metodoma, koje nisu vezane isključivo za logističku regresiju, kao što su metoda maksimalne verodostojnosti, Wald test, Hosmer - Lemeshow test, Pirsonova hi-kvadrat statistika, ROC kriva, tabele klasifikacije.

U poslednjem poglavlju rada smo prikazali evaluaciju modela sa zavisnom promenljivom gojaznost, gde su se kao značajni rizični faktori pokazali: starost, stanje struka, obim kukova, sprema majke, zbir DKN, dužina rekreacije, količina konzumiranog belog hleba, proizvoda od mesa i gaziranih pića nedeljno.

Logistička regresija se pokazala kao moćan alat za pronalaženje zakonitosti u skupu nominalnih varijabli i kao takva ima široku primenu u ekonomskim, populacionim, biološkim, marketinškim, medicinskim i mnogim drugim istraživanjima.

Literatura

[1] Chatterjee Samprit, Hadi S. Ali, Regression Analysis by Example-fourth edition, John Wiley & Sons, Inc., 2006.

[2] Dobson J. Annette, An Introduction to Generalized Linear Models-second edition, Chapman & Hall/CRC, 2002.

[3] Efremov Alexander, Stepwise Logistic Reression, Experian, 2010.

[4] Field Andy, Discovering Statistics Using SPSS-second edition, Sage publications, 2000

[5] Hallett C. David, Goodness of Fit Test in Logistic Regression, University of Toronto, 1999

[6] Hosmer W. David, Lemeshow Stanley, Applied Logistic Regression-second edition, Wiley Series, 2000

[7] Jaccard James, Interaction Effects in Logistic Regression, Sage publications, 2001.

[8] Kleinbaum G. David, Klein Mitchel, Logistic regression: A Self Learning Text – third edition, Springer, 2010.

[9] Maltus Thomas, An Essey on the Principle of Population, Paul's church-yard, 1798.

[10] Menard Scott, Applied Logistic Regression Analysis-second edition, Sage publications, 2001.

[11] Pampel C. Fred, Logistic regression: A Primer, Sage publications, 2000

[12] Pohar Maja, Blas Mateja, Turk Sandra, Comparison Logistic Reression Models and Linear Discriminant Analysis: A Simulation Study, Metodološki zvezki, Vol 1, No. 4, pp. 143-161, 2004.

[13] Stevenson Mark, An Introduction to Logistic Regression, EpiCentre, 2008

[14] Uusipaikka Esa, Confidence Intervals in Generalized Regression Models, Chapman & Hall/CRC, 2009.

[15] http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore16.pdf

[16] http://core.ecu.edu/psyc/wuenschk/MV/MultReg/Logistic-SPSS.pdf

[17] http://pat-thompson.net/PDFversions/Theses/2010CastilloGarsow.pdf

[18] http://statmaster.sdu.dk/courses/st111/module14/module.pdf

[19] http://www.childrensmercy.org/stats/journal/confidence.aspx

[20] http://www.czep.net/stat/mlelr.pdf

[21] http://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/confounding-interactions-methods

[22] http://www.jeremydawson.co.uk/slopes.htm

[23] http://www.jeremymiles.co.uk/regressionbook/data/index.html

[24] http://www.medcalc.org/manual/roc-curves.php

[25] http://www.myoops.org/twocw/jhsph/courses/StatisticalReasoning2/PDFs/Lecture8.pdf

[26] http://www.stat.ufl.edu/~aa/sta6127/ch15.pdf

[27] http://www.statistical-solutions-software.com/BMDP-documents/BMDP-PR.pdf

[28] http://www.statistical-solutions-software.com/BMDP-documents/BMDP-LR.pdf

[29] http://www.utdallas.edu/~pkc022000/6390/SP06/NOTES/Logistic_Regression_4.pdf

UNIVERZITET U NOVOM SADU
PRIRODNO MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: Monografska dokumentacija
TD

Tip zapisa: Tekstualni štampani materijal
TZ

Vrsta rada: Master rad
VR

Autor: Gorica Gvozdić
AU

Mentor: dr Zagorka Lozanov-Crvenković
ME

Naslov rada: Primenjena logistička regresija
NR

Jezik publikacije: srpski (latinica)
JP

Jezik izvoda: s/en
JI

Zamlja publikovanja: Republika Srbija
ZP

Uže geografsko područje: Vojvodina
UGP

Godina: 2011.
GO

Izdavač: autorski reprint
IZ

Mesto i adresa: Novi Sad, Trg D. Obradovića 4

MA

Fizički opis rada: **FOR**	(8/70/29/28/10/0/8) (broj poglavlja/strana/lit.citata/tabela/slika/grafika/priloga)
Naučna oblast: **NO**	matematika
Naučna disciplina: **ND**	statistika
Predmetne odrednica, ključne reči:(**PO**, **UDK**)	logistička regresija, metod maksimalne verodostojnosti, Wald test, Hosmer –Lemeshow test, tabele klasifikacije, ROC kriva
Čuva se: **ČS**	u biblioteci Departmana za matematiku i informatiku
Važna napomena: **VN**	nema
Izvod (**IZ**):	U ovom radu su izloženi osnovni pojmovi vezani logističku regresiju. Objašnjene su sve faze građenja logističkog regresionog modela, počev od slaganja logističkog regresionog modela sa podacima, preko interpretacije fitovanog logističkog modela do procene slaganja modela sa podacima.
Datum prihvatanja teme od strane NN veća: **DP**	decembar 2011.
Datum odbrane: **DO**	decembar 2011.
Članovi komisije: **KO**	
Predsednik:	dr Ljiljana Gajić, redovni profesor Prirodno-matematičkog fakulteta u Novom Sadu
Član:	dr Ivana Štajner-Papuga, vanredni profesor Prirodno-matematičkog fakulteta u Novom SaduSadu
Mentor:	dr Zagorka Lozanov-Crvenković, redovni profesor Prirodno-matematičkog fakulteta u Novom Sadu

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
KEY WORDS DOCUMENTATION

Accession number:
ANO

Identification number:
INO

Document type: Monograph type
DT

Type of record: Textual printed material
TR

Contents code: Master's thesis
CC

Author: Gorica Gvozdić
AU

Mentor: Dr Zagorka Lozanov-Crvenković
ME

Title: Applied logistic regression
TI

Language of text: Serbian (Latin)
LT

Language of abstract: s /en
LT

Country of publication: Republic of Serbia
CP

Locality of publication: Vojvodina
LP

Publication year: 2011.
PY

Publisher: author's reprint
PU

Publ. place: Novi Sad, Trd D. Obradovića 4
PP

Physical description: (8/70/29/28/10/0/8)
PD

Scientific field: Mathematics
SF

Scientific discipline: Statistics
SD

Subject Key words: Logistic regression, maximum likelihood, Wald test, Hosmer –Lemeshow test, Wald test, classification tables, ROC curve
SKW

Holding data: In the library of Department of Mathematics and Informatics
HD

Note:
N

Abstract (**AB**): This paper presents the basic concepts related to logistic regression. All stages of building logistic regression model are explained here, starting from fitting logistic regression model via interpretation of the fitted logistic regression model to assessing the fit of model.

Accepted on Scientific board on: December 2011
AS

Defended: December 2011
DE

Thesis Defend board:
DB

President: Dr Ljiljana Gajić, full profesor, Faculty of Sciences, University of Novi Sad

Member: Dr Ivana Štajner-Papuga, associate professor, Faculty of Sciences, University of Novi Sad

Mentor: Dr Zagorka Lozanov-Crvenković, full profesor, Faculty of Sciences, University of Novi Sad