Univerzitet u Novom Sadu

Prirodno-matematički fakultet

Departman za matematiku i informatiku

Milica Pejović

Logistička regresija sa primenom na određivanje optimalnog tretmana u zaštiti bilja

- Master rad -

Mentor

Prof. dr Zagorka Lozanov-Crvenković

Novi Sad, 2021.

Sadržaj

Uvod

U fokusu mnogih statističkih istraživanja jeste i ispitivanje postojanja zavisnosti između određenih pojava i opisivanje veze kojom su posmatrane pojave povezane. Regresija je jedna od statističkih metoda koje se koriste za modeliranje veze. Ukoliko pojava za koju se ispituje zavisnost od drugih faktora ima takvu prirodu da se modelira slučajnom promenljivom koja je prekidna ili kategorijalna, koristi se specijalna vrsta regresije – logistička regresija. Izlaz regresije u slučajevima kada je zavisna promenljiva neprekidna jeste njena predviđena vrednost. Međutim, kada je reč o logističkoj regresije vrši se modeliranje verovatnoće da zavisna promenljiva uzme određenu kategorijalnu vrednost usled fiksiranih vrednosti nezavisnih promenljivih. Upravo je logistička regresija tema ovog master rada. U radu je prvo opisan teorijski okvir logističke regresije koji je kasnije primenjen praktično, kako bi se odredio optimalan insekticidni tretman u usevu suncokreta u cilju smanjenja oštećenja biljaka izazvanih najznačajnim zemljišnim štetočinama - žičarama.

U prvom poglavlju rada analizirana je binarna logistička regresija kod koje zavisna promenljiva uzima dve kategorijalne vrednosti. Pre svega je dat model logističke distribucije po kojem se vrši modeliranje verovatnoće da zavisna promenljiva uzme određenu kategorijalnu vrednost. Sledeći korak je bio fitovanje logističkog regresijskog modela kod kog se na osnovu datog uzorka vrši ocena nepoznatih parametara. Dobijanjem ocena nepoznatih parametara i njihovim ubacivanjem u model izvodi se ocena logističkih verovatnoća. Kako postoji veliki broj promenljivih koji mogu uticati na zavisnu promenljivu, analizirana su dva testa kojima se ispituje statistička značajnost nezavisnih promenljivih – Test količnika verodostojnosti i Wald test. Jedna od veoma značajnih stavki prilikom izvođenja logističke regresije jeste pravilna interpretacija ocenjenih koeficijenata modela i ostalih dobijenih pokazatelja. Sama interpretacija zavisi od vrste nezavisne promenljive, odnosno da li je reč o kategorijalnoj ili neprekidnoj nezavisnoj promenljivoj. Na kraju prvog poglavlja je objašnjen Hosmer-Lemeshow test za procenu slaganja modela sa podacima u smislu slaganja registrovanih vrednosti zavisne promenljive i predviđenih vrednosti dobijenih na osnovu logističkog modela.

U drugom poglavlju, koje predstavlja uopštenje binarne logističke regresije, prelazi se na analiziranje logističkog regresijskog modela kod kojeg zavisna promenljiva ima više od dve kategorijalne vrednosti. Ovakav logistički regresijski model se naziva multinomni logistički

regresijski model. Ocena nepoznatih parametara modela se vrši istom metodom korišćenom u slučaju binarne regresije. Izvodi se generalizacija testa količnika verodostojnosti i Hosmer-Lemesow testa opisanih u prvom poglavlju.

U trećem poglavlju je izvršena analiza logističkog regresijskog modela u slučaju kada je merna skala zavisne promenljive ordinalna, odnosno postoji mogućnost rangiranja kategorijalnih vrednosti zavisne promenljive.

Četvrto poglavlje predstavlja praktični deo rada u kojem će se teorijski deo logističke regresije primeniti u izboru optimalnog insekticida u cilju zaštite useva suncokreta (hibrid Romeo) od najznačajnijih zemljišnih štetočina - žičara. Do potrebnog uzorka za sprovođenje logističke regresije došlo se eksperimentima izvedenih u toku 2021. godine, na dva lokaliteta na oglednim poljima Instituta za ratarstvo i povrtarstvo (kod Bačkog jarka - „Polje 1“ i kod Novog Sada - „T-12“), na Rimskim šančevima, Novi Sad, Srbija. Svaka vrednost zavisne promenljive u uzorku je dobijana od biljaka koje su bile tretirane nekim od šest različitih insekticida. Prvo je izvedena binarna logistička regresija gde se posmatra uticaj lokaliteta i tretmana na prisustvo oštećenja kod biljke. Zatim je definisan nivo oštećenja kod biljaka pomoću pet kategorija nivoa oštećenja kod biljaka i potom je izvedena multinomna regresija u cilju ispitivanja uticaja lokaliteta i tretmana na nivo oštećenja kod suncokreta. Na kraju je dat zaključak sprovedenih logističkih modela u smislu izbora najboljeg insekticida u zaštiti suncokreta od oštećenja izazvanih žičarama.

1. Logistička regresija

U procesu statističkog istraživanja značajan deo zauzima ispitivanje i opisivanje veze između različitih pojava. Opisivanje veze se vrši pronalaženjem odgovarajuće funkcionalne veze koja povezuje posmatrane pojave i upravo se time bavi regresiona analiza. Regresijska analiza je metod koji utvrđuje koje pormenljive imaju uticaj na promenljivu koja je u fokusu našeg interesovanja. Promenljiva koja je u fokusu našeg interesovanja naziva se zavisna promenljiva, dok su nezavisne promenljive one koje imaju uticaj na posmatranu pojavu. Obično u stvarnom životu postoji veliki broj promenljivih koje utiču na ishod naše posmatrane pojave, međutim treba utvrditi koje promenljive statistički značajno utiču na tu pojavu. Sam postupak određivanja regresije nam omogućava da utvrdimo koje promenljive su najvažnije za ishod zavisne promenljive, a koje se mogu zanemariti i na taj način stižemo do najekonomičnijeg modela koji opisuje vezu između zavisne promenljive i značajnih nezavisnih promenljivih. Regresija jeste jedna od velikog broja statističkih metoda za modeliranje i analizu kompleksnih skupova podataka. Sve te metode se jednim imenom nazivaju statistical learning. Specifičnost regresije u odnosu na druge statističke tehnike jeste u tome što kod nje postoji izlaz, odnosno zavisna promenljiva. Najjednostavniji oblik regresije jeste linearni regreijski model. Ukoliko je zavisna promenljiva diskretna ili kategorijalna u tom slučaju ne možemo koristiti linearni regresijski model jer se za njegovu upotrebu zahteva da zavisna promenljiva bude neprekidna. U slučaju diskretne ili kategorijalne zavisne promenjljive koristi se logistički regresijski model. Upravo zbog ove razlike između logističke i linearne regresije, kod logističke regresije ne postoje pretpostavke o raspodeli promenljivih i postoji razlika u izboru parametara. Međutim, metode koje se koriste u logističkoj regresiji motivisani su opštim principima koji se koriste u linaranoj regresiji i to se ilustruje sledećim primerom.

Primer *(videti u [1]):*

Posmatramo tabelu 1 u kojoj su prikazane godine stotinu subjekata i da li oni boluje od koronarnog srčanog oboljenja. Promenljiva GOD predstavlja starost subjekta, dok promenljiva KSO jeste dihotomna promenljiva koja uzima vrednost 0 ukoliko osoba ne boluje od koronarnog srčanog oboljenja, a u suprotnom uzima vrednost 1. Cilj jeste da utvrdimo u kojoj meri starost utiče na pojavu koronarnog srčanog oboljenja, dakle naša zavisna promenljiva je kategorijalna. Ukoliko zavisnu promenljivu posmatramo kao neprekidnu i sprovedemo prvi korak u linearnoj regresiji – crtanje dijagrama rasipavanja, taj dijagram bi bio prikazan na slici 1.

Tabela 1: Starost subjekata i prisustvo KSO

ID	GOD	KSO
1	20	0
2	23	0
3	24	0
4	25	0
5	25	1
6	26	0
7	26	0
8	28	0
9	28	0
10	29	0
11	30	0
12	30	0
13	30	0
14	30	0
15	30	0
16	30	1
17	32	0
18	32	0
19	33	0
20	33	0

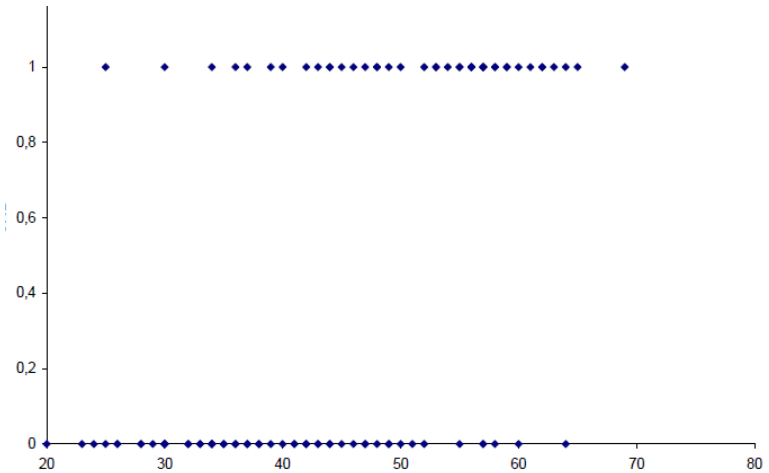
ID	GOD	KSO
21	34	0
22	34	0
23	34	1
24	34	0
25	34	0
26	35	0
27	35	0
28	36	0
29	36	1
30	36	0
31	37	0
32	37	1
33	37	0
34	38	0
35	38	0
36	39	0
37	39	1
38	40	0
39	40	1
40	41	0

ID	GOD	KSO
41	41	0
42	42	0
43	42	0
44	42	0
45	42	1
46	43	0
47	43	0
48	43	1
49	44	0
50	44	0
51	44	1
52	44	1
53	45	0
54	45	1
55	46	0
56	46	1
57	47	0
58	47	0
59	47	1
60	48	0

ID	GOD	KSO
61	48	1
62	48	1
63	49	0
64	49	0
65	49	1
66	50	0
67	50	1
68	51	0
69	52	0
70	52	1
71	53	1
72	53	1
73	54	1
74	55	0
75	55	1
76	55	1
77	56	1
78	56	1
79	56	1
80	57	0

ID	GOD	KSO
81	57	0
82	57	1
83	57	1
84	57	1
85	57	1
86	58	0
87	58	1
88	58	1
89	59	1
90	59	1
91	60	0
92	60	1
93	61	1
94	62	1
95	62	1
96	63	1
97	64	0
98	64	1
99	65	1
100	69	1

Kada bi se govorilo o linearnoj regresiji, dijagrama rasipanja bi pomogao u pretpostavljanju funkcionalne veze između zavisne i nezavisne promenljive, odnosno odredila bi se glatka kriva koja će najbolje aproksimirati date tačke na dijagramu rasipanja. Kod datog primera nailazi se na problem velike varijabilnosti zavisne promenljive KSO za sve godine starosti (slika 1), što dovodi do komplikacije prilikom željenje aproksimacije glatkom krivom.



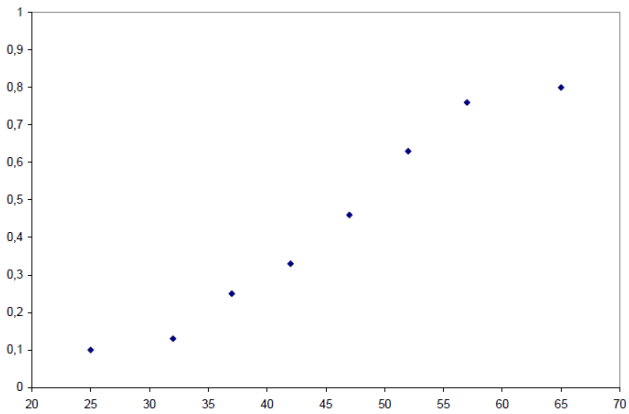
Slika 1: Grafički prikaz starosti i prisustva KSO

Jedino što se može zaključiti sa datog dijagrama rasipanja jeste pravilo da su subjekti bez koronarnog srčanog oboljenja mlađi od onih kod koji je zabeleženo prisustvo oboljenja i da se sve tačke nalaze na dve paralelne prave: $y = 0$ i $y = 1$, što ukazuje na postojanje dve vrednosti za zavisnu promenljivu. U cilju odražavanja sturkturne veze između prisustva CSO i starosti vrši se pravljenje intervala za nezavisnu promenljivu. Zatim se za svaku grupu računa verovatnoća prisustva koronarnog srčanog oboljenja u okviru posmatrane grupe.

Tabela 2: Verovatnoća prisustva KSO unutar svake starosne grupe

starosna grupa	broj osoba unutar grupe	KSO =0	KSO=1	verovatnoća (sredina)
20-29	10	9	1	0.10
30-34	15	3	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Ukupno	100	57	43	0.43

Ukoliko se sada grafički za svaku starosnu grupu prikaže verovatnoća prisustva koronarnog srčanog oboljenja, dobija se grafik (slika 2), odakle se izvodi zaključak da povećanjem starosne grupe dolazi do rasta proporcije subjekata koji imaju koronarno srčano oboljenje.



Slika 2: Verovatnoća prisustva KSO unutar svake starosne grupe

Ukoliko se sa Y označi zavisna promenljiva, sa X nezavisna promenljiva i sa x njena konkretna vrednost, ključna stavka svakog regresijskog modela jeste odrediti očekivanu vrednost zavisne promenljive za zadatu vrednost nezavisne promenljive, odnosno odrediti $E(Y|X=x)$. Kod linearne regresije je očekivanje za zavisnu promenljivu linarna fukcija po nezavisnoj promenljivoj, odnosno $E(Y|x)=\alpha+\beta x$ i dobijena prava se naziva populaciona linija. Dalje su se određivale ocene parametara α i β, u oznaci $\hat{\alpha}$ i $\hat{\beta}$, pomoću kojih se izvode ocene za očekivanu vrednost zavisne promenljive $\hat{E}(Y|x)=\hat{\alpha}+\hat{\beta}x$. Dobijena prava se naziva uzoračka linija. Isti postupak se sprovodi i kod logističke regresije. Razlika jeste u tome što E(Y|x) kod linarne regresije može uzeti bilo koji realan broj, dok kod logističke regresije to nije slučaj. Kako je u posmatranom primeru KSO dihotomna promenljiva ona može uzeti dve vrednosti sa određenim verovatnoćama. Ako se pretpostavi da promenljiva KSO uzima vrednost 1 sa verovatnoćom π, tada KSO ima raspodelu datu sa KSO: $\begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$. Iz ove pretpostavke sledi da će i slučajna promenljiva $KSO|x$ takođe uzimati vrednosti 0 i 1, pri tome neka uzima vrednos 1 sa verovatnoćom $\pi(x)$, odnosno neka je raspodela slučajne promenljive $KSO|x$: $\begin{pmatrix} 0 & 1 \\ 1-\pi(x) & \pi(x) \end{pmatrix}$. Tada je

$$E(KSO|x)=0*\big(1-\pi(x)\big)+1*\pi(x)=\pi(x),$$

te kako je $\pi(x)$ verovatnoća sledi da $0\leq E(KSO|x)\leq 1$. Ukoliko kolonu verovatnoća (sredina) iz tebele 2 posmatramo kao ocena očekivane vrednosti za KSO usled date grupe starosti x, može se zaključiti da što je uslovno očekivanje bliže 0 ili 1 to promena u $E(KSO|x)$ postaje progresivno manja. Odavde prirodno dolazi da se uslovno očekivanje u slučaju dihotomne zavisne promenljive, odnosno verovatnoću $\pi(x)$, može modelirati logističkom distribucijom. Dakle, logistički model koji se analizira će biti:

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

Ukoliko se pređe na k nezavinih promenljivih, logistički model se može uopštiti:

$$\pi(x)=P(Y=1|X_1=x_1,X_2=x_2,\dots,X_k=x_k)=\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_j}}.$$

Kod linearne regresije vrednost zavisne promenljive kada nezavisna promenljiva uzme konkretnu vrednost modelirali se sa

$$Y|x = E(Y|x) + \varepsilon,$$

pri čemu je ε slučajna greška. Prilikom izvođeja zaključaka kod linearne regresije često se koristi pretpostavka da je ε normalno raspodeljena, sa očekivanjem 0 i disperzijom istom za sve vrednosti x nezavisne promenljive, što je impliciralo da je i slučajna promenljiva $Y|x$ sa normalnom raspodelom i konstantnom varijansom. S druge strane, u slučaju dihotomne slučajne promenljive pretpostavlja se da je

$$KSO|x = \pi(x) + \varepsilon.$$

Obzirom da slučajna promenljiva $KSO|x$ ima raspodelu $KSO|x$: $\begin{pmatrix} 0 & 1 \\ 1-\pi(x) & \pi(x) \end{pmatrix}$, sledi da slučajna promenljiva ε može uzeti vrednosti $-\pi(x)$ i $1-\pi(x)$ sa verovatnoćama $1-\pi(x)$ i $\pi(x)$, respektivno. Odnosno, ε ima binomnu raspodelu ε: $\begin{pmatrix} -\pi(x) & 1-\pi(x) \\ 1-\pi(x) & \pi(x) \end{pmatrix}$. Očekivanje greške ε je $E(\varepsilon) = 0$, dok je disperzija $D(\varepsilon) = \pi(x)(1-\pi(x))$.

U verovatnoći se šansa uspeha događaja definiše kao količnik verovatnoće da se događaj desi i verovatnoće da se taj događaj ne realizuje. U tom slučaju, šansa da posmatrana zavisna promenljiva KSO uzme vrednost 1 ukoliko je nezavisna promenljiva uzela vrednos x je

$$\frac{P(KSO = 1|X = x)}{1 - P(KSO = 1|X = x)} = \frac{\pi(x)}{1-\pi(x)}.$$

Pojam šanse se uvodi kako bi se definisala logit funkcija u oznaci *logit*:

$$logit(x) \coloneqq \ln\frac{x}{1-x}, x \in (0,1).$$

Domen funkcije $logit$ je $(0,1)$, pa ukoliko se input funkcije $logit$ posmatra kao verovatnoća da se određeni događaj desio, $logit$ funkcija zapravo predstavlja prirodni logaritam šanse uspeha posmatranog događaja. Logit šanse događaja da zavisna diihotomna promenljiva zabeleži prisustvo posmatrane osobine će biti linearan po koeficijentima $\beta_j, j = 0, \dots, k$ (formula 1.1).

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}}$$

$$1 - \pi(x) = \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}}$$

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln\left(e^{\beta_0 + \sum_{j=1}^{k} \beta_j x_j}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j \, x_j . \quad (1.1)$$

Ukoliko je merna skala zavisne promenljive nominalna govorimo o nominalnoj logističkoj regresiji. Prilikom merenja zavisne promenljive nominalnom skalom vrši se samo imenovanje, kategorizacija ili klasifikacija mogućih vrednosti obeležja. U tom slučaju možemo reći samo kada su dve vrednosti obeležja iste, ali ne možemo vršiti poređenje vrednosti obeležja. S druge strane, ukoliko je merna skala nezavisne promenljive ordinalna, radi se o ordinalnoj logističkoj regresiji. Prilikom merenja zavisne promenljive ordinalnom skalom vrši se i rangiranje mogućih vrednosti obeležja, pa se ona mogu i upoređivati. Promenljiva koja predstavlja pol meriće se nominalnom skalom (kategorije: muško i žensko), kao i promenljiva koja predstavlja krvnu grupu (kategorije: A, B, AB, 0). Naspram toga, za merenje promenljive koja predstavlja zadovoljstvo korisnika upotrebiće se ordinalna skala (moguće vrednosti: nezadovoljan, indiferentan, zadovoljan).

1.1 Fitovanje logističkog regresionog modela

Posmatra se logistički model sa jedom zavisnom promenljivom X i kod kojeg je nezavisna promenljiva Y dihotomna, kodirana sa 0 i 1:

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

Pod fitovanjem logističkog regresionog modela podrazumeva se ocenu vrednosti nepoznatih parametara $\beta_0 \, i \, \beta_1$ na osnovu datog uzorka. Pretpostavlja se da postoji uzorak veličine n: $(x_i, \, y_i), i = 1, \dots, n$, gde y_i predstavja rezultirajuću vrednos zavisne promenljive za registrovanu vrednost x_i nezavisne promenljive i neka su dati parovi međusobno nezavisni. Za ocenjivanje nepoznatih parametara koristi se Metod maksimalne verodostojnosti, kod kojeg se na osnovu datog uzorka određuju nepoznati parametri $\beta_0 \, i \, \beta_1$ tako da se maksimizira verovatnoća da dati uzorak bude izabran. Literatura korišćena za izradu ovog dela rada je [1], [3], [7] i [15].

Neka raspodela obeležja Z pripada familiji raspodela $\{F(z,\theta), z \in \boldsymbol{R}, \theta \in \Theta\}$, pri čemu je Θ jednodimenzionalan ili višedimenzionalan skup parametara. Familija raspodela se naziva dopustiva familija raspodela za obeležje Z, dok je skup Θ dopustiv skup parametara. U slučaju diskretne slučajne promenljive familija raspodela je zadata sa

$$\{p(z_k,\theta), z_k \in \boldsymbol{R}, k = 1,2,\dots,\theta \in \Theta\},$$

pri čemu je $p(z_k,\theta) = P_\theta(Z = z_k)$, odnosno $p(z_k,\theta)$ predstavlja verovatnoću da slučajna promeljiva Z uzme vrednost z_k ako je nepoznati parametar baš θ. Ukoliko je $(Z_1, Z_2, \dots, Z_n)$ prost slučajan uzorak, odnosno $Z_1, Z_2, \dots, Z_n$ su međusobno nezavisne i imaju istu raspodelu, jedan od načina za ocenjivanje nepoznatog parametra θ je metod maksimalne verodostojnosti.

Definicija 1.1. Neka je Z diskretna slučajna promenljiva. Funkcija verodostojnosti za prost slučajan uzorak $(Z_1, Z_2, \dots, Z_n)$ na osnovu realizovanog uzorka $(z_1, z_2, \dots, z_n)$ obima n je:

$$l(\theta, z_1, z_2, \dots, z_n) = p(z_1,\theta)p(z_2,\theta)\dots p(z_n,\theta).$$

Definicija 1.2. Neka je

$$l(\theta) = l(\theta, z_1, z_2, \dots, z_n)$$

funkcija verodostojnosti za prost slučajan uzorak $(Z_1, Z_2, \dots, Z_n)$ na osnovu realizovanog uzorka $(z_1, z_2, \dots, z_n)$. Ako je $\hat{\theta} = u(z_1, z_2, \dots, z_n)$ u dopustivom skupu parametara Θ i maksimizira $L(\theta)$, tada je statistika $\hat{\theta} = u(Z_1, Z_2, \dots, Z_n)$ ocena nepoznatog parametra θ dobijena metodom maksimalne verodostojnosti. Realizovana vrednost $\hat{\theta} = u(z_1, z_2, \dots, z_n)$ je ocena maksimalne verodostojnosti za θ na osnovu uzorka $(z_1, z_2, \dots, z_n)$.

Kako funkcija verodostojnosti predstavlja zapravo verovatnoću da dati uzorak bude izabran, metod maksimalne verodostojnosti se dobija ocena $\hat{\theta} = u(z_1, z_2, \dots, z_n)$ za koju je spomenuta verovatnoća maksimalna. Prilikom posmatranja uzorka iz logističke regresije $(x_i,\ y_i)$ zaključuje se sledeće:

- ako je $y_i = 1$ verovatnoća da slučajna promenljiva X uzme vrednost x_i modelirana je sa $\pi(x_i)$;
- ako je $y_i = 0$ verovatnoća da slučajna promenljiva X uzme vrednost x_i modelirana je sa $1 - \pi(x_i)$.

Odnosno, verovatnoća da slučajna promenljiva X uzme vrednost x_i kada je dihotomna promenljiva Y uzela vrednost y_i data je sa $\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$. Uzimajući u obzir da je uzorak iz logističke regresije prost slučajan uzorak, njegova funkcija verodostojnosti na osnovu realizovanog uzorka $(x_i,\ y_i), i = 1, \dots, n$, prema definiciji 1.2. je:

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\big(1-\pi(x_i)\big)^{1-y_i}.$$

Određuje se tačka u kojoj se dostiže maksimum funkcije $l(\beta_0, \beta_1)$ i ta tačka predstavlja ocenu nepoznatih parametara $\beta_0, \beta_1,$ u oznaci $\hat{\beta}_0, \hat{\beta}_1$. Nezavisno od konkretne vrednosti nepoznatih parametara, svi članovi u proizvodu iz funkcije verodostojnosti su manji od jedan jer oni predstavljaju određenu verovatnoću, te će njihov količnik biti mali broj. Iz navedenog razloga umesto maksimuma funkcije $l(\beta_0, \beta_1)$, traži se maksimum funkcije $L(\beta_0, \beta_1) \coloneqq ln\big(l(\beta_0, \beta_1)\big)$

$$L(\beta_0, \beta_1) \coloneqq ln\big(l(\beta_0, \beta_1)\big) = \sum_{i=1}^{n}\big[y_i\, ln\, \pi(x_i) + (1-y_i)\, ln\big(1-\pi(x_i)\big)\big].$$

Kako je funkcija ln monotono rastuća, funkcije $l(\beta_0, \beta_1)$ i $L(\beta_0, \beta_1)$ dostižu maksimum u istoj tački. Znajući da je $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$, sledi

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n}\left[y_i\big(\beta_0 + \beta_1 x_i - ln\big(1+e^{\beta_0+\beta_1 x_i}\big)\big) + (1-y_i) ln\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)\right]$$

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n}\big[y_i(\beta_0 + \beta_1 x_i) - ln\big(1+e^{\beta_0+\beta_1 x_i}\big)\big].$$

Kako bi se odredio maksimum funkcije $L(\beta_0, \beta_1)$ parcijalni izvodi po β_0 i β_1 se izjednačavaju sa nulom:

$$\frac{\partial L}{\partial \beta_0} = 0 \iff \sum_{i=1}^{n}\left[y_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right] = 0 \iff \sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

$$\frac{\partial L}{\partial \beta_1} = 0 \iff \sum_{i=1}^{n}\left[y_i x_i - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} x_i\right] = 0 \iff \sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0.$$

Ukoliko je broj nezavisnih promenljivih $k > 1$ jednostavnije je preći na matrični zapis sistema za određivanje nepoznatih parametara. Uvode se sledeće oznake:

$\boldsymbol{\beta} = (\beta_{0,}\, \beta_1, \dots, \beta_k)^T$ – vektor nepoznatih parametara;

$\boldsymbol{x_i} = (1, x_{1i}, x_{2i}, \dots, x_{ki})^T$ – gde x_{ji}, $j = 1,2, \dots, k$ predstavlja registrovanu vrednost nezavisne promenljive X_j u i-tom članu uzorka, $i = 1, \dots, n$;

$\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$;

$\boldsymbol{p} = \left(\pi(\boldsymbol{x_1}), \pi(\boldsymbol{x_2}) \dots, \pi(\boldsymbol{x_n})\right)^T$;

$$\mathrm{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}_{n*(k+1)} .$$

Tada je $\pi(\boldsymbol{x_i}) = \frac{e^{{x_i}^T \beta}}{1+e^{{x_i}^T \beta}}$ i sistem ima $k + 1$ jednačinu:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} [y_i - \pi(\boldsymbol{x_i})] = 0 \quad (1.2)$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = 0, j = 1, \dots, k \;\Leftrightarrow\; \sum_{i=1}^{n} x_{ji}[y_i - \pi(\boldsymbol{x_i})] = 0, j = 1, \dots k. \quad (1.3)$$

Gornji sistem se može napisati u ekvivalentnom matričnom obliku:

$$X^T(\boldsymbol{y} - \boldsymbol{p}) = \boldsymbol{0}. \quad (1.4)$$

Sistem (1.4) ima $k + 1$ nelinearnu jednačinu i rešava se nekim iterativnim postupakom. Rešenja $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ predstavljaju ocene nepoznatih parametara $\beta_0, \beta_1, \dots, \beta_k$. $\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}}}$ se naziva očekivana (predviđena) vrednost promenljive Y ako su nezavisne promenljiva $X_1, \dots, X_k$ uzele vrednost $x_{1i}, \dots, x_{ki}$, respektivno. Prva jednačina sistema obezbeđuje da su ocene parametra takve da je suma realizovanih vrednosti za Y jednaka sumi predviđenih vrednosti na osnovu modela. Kako bi se odredili drugi parcijalni izvodi, računa se $\frac{\partial \pi(\boldsymbol{x_i})}{\partial \beta_j}$:

$$\frac{\partial \pi(\boldsymbol{x_i})}{\partial \beta_j} = \frac{\partial\left(\frac{e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}}{1+e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}}\right)}{\partial \beta_j} = \frac{x_{ji} e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}} \left(1+e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}\right) - x_{ji}\left(e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}\right)^2}{\left(1+e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}\right)^2}$$

$$\frac{\partial \pi(\boldsymbol{x_i})}{\partial \beta_j} = \frac{x_{ji} e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}}{\left(1+e^{\beta_0 + \sum_{l=1}^{k} \beta_l x_{li}}\right)^2} = x_{ji}\pi(x_i)(1 - \pi(x_i)) \quad (1.5)$$

Iz jednačina (1.2) i (1.3), uz korišćenje (1.5), sledi da su drugi parcijalni izvodi:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j{}^2} = -\sum_{i=1}^{n} x_{ji}^2 \pi(\boldsymbol{x_i})\big(1-\pi(\boldsymbol{x_i})\big), j = 0, \dots k$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \beta_p} = -\sum_{i=1}^{n} x_{ji} x_{pi} \pi(\boldsymbol{x_i})\big(1-\pi(\boldsymbol{x_i})\big), j = 0, \dots k, p = 0, \dots k.$$

Ukoliko se definiše matrica $V = diag\left(\pi(\boldsymbol{x_i})\big(1-\pi(\boldsymbol{x_i})\big)\right)_{n*n}$ iz izvedenih drugih parcijalnih izvoda sledi da je Hesijan funkcije $L(\boldsymbol{\beta})$:

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -X^T V X.$$

Da bi rešenja sistema bile ocene parametara Hesijan funkcije $L(\boldsymbol{\beta})$ mora bit negativno definitna matrica. Matrica $X^T V X$ naziva se informaciona matrica i ona se može dobiti direktno iz uzorka. Za dovoljno veliko n važi da $\widehat{\boldsymbol{\beta}}: N\left(\boldsymbol{\beta}, \left(-\frac{\partial^2 L(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}^2}\right)^{-1}\right)$, odnosno inverzna matrica informacione matrice na dijagonali sadrži standardne devijacije $\sigma^2\big(\widehat{\beta_\iota}\big), i = 0, \dots, k,$ dok su vandijagonalni elementi kovarijacije $cov\big(\widehat{\beta_\iota}, \widehat{\beta_j}\big), i = 0, \dots k, j \epsilon \{0, \dots, k\} \backslash \{i\}$. Više o navedenom se može naći u [1].

1.2 Testiranje značajnosti koeficijenata

U cilju stvaranja najekonomičnijeg modela za predviđanje kategorijalne promenljive na osnovu skupa nezavisnih promenljivih koriste se testovi značajnosti koeficijenata. Pomoću testova se ocenjuje da li je nezavisna promenljiva u modelu statistički značajna, odnosno da li model koji sadrži nezavisnu promenljivu bolje opisuje rezultirajuću promenljivu od modela koji je ne sadrži. Ukoliko se utvrdi da promenljiva nije statistički značajna izbacuje se iz modela i na taj način se model rasterećuje od nepotrebnih nezavisnih promenljivih. Kako bi se utvrdilo da li je neka promenljiva statistički značajna, prvi korak jeste napraviti dva modela. Model koji sadrži promenljivu i model koji je ne sadrži. Sledeći korak je upoređivanje registrovanih vrednosti sa predviđenim vrednostima na osnovu dva kreirana modela. Ukoliko su predviđene vrednosti modela koji sadrži promenljivu bolje od onih dobijenih na osnovu modela koji je ne sadrži,

zaključujemo da je promenljiva statistički značajna. Literatura korišćena za izradu ovog dela rada je [2], [3] i [13].

1.2.1 Test količnika verodostojnosti

Sam naziv Test količnika verodostojnosti ukazuje na to da ovaj test koristi verodostojnost za poređenje registrovanih i predviđenih vrednosti na osnovu modela, tačnije koristi sledeći izraz:

$$D = -2\ln\left(\frac{verodostojnost\ fitovanog\ modela}{verodostojnost\ zasićenog\ modela}\right). \quad (1.6)$$

Pod izrazom verodostojnost modela podrazumeva se vrednost funkcije verodostojnosti datog modela u ocenjenim vrednostima nepoznatih parametara, dok je zasićen model onaj koji sadrži sve dostupne nezavisne promenljive. U prethodnom delu teksta izedeno je da funkcija verodostojnosti ima oblik $l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}$, odakle sledi da je:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}(x_i)}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}(x_i)}{1-y_i}\right)\right] \quad (1.7)$$

i statistika (1.7) se naziva odstupanje. Kako bi se odredilo da li je promenljiva u modelu značajna, računa se promena u odstupanju izazvana uključivanjem nezavisne promenljive u model (G):

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

Sređivanjem gornjeg izraza, dolazi se do statistike:

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

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

$$G = -2ln\left(\frac{verodostojnost\ modela\ bez\ nezavisne\ promenljive}{verodostojnost\ modela\ sa\ nezavisnom\ promenljivom}\right) \quad (1.8)$$

Posmatra se slučaj kada je boj nezavisnih promenljivih u modelu jedan. Model bez nezavisne promenljive će tada biti $\pi_*(x_i) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ i postoji samo jedan nepoznat parametar. Ocena parametra β_0 metodom maksimalne verodostojnosti se dobija rešavanjem jednačine:

$$\sum_{i=1}^{n}\left[y_i - \frac{e^{\beta_0}}{1+e^{\beta_0}}\right] = 0 \Leftrightarrow\ e^{\beta_0}(n - \sum_{i=1}^{n} y_i) = \sum_{i=1}^{n} y_i\ .$$

Sledi da je $\hat{\beta}_0 = ln\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}$ i $\hat{\pi}_*(x_i) = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}$, odnosno predviđena vrednost je konstantna. Ukoliko se ovi rezultati ubace u statistiku 1.8, dobija se:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_i}(1-\hat{\pi}(x_i))^{1-y_i}}\right], \text{ gde } n_0 = \sum y_i\,, n_1 = \sum_{i=1}^{n}(1-y_i)\,,$$

odnosno

$$G = 2[\sum_{i=1}^{n}[y_i\, ln\,\hat{\pi}(x_i) + (1-y_i)\, ln\,\hat{\pi}(x_i)] - [n_1\, ln(n_1)\, + n_0\, ln(n_0) - nln(n)]].$$

Nulta raspodela stastistike G je G: χ_1^2. Kod definisanja veličine odstupanja u formuli 1.6 posmatrao se logaritam količnika verodostojnosti pomnožen sa konstantom -2 upravo iz razloga da bi se dobila statistika G čija je raspodela poznata.

Sumirano, posmatra se hipoteza $H_0(\beta_1 = 0)$ protiv $H_A(\beta_1 \neq 0)$, na osnovu datog uzorka $(x_i,\, y_i), i = 1, \dots, n$. Test statistika je G koja kao nultu raspodelu ima χ_1^2. Velika vrednost test statistike ukazuje da postoji značajna razlika među modelima i da hipoteza H_0 nije tačna, te je test

$$\alpha = P_{H_0}\{G > c\}.$$

α predstavlja verovatnoću greške prve vrste i njenu vrednost bira sam istraživač, dok je c kvantil reda $1-\alpha$ raspodele χ_1^2. Može se desiti da zbog slučajnosti uzorka odbacimo hipotezu H_0, ali je hipoteza H_0 stvarno tačna i ova greška se naziva greška prve vrste. Neka je g_{reg} registrovana vrednost statistike G, tada se zaključak donosi na sledeći način:

- ako je registrovana vrednost g_{reg} statistike G van interval $[c, \infty)$, ne postoji razloga za odbacivanje hipoteze H_0;
- ako je registrovana vrednost g_{reg} statistike G element interval $[c, \infty)$, odbacuje se hipoteza H_0.

1.2.2 Wald test

Pored testa količnika verodostojnosti, Wald test je još jedan od testova kojima se određuje da li je nezavisna promenljiva značajna. U slučaju univarijantne logističke regresije posmatra se hipoteza $H_0(\beta_1 = 0)$ protiv $H_A(\beta_1 \neq 0)$. Test statistika kod Wald testa je kvadrat količnika ocene za

parametar β_1 dobijenog metodo maksimalne verodostojnosti $(\hat{\beta}_1)$ i ocene standardne disperzije od $\hat{\beta}_1$:

$$Z^2 = \left(\frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}\right)^2$$

Ako je hipoteza $H_0(\beta_1 = 0)$ tačna statistika $\frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}$ ima približno $N(0,1)$ raspodelu, pa statistika Z^2 ima približno hi-kvadratnu raspodelu sa jednim stepenom slobode. Test se sprovodi tako što se računa p-vrednost test statistike i ukoliko je ta vrednost manja od greške prve vrste H_0 se odbacuje, odnosno zaključuje se da je nezavisna promenljiva u modelu značajna. Međutim, nedostatak Wald testa je što se kod manjih uzorka često prihvata nulta hipoteza iako je nezavisna promenljiva značajna. Zbog toga je preporučljivo koristiti ipak test količnika verodostojnosti.

U slučaju multivarijabilne logističke regresije testrira se nulta hipoteza $H_0(\beta_1 = 0, \dots, \beta_k = 0)$ protiv alternativne hipoteze da postoji koeficijent u modelu koji je različit od nule. Test statistika koja se koristi je

$$W = \overline{\boldsymbol{\beta}}(\bar{X}^T V \bar{X}), pri\ čemu$$

$\overline{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_k)^T$

$$\bar{X} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}_{n*k}.$$

Ako je hipoteza $H_0(\beta_1 = 0, \dots, \beta_k = 0)$ tačna, statistika W ima približno hi-kvadrat raspodelu sa p stepeni slobode.

1.3 Interpretacija logističkog regresionog modela

Nakod određivanja najekonomičnijeg modela, odnosno ocenjivanja nepoznatih parametara i ostavljanja u modelu onih nezavisnih promenljivih koje su statistički značajne, prelazi se na interpretaciju dobijenog modela. Jedan od najznačajnih koraka u sprovođenju statističke analize jeste izvođenje pravilnih zaključaka na osnovu dobijenih ocena koeficijenata u modelu, što

zapravo predstavlja njegovu interpretaciju. Prilikom interpretacije logističkog modela treba rešiti sledeća pitanja:

1. Koja je funkcionalna veza između zavisne i nezavisne promenljive?
2. Koja je odgovarajuća jedinica promene za nezavisnu promenljivu?

Odgovor na prvo pitanje je logit funkcija. Pokazano je da ukoliko se posmatra logaritam šanse da zavisna prmenljiva uzme vrednost 1 ukoliko su nezavisne promenljive uzele vrednost **x** jeste linearna funkcija, tj.

$$g(\boldsymbol{x}) = \ln \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

Radi jednostavnije interpretacije, na dalje se pretpostavlja da se radi o univarijantnom logističkom modelu. Odgovor na drugo pitanje zavisi od toga da li je nezavisna promenljiva kategorijalna ili je neprekidna. Shodno tome, deficisaće se odgovarajuća promena za nezavisnu promenljivu.

1.3.1 Dihotomna nezavisna promenljiva u univarijantnom logističkom modelu

Dihotomnu nezavisnu prmenljivu kodiramo sa 0 i 1. Ukoliko se posmatra razlika logit funkcije u ove dve vrednosti dobija se:

$$g(1) - g(0) = \beta_1,$$

odnosno koeficijent β_1 predstavlja promenu logita po jedinici promene nazavisne promenljive. Kada je nezavisna promenljiva dihotomna, tada se pomoću logističkog modela modeliraju sledeće verovatnoće:

- $P\{Y = 1|X = 0\} = \pi(0) = \frac{e^{\beta_0+\beta_1*0}}{1+e^{\beta_0+\beta_1*0}}$
- $P\{Y = 0|X = 0\} = 1 - \ P\{Y = 1|X = 0\} = \frac{1}{1+e^{\beta_0}}$
- $P\{Y = 1|X = 1\} = \pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$
- $P\{Y = 1|X = 0\} = 1 - \ \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$

Šansa da zavisna promenljiva uzme vrednost 1 ako nazavisna promenljiva uzme vrednost 0 je $\frac{P\{Y=1|X=0\}}{1-P\{Y=1|X=0\}} = \frac{\pi(0)}{1-\pi(0)}$. S druge strane, šansa da zavisna promenljiva uzme vrednost 1 ako nezavisna promenljiva uzme vrednost 1 je $\frac{P\{Y=1|X=1\}}{1-P\{Y=1|X=1\}} = \frac{\pi(1)}{1-\pi(1)}$. Odnos šansi (OR) se definiše

kao odnos šanse da zavisna promenljiva uzme vrednost 1 ukoliko je nezavisna uzela 1 i šanse da zavisna promenljiva uzme vrednost 1 ukoliko je nezavisna uzela 0:

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

Odnos šansi pokazuje koliko je puta veća (ili manja) šansa da zavisna promenljiva uzme vrednost 1 među pojavama sa vrednošću nezavisne promenljive 1 od šanse da zavisna promenljiva uzme vrednost 1 među pojavama sa vrednošću nezavisne promenljive 0.

Daljim sređivanjem izraza OR dobija se povezanost između odnosa šansi i koeficijenta β_1:

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}}}{\frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}}} = e^{\beta_1}.$$

Odnosno, ocenjen odnos šansi je

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

Neka Y predstavlja prisustvo ili odsustvo alergije na polen, a X označava da li je osoba sa sela ili nije i neka je ocenjen odnos šansi 2. Tada bi se ovakav rezultat interpretirao na sledeći način: dvostruko je veća šansa da osoba ima alergiju na polen među stanovnicima sela nego među stanovnicima grada u posmatranoj populaciji. Kao referentna vrednost je uzeto prisustvo alergije kod stanovnika grada. Kada bi se za referentnu vrednost uzelo prisustvo alergije kod stanovnika sela, dobilo bi se da je odnos šansi 0.5, što se interpretira da je duplo manja šansa da osoba ima alergiju na polen među stanovnicima grada nego među stanovnicima sela u posmatranoj populaciji. Dakle, bez obzira šta se uzima za referentnu vrednost zaključak je isti.

Relativni rizik (RR) se definiše kao odnos verovatnoće da zavisna promenljiva uzme vrednost 1 kada nezavisna uzme vrednost 1 i verovatnoće da zavisna promenljiva uzme vrednost 1 kada nezavisna uzme vrednost 0:

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

Relativni rizik pokazuje koliko je puta veća (ili manja) verovatnoća uspeha među pojavama sa vrednošću nezavisne promenljive 1 od verovatnoće uspeha među pojavama sa vrednošću nezavisne promenljive 0. Lako se iz definicija za odnos šansi i relativnog rizika zaključuje da je:

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

Kako $1 - \pi(0)$ predstavlja verovatnoću neuspeha kada nezavisna promenljiva uzme vrednost 0, a $1 - \pi(1)$ verovatnoću neuspeha kada nezavisna promenljiva uzme vrednost 1, sledi da ukoliko su verovatnoće neuspeha u okviru obe grube približno jednake odnos šansi je približno jednak relativnom riziku. Podjednaka verovatnoću neuspeha u obe grupe se pojavljuje kod ispitivanja prisustva bolesti koje su same po sebi retke.

Primer *(videti u [2]):*

U tabeli 3 su dati podaci o broj preživelih putnika na Titanku prema polu. Ukupno je bilo 1313 putnika, od toga 462 žene i 851 muškarac.

Tabela 3: Podaci o broju preživelih putnika na titaniku prema polu

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

Računa se odnos šansi i relativni rizik, posle čega se daje tumačenje dobijenih vrednosti.

Odnos šansi:

$$\widehat{OR} = \frac{\text{šansa za smrt kod muškaraca}}{\text{šansa za smrt kod žena}} = \frac{(708/851)/(142/851)}{(154/462)/(308/462)} = \frac{4.993}{0.5} = 9.986.$$

Šansa da muškarac pogine je skoro deset puta veća od šanse da žena pogine na Titaniku. S druge strane, relativni rizik je odnos verovatnoće za smrt kod muškaraca i žena:

$$RR = \frac{verovatnoća\ smrti\ kod\ muškaaca}{verovatnoća\ smrti\ kod\ žena} = \frac{709/851}{154/462} = \frac{0.8333}{0.3333} = 2.5.$$

Dakle, verovatnoća smrti kod muškaraca je 2.5 puta veća od verovatnoće smrti kod žena.

Zaključuje se da su i šansa i verovatnoća za smrt muškaraca veće nego kod žena. Mera relativnog rizika je približnija ljudskoj prirodi i intuitivno prihvatljiva. Ukoliko u prvoj grupi šansa za smrt iznosi 20% i u drugoj 80%, relativni rizik je 4 što je i intuitivno jasno. Međutim, odnos šansi iznosi čak $\frac{4}{0.25} = 16.$

1.3.2 Polihotomna nezavisna promenljiva u univarijantnom logističkom modelu

Posmatra se slučaj kada nezavisna promenljiva uzima više od dve vrednosti. Neka je zavisna promenljiva Oštećenje koja uzima vrednost 0 kada biljka nije oštećenja, a u suprotnom ima vrednost 1. Analizira se uticaj promenljive Tretman koja predstavlja vrstu insekticida koji je korišćen u zaštiti biljaka od žičara i ima šest kategorija (tabela 4). Da bi se izvršila pravilna interpretacija logističkog modela potrebno je odrediti referentnu kategoriju nezavisne promenljive. Neka je uzeta kategorija insekticida Attracap kao referentna, tada se kodiranje tretmana vrši na način prikazan u tabeli 5.

Tabela 4: Polihotomna nezavisna promenljiva

	Tretman					
Oštećenje	Attracap	Force 20CS	Force 1.5G	Buteo Start	Lumiposa	Sonido
0	99	111	104	89	95	25
1	41	29	36	51	45	25
Ukupno	140	140	140	140	140	50
OR	0	0.63	0.84	1.4	1.15	2.44

Tabela 5: Kodiranje promenljive Tretman

	Kodiranje Tretmana				
	1(Force 20CS)	2(Force 1.5G)	3(Buteo Start)	4(Lumiposa)	5(Sonido)
Attracap	.000	.000	.000	.000	.000
Force 20CS	1.000	.000	.000	.000	.000
Force 1.5G	.000	1.000	.000	.000	.000
Buteo Start	.000	.000	1.000	.000	.000
Lumiposa	.000	.000	.000	1.000	.000
Sonido	.000	.000	.000	.000	1.000

Na osnovu tabele 4 može se izračunati šansa za pojavom oštećenja prilikom korišćenja insekticida Attracap:

$$\frac{P\{Oštećenje=1|Attracap\}}{P\{Oštećenje=0|Attracap\}} = \frac{41}{99} = 0.41.$$

Odnos šansi za tretman Force 20CS predstavlja količnik šanse da se zabeleži oštećenje prilikom korišćenja Force 20CS i šanse oštećenja prilikom korišćenja Attracap:

$$OR(Force\ 20CS, Attracap) = \frac{\frac{P\{Oštećenje=1|Force\ 20CS\}}{P\{Oštećenje=0|Force\ 20CS\}}}{0.41} = \frac{\frac{29}{111}}{0.41} = 0.63.$$

Dakle, Force 20CS ima 1.59 puta manju šansu za pojavom oštećenja nego Attracap. Za svaki nereferentni tretman je u tabeli 4 prikazano koliko je puta veća šansa za pojavom oštećenja prilikom njegovog korišćenja nego primenom insekticida Attracap. Slično kao i kod dihotomne nezavisne promenljive, prirodni logaritam ocenjenog odnosa šansi posmatranog nereferentnog tretmana jednak je ocenjenom koeficijentu uz posmatran tretman.

$$\begin{aligned}\ln\left(\widehat{OR}(Force\ 20CS, Attracap)\right) &= \hat{g}(Force\ 20CS) - \hat{g}(Attracap)\\ &= \widehat{\beta_0} + \widehat{\beta_1}(Force\ 20CS = 1) + \widehat{\beta_2}(Force\ 1.5G = 0) + \widehat{\beta_3}(Buteo\ Start = 0)\\ &+ \widehat{\beta_4}(Lumiposa = 0) + \widehat{\beta_5}(Sonido = 0)\\ &- \left(\widehat{\beta_0} + \widehat{\beta_1}(Force\ 20CS = 0) + \widehat{\beta_2}(Force\ 1.5G = 0) + \widehat{\beta_3}(Buteo\ Start = 0)\right.\\ &\left. + \widehat{\beta_4}(Lumiposa = 0) + \widehat{\beta_5}(Sonido = 0)\right) = \widehat{\beta_1}\end{aligned}$$

Odnosno, $\widehat{\beta_1} = ln(0.63) = -0.462$. Analogno, dolazi se do sledećih zaključaka:

$$\widehat{\beta_2} = ln(0.84) = -0.17, \widehat{\beta_3} = ln(1.4) = 0.34, \widehat{\beta_4} = ln(1.15) = 0.14, \widehat{\beta_5} = ln(2.44) = 0.89.$$

1.3.2 Neprekidna nezavisna promenljiva u univarijantnom logističkom modelu

U slučaju neprekidne nezavisne promenljive nije dobro definisati promenu nezavisne promenljive od jedne jedinice. U slučajevima kada nezavisna promenljiva uzima vrednosti između 0 i 1, posmatranje promene od jedne jedinice je suviše veliko. Nasuprot tome, ukoliko nezavisna promenljiva ima veoma veliki raspon posmatranje njene promene od jedne jedinice nije od koristi. Zbog toga se u slučaju neprekidne nezavisne promenljive posmatra promena logita prilikom promene nezavisne promenljive za c jedinica:

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

Sada se odnos šansi (OR) definiše kao odnos šanse da zavisna promenljiva uzme vrednost 1 ukoliko je nezavisna uzela $x + c$ i šanse da zavisna promenljiva uzme vrednost 1 ukoliko je nezavisna uzela x:

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

Odnosno, ocenjen odnos šansi je

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

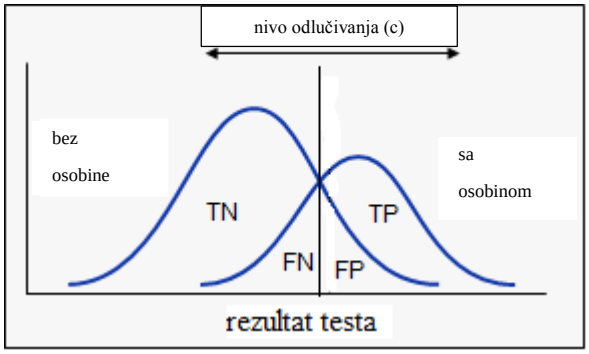
Dakle, odnos šansi zavisi od definisane promene nezavisne promenljive. Neka zavisna promenljiva Y predstavlja prisustvo ili odsustvo bolesti KSO i X predstavlja starost. Ako je ocenjen odnos šansi $\widehat{OR} = e^{0.038c}$, pitanje je kako se sada interpretira dobijeni rezultat. Ukoliko je interesovanje kako povećanje starosti za 5 godina utiče na prisustvo bolesti KSO, računamo odnos šansi za $c = 5$: $\widehat{OR}(5) = e^{0.038*5} = 1.21$. Dobijenu vrednost se interpretira na sledeći način: svako povećanje starosti za 5 godina povećava i šansu oboljenja od KSO za 1.21 put.

1.4 Procena slaganja modela sa podacima

1.4.1 Tabela klasifikacije i ROC kriva

Tabela klasifikacije se koristi kada je zavisna promenljiva dihotomna. U njoj se vrši upoređivanje registrovane vrednosti zavisne promenljive i njene predviđene vrednosti na osnovu modela. Logističkim modelom dobija se ocena $\hat{\pi}(\boldsymbol{x_i})$, odnosno verovatnoća da zavisna promenljiva uzme vrednost 1 ako su nezavise uzele vrednost $\boldsymbol{x_i}$. Da bi se izvršilo predviđanje vrednosti zavisne promenljive definiše se cut-off verovatnoća c, koju bira sam istraživač i najčešća njena vrednost je 0.5. Ukoliko je logistička verovatnoća veća od c predviđa se da je vrednost zavisne promenljive (u oznaci $\hat{y}_i$) 1, a u suprotnom $\hat{y}_i = 0$. Odnosno, ako je $\hat{\pi}(x_i) > c$ uzima se da je $\hat{y}_i = 1$, a ukoliko

je $\hat{\pi}(x_i) \leq c$ uzima se da je $\hat{y}_i = 0$. Ukoliko se posmatraju rezultati klasifikacije samo na deo uzorka sa prisutnom osobinom, neće se dobiti idealno klasifikovanje i postojaće opservacije koje su prema testu bez prisustva osobine. Slično, ukoliko se posmatraju rezultati testa samo na deo uzorka bez posmatrane osobine, postojaće instance koje su prema testu sa prisutnom osobinom. Odnosno, ukoliko se na osnovu logističkog modela klasifikuju subjekti na one sa posmatranom osobinom i bez nje, biće pojava koje su na osnovu modela pogrešno klasifikovane. Postojaće subjekti kod kojih se na osnovu modela beleži pojava osobine iako kod njih ta osobina nije registrovana, ali će takođe i postojati subjekti koji će na osnovu modela biti svrstani u grupu bez prisutne osobine iako je kod njih ona registrovana (slika 3).



Slika 3:Pogrešna klasifikacija zavisne promenljive na osnovu modela

Deo označen sa TN (True Negative fraction) predstavlja one slučajeve među observacijama bez prisustva osobine koji su tačno klasifikovni pomoću testa. FN (False Negative fraction) je deo pojava sa prisutnom osobinom koji je netačno klasifikovan. FP (false positive fraction) predstavlja slučajeve među pojavama bez osobine koji su pogrešno klasifikovani, dok je TP (True Positive fraction) deo pojava sa prisutnom osobinom koji je tačno klasifikovan.

Dat je model za ocenjivanje verovatnoće da je osoba gojazna. Ukoliko se odabere $c = 0.5$ i za svaku osobu od 2104 posmatranih proceni prisustvo ili odsustvo gojaznosti na osnovu modela, dobija se sledeća tabela klasifikacije (videti u [1]):

Tabela 6: Tabela klasifikacije

Test	Registrovano		
	Nisu gojazni ($y_i = 0$)	Gojazni ($y_i = 1$)	Ukupno
Nisu gojazni ($\hat{y}_i = 0$)	530	250	780
Gojazni ($\hat{y}_i = 1$)	385	939	1324
Ukupno	915	1189	2104

Ukupan broj gojaznih ljudi u posmatranoj populaciji je 1189, od toga je model tačno klasifikovao njih 939. Broj ljudi koji nije gojazan je 915, i njih 530 je tačno model klasifikovao. Prirodno se nameće pitanje kolika je verovatnoća da je predviđena vrednost za registrovano gojazne ljude da su oni gojazni, a sa druge strane kolika je verovatnoća da je predviđena vrednost negojaznost za ljude koji zaista nisu gojazni. Ove verovatnoće imaju poseban naziv:

- Verovatnoća da je predviđena vrednost zavisne promenljive 1 ukoliko je njena registrovana vrednost 1 naziva se senzitivnost testa. Matematički zapisano, $P\{\hat{y} = 1|y = 1\}$ je senzitivnost testa;
- Verovatnoća da je predviđena vrednost zavisne promenljive 0 ukoliko je njena registrovana vrednost 0 naziva se specifičnost testa. Matematički zapisano, $P\{\hat{y} = 0|y = 0\}$ je specifičnost testa.

Analizira se sada specifičnost i senzitivnost za datu tabelu klasifikacije. Senzitivnost testa iznosi:

$$P\{osoba\ je\ klasifikovana\ kao\ gojazna|osoba\ je\ gojazna\} = \frac{939}{1189} = 78.97\%.$$

Model će 78.97% gojaznih ljudi i klasifikovati kao gojazne, dok će među gojaznima njih 21.03% pogrešno klasifikovati (kao negojazne). Specifičnost testa iznosi:

$$P\{osoba\ je\ klasifikovana\ kao\ negojazna|osoba\ je\ negojazna\} = \frac{530}{915} = 57.92\%.$$

U grupi negojaznih ljudi, model će njih 57.92% tačno klasifikovati. Kada se izračuna tačnost klasifikacije po grupa, računa se i ukupna stopa tačne klasifikacije modela kao:

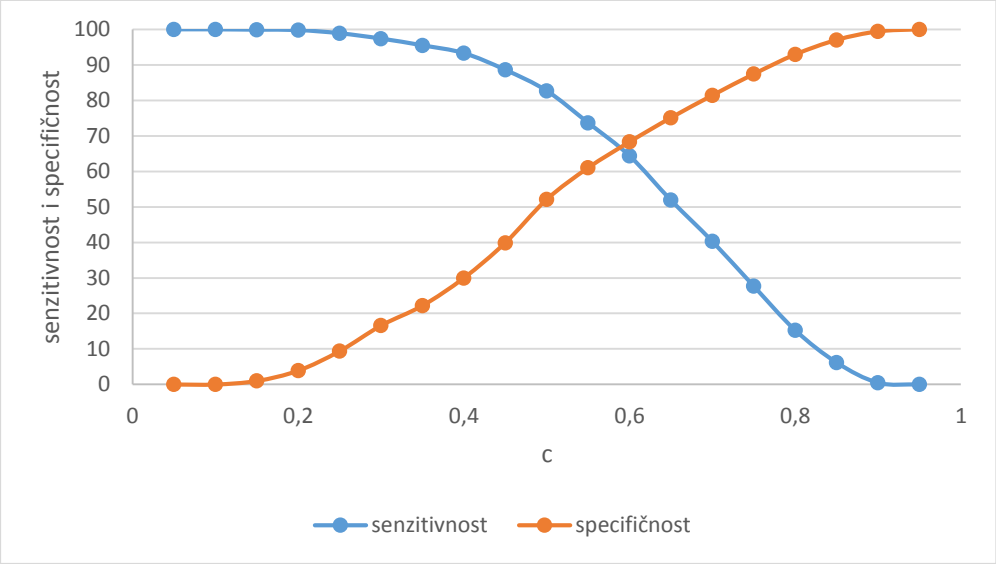
$$\frac{939+530}{2104} = 69.82\%,$$

odnosno ukupno je pogrešno klasifikovano 30.18% posmatranih osoba.

Značajno je napomenuti da ocenjene verovatnoće na osnovu model zavise od cut-off verovatnoće. Samim tim za različite vrednosti cut-off verovatnoće dobiće se različite vrednosti senzitivnnosti i specifičnosti. Ukoliko je cilj sprovođenja analize što uspešnija klasifikacija, izabraće se cut-off verovatnoća koja maksimizira oba pokazatelja, i senzitivnost i specifičnost. Za model gojaznosti vrednosti senzitivnosti i specifičnosti za različite izbore cut-off verovatnoće prikazane su u tabeli 7 i grafički na slici 4. Na osnovu grafika zaključak je da se maksimum senzitivnost i specifičnost dostiže prilikom izbora nivoa odlučivanja od 0.6.

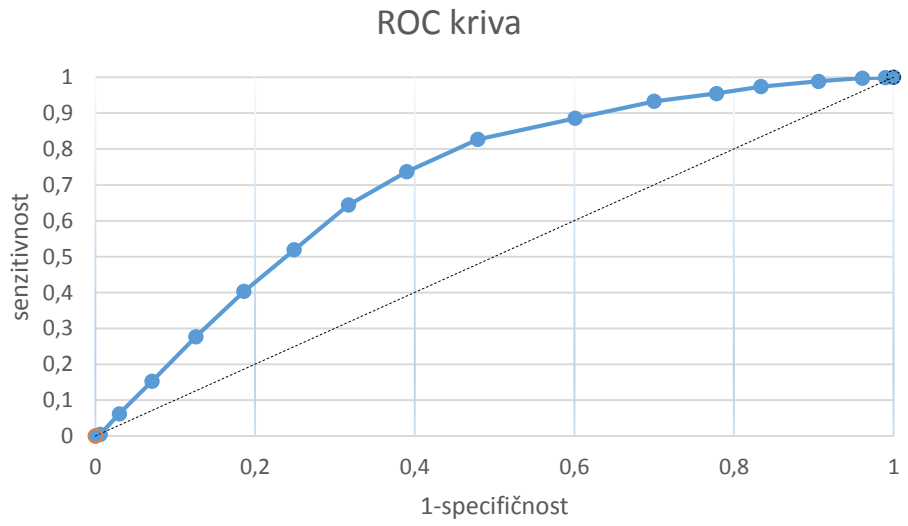
Tabela 7: Zavisnost senzitivnosti i specifičnosti od cut-off verovatnoće

c	senzitivnost (%)	specifičnost (%)	1 - specifičnost (%)
0.05	100	0	100
0.1	100	0	100
0.15	99.9	1	99
0.2	99.8	3.9	96.1
0.25	98.9	9.4	90.6
0.3	97.4	16.6	83.4
0.35	95.5	22.2	77.8
0.4	93.3	30	70
0.45	88.6	39.9	60.1
0.5	82.7	52.1	47.9
0.55	73.7	61	39
0.6	64.4	68.3	31.7
0.65	51.9	75.1	24.9
0.7	40.3	81.4	18.6
0.75	27.7	87.4	12.6
0.8	15.3	92.9	7.1
0.85	6.2	97	3
0.9	0.5	99.4	0.6
0.95	0	100	0



Slika 4: Grafički prikaz zavisnosti senzitivnosti i specifičnosti od cut verovatnoće

Na osnovu podataka iz tabele 7 može se doći do krive na slici 5. Ta kriva se naziva Receiver Operating Characteristic (ROC) kriva i povezuje verovatnoću tačne klasifikacije osoba sa posmatranom osobinom ($P\{\hat{y} = 1|y = 1\}$) i verovatnoću netačne klasifikacije osoba bez prisustva osobine ($P\{\hat{y} = 1|y = 0\}$), odnosno povezuje $senzitivnost$ i $1 - specifičnost$.



Slika 5: ROC kriva

Površina ispod ROC krive (AUC – The area under the curve) je mera sposobnosti modela u razdvajanju subjekata koji imaju posmatranu osobinu od onih koji je nemaju. U konkretnom slučaju modela gojaznosti ta površina predstavlja verovatnoću da osobe koje su gojazne imaju veću ocenjenu verovatnoću od onih kod kojih to nije slučaj. Prilikom ocene uspešnosti klasifikacije modela koristi se AUC (videti u [2]) po pravilu datom u tabeli 8.

Tabela 8: Ocena uspešnosti klasifikacije na osnovu ROC krive

AUC	Ocena uspešnosti klasifikacije
0.5	nema diskriminacije
(0.5, 0.7)	loše razdvajanje
[0.7, 0.8)	prihvatljivo razdvajanje
[0.8,0.9)	odlično razdvajanje
≥ 0.9	izvanredno razdvajanje

Test senzitivnosti i specifičnosti zavisi od raspodele verovatnoća u uzorku, te ukoliko se primeni isti model na dve različite populacije, pomoću pomenutih testova bi se mogao dobiti različit zaključak o dobrobiti modela. Takođe, pokazatelji koji se dobijaju iz tabele klasifikacije su veoma osetljivi na relative veličinu dve grupe (grupa sa prisustvom i bez prisustva posmatrane pojave) i uvek je u prednosti grupa kojaima veći broj članova. Korišćena literatura za izradu podnaslova 1.4.1 je [1] i [2].

1.4.2 Hosmer-Lemeshow test

Kod procene slaganja modela sa podacima posmatra se razlika registrovanih i predviđenih vrednosti zavisne promenljive. Neka je $y = (y_1, \dots, y_n)^T$ vektor registrovanih i $\hat{y} = (\hat{y}_1, \dots, \hat{y}_n)^T$ vektor predviđenih vrednosti promenljive na osnovu modela. Cilj je da rastojanje između vektora y i $\hat{y}$ bude malo i da svaki par $(y_i, \hat{y}_i)$ nema veliki doprinos posmatranoj razlici. Kod test statistike za Hosmer-Lemeshow test koristi se statistika koja je m-asimptotska, te se u nastavku pre izvođenja testa definiše m-asimptotska raspodela statistike. Covariate patterns su opservacije sa istim vrednostima za sve nezavisne promenljive. Na primer, ako promenljiva X_1 predstavlja pol, a promenljiva X_2 visinu i posmatramo 10 subjekata:

$$(\check{z}enski, 165), (\check{z}enski, 165), (\check{z}enski, 167), (\check{z}enski, 168), (\check{z}enski, 172),$$

$$(muški, 173), (muški, 175), (muški, 179), (muški, 187), (muški, 190),$$

tada postoji devet covariate patterns jer postoje dve opservacije sa istim vrednostima nezavisnih promenljivih $(\check{z}enski, 165), (\check{z}enski, 165)$ i te dve opservacije čine jedan covariate pattern. Neka je broj nezavisnih promenljivih k i J broj kovarijantnih obrazaca. Jasno je da je broj kovarijantih obrazaca jedank broju različitih registrovanih vrednosti vektora nezavisnih promneljivih. Uvodi se oznaka $m_j, j = 1, \dots, J$, koja označava broj subjekata u kovarijantnom obrascu j. Opservacije su grupisane tako da se u jednoj grupi nalaze svi subjekti sa istim vrednostima nezavisnih promenljivih i za svaku grupu se odredio broj članova, jasno je da je zbir broja članova u svakoj grupi jednak broju opservacija: $\sum_{j=1}^{J} m_j = J$. Ako se fiksira broj kovarijantih obrazaca i povećava obim uzoraka tada će se broj članova svake grupe povećavati. M-asimptotske raspodele jesu one kod kojih $m_j, j = 1, .., n$, postaje veliko.

Kod Hosmer-Lemeshovog testa vrši se grupisanje uzorka na osnovu ocenjenih verovatnoća $\hat{\pi}(\boldsymbol{x_1}), \dots, \hat{\pi}(\boldsymbol{x_n})$. Neka G predstavlja fiksiran broj grupa (najčešće $G = 10$) i n obim uzorka. Prva grupa sadrži najmanjih n/G vrednosti ocenjenih verovatnoća, druga grupa sadrži sledećih najmanjih n/G ocenjenih verovatnoća i tako se forimara G grupa. Za svaku grupu računamo:

- $o_{1g} = \sum_{k=1}^{n_g} y_k$ – broj subjekata sa prisutnom osobinom $(Y = 1)$;
- $o_{0g} = \sum_{k=1}^{n_g} (1 - y_k)$ – broj subjekata u grupi sa bez posmatrane osobine $(Y = 0)$;
- $e_{1g} = \sum_{k=1}^{n_g} \hat{\pi}(\boldsymbol{x_k})$ – očekivani broj subjekata u grupi sa posmatranom osobinom;
- $e_{0g} = \sum_{k=1}^{n_g} \left(1 - \hat{\pi}(\boldsymbol{x_k})\right)$ – očekivani broj subjekata u grupi bez posmatrane osobine,

gde n_g predstavlja broj subjekata u posmatranoj g-toj grupi, $g = 1, \dots, G$ (tabela 9).

Tabela 9: Broj registrovanih i očekivanih subjekata u okviru svake grupe

Posmatrana osobina		Grupa				Ukupno
		1	2	…	G	
		Presečne verovatnoće				
		$\hat{\pi}(\boldsymbol{x_k}) \epsilon [\hat{\pi}_{L1}, \hat{\pi}_{U1}]$	$\hat{\pi}(\boldsymbol{x_k}) \epsilon [\hat{\pi}_{L2}, \hat{\pi}_{U2}]$	…	$\hat{\pi}(\boldsymbol{x_k}) \epsilon [\hat{\pi}_{LG}, \hat{\pi}_{UG}]$	
Pristutna $(Y = 1)$	Registrovano	o_{11}	o_{12}	…	o_{1G}	o_1
	Očekivano	e_{11}	e_{12}	…	e_{1G}	e_1
Odsutna $(Y = 0)$	Registrovano	o_{01}	o_{02}	…	o_{0G}	o_0
	Očekivano	e_{01}	e_{02}	…	e_{0G}	e_0
Ukupno (φ)		$\approx n/G$	$\approx n/G$	…	$\approx n/G$	n

$\varphi = Registrovano\ (Y = 1) + Registrovano\ (Y = 0) = Očekivano\ (Y = 1) + Očekivano\ (Y = 1)$

$n/G = o_{1g} + o_{0g} = e_{1g} + e_{0g}, n = o_0 + o_1 = e_0 + e_1$

Na osnovu tabele o broju registorvanih i očekivanih subjekata dobija se test statistika $\hat{C}_G$ Hosmer-Lemeshow testa za procenu slaganja modela sa podacima:

$$\hat{C}_G = \sum_{p=0}^{1} \sum_{g=1}^{G} \frac{(o_{pg} - e_{pg})^2}{e_{pg}}.$$

Statistika $\hat{C}_G$ naziva se Pirsonova hi-kvadrat statistika. Neka je J broj različitih registorvanih vrednosti vektora nezavisih promenljivih. Kada je $J = n$ i $k \leq G$ statistika $\hat{C}_G$ ima približno χ^2 raspodelu sa $G - 2$ stepena slobode. U slučaju kada je $J \approx n$ Hosmer i Lemesh smatraju da se $\hat{C}_G$ takođe može aproksimirati raspodelom χ^2_{G-2}. Statistika $\hat{C}_G$ ima m-asimptotsku raspodelu i ukoliko je njena p-vrednost veća od 0.05 smatra se da se model dobro slaže sa podacima. Više o Hosmer Lemeshow testu se može naći u [6].

2. Multinomna logistička regresija

U prvom delu rada, prikazana je analiza binomne logističke regresije, odnosno logističke regresije kod koje je zavisna promenljiva uzimala dve vrednosti. U ovom delu, cilj je model uopštiti za slučaj kada zavisna promenljiva uzima više od dve kategorije. Logistički regresioni model u kojem zavisna promenljiva ima više od dve kategorijalne vrednosti naziva se multinomna, polihotomna ili polihona logistička regresija. U daljem radu pretpostavlja se da zavisna promenljiva ima tri kategorijalne vrednosti kodirane sa 0, 1 i 2. Problem sa više od tri kategorijalne vrednosti podrazumeva istu analizu, dok se jedino nailazi na problem komplikovanije notacije. Kod binomne logističke regresije definisala se jedna logit funkciju, dok se kod modela sa tri kategorije definišu dve logit funkcije, pri čemu se jedna kategorijalna vrednost zavisne promenljive uzima kao referentna. Pretpostavlja se da postoji k nezavisnih promenljivih $(X_1, X_2, \dots, X_k)$, da se $Y = 0$ uzima kao refernta vrednost zavisne promenljive i definišu se sledeće logit funkcije za konkretne vrednosti zavisnih promenljivih $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$:

$$g_1(\boldsymbol{x}) = \ln\left(\frac{P\{Y=1|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}}\right) = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1k}x_k = \boldsymbol{x^T\beta_1}; \quad (2.1)$$

$$g_2(\boldsymbol{x}) = \ln\left(\frac{P\{Y=2|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}}\right) = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2k}x_k = \boldsymbol{x^T\beta_2}, \quad (2.2)$$

gde $\boldsymbol{x} = (1, x_1, x_2, \dots, x_k)_{k+1}, \boldsymbol{\beta_1} = (\beta_{10}, \beta_{11}, \dots., \beta_{1k})_{k+1}$ i $\boldsymbol{\beta_2} = (\beta_{20}, \beta_{21}, \dots., \beta_{2k})_{k+1}$. Na osnovu logit funkcija (2.1) i (2.2), izvode se uslovne verovatnoće $\pi_j(\boldsymbol{x}) = P\{Y = j|\boldsymbol{x}\}, j = 0, 1, 2,$ kao funkcije od vektora **x** i u kojima konfigurišu vektori nepoznatih parametara $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}$.

$$g_1(\boldsymbol{x}) = \ln\left(\frac{P\{Y=1|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}}\right) \implies e^{g_1(x)} = \frac{P\{Y=1|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}} \quad (2.3)$$

$$g_2(\boldsymbol{x}) = \ln\left(\frac{P\{Y=2|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}}\right) \implies e^{g_2(x)} = \frac{P\{Y=2|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}} \quad (2.4)$$

Sabiranjem levih i desnih strana jednakosti (2.3) i (2.4), dobija se:

$$e^{g_1(x)} + e^{g_2(x)} = \frac{P\{Y=1|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}} + \frac{P\{Y=2|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}} = \frac{1-P\{Y=0|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}} = \frac{1}{P\{Y=0|\boldsymbol{x}\}} - 1,$$

odakle je:

$$\pi_0(\boldsymbol{x}) = P\{Y=0|\boldsymbol{x}\} = \frac{1}{1+e^{g_1(x)}+e^{g_2(x)}}. \quad (2.5).$$

Iz jednakosti (2.3), (2.4), (2.5), sledi:

$$\pi_1(\boldsymbol{x}) = P\{Y = 1|\boldsymbol{x}\} = \frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}} \quad (2.6)$$

$$\pi_2(\boldsymbol{x}) = P\{Y = 2|\boldsymbol{x}\} = \frac{e^{g_2(x)}}{1+e^{g_1(x)}+e^{g_2(x)}} \quad (2.7)$$

Na osnovu izvedenih logit funkcija definisanih u (2.6) i (2.7) može se izvesti i prirodni logaritam odnosa verovatnoće da prilikom fiksiranih vrednosti nezavisne promenljive zavisna promenljiva uzme kategoriju 1 i verovatnoće da primi kategoriju 2.

$$ln\left(\frac{P\{Y = 1|\boldsymbol{x}\}}{P\{Y = 2|\boldsymbol{x}\}}\right) = ln\left(\frac{\frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}}{\frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}}\right) = g_1(x) - g_2(x).$$

Iz navedenog izvođenja se može ustanoviti da je definisanje dve logit funkcije dovoljno, jer na osnovu njih možemo dobiti za svaki par kategorijalnih vrednosti zavisne promenljive odnos verovatnoća da zavsina promenljiva primi jednu kategoriju i verovatnoće da ona uzme drugu kategoriju.

2.1 Fitovanje multinomnog logističkog regresijog modela

Kao i kod binomne logističke regresije, nepoznati parametri određuju se metodom maksimalne verodostojnosti. Neka je dat uzorak veličine n: $(\widetilde{\boldsymbol{x}_{\boldsymbol{i}}}, y_i), i = 1, \dots, n$, gde y_i predstavlja registrovanu vrednost zavisne promenljive za vektor $\widetilde{\boldsymbol{x}_{\boldsymbol{i}}}$ registorvanih vrednosti nezavisnih promenljivih, neka su dati parovi međusobno nezavisni i neka je $\boldsymbol{x_i} = (1, \widetilde{\boldsymbol{x}_{\boldsymbol{i}}})$. Kako bi se definisala funkcija verodostojnosti uvode se tri binarne promeljive Y_0, Y_1, Y_2 definisane na sledeći način:

$$Y_0 = \begin{cases} 0, ako\ je\ Y = 1\ ili\ Y = 2 \\ 1, ako\ je\ Y = 0 \end{cases}, \quad Y_1 = \begin{cases} 0, ako\ je\ Y = 0\ ili\ Y = 2 \\ 1, ako\ je\ Y = 1 \end{cases}, \quad Y_2 = \begin{cases} 0, ako\ je\ Y = 0\ ili\ Y = 1 \\ 1, ako\ je\ Y = 2 \end{cases}.$$

Dakle, na osnovu vrednosti zavisne slučajne promenljive y_i određuje se vrednost promenljivih Y_0, Y_1, Y_2 u oznaci y_{0i}, y_{1i}, y_{2i}, respektivno. Posmatra se par $(\widetilde{\boldsymbol{x}_{\boldsymbol{i}}}, y_i)$ iz datog uzorka i zaključuje sledeće:

- ako je $y_i = 0$ verovatnoća da nezavisne promenljive uzmu vrednost $\widetilde{\boldsymbol{x}_\iota}$ modelirana je sa $\pi_0(\boldsymbol{x_i})$;
- ako je $y_i = 1$ verovatnoća da nezavisne promenljive uzmu vrednost $\widetilde{\boldsymbol{x}_\iota}$ modelirana je sa $\pi_1(\boldsymbol{x_i})$;
- ako je $y_i = 2$ verovatnoća da nezavisne promenljive uzmu vrednost $\widetilde{\boldsymbol{x}_\iota}$ modelirana je sa $\pi_2(\boldsymbol{x_i})$.

Odnosno, u opštem slučaju $P\{Y = j|\widetilde{\boldsymbol{x}_\iota}\} = \pi_0(\boldsymbol{x_i})^{y_{0i}}\pi_1(\boldsymbol{x_i})^{y_{1i}}\pi_2(\boldsymbol{x_i})^{y_{2i}}, j = 0,1,2$. Obzirom da je reč o prostom slučajanom uzorku, njegova funkcija verodostojnosti na osnovu realizovanog uzorka $(\widetilde{\boldsymbol{x}_\iota}, y_i), i = 1, \dots, n$ je

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_0(\boldsymbol{x_i})^{y_{0i}}\pi_1(\boldsymbol{x_i})^{y_{1i}}\pi_2(\boldsymbol{x_i})^{y_{2i}}, \text{ gde } \boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2}).$$

Kod metoda maksimalne verodostojnosti traži se maksimum funkcije $l(\boldsymbol{\beta})$ po vektoru nepoznatih parametara $\boldsymbol{\beta}$, odnosno traži se vektor $\boldsymbol{\beta}$ koji maksimizira verovatnoću da dati uzorak bude izabran. Funkcije $l(\boldsymbol{\beta})$ i $L(\boldsymbol{\beta}) = ln(l(\boldsymbol{\beta}))$ dostižu maksimum u istoj tački, pa se zbog lakšeg računa traži maksimum funkcije $L(\boldsymbol{\beta})$ po vekoru $\boldsymbol{\beta}$.

$$L(\boldsymbol{\beta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_{0i} ln\big(\pi_0(\boldsymbol{x_i})\big) + y_{1i} ln\big(\pi_1(\boldsymbol{x_i})\big) + y_{2i} ln\big(\pi_2(\boldsymbol{x_i})\big) \quad (2.8)$$

U nastavku se raspisuje funkcija $L(\boldsymbol{\beta})$ i pronalaze parcijalni izvodi.

$$L(\boldsymbol{\beta}) = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_{0i} ln\big(1 + e^{g_1(x_i)} + e^{g_2(x_i)}\big) + y_{1i}\left(ln\big(1 + e^{g_1(x_i)} + e^{g_2(x_i)}\big) - g_1(\boldsymbol{x_i})\right)$$
$$+ y_{2i}\left(ln\big(1 + e^{g_1(x_i)} + e^{g_2(x_i)}\big) - g_2(\boldsymbol{x_i})\right) \quad (2.9)$$

$$L(\boldsymbol{\beta}) = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_{0i} ln\big(1 + e^{\beta_{10}+\cdots+\beta_{1k}x_{ki}} + e^{\beta_{20}+\cdots+\beta_{2k}x_{ki}}\big)$$
$$+ y_{1i}\big(ln\big(1 + e^{\beta_{10}+\cdots+\beta_{1k}x_{ki}} + e^{\beta_{20}+\cdots+\beta_{2k}x_{ki}}\big) - \beta_{10} + \beta_{11}x_{1i} + \cdots + \beta_{1k}x_{ki}\big)$$
$$+ y_{2i}\big(ln\big(1 + e^{\beta_{10}+\cdots+\beta_{1k}x_{ki}} + e^{\beta_{20}+\cdots+\beta_{2k}x_{ki}}\big) - \beta_{20} + \beta_{21}x_{1i} + \cdots + \beta_{2k}x_{ki}\big)$$

Odakle sledi da su su parcijalni izvodi:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{1p}} = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_{0i}\pi_1(\boldsymbol{x_i})x_{pi} + y_{1i}\big(\pi_1(\boldsymbol{x_i})x_{pi} - x_{pi}\big) + y_{2i}\pi_1(\boldsymbol{x_i})x_{pi}, p = 0,1 \dots, k$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{1p}} = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \pi_1(\boldsymbol{x_i})x_{pi}(y_{0i} + y_{1i} + y_{2i}) - y_{1i}x_{pi}, p = 0,1 \dots, k.$$

Analogno,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{2p}} = -\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \pi_2(\boldsymbol{x_i}) x_{pi}(y_{0i} + y_{1i} + y_{2i}) - y_{2i} x_{pi}, p = 0{,}1 \ldots, k.$$

Iz samih definicija binarnih promenljivih Y_0, Y_1, Y_2 zaključuje se da bez obzira na vrednost slučajne promenljiva Y zbir njihovih vrednosti će biti jedan, odnosno $y_{0i} + y_{1i} + y_{2i} = 1, i = 1, \ldots, n.$, odakle sledi da je:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_{jp}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_{pi}(\, y_{ji} - \pi_j(\boldsymbol{x_i})),\ j = 1, 2, p = 0{,}1 \ldots, k. \quad (2.10)$$

Parcijalni izvodi iz (2.10) izjednačavaju se sa nulom i rešava se sistem po nepoznatim parametrima $\boldsymbol{\beta}$. Rešenje sistema označava se sa $\widehat{\boldsymbol{\beta}}_{2(k+1)}$ i naziva vektor ocena nepoznatih parametara dobijenih na osnovu metoda maksimalne verodostojnosti. Kao i kod binarnog slučaja, dobijen sistem nije linearan pa se on rešava iterativnim postupkom. U nastavku su određeni parcijalni izvodi drugog reda funkcije $L(\boldsymbol{\beta})$**,** (2.11) i (2.12).

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jp} \partial \beta_{jq}} = -\sum_{i=1}^{n} x_{pi} \frac{\partial \pi_j(\boldsymbol{x_i})}{\partial \beta_{jq}} = -\sum_{i=1}^{n} x_{pi} \frac{\partial \left(\dfrac{e^{g_j(\boldsymbol{x_i})}}{1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})}} \right)}{\partial \beta_{jq}}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jp} \partial \beta_{jq}} = -\sum_{i=1}^{n} x_{pi} \frac{x_{qi} e^{g_j(\boldsymbol{x_i})} \left(1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})}\right) - e^{g_j(\boldsymbol{x_i})} e^{g_j(\boldsymbol{x_i})} x_{qi}}{(1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})})^2}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jp} \partial \beta_{jq}} = -\sum_{i=1}^{n} x_{pi} x_{qi} \pi_j(\boldsymbol{x_i}) \left(1 - \pi_j(\boldsymbol{x_i})\right), j = 1{,}2, p = 0{,}1, \ldots, k, q = 0{,}1, \ldots, k \quad (2.11)$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jp} \partial \beta_{j'q}} = -\sum_{i=1}^{n} x_{pi} \frac{\partial \pi_j(\boldsymbol{x_i})}{\partial \beta_{j'q}} = -\sum_{i=1}^{n} x_{pi} \frac{\partial \left(\dfrac{e^{g_j(\boldsymbol{x_i})}}{1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})}} \right)}{\partial \beta_{j'q}}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jp} \partial \beta_{j'q}} == -\sum_{i=1}^{n} x_{pi} \frac{-e^{g_j(\boldsymbol{x_i})} e^{g_{j'}(\boldsymbol{x_i})} x_{qi}}{(1 + e^{g_1(\boldsymbol{x_i})} + e^{g_2(\boldsymbol{x_i})})^2}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_{jp} \partial \beta_{j'q}} = \sum_{i=1}^{n} x_{pi} x_{qi}\, \pi_j(\boldsymbol{x_i}) \pi_{j'}(\boldsymbol{x_i}), j = 1{,}2, j' = 1{,}2, j \neq j', p = 0{,}1, \ldots, k, q = 0{,}1, \ldots, k \quad (2.12)$$

Hesijan funkcije $L(\boldsymbol{\beta})$ je matrica formata $2(k+1) * 2(k+1)$ i ukoliko je ona negativno definitna sledi da je $\widehat{\boldsymbol{\beta}}_{2(k+1)}$ zaista vektor u kojem posmatrana funkcija dostiže maksimum. Kao i u slučaju binarne logističke regresije važi: $\left(-\frac{\partial^2 L(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}^2}\right)^{-1} = \sigma^2(\widehat{\boldsymbol{\beta}})$. Matrica $-\frac{\partial^2 L(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}^2}$ se naziva informaciona matrica i obeležava sa $\hat{I}(\widehat{\boldsymbol{\beta}})$. U slučaju modela sa tri kategorije, informaciona matrica se može predstaviti kao blok matrica formata 2x2 (videti u [1]). Neka je $V_j = diag\left(\hat{\pi}_j(\boldsymbol{x_i})\left(1 - \hat{\pi}_j(\boldsymbol{x_i})\right)\right)_{n*n}, j = 1,2$, i $V_3 = diag(\hat{\pi}_1(\boldsymbol{x_i})\hat{\pi}_2(\boldsymbol{x_i}))_{n*n}$, tada je

$$\hat{I}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \hat{I}(\widehat{\boldsymbol{\beta}})_{11} & \hat{I}(\widehat{\boldsymbol{\beta}})_{12} \\ \hat{I}(\widehat{\boldsymbol{\beta}})_{21} & \hat{I}(\widehat{\boldsymbol{\beta}})_{22} \end{bmatrix},$$

pri čemu je $\hat{I}(\widehat{\boldsymbol{\beta}})_{11} = X^T V_1 X,\ \hat{I}(\widehat{\boldsymbol{\beta}})_{12} = \hat{I}(\widehat{\boldsymbol{\beta}})_{21} = X^T V_3 X,\ \hat{I}(\widehat{\boldsymbol{\beta}})_{22} = X^T V_2 X.$

Primer** (videti u [1])**:

Posmatra se model u kojem zavisna promenljiva MESTO predstavlja mesto pružanja nege adolescentima nakon sprovedenog određenog psihološkog tretmana, dok je nezavisna promenljiva NASILJE istorija nasilja i uzima vrednost 1 ukoliko postoji nasilje, u suprotnom je njena vrednost 0. Kategorije zavisne promenljive kodirane su na sledeći način: 0 – ambulanta, 1 – delimično stambeno, 2 – stambeno. Na osnovu dela uzorka prikupljenog za potrebe istraživanja u akademskom članku Fontanella et al. (2008) fitovan je logsitički model i dobijena sledeća matrica ocenjenih kovarijansi za ocenjene koeficijente modela:

Tabela 10: Tabela ocena kovarijansi za ocene koeficijenata modela

		$g_1(x)$		$g_2(x)$	
		X	Slobodan član	X	Slobodan član
$g_1(x)$	X	0.06616			
	Slobodan član	-0.05096	0.05096		
$g_2(x)$	X	0.01809	-0.01250	0.09437	
	Slobodan član	-0.01250	0.01250	-0.07917	0.07917

Dijagonalna matrica sadrži ocene kovarijansi i varijansi za ocenjene koeficijente posmatrane logit fukcije, dok vandijagonalna matrica sadrži ocene kovarijansi za ocenjene koeficijente različitih logit funkcija.

2.2 Testiranje značajnosti koeficijenata

Neka zavisna promenliva ima $P+1$ kategorijalnu vrednost kodirane sa $0, 1, \dots, P$. U slučaju multinomnog logističkog regeresijskog modela sa k nezavisnih promenljivih, za svaku nezavisnu promenljivu testira se nulta hipoteza da je koeficijent (ili koeficijenti ukoliko je reč o kategorijalnoj nezavisnoj promenljivoj sa više od dve kategorije) koji joj odgovara jednak nuli. Test statistika je:

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

i njena nulta raspodela je hi-kvadrat raspodela sa stepenom slobode P ukoliko je reč o neprekidnoj nezavisnoj promenljivi, a $P * (Q-1)$ ukoliko je posmatrana nezavisna promenljiva kategorijalna sa Q kategorijalnih vrednosti. Ukoliko je p-vrednost test statistike manja od 0.05 nulta hipoteza se odbacuje i zaključak je da je posmatrana nezavisna promenljiva značajna za model.

Primer:

Koristimo ovaj test kako bismo odredili da li je koeficijent uz promenljivu NASILJE statistički značajan. Posmatra se model bez nezavisne promenljive i računa logaritam funkcije verodostojnosti tog modela: $L_0 = -524.37$, zatim se računa logaritam verodostojnosti modela sa nezavisnom promneljivom: $L_1 = -515.73$, odakle sledi da je vrednost test statistike:

$$G = -2 * L_0 + 2 * L_1 = 2 * 524.37 - 2 * 515.73 = 17.28.$$

Kako je nulta raspodela test statistike χ_2^2, njena p-vrednost je 0.002. Kako je p-vrednost manja od 0.05 sledi da je koeficijent uz promenljivu NASILJE statistički značajan.

Kod binarne logističke regresije odnos šansi je definisan kao odnos šanse da zavisna promenljiva uzme vrednost 1 ukoliko je nezavisna uzela 1 i šanse da zavisna promenljiva uzme vrednost 1 ukoliko je nezavisna uzela 0. Kako kod mulitnomne logističke regresije postoji više od dve vrednosti zavisne promenljive, prvo se određuje njena referentna vrednost. Ukoliko se izabere

referentna vrednost $Y = 0$, odnos šansi ishoda $Y = j$ u odnosu na referentni ishod za vektore nezavisnih vrednosti $\boldsymbol{a}$ i $\boldsymbol{b}$ je definisan sa:

$$OR_j(\boldsymbol{a}, \boldsymbol{b}) = \frac{P\{Y=j|X=a\}/P\{Y=0|X=a\}}{P\{Y=j|X=b\}/P\{Y=0|X=b\}} = \frac{\pi_j(a)/\pi_0(a)}{\pi_j(b)/\pi_0(b)},\ j\epsilon\{1,2\}.$$

Ubacivanjem ocenjenih logističkih verovatnoća dobija se ocenjeni odnos šansi ishoda $Y = j$ u odnosu na referentni ishod za vektore nezavisnih vrednosti $\boldsymbol{a}$ i $\boldsymbol{b}$, u oznaci $\widehat{OR}_j(\boldsymbol{a}, \boldsymbol{b})$.

Primer:

Dati su podaci o broju prisutnog nasilja u okviru svakog mesta pružanja nege za 508 adolescenata nakon podvrgnutog psihološkog tretmana:

Tabela 11: Broj nasilja u okviru svakog mesta pružanja nege

MESTO	Nasilje		Ukupno
	Da	Ne	
Ambulanta (0)	179	80	259
Delimično stambeno (1)	104	26	130
Stambeno (2)	104	15	119
	387	121	508

Kako bi se odredilo koliko je veća šansa za smeštajem u delimično stambenim uslovima lečenja kod adolescenata sa istorijom nasilja nego kod onih bez istorije nasilja, uzima se $Y = 0$ kao referentna vrednost i računa se $\widehat{OR}_1(1,0)$:

$$\widehat{OR}_1(1,0) = \frac{\frac{P\{Y=1|X=1\}}{P\{Y=0|X=1\}}}{\frac{P\{Y=1|X=0\}}{P\{Y=0|X=0\}}} = \frac{\frac{104}{179}}{\frac{26}{80}} = 1.79.$$

Šansa za smeštanjem u delimično stabeno lečenje je veća 1.79 puta među adolescentima sa istorijom nasilja nego kod onih bez istorije. Slično,

$$\widehat{OR}_2(1,0) = \frac{\frac{P\{Y=2|X=1\}}{P\{Y=0|X=1\}}}{\frac{P\{Y=2|X=0\}}{P\{Y=0|X=0\}}} = \frac{\frac{104}{179}}{\frac{15}{80}} = 3.1,$$

odnosno šansa za smeštanjem u stambeno lečenje je veća 3.1 put među adolescentima sa istorijom nasilja u odnosu na adolescente koji nemaju zabeleženu istoriju nasilja.

2.3 Uopšten Hosmer-Lemeshow test

Hosmer-Lemeshow test za procenu slaganja binarnog logističkog regresijskog modela se uopštava na multinomni slučaj zavisne promenljive. Neka je G broj fiksiranih grupa. Uzorak se grupiše na osnovu ocenjenih verovatnoća $\hat{\pi}_j(\boldsymbol{x_i})$. Prva grupa sadrži najmanjih n/G vrednosti ocenjenih verovatnoća, druga grupa sadrži sledećih najmanjih n/G ocenjenih verovatnoća i tako se forimara G grupa. Unutar svake grupe računa se registrovan i predviđen broj subjekata za svaku kategoriju zavisne promenljive, za šta će biti potrebne sledeće binarne slučajne promenljive:

$$Y_{ij} = \begin{cases} 1, & Y_i = j \\ 0, & inače \end{cases}, \; i = 1, \dots, n, \; j = 0, \dots, P.$$

Pomoću iznad definisanih slučajnih promenljivih izražavamo broj registrovanih (o_{jg}) i predviđenih (e_{jg}) subjekata unutar grupe g iz kategorije $Y = j$:

$$o_{jg} = \sum_{l \in \Omega_g} y_{lj} \;\; i \; e_{jg} = \sum_{l \in \Omega_g} \hat{\pi}_j(x_l),$$

pri čemu $g = 1, \dots, G, j = 0, 1, \dots, P$ i Ω_g predstavlja skup indeksa opservacija koje se javljaju u grupi g. Kod binarne logističke regresije unutar svake grupe su računate po dve registrovane i predviđene vrednosti. Sada kada zavisna promenljiva ima $P + 1$ kategoriju, unutar svake grupe se računa $P + 1$ registrovanih vrednosti i $P + 1$ predviđenih vrednosti. Shodno tome, dobija se tabela formata $(P + 1) * G$ registrovanih i predviđenih vrednosti.

Tabela 12: Broj predviđenih i registrovanih vrednosti po grupama unutar svake kategorije zavisne promenljive

Kategorija		Grupa				Ukupno
		1	2	…	G	
		Presečne verovatnoće				
		$\hat{\pi}(\boldsymbol{x_k}) \epsilon [\hat{\pi}_{L1}, \hat{\pi}_{U1}]$	$\hat{\pi}(\boldsymbol{x_k}) \epsilon [\hat{\pi}_{L2}, \hat{\pi}_{U2}]$	…	$\hat{\pi}(\boldsymbol{x_k}) \epsilon [\hat{\pi}_{LG}, \hat{\pi}_{UG}]$	
$Y = 0$	Registrovano	o_{01}	o_{02}	…	o_{0G}	o_0
	Očekivano	e_{01}	e_{02}	…	e_{0G}	e_0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$Y = j$	Registrovano	o_{j1}	o_{j2}	…	o_{jG}	o_j
	Očekivano	e_{j1}	e_{j2}	…	e_{jG}	e_j
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$Y = P$	Registrovano	o_{P1}	o_{P2}	…	o_{PG}	o_P
	Očekivano	e_{P1}	e_{P2}	…	e_{PG}	e_P
Ukupno (φ)		$\approx n/G$	$\approx n/G$	…	$\approx n/G$	n

Na osnovu tabele dobija se Pirsonova hi-kvadrat test statistika Hosmer-Lemeshow testa:

$$\hat{C}_G = \sum_{j=0}^{P} \sum_{g=1}^{G} \frac{(o_{jg} - e_{jg})^2}{e_{jg}}.$$

Pokazano je da se statistika $\hat{C}_G$ dobro aproksimira hi-kvadrat raspodelom sa stepenom slobode $(G-2) * P$. Ukoliko je p-vrednost test statistike manja od 0.05 smatra se da se model ne slaže dobro sa podacima. Literatura korišćenja za generalizaciju Hosmer Lemeshow testa je [12] i [14].

3. Ordinalna logistička regresija

Često se nailazi na slučaj da zavisna promenljiva ima ordinalnu mernu skalu, odnosno uzima kategorijalne vrednosti koje se mogu rangirati, od najnižeg do najvišeg ranga. Ordinalna logistička regresija analizira slučaj kada je merna skala zavisne promenljive upravo ordinalna. Na primer, zadovoljstvo korisnika određenog mobilnog operatera može biti predstavljeno promenljivom koja uzima sledeće kategorije:

- izuzetno nezadovoljan korisnik
- nezadovoljan korisnik
- indiferentan korisnik
- zadovoljan korisnik
- izuzetno zadovoljan korisnik

Kod multinomne logističke regresije za nominalnu zavisnu promenljivu Y sa $P+1$ kategorijom za referentnu vrednost zavisne promenljive uzeta je kategorija $Y=0$ i posmatrane su sledeće logit funkcije:

$$g_p(\boldsymbol{x}) = ln\left(\frac{P\{Y=1|\boldsymbol{x}\}}{P\{Y=0|\boldsymbol{x}\}}\right), p = 1, \dots, P.$$

Kako je kod ordinalne logističke regresije uključeno i rangiranje mogućih vrednosti zavisne promenljive, treba odrediti koji će se verovatnoće upoređivati, kao i koji je najprikladniji model za logit. Shodno tome, izdvajaju se sledeći ordinalni logistički modeli:

- osnovni logit model,
- logit model susedne kategorije,
- continuation-ratio logit model,
- proporcionalan logit model.

Neka ordinalna zavisna promenljiva ima $P+1$ kategoriju, kodirane sa $0,1,\dots,P$, pri čemu 0 označava kategoriju najnižeg ranga, sledeći rang je kodiran sa 1 i sve tako do kategorije najvišeg ranga kodirane sa P. Verovatnoća da zavisna promenljiva uzme vrednost j ukoliko su nezavisne promenljive uzele vrednost $\boldsymbol{x}$ označava se sa $\phi_j(\boldsymbol{x})$, odnosno

$$P\{Y=j|\boldsymbol{x}\} = \phi_j(\boldsymbol{x}), \;\; j = 0,1,\dots,P.$$

1) Kod baznog logit modela upoređuje se svaka kategorija sa referentnom kategorijom $Y = 0$, te se definišu sledeće logit funkcije:

$$g_j(\boldsymbol{x}) = ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) = \beta_{j_0} + \boldsymbol{x}^T\boldsymbol{\beta_j}, j = 1, \dots, P,$$

gde $\boldsymbol{x} = (x_1, \dots, x_{\boldsymbol{k}})^{\boldsymbol{T}}$ i $\boldsymbol{\beta_j} = (\beta_1, \dots, \beta_{\boldsymbol{k}})^{\boldsymbol{T}}, j = 1, \dots, P$.

2) Kada se porede dve susedne kategorije i to viša sa nižom, reč je o logističkom modelu susedne kategorije i susedni logiti se modeliraju sa:

$$a_j(\boldsymbol{x}) = ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_{j-1}(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{x}^T\boldsymbol{\beta}, j = 1, \dots, P,$$

pri čemu se pretpostavlja da logaritmi odnosa šansi ne zavise od vrednosti koju je zavisna promenljiva uzela i da su linearni po koeficijentima.

3) U slučaju continuation-ration logističkom modela vrši se upoređivanje kategorije sa svim niže rangiranim kategorijama. Logit ovog modela je:

$$r_j(\boldsymbol{x}) = ln\left(\frac{P\{Y = j|\boldsymbol{x}\}}{P\{Y < j|\boldsymbol{x}\}}\right) = ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_0(\boldsymbol{x}) + \cdots + \phi_{j-1}(\boldsymbol{x})}\right) = \theta_j + \boldsymbol{x}^T\boldsymbol{\beta}, j = 1, \dots, P.$$

Logit funkcija binarnog logističkog modela predstavlja logaritam odnosa verovatnoće da zavisna promenljiva uzme vrednost 1 i verovatnoće da zavisna promenljiva uzme vrednost 0, te se ona može dobiti iz logita svakog od tri modela (1-3) za $P = 1$.

4) Proporcionalan logit model je onaj kod kojeg se upoređuje verovatnoća da zavisna promenljiva primi kategorije manje ili jednake zadatoj i verovatnoća da ona uzme više rangirane kategorije. Logit modela je dat sa:

$$c_j(\boldsymbol{x}) = ln\left(\frac{P\{Y \leq j|\boldsymbol{x}\}}{P\{Y > j|\boldsymbol{x}\}}\right) = ln\left(\frac{\phi_0(x)+\cdots+\phi_{j-1}(x)+\phi_j(x)}{\phi_{j+1}(x)+\phi_{j+2}(x)+\cdots+\phi_P(x)}\right) = \tau_j + \boldsymbol{x}^T\boldsymbol{\beta}, j = 1, \dots, P.$$

U slučaju modela 4) za $P = 1$ posmatra se šansa da zavisna promenljiva uzme vrednost 0, šansa suprotnog događaja u odnosu na događaj posmatran u binarnom logističkom modelu.

Literatura korišćena za izradu dela rada o ordinalnoj logističkoj regresiji je [1], [8] i [16].

3.1 Fitovanje ordinalnog logističkog regresijskog modela

Neka je da uzorak veličine n: $(\boldsymbol{x_i}, y_i), i = 1, \dots, n$, gde y_i predstavlja registrovanu vrednost zavisne promenljive za vektor $\boldsymbol{x_i}$ registorvanih vrednosti nezavisnih promenljivih i neka su dati parovi međusobno nezavisni. Uvodi se $P + 1$ slučajnih promenljivih $Z_0, Z_1, \dots, Z_P$ definisanih na sledeći način:

$$Z_j = \begin{cases} 1, ako\ Y = j \\ 0, inače \end{cases}, \quad j = 0,1, \dots, P.$$

Vrednost ovih slučajnih promenljivih zavisi od ranga koji je zavisna promenljiva registrovala. Na osnovu vrednosti zavisne promenljive y_i određuje se vrednost promenljivih $Z_0, Z_1, \dots, Z_P$ u oznaci $z_{0i}, z_{1i}, \dots, z_{Pi}$, respektivno. Na osnovu definicije slučajnih promenljivih $Z_j, j = 0,1, \dots, P$ sledi da za proizvoljnu vrednost y_i zavisne promenljive važi $\sum_{j=0}^{P} z_{ji} = 1$. Kod multinomne logističke regresije verovatnoće $\pi_j(\boldsymbol{x}), j = 0,1, \dots, P$, su se preko logit funkciija izrazile kao funkcije nepoznatih parametara. Isti postpuka se sprovodi i kod ordinalnog logističkog regresijskog modela, s tim što se vodi računa kako je definisana logit funkcija. Kod baznog modela logit funkcije su $g_j(\boldsymbol{x}), j = 0,1, \dots P$, u slučaju logističkog modela dve susedne kategorije to su funkcije $a_j(\boldsymbol{x}), j = 0,1, \dots P$, kod continuation-ratio logističkog modela logit funkcije $r_j(\boldsymbol{x}), j = 0,1, \dots P$, dok su $g_j(\boldsymbol{x}), j = 0,1, \dots P$ logit funkcije proporcionalnog logit modela. Na osnovu logit funkcija posmatranog modela, verovatnoće $\phi_p(\boldsymbol{x}), p = 0,1, \dots P$ se izražavaju preko nepoznatih parametara i definiše se funkcija verodostojnosti:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \phi_0(\boldsymbol{x_i})^{z_{0i}} \phi_1(\boldsymbol{x_i})^{z_{1i}} \dots \phi_P(\boldsymbol{x_i})^{z_{Pi}}.$$

$\boldsymbol{\beta}$ predstavlja vektor nepoznatih parametara. Dimenzija vektorka je $P + (P * k)$ kod baznog model, jer za svaki logit imamo jedan nepoznat koeficijent preseka i k koeficijenata nagiba. U preostala tri slučaja, za svaki logit imamo jedan nepoznat koeficijent preseka dok su koeficijeti nagiba isti za sve logite, odakle sledi da dimenzija vektora $\boldsymbol{\beta}$ iznosi $P * k$. Posmatra se funkcija $L(\boldsymbol{\beta}) = ln(l(\boldsymbol{\beta}))$ i ukoliko je njen Hesijan negativno definitna matrica, maksimum funkcije $L(\boldsymbol{\beta})$ se pronalazi rešavanjem sistema koji se dobija izjednačavanjem parcijalnih izvoda prvog reda po svakoj komponenti vektora $\boldsymbol{\beta}$ sa nulom. Maksimizator funkcije je vektor ocena nepoznatih parametara dobijen metodom maksimalne verodostojnosti i označava se sa $\widehat{\boldsymbol{\beta}}$. Kao i kod slučaja

binomne i multinomne logističke regresije, inverz informacione matrice sadrži varijanse i kovarijanse ocenjivača.

U nastavku se uspostavlja veza između baznog i logističkog modela susednih kategorija. Odnosno, pretpostavlja se da su date logit funkcije modela susednih kategorija i izvodi se logit funkcija baznog modela.

$$ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) = ln\left(\frac{\phi_1(\boldsymbol{x})}{\phi_0(\boldsymbol{x})} * \frac{\phi_2(\boldsymbol{x})}{\phi_1(\boldsymbol{x})} * \frac{\phi_3(\boldsymbol{x})}{\phi_2(\boldsymbol{x})} * \ldots * \frac{\phi_j(\boldsymbol{x})}{\phi_{j-1}(\boldsymbol{x})}\right)$$

$$ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) = ln\left(\frac{\phi_1(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) + ln\left(\frac{\phi_2(\boldsymbol{x})}{\phi_1(\boldsymbol{x})}\right) + ln\left(\frac{\phi_3(\boldsymbol{x})}{\phi_2(\boldsymbol{x})}\right) + \cdots + ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_{j-1}(\boldsymbol{x})}\right)$$

$$ln\left(\frac{j(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) = a_1(\boldsymbol{x}) + a_2(\boldsymbol{x}) + a_3(\boldsymbol{x}) + \cdots + a_j(\boldsymbol{x})$$

$$ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) = (\alpha_1 + \boldsymbol{x}^T\boldsymbol{\beta}) + (\alpha_2 + \boldsymbol{x}^T\boldsymbol{\beta}) + (\alpha_3 + \boldsymbol{x}^T\boldsymbol{\beta}) + \cdots + (\alpha_j + \boldsymbol{x}^T\boldsymbol{\beta})$$

$$ln\left(\frac{\phi_j(\boldsymbol{x})}{\phi_0(\boldsymbol{x})}\right) = (\alpha_1 + \alpha_2 + \cdots + \alpha_j) + j\boldsymbol{x}^T\boldsymbol{\beta} \quad (3.1)$$

Koeficijenti preseka baznog modela će biti $\alpha_1 + \alpha_2 + \cdots + \alpha_j$, $j = 1, \ldots, P$, dok su koeficijenti nagiba $j\boldsymbol{\beta}$, $j = 1, \ldots, P$. Ova činjenica je korisna prilikom izvođenja ocena koeficijenata modela susednih kategorija na osnovu ocenjenih koeficijenata baznog modela. Ilustracija je data u sledećem primeru.

Primer** (videti u [1])**:

Posmatra se uticaj konzumiranja cigareta majki tokom trudnoće na težinu rođene bebe. Promenljiva CIGARETE uzima vrednost 1 ukoliko majka tokom trudnoće konzumira cigarete, dok u suprotnom uziva vrednost 0. Zavisna promenljiva T predstavlja kategorije težine beba i kodirana je na sledeći način:

- 0 – težina bebe veća od $3500g$
- 1 – težina bebe je iz intervala $(3000g, 3500g]$
- 2 – težina bebe je iz interval $(2500g, 3000g]$
- 3 – težina bebe manja od $2500g$ ili jednaka.

Posmatra se uzorak dat u tabeli 13 i na osnovu njega su izvedene ocene baznog modela oblika: $g_j(x) = \beta_{j0} + \beta_{j1}x, \beta_{j1} = j\beta, j = 1, 2, 3$, prikazane u tabeli 14.

Tabela 13:Uzorak za ordinalnu logističku regresiju

T	CIGARETE		
	0	1	
0	35	11	46
1	29	17	46
2	22	16	38
3	29	30	59
	115	74	189

Tabela 14: Ocenjeni koeficijenti baznog modela

$g_p(x)$	Koeficijent	Ocena koeficijenta
$g_1(x)$	β_{11}	0.370
	β_{10}	-0.110
$g_2(x)$	β_{21}	0.739
	β_{20}	-0.441
$g_3(x)$	β_{31}	1.109
	β_{30}	-0.175

Ocene koeficijenata baznog modela su rešenja nelinearnog sistema dobijenog izjednačavanjem parcijalnih izvoda funkcije verodostojnosti baznog modela sa nulom. Ako bi se na isti način određivale ocene koeficijenta modela susednih kategorija, rešavao bi se nov sistem nelinearnih jednačina dobijen koristeći funkciju verodostojnosti modela susednih kategorija. Kako bi se izbeglo dvostruko rešavanje sistema nelinearnih jednačica, odnosno primena dva puta iterativnog postupka koristi se izvedena veza između baznog i modela susednih kategorija. Na taj način jednostavnijim postupkom se može doći do ocena koeficijenata modela susednih kategorija. Na osnovu samih definicija logit funkcija posmatrana dva modela sledi da su njihove prve logit

funkcije jednake, odnosno ukoliko sa $a_j(x) = \alpha_j + \gamma x, j = 1, 2, 3$, obeležimo model susednih kategorija sledi:

$$\hat{a}_1(x) = \hat{g}_1(x) = \hat{\beta}_{10} + \hat{\beta}_{11}x = -0.11 + 0.37x.$$

Dakle, dobijaju se ocene $\hat{\alpha}_1 = -0.11$ i $\hat{\gamma} = 0.37$. Iz (3.1) za $j = 2$ sledi:

$$ln\left(\frac{\boldsymbol{\phi_2(x)}}{\boldsymbol{\phi_0(x)}}\right) = \hat{\beta}_{20} + \hat{\beta}_{21}x = (\hat{\alpha}_1 + \hat{\alpha}_2) + 2x\hat{\gamma},$$

odnosno

$$\hat{\alpha}_1 + \hat{\alpha}_2 = \hat{\beta}_{20} \Leftrightarrow \hat{\alpha}_2 = \hat{\beta}_{20} - \hat{\alpha}_1 = -0.441 + 0.11 = -0.331.$$

Kako je koeficijent uz zavisnu promenljivu modela susednih kategorija isti u svakom logitu važi

$$\hat{a}_2(x) = -0.331 + 0.37x, \text{ odnosno } \hat{\alpha}_2 = -0.331.$$

Iz formule (3.1) za $j = 3$ sledi:

$$ln\left(\frac{\boldsymbol{\phi_3(x)}}{\boldsymbol{\phi_0(x)}}\right) = \hat{\beta}_{30} + \hat{\beta}_{31}x = (\hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3) + 3x\hat{\gamma},$$

odnosno

$$\hat{\alpha}_3 = \hat{\beta}_{30} - (\hat{\alpha}_1 + \hat{\alpha}_2) = -0.175 - (-0.11 - 0.331) = 0.226.$$

Koristi se odnos šansi kod modela susednih kategorija u cilju ragistrovanja koliko puta je veća šansa da se rodi beba u sledećoj nižoj telesnoj kategoriji među pušačima u odnosu na nepušače:

$$OR_{j,j-1} = \frac{\frac{P\{Y = j|X = 1\}}{P\{Y = j-1|X = 1\}}}{\frac{P\{Y = j|X = 0\}}{P\{Y = j-1|X = 0\}}} = \frac{\frac{\phi_j(1)}{\phi_{j-1}(1)}}{\frac{\phi_j(0)}{\phi_{j-1}(0)}} = \frac{e^{\alpha_j + \beta_{11}}}{e^{\alpha_j}} = e^{\beta_{11}}, j = 1, \dots, P.$$

Ocena odnosa šansi je $\widehat{OR}(j, j-1) = e^{\hat{\beta}_{11}} = e^{0.37} = 1.45, j = 1, 2, 3$, odnosno šansa da se rodi beba u sledećoj nižoj telesnoj kategoriji među pušačima je 1.45 puta veća od šanse među nepušačima..

3.2 Ordinalna verzija Hosmer-Lemeshow testa

Posmatra se proporcionalni logit model i analizira se ordinalna verzija Hosmer-Lemeshow testa, međutim analogna analiza se sprovodi i u slučaju preostalih logit modela ordinalne logističke regresije. Na osnovu ocenjenih proporcionalnih logit modela lako se određuju kumulativne verovatnoće:

$$\hat{P}\{Y \leq j|x\} = \frac{e^{\hat{c}_j(x)}}{1 + e^{\hat{c}_j(x)}}, j = 0,1, \dots, P.$$

Dalje, na osnovu ocenjenih kumulativnih verovatnoća ocenjujemo verovatnoće $\boldsymbol{\phi_j(x)}$:

$$\hat{\boldsymbol{\phi}}_j(\boldsymbol{x}) = \begin{cases} \hat{P}\{Y \leq 0|\boldsymbol{x}\}, j = 0 \\ \hat{P}\{Y \leq p|\boldsymbol{x}\} - \hat{P}\{Y \leq p-1|\boldsymbol{x}\}, j = 1, \dots, P-1 \\ 1 - \hat{P}\{Y \leq P-1|\boldsymbol{x}\}, j = P \end{cases}$$

U slučaju ordinalne logističke regresije grupisanje se vrši prema ordinalnim brojevima (Lipsitz, Fitzmaurice and Molenberghs, 1996) defiisanih na sledeći način:

$$s_i = \sum_{j=0}^{P} j\, \hat{\boldsymbol{\phi}}_j(\boldsymbol{x_i}).$$

Neka je G broj fiksiranih grupa. Prva grupa sadrži sadrži n/G opservacija sa najmanjim ordinalnih brojevima, druga grupa sadrži sledećih n/G opservacija sa najmanjim ordinalnim brojevima i tako se formira G grupa, pri čemu se opservacije sa istim ordinalnim brojem sortiraju prema registrovanoj vrednosti zavisne promenljive. Unutar svake grupe računa se registrovan i predviđen broj subjekata za svaku kategoriju zavisne promenljive, za šta će biti potrebne sledeće binarne slučajne promenljive:

$$Y_{ij} = \begin{cases} 1, & Y_i = j \\ 0, & inače \end{cases}, \; i = 1, \dots, n, \; j = 0, \dots, P.$$

Pomoću iznad definisanih slučajnih promenljivih izražavamo broj registrovanih (o_{jg}) i predviđenih (e_{jg}) subjekata unutar grupe g iz kategorije $Y = j$:

$$o_{jg} = \sum_{l \in \Omega_g} y_{lj} \;\; i \; e_{jg} = \sum_{l \in \Omega_g} \hat{\pi}_j(x_l),$$

pri čemu $g = 1, \dots, G, j = 0, 1, \dots, P$ i Ω_g predstavlja skup indeksa opservacija koje se javljaju u grupi g. Test statistika kod ordinalne verzije Hosmer-Lemeshow testa je:

$$\hat{C}_G = \sum_{p=0}^{P} \sum_{g=1}^{G} \frac{(o_{pg} - e_{pg})^2}{e_{pg}}.$$

Pokazano je da se statistika $\hat{C}_G$ dobro aproksimira hi-kvadrat raspodelom sa stepenom slobode $(G - 2) * P + (P - 1)$. Razlika Hosmer-Lemeshow testa kod multinomne regresije u odnosu na onaj kod ordinalne regresije je u načinu grupisanja opservacija i u broju stepeni slobode test statistike. Obično se uzima da je $G \approx 10$. Nije preporučljivo da je broj grupa mali jer će se na taj način smanjiti moć testa usled heterogenosti unutar grupa. S druge strane, ukoliko je broj grupa preveliki aproksimiranje test statistike hi-kvadrat raspodelom neće biti dobro usled malog broja opservacija u grupama.

4. Primena logističke regresije u određivanju optimalnog insekticidnog tretmana u cilju zaštite useva suncokreta od žičara

U ovom delu rada, primenjuje se logistička regresija (binomna i multinomna) u cilju analiziranja nivoa oštećenja biljaka suncokreta od žičara, kao najznačajnijih zemljišnih štetočina, u zavisnosti od lokaliteta i primenjenog insekticida. U nastavku se opisuje protokol ogleda za ispitivanje biološke efikasnosti insekticida u suzbijanju žičara (Coleoptera: fam. Elateridae)

Lokalitet

Eksperimenti su izvedeni u toku 2021. godine, na dva lokaliteta na oglednim poljima Instituta za ratarstvo i povrtarstvo („Polje 1“ – kod Bačkog jarka i „T-12“ – kod Novog Sada), na Rimskim šančevima, Novi Sad, Srbija. Odabrana su dva lokaliteta, prostorne udaljenosti 4 km vazdušnom linijom, koja se značajno razlikuju po brojnosti populacije žičara i on prednjači u Bačkom jarku. Inicijalna brojnost žičara se na pomenutim lokalitetima kontinuirano prati poslednje četiri godine. S obzirom malu prostornu udaljenost, može se reći da se klimatski uslovi na oba lokaliteta ne razlikuju. U pogledu pedoloških karakteristika, na oba lokaliteta zemljište je černozem sa procentualnim udelom humusa oko 3%.

Usev i tehnologija gajenja

Efikasnost insekticida u suzbijanju žičara je ispitana u usevu suncokreta, hibrid Romeo. Setva je izvedena u prvoj polovini aprila 2021. godine, mašinski, na međuredni razmak od 70 cm, i 21 cm rastojanje između biljaka u redu, što je iznosilo oko 60.000 biljaka/ha. Primenjene su sve preporučene agrotehničke mere za gajenje suncokreta, kao što su zimsko duboko oranje, predsetvena priprema, inkorporacija herbicida, kako bi se smanjila pojava korova na početku vegetacije i prihrana.

Eksperimentalni protokol

Ogledi su postavljeni po pravilima preliminarnih ogleda po dizajnu potpuno slučajnog blok sistema (Hadživuković, 1991). Biološki eksperimenti su izvedeni po metodi PP 1/46 (3) (EPPO Standards PP1, Vol. 35/1, 2005), metodi specifičnoj za ispitivanje efikasnosti preparata za suzbijanje žičara (*Agriotes* spp). Veličina osnovne parcelice je iznosila 28 m^2, odnosno 10 m

dužine sa po četiri reda biljaka suncokreta. U ogledu je primenjeno šest insekticidi na dva načina: **a)** inkorporacijom (unošenjem u zemljište uporedo sa setvom) i to **Attracap granule** (aktivna materija gljiva *Metarhizium brunneum* Cb-15-III) u količini 30 kg/ha i **Force 1.5 G** (aktivna materija teflutrin) u količini od 8 kg/ha semena i **b)** tretiranjem semena pre setve i to **Force 20 CS** (aktivna materija teflutrin), nanošenjem 250 mL/100 kg semena, **Buteo Start** (aktivna materija flupirdifuron) u količini 1,1 L/100 kg semena, **Lumiposa** (aktivna materija ciantraniliprol) u količini 0,8 L/100 kg semena i **Sonido** (aktivna materija tiakloprid) 2,5 L/100 kg semena. **Kontrola** je podrazumevala netretiano seme.
Ogled je izveden u pet ponavljanja na „Polju 1“ i devet ponavljanja na lokalitetu „T-12“.

Obeležja posmatranja

Prema pomenutoj metodi (EPPO PP 1/46 (3)), utvrđuje se broj niklih i broj oštećenih biljaka u fazi nicanja i u fazi 3 para listova u sva četiri reda. Pored pomenutog obeležja, u fazi 3 para listova, sa svake eksperimentalne parcelice uzorkovano je po 10 biljaka na kojima je ocenjen nivo oštećenja prema skali od 0 do 4:
0 - nema oštećenja
1 - jedva viljivo oštećenje
2 - vidljiva oštećenja koja ne utiču na vitalnost biljke
3- biljka vidno oštećena, ali ima šanse da se oporavi
4- biljka vidno oštećenja, uvenula bez mogućnosti oporavka ili je u potpunosti uništena

Ocena nivoa oštećenja biljaka (prema navedenoj skali) korišćena je kao dopunska metoda u okviru standardizovane EPPO procedure PP 1/46 (3) (EPPO Standards PP1, Vol. 35/1, 2005) za ispitivanje biološke efikasnosti insekticida za suzbijanje žičara. Pomenuta procedura se zasniva samo na proceni sklopa biljaka, što često nije adekvatni pokazatelj usled uticaja i drugih faktora. Zbog toga, dopuna pomenute procedure primenom skale oštećenja daje znatno preciznije informacije o efikasnosti i nivou zaštite koje pružaju pojedini insekticidi. U tom smislu, logistička regresija ima izuzetan značaj kao najvalidnija statistička metoda koja će omogućiti pravilnu interpretaciju rezultata bioloških ogleda ovog tipa.

Statistička obrada podataka

Statistička obrada podataka rađena je u programskom paketu Satistical Package for Social Sciences (SPSS). Uvodimo sledeće promenljive:

Nivo oštećenja – kategorijalna promenljiva koja opisuje nivo oštećenja biljke i to za obradu primenom binomne regresije (0 - nema oštećenja, 1 - ima oštećenja), a za analiziranje primenom multinomne regresije (0 - nema oštećenja, 1 - jedva viljivo oštećenje, 2 - vidljiva oštećenja koja ne utiču na vitalnost biljke, 3- biljka vidno oštećena, ali ima šanse da se oporavi, 4- biljka vidno oštećenja, uvenula bez mogućnosti oporavka ili u potpunosti uništena);

Lokalitet – kategorijlna promenljiva koja predstavlja lokaciju (polje) na kojoj se nalazio usev. Postoje dve kategorije, Bački jarak i Novi Sad.

Tretman – kategorijalna promenljiva koja predstavlja vrstu insekticida korišćenog prilikom tretmana biljke. Kategorije promenljive Tretman su: Attracap, Force 20CS, Force 1.5G, Buteo Start, Lumiposa, Sonido, kontrola.

Ukupan broj posmatranih biljaka je 890. Unutar svake promenljive za svaku kategoriju je određen procenat biljaka koji joj pripada. Najmanje je biljaka koje imaju nivo oštećenja 4 (2%), Sonido je korišćn kao tretman zaštite kod najmanje biljaka (5.6%) i u Bačkom jarku je manje uzgajano biljaka nego u Novom Sadu.

Tabela 15: Udeo sveke kategorije promenljive u ukupnom uzorku

		N	Učešće
Nivo ostecenja	,0	251	28,2%
	1,0	318	35,7%
	2,0	210	23,6%
	3,0	93	10,4%
	4,0	18	2,0%
Tretman	Attracap (1)	140	15,7%
	Force 20CS (2)	140	15,7%
	Force 1.5G (3)	140	15,7%
	Buteo Start (4)	140	15,7%
	Lumiposa (5)	140	15,7%
	Sonido (6)	50	5,6%
	Kontrola (7)	140	15,7%
Lokalitet	Bački jarak (1)	350	39,3%
	Novi Sad (2)	540	60,7%
Total		890	

Kako bi se prezentovao primer upotrebe binarne logističke regresije, definiše se nova promenljiva Oštećenje na sledeći način

$$\boldsymbol{Oštećenje} = \begin{cases} 0, ako\ Nivo\ oštećenja\ \in \{0,1\} \\ 1, ako\ Nivo\ oštećenja\ \in \{2,3,4\} \end{cases}.$$

Odnosno, ukoliko biljka nema oštećanja ili su ona jedva vidna smatramo se da biljka nema oštećenje. Ukoliko je oštećenje vidljivo ali ne utiče na vitalnost biljke, vidljivo ali ima šanse da se oporavi, vidljivo i biljka je bez mogućnosti za oporavak ili je biljka u potpunosti uništena smatramo da postoji oštećenje biljke. Zavisna promenljiva binarne regresije će biti Oštećenje i analizira se uticaj promenljivih Lokalitet i Tretman. Nakon sprovođenja binarne regresije prelazi se na multinomnu i analizira se uticaj promenljivih Lokalitet i Tretman na Nivo oštećenja.

4.1 Binarna logistička regresija

U programskom paketu SPSS, primenjena je binarna logistička regresija sa zavisnom promenljivom **Oštećenje** i nezavisnim promenljivima **Lokalitet** i **Tretman**. Za referentnu vrednost lokaliteta uzet je Novi Sad. Kako bi se analiza uticaja tretmana na oštećenje biljke vršila u odnosu na biljke koje nisu tretirane insekticidima, za referentnu vrednost tretmana uzeta je kategorija kontrola. Vrlo je bitno prilikom samog procesa sprovođenja binarne regresije jasno definisati referentne kategorije kako bi se kasnije izvršilo pravilno tumačenje dobijenih odnosa šansi (OR). Nakon pokretanja binarne logističke regresije u SPSSu dobija se tabela kodiranja za nezavisne kategorijalne promenljive iz koje se vidi koja je referentna kategorija svake promenljive.

Tabela 16: Kodiranje nezavisnih promenljivih

		Frequency	Parameter coding					
			(1)	(2)	(3)	(4)	(5)	(6)
Tretman	Attracap	140	1,000	,000	,000	,000	,000	,000
	Force 20CS	140	,000	1,000	,000	,000	,000	,000
	Force 1.5G	140	,000	,000	1,000	,000	,000	,000
	Buteo Start	140	,000	,000	,000	1,000	,000	,000
	Lumiposa	140	,000	,000	,000	,000	1,000	,000
	Sonido	50	,000	,000	,000	,000	,000	1,000
	kontrola	140	,000	,000	,000	,000	,000	,000
Lokalitet	Bački jarak	350	1,000					
	Novi Sad	540	,000					

Unutar svake nezavisne promenljive kategorija čije su sve vrednosti parametara kodiranja u tabeli 16 jednake nuli je referentna vrednost posmatrane promenljive. Kodiranje lokaliteta je izvršeno na sledeći način: 1 – Bački jarak, dok je referentna kategorija Novi Sad. Kategorije tretmana su kodirane na sledeći način: 1 – Attracap, 2 – Force 20CS, 3 – Force 1.5G, 4 – Buteo Start, 5 – Lumiposa, 6 – Sonido, dok je kategorija kontrola referentna.

Iz tabele 17 očitava se da je vrednost test statistike za Hosmer-Lemeshow test 9.627 sa 8 stepeni slobode. Najbitnija je p-vrednost test statistike i ona iznosi 0.292. Kako je signifikantnost veća od 0.05 na osnovu Hosmer-Lemeshow testa zaključujemo da se model dobro slaže sa podacima.

Tabela 17: Hosmer-Lemeshow test

Hosmer and Lemeshow Test

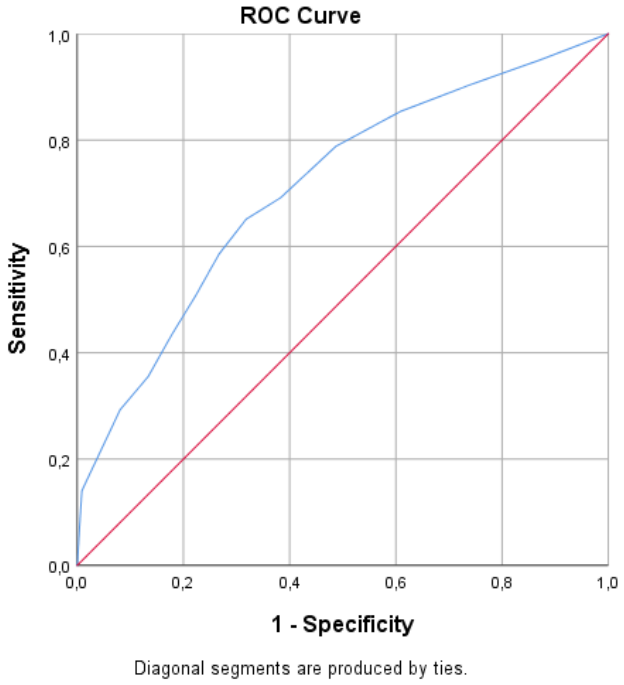
Step	Chi-square	Df	Sig.
1	9,627	8	,292

Na osnovu tabele klasifikacije dobijene prilikom pokretanja binarne logističke regresije u programskom paketu SPSS računa se specifičnost i senzitivnost.

Tabela 18: Tabela klasifikacije

			Predicted		
			Ostecenje		Percentage Correct
	Observed		,0	1,0	
Step 1	Ostecenje	,0	468	101	82,2
		1,0	182	139	43,3
	Overall Percentage				68,2

Senzitivnost iznosi $\frac{139}{240} = 57.92\%$. Ovaj rezultat pokazuje da 57.92% oštećenih biljaka iz uzorka model i klasifikuje kao oštećenje, dok će 42.08% oštećenih biljaka proglasiti za neoštećene. Specifičnost je $\frac{468}{650} = 72\%$, dakle 72% neoštećenih biljaka uzorka model dobro klasifikuje, dok je za 28% pogrešno predviđanje. Ukupna stopa tačne klasifikacije model je $\frac{139+468}{890} = 68.2\%$. Ovi razultati su ukoliko se uzme vrednost cut-off verovatnoće 0.05. Ukoliko se na osnovu ocenjenih verovatnoća pomoću modela odredi ROC kriva dolazi se do zaključka da model prihvatljivo razdvaja biljke na oštećene i neoštećene (površina ispod ROC krive je 71.1%).



Slika 6: ROC kriva

Iz tabele 19 donose se zaključci o značajnosti pojedinih kategorija nezavisnih promenljivih u modelu i interpretiraju se dobijeni odnosi šansi.

Tabela 19: Promenljive u modelu

Variables in the Equation

	B	S.E.	Wald	Df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Lokalitet(1)	1,061	,1651	43,568	1	,000	2,890	2,109	3,961
Tretman			78,215	6	,000			
Attracap(1)	-1,695	,267	40,209	1	,000	,184	,109	,310
Force 20Cs(2)	-2,181	,285	58,681	1	,000	,113	,065	,197
Force 1.5G(3)	-1,885	,273	47,624	1	,000	,152	,089	,259
ButeoStart(4)	-1,350	,260	27,038	1	,000	,259	,156	,431
Lumiposa(5)	-1,553	,264	34,660	1	,000	,212	,126	,355
Sonido(6)	-1,445	,358	16,312	1	,000	,236	,117	,475
Constant	,384	,189	4,121	1	,042	1,468		

Koeficijent uz promenljivu Lokalitet(1) je pozitivan i statistički značajno različit od nule jer je p-vrednost Waldove test statistike 0. Prema tome, Lokalitet(1) utiče na pojavu oštećena. Kako je odnos šansi $OR = 2.89$ sledi da je šansa za pojavom oštećenja u Bačkom jarku 2.89 puta veća u odnosu na šansu pojave oštećenja u Novom Sadu. 95% interval poverenja ne sadrži jedinicu, prema tome $OR=2.89$ je statistički značajno različit od jedinice. Obzirom da je brojnost žičara dosta veća u Bačkom jarku, izvedeni zaključak se i prirodno nameće.

Koeficijenti uz sve vrste tretmana (osim kontrolnog) su negativni, što znači da svi insekticidi utiču na smanjenje verovatnoće za nastankom oštećenja biljke. Takođe, na osnovu Wald testa svaka primena insekticida je značajna za model i koeficijent uz njih je statistički značajno različit od nule. Interval poverenja za odnos šansi bilo kog tretmana je statistički značajno različit od jedinice jer njegov interval poverenja ne sadrži jedinicu. Odnos šansi koji odgovara tretmanu j označava se sa OR_j.

$OR_{Attracap} = 0.184$, odnosno šansa za pojavom oštećenja je 5.43 $\left(\frac{1}{0.184} = 5.43\right)$ puta manja prilikom primene insekticida Attracap u odnosu na šansu za oštećenjem kod netretiranih biljaka. Tabelarno prikazujemo za svaku vrstu tretmana koliko je šansa za pojavom oštećenja prilikom njegovog korišćenja manja u odnosu na šansu za pojavom oštećenja unutra kontrolne grupe (tabela 20).

Tabela 20: Odnos šanse za oštećenjem primenom inekticitida i šanse oštećenja kontrolne grupe

Tretman	OR	Koliko je puta šansa za pojavom oštećenja manja prilikom korišćenja određenog insekticida u odnosu na šansu za oštećenjem u kontrolnoj grupi?
Unošenje u zemljište		
Attracap	0.184	5.44
Force 1.5G	0.152	***6.57***
Tretman semena pre setve		
Force 20CS	0.113	***8.84***
Buteo Start	0.259	3.86
Lumiposa	0.212	4.72
Sonido	0.236	4.23

Iznosi se krajnji zaključak posmatranog binarnog logističkog modela. Unutar grupe insekticida koji se unose u zemljište, najveće smanjenje šanse za oštećenjem biljke u odnosu na šansu za oštećenjem netretiranih biljka ima Force1.5G. Među insekticidima kojima se seme tretira pre setve najveće smanjenje šanse za oštećenje biljaka u odnosu na šansu za oštećenjem biljaka unutar kontrolne grupe ima Force 20CS. Gledano u globalu, Force 20CS ima najveći doprinos smanjenju šanse za oštećenjem biljaka u poređenju sa šansom za oštećenje netretiranih biljaka.

Sprovodi se binarna logistička regresija uz odabir Attracap insekticida kao referentne kategorije tretmana umesto kontrolne, dok je za referentnu vrednost Lokaliteta zadržan Novi Sad (tabela 21).

Tabela 21: Attracap kao referentna kategorija promenljive Tretman

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)	(5)	(6)
Tretman	Attracap	140	,000	,000	,000	,000	,000	,000
	Force 20Cs	140	1,000	,000	,000	,000	,000	,000
	Force 1.5G	140	,000	1,000	,000	,000	,000	,000
	Buteo Start	140	,000	,000	1,000	,000	,000	,000
	Lumiposa	140	,000	,000	,000	1,000	,000	,000
	Sonido	50	,000	,000	,000	,000	1,000	,000
	Kontrola	140	,000	,000	,000	,000	,000	1,000
Lokalitet	Bački jarak	350	1,000					
	Novi Sad	540	,000					

Dobijaju se isti rezultati tabele klasifikacije, Hosmer-Lemeshow testa, ROC krive i predviđene verovatnoće oštećenja kao i u slučaju odabira kontrole kao referentne kategorije tretmana. Jedina razlika će biti u delu tabele ocenjenih koeficijenata modela koji odgovara promenljivoj Tretman. Izborom Attracap-a kao referentne kategorije omogućava se da program izbaci pokazatelje pomoću kojih ćemo upoređivati šansu za oštećenjem unutar svake nereferentne kategorije i šansu za oštećenjem biljaka pri korišćenju Attracapa.

Tabela 22: Koeficijenti modela kod kojeg je referentni tretman Attracap

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for Lower	95% C.I.for Upper
Lokalitet(1)	1,061	,161	43,568	1	,000	2,890	2,109	3,961
Tretman			78,215	6	,000			
Force 20CS(1)	-,486	,287	2,870	1	,090	,615	,351	1,079
Force 1.5G(2)	-,190	,276	,473	1	,492	,827	,482	1,420
Buteo Start(3)	,345	,264	1,715	1	,190	1,412	,842	2,367
Lumiposa(4)	,143	,267	,285	1	,594	1,153	,683	1,947
Sonido(5)	,250	,353	,502	1	,479	1,284	,643	2,563
kontrola(6)	1,695	,267	40,209	1	,000	5,448	3,226	9,199
Constant	-1,311	,204	41,141	1	,000	,270		

Na osnovu Wald testa ponovo se dobija da promenljive Lokalitet i Tretman imaju značajnu ulogu u predviđanju oštećenja biljke i treba ih ostaviti u model. $OR_{Force\ 20CS} = 0.615$, odnosno šansa za pojavom oštećenja je 1.63 $\left(\frac{1}{0.615} = 1.63\right)$ puta manja prilikom primene insekticida Force 20CS u odnosu na šansu za oštećenjem kod biljaka tretiranih Attracap-om. Kako 95% interval poverenja za $OR_{Force\ 20CS}$ sadrži jedinicu sledi da on statistički nije značajno različit od jedinice. Tabelarno se prikazuje za svaku vrstu nereferentnog tretmana koliko je šansa za pojavom oštećenja prilikom korišćenja posmatranog tretmana manja u odnosu na šansu za pojavom oštećenja unutra grupe biljaka tretirane Attracap-om (tabela 23).

Tabela 23: Odnos šanse za oštećenjem primenom nereferentnog tretmana i šanse oštećenja pri korišćenju insekticida Attracap

Tretman	OR	OR statistički značajno različit od jedinice?	Koliko je puta šansa za pojavom oštećenja manja/veća prilikom korišćenja određenog insekticida u odnosu na šansu za oštećenjem u grupi biljaka sa primenom Attracapa?
Unošenje u zemljište			
Force 1.5G	0.827	**NE**	manja 1.21 puta
Tretman pre setve			
Force 20CS	0.615	***NE***	manja 1.63 puta
Buteo Start	1.412	**NE**	veća 1.412 puta
Lumiposa	1.153	**NE**	veća 1.153 puta
Sonido	1.284	**NE**	veća 1.284 puta
Bez primene insekticida			
Kontrola	5.448	**DA**	veća 5.448 puta

Dakle, šansa za oštećenjem kada je biljka tretirana sa Force 1.5G je manja 1.21 put u onosu na šansu za njenim oštećenjem prilikom primene referentnog insekticida Attracap. U odnosu na Attracap bolju efikasnost je zabeležio i Force 20CS čija je šansa za oštećenjem 1.63 puta manja u odnosu na šansu za oštećenjem primenom Attracapa. Zaključak koji se može doneti korišćenjem binarnog logističkog regresijskog modela prilikom izbora kontorlne grupe kao referentne je da su od Attracap-a efikasniji Force 20CS i Force 1.5G (tabela 20), što isto važi i u slučaju izbora Attracap-a kao referentne kategorije. Lošiju efikasnost zabeležili su ostali insekticidi: Buteo Start, Lumiposa, Sonido, kao i tretman koji isključuje korišćenje bilo kog insekticida.

4.2 Multinomna logistička regresija

Kod multinomne logističke regresije posmatra se uticaj promenljivih Tretman i Lokalitet na verovatnoću da promenljiva Nivo oštećenja uzme određenu kategorijalnu vrednost. Za referentu kategoriju zavisne promenljive uzet je nivo oštećenja 0 - nema oštećenja. Kako Nivo oštećenja ima pet kategorija, broj logit funkcija će biti 4: $g_j(x_1, x_2) = \ln\left(\frac{P\{Nivo\ oštećenja = j|x_1, x_2\}}{P\{Nivo\ oštećenja = 0|x_1, x_2\}}\right)$, $j = 1, 2, 3, 4$ i na osnovu njih se računaju verovatnoće

$$\pi_j(\boldsymbol{x}) = P\{Y = j|\boldsymbol{x}\} = \frac{e^{g_j(x)}}{1+e^{g_1(x)}+e^{g_2(x)}+e^{g_3(x)}+e^{g_4(x)}}, j = 1, 2, 3, 4.$$

Kako je reč o kategorijalnim nezavisnim promenljivima, uvode se dizajn promenljive.

U tabeli 24 prikazani su rezultati testa količnika verodostojnosti za testiranje značajnosti koeficijenata u modelu. Kako su signifikantne vrednosti za Lokalitet i Tretman jednake nuli, što je manje od 0.05, na osnovu testa količnika verodostojnosti zaključuje se da su obe promenljive statistički značajne i treba ih ostaviti u modelu.

Tabela 24: Test količnika verodostojnosti

Likelihood Ratio Tests

	Model Fitting Criteria	Likelihood Ratio Tests		
Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	204,318[a]	,000	0	.
Lokalitet	285,245	80,927	4	,000
Tretman	359,243	154,925	24	,000

Za procenu slaganja modela sa podacima koristi se uopšten Hosmer-Lemeshow test binarne logističke regresije. Na osnovu tabele 25 vidimo da je vrednost Pirsonove test statistike 36.98 sa 20 stepeni slobode i da je njena p-vrednost 0.012. Kako je p-vrednost manja od 0.05 zaključak je da se model ne slaže dobro sa podacima.

Tabela 25: Pirsonova test statistika

Goodness-of-Fit

	Chi-Square	Df	Sig.
Pearson	36,980	20	,012
Deviance	39,890	20	,005

U tabeli 26 prikazane su ocene koeficijenata za svaki od četiri logit modela. Kao referentna vrednost nezavisne promenljive Lokalitet izabran je Novi Sad, dok je kod promenljive Tretman odabrana kontrolna grupa. Kao i kod binarne logističke regresije, vrlo je bitno i kod multinomne jasno definisati referentne vrednosti kako bi se dobijeni pokazatelji u tabeli 26 pravilno interpretirali. Referentna vrednost svake nezavisne promenljive u tabli 26 koju prikazuje SPSS se prepoznaje po tome što je ocenjen koeficijent uz referentnu kategoriju 0[b]. Referentna vrednost nezavisne promenljive bira sam korisnik programskog paketa prilikom poziva naredbe za multinomnu logističku regresiju i odabira zavisne promenljive modela.

Tabela 26: Koeficijenti multinomnog logističkog modela

Parameter Estimates

Nivo ostecenja		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower	Upper
1	Intercept	,623	,324	3,693	1	,055			
	[Lokalitet=Bački jarak]	1,171	,212	30,431	1	,000	3,225	2,127	4,889
	[Lokalitet=Novi Sad]	0[b]	.	.	0	.	.	.	.
	[Tretman=Attracap]	-,662	,383	2,987	1	,084	,516	,243	1,093
	[Tretman=Force 20CS]	-1,214	,379	10,288	1	,001	,297	,141	,624
	[Tretman=Force 1.5G]	-1,193	,381	9,783	1	,002	,303	,144	,641
	[Tretman=Buteo Start]	-,470	,390	1,451	1	,228	,625	,291	1,343
	[Tretman=Lumiposa]	-,369	,388	,904	1	,342	,691	,323	1,479
	[Tretman=Sonido]	-,135	,659	,042	1	,837	,873	,240	3,179
	[Tretman=kontrola]	0[b]	.	.	0	.	.	.	.
2	Intercept	,707	,318	4,946	1	,026			
	[Lokalitet=Bački jarak]	1,744	,232	56,706	1	,000	5,719	3,633	9,005
	[Lokalitet=Novi Sad]	0[b]	.	.	0	.	.	.	.
	[Tretman=Attracap]	-1,661	,402	17,070	1	,000	,190	,086	,418
	[Tretman=Force 20CS]	-2,261	,403	31,469	1	,000	,104	,047	,230
	[Tretman=Force 1.5G]	-2,017	,396	25,888	1	,000	,133	,061	,289
	[Tretman=Buteo Start]	-1,419	,408	12,117	1	,000	,242	,109	,538
	[Tretman=Lumiposa]	-1,238	,401	9,542	1	,002	,290	,132	,636
	[Tretman=Sonido]	-,792	,658	1,451	1	,228	,453	,125	1,644
	[Tretman=kontrola]	0[b]	.	.	0	.	.	.	.
3	Intercept	,661	,323	4,184	1	,041			
	[Lokalitet=Bački jarak]	1,797	,295	37,219	1	,000	6,032	3,386	10,744
	[Lokalitet=Novi Sad]	0[b]	.	.	0	.	.	.	.
	[Tretman=Attracap]	-2,709	,479	31,919	1	,000	,067	,026	,170
	[Tretman=Force 20CS]	-4,079	,614	44,202	1	,000	,017	,005	,056
	[Tretman=Force 1.5G]	-3,577	,541	43,687	1	,000	,028	,010	,081
	[Tretman=Buteo Start]	-1,723	,426	16,385	1	,000	,179	,078	,411
	[Tretman=Lumiposa]	-2,829	,527	28,854	1	,000	,059	,021	,166
	[Tretman=Sonido]	-2,458	,807	9,274	1	,002	,086	,018	,416
	[Tretman=kontrola]	0[b]	.	.	0	.	.	.	.
4	Intercept	-1,117	,513	4,749	1	,029			
	[Lokalitet=Bački jarak]	2,693	,551	23,918	1	,000	14,776	5,022	43,477
	[Lokalitet=Novi Sad]	0[b]	.	.	0	.	.	.	.
	[Tretman=Attracap]	-3,113	,845	13,562	1	,000	,044	,008	,233
	[Tretman=Force 20CS]	-	7299,607	,000	1	,998	2,533E-	,000	,000
	[Tretman=Force 1.5G]	-4,194	1,101	14,497	1	,000	,015	,002	,131
	[Tretman=Buteo Start]	-3,541	1,103	10,301	1	,001	,029	,003	,252
	[Tretman=Lumiposa]	-2,453	,742	10,919	1	,001	,086	,020	,369
	[Tretman=Sonido]	-	,000	.	1	.	2,470E-	2,470E-10	2,470E-10
	[Tretman=kontrola]	0[b]	.	.	0	.	.	.	.

Za svaku logit funkciju se posebno analiziraju njeni koeficijenti uz nezavisne promenljive. Posmatraju se informacije o prvom logitu koje su date u tabeli 26 u delu koji odgovara nivou oštećenja 1. Broj 3,225 pokazuje da je šansa za pojavom jedva vidljivog oštećenja u odnosu na neoštećenje unutrar grupe biljaka uzgajanih u Bačkom jarku veća za 3.225 puta nego u Novom Sadu. Što se tiče tretmana, najefikasniji je Force 20 CS. Šansa za pojavom jedva vidljivog oštećenja u odnosu na neoštećenje je 3.37 $\left(\frac{1}{0.297} = 3.37\right)$ puta manja kod biljaka tretiranih insekticidom Force 20 CS nego kod netretiranih biljaka. Kako interval poverenja posmatrnog odnosa šansi ne sadrži jedinicu sledi da je on statistički značajan. Najmanje efikasan tretman je Sonido i odnos šansi koji njemu odgovara nije statistički značajno različit od jedinice, jer interval poverenja (0.240, 3.179) sadrži jedan. Rangirani tretmani po efikasnosti su dati u tabeli 27.

Tabela 27: Rangiranje tretmana prema efikasnosti u zaštiti od nivoa oštećenja 1

Tretman	OR	Koliko je puta šansa za pojavom nivoa oštećenja 1 – jedva vidljivog oštećenja u odnosu na neoštećenje manja prilikom korišćenja određenog insekticida u odnosu na posmatranu šansu kod kontrolne grupe? (1/OR)	Da li je OR statistički značajno različit od jedinice?
Force 20CS	0.297	***3.37***	*DA*
Force 1.5G	0.303	3.3	DA
Attracap	0.516	1.94	NE
Buteo Start	0.625	1.6	NE
Lumiposa	0.691	1.45	NE
Sonido	0.873	***1.14***	*NE*

Na osnovu informacija u delu tabele 26 koji odgovara drugom logitu zaključujemo da je šansa za pojavom oštećenja nivoa 2, odnosno vidljivim oštećenjima koja ne utiču na vitalnost biljke, u odnosu na nepostojanje oštećenja u Bačkom jarku veća za 5.719 puta nego u Novom Sadu. U okviru nivoa oštećenja 2 takođe se dobija da je najefikasniji tretman Force 20CS jer je šansa za pojavu vidljivih oštećenja koja ne utiču na vitalnost biljke u odnosu na neoštećenje 9.61 $\left(\frac{1}{0.104} = 9.61\right)$ puta manja kod biljaka tretiranih ovih insekticidom nego kod netretiranih biljaka. Ostali tretmani manje smanjuju šansu za pojavom vidljivih oštećenja bez uticaja na vitalnost biljke u odnosu na netretirane stabljike suncokreta: Force 1.5G smanjuje $\frac{1}{0.133} = 7.52$ puta, Attracap $\frac{1}{0.190} = 5.26$, Buteo Start $\frac{1}{0.242} = 4.13$, Lumiposa $\frac{1}{0.29} = 3.45$, Sonido $\frac{1}{0.453} = 2.21$ put. Sonido je

najmanje efikasan i njegov odnos šansi nije statistički značajan, što je posledica malog udela biljaka tretiranih Sonido insekticidom u ukupnom uzorku.

Tabela 28: Rangiranje tretmana prema efikasnosti u zaštiti od nivoa oštećenja 2

Tretman	OR	Koliko je puta šansa za pojavom nivoa oštećenja 2 u odnosu na neoštećenje manja prilikom korišćenja određenog insekticida u odnosu na posmatranu šansu kod kontrolne grupe? (1/OR)	Da li je OR statistički značajno različit od jedinice?
Force 20CS	0.104	***9.61***	*DA*
Force 1.5G	0.133	7.52	DA
Attracap	0.190	5.26	DA
Buteo Start	0.242	4.13	DA
Lumiposa	0.290	3.45	DA
Sonido	0.453	***2.21***	*NE*

Analizira se logit koji odgovara nivou oštećenja 3 – biljka vidno oštećenja, ali ima šanse da se oporavi. Šansa za pojavom nivoa oštećenja 3 u odnosu na neoštećenje unutrar grupe biljaka uzgajanih na lokalitetu Bački jarak veća je za 6.032 puta nego na lokalitetu Novi Sad. Šansa za nivoa oštećenja 3 u odnosu na neoštećenje je 58.82 puta manja prilikom korišćenja Force 20CS nego kod netretiranih biljaka. Iz tabele 29 se zaključuje da je insekticidit Lumiposa dosta poboljšao svoju efikasnost u grupi nivoa oštećenja 3 u odnosu na nivo oštećenja 2 i sada se nalazi na trećem mestu, međutim i dalje je ta efikasnost dosta manja u odnosu na prvorangirani insekticid. Najmanju efikasnost u smanjenju šanse za oštećenjem nivoa 3 ima Buteo Start.

Tabela 29: Rangiranje tretmana prema efikasnosti u zaštiti od nivoa oštećenja 3

Tretman	OR	Koliko je puta šansa za pojavom nivoaoštećenja 3 u odnosu na neoštećenje manja prilikom korišćenja određenog insekticida u odnosu na posmatranu šansu kod kontrolne grupe? (1/OR)	Da li je OR statistički značajno različit od jedinice?
Force 20CS	0.017	***58.82***	*DA*
Force 1.5G	0.028	35.71	DA
Lumiposa	0.059	16.95	DA
Attracatp	0.067	14.92	DA
Sonido	0.086	11.63	DA
Buteo Start	0.179	***5.59***	DA

Konačno, analiziranjem informacija iz tabele 26 o četvrtom logitu, zaključak je da je šansa za pojavom nivoa oštećenja 4 (biljka uvenula ili u potpunosti uništena) u odnosu na neoštećenje unutrar grupe biljaka uzgajanih u Bačkom jarku veća za 14.776 puta nego u Novom Sad. Kako u datom uzorku ne postoji oštećenje nivoa 4 kada se primenjuje Force 20CS i Sonido, dobijaju se veoma veliki negativni koeficijenti uz posmatrane tretmane. Ova činjenica dalje implicira da će šansa za pojavom potpunog uništenja u odnosu na neoštećenje biti mnogo puta manja prilikom korišćenja tretmana Force 20CS ili Sonido nego posmatrana šansa kod kontrolne grupe. Među ostalim tretmanima najefikasniji je Force 1.5G, Buteo Start je dosta poboljšao efikasnost u odnosu na grupu nivoa oštećenja 3, dok je Lumiposa zauzela poslednje mesto rangiranja (tabela 30).

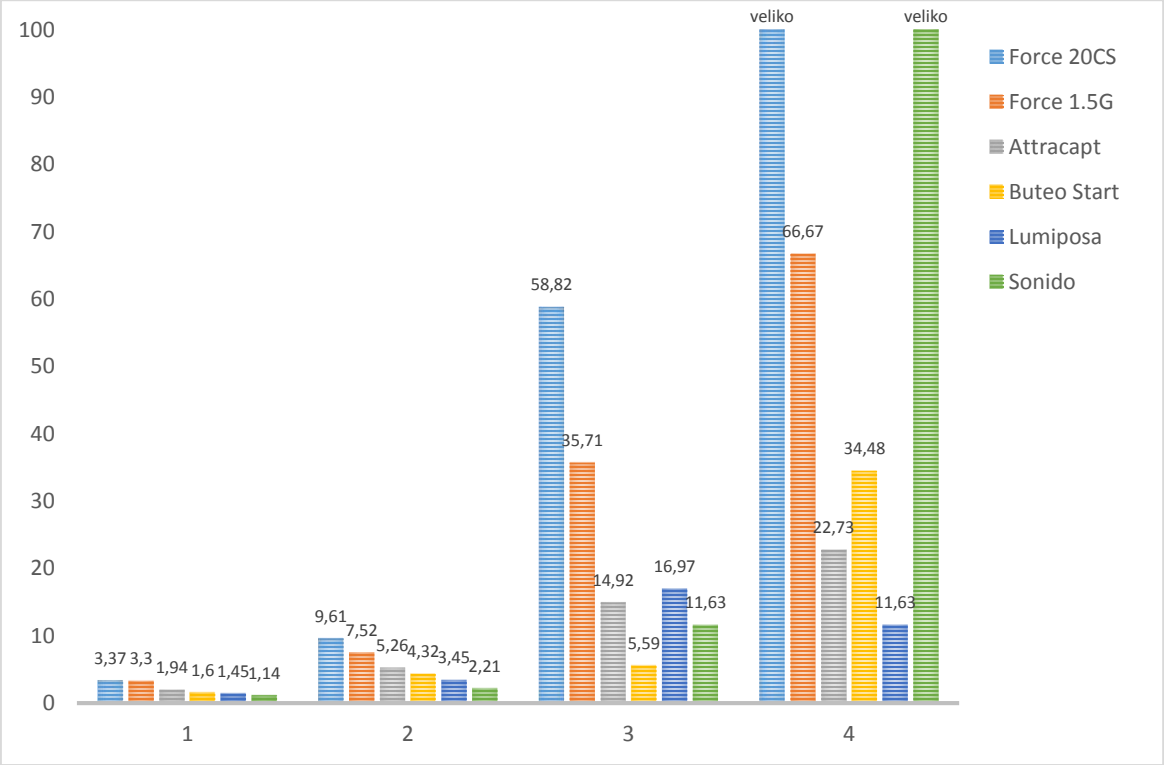
Tabela 30: Rangiranje tretmana prema efikasnosti u zaštiti od nivoa oštećenja 4

Tretman	OR	Koliko je puta šansa za pojavom nivoa oštećenja 4 (biljka uvenulli ili u potpunosti uništena) u odnosu na neoštećenje manja prilikom korišćenja određenog insekticida u odnosu na posmatranu šansu kod kontrolne grupe? (1/OR)	Da li je OR statistički značajno različit od jedinice?
Force 1.5G	0.015	66.67	DA
Buteo Start	0.029	34.48	DA
Attracatp	0.044	22.73	DA
Lumiposa	0.086	***11.63***	DA

Na osnovu sprovedene multinomne logističke regresije, šansa za svakim nivoom oštećenja je statistički značajno puta veća u Bačkom jarku nego u Novom Sadu. Ovaj rezultat je i očekivan obzirom da su ova dva lokaliteta oglednih polja Instituta za ratarstvo i povrtarstvo slična po pitanju klimatskih uslova i sastava zemljišta, ali brojnost žičara je dosta veće na polju u Bačkom jarku. Unutar svakog nivoa oštećenja kao najbolji tretman u cilju zaštite suncokreta od žičara pokazao se Force 20CS. Šansa za pojavom oštećenja nivoa 1 u odnosu na neoštećenje je 3.37 puta manja prilikom korišćenja Force 20CS nego kod netretiranih biljaka. Kako se nivo oštećenja povećava tako se povećava broj koji pokazuje koliko je puta šansa za pojavom posmatranog nivoa oštećenja u odnosu na neoštećenje manja prilikom korišćenja Force 20CS nego kod netretiranih biljaka. Ova činjenica nije prisutna samo kod tretmana Force 20CS. Na osnovu tabele 31 i njenog grafičkog prikaza na slici 7, zapaža se povećanje značajnosti korišćenja svakog insekticida usled viših kategorija nivoa oštećenja.

Tabela 31: Povećanje značajnosti u smanjivanju šanse za oštećenjem prilikom rasta nivoa oštećenja

Tretman	Koliko je manja šansa za pojavom posmatranog nivoa oštećenja u odnosu na neoštećenje prilikom korišćenja određenog tretmana nego kod biljaka bez tretmana?			
	Nivo oštećenja 1	Nivo oštećenja 2	Nivo oštećeja 3	Nivo oštećenja 4
Force 20CS	3.37	9.61	58.82	veoma puno puta
Force 1.5G	3.3	7.52	35.71	66.67
Attracap	1.94	5.26	14.92	22.73
Buteo Start	1.6	4.32	5.59	34.48
Lumiposa	1.45	3.45	16.97	11.63
Sonido	1.14	2.21	11.63	veoma puno puta



Slika 7: Povećanje značajnosti u smanjivanju šanse za oštećenjem prilikom rasta nivoa oštećenja

5. Zaključak

Logistička regresija ima veoma široku primenu u različitim sferama ljudskih delatnosti obzirom da je veliki broj pojava predstavljen kategorijalnom promenljivom. Kao primer, navedena je njena primena u poljoprivredi usled izbora optimalnog insekticida u zaštiti biljaka suncokreta hibrid Romeo. Prilikom ogleda za ispitivanje biološke efikasnosti insekticida u suzbijanju žičara posmatrano je 890 biljaka suncokreta na dva polja Instituta za poljoprivredu i ratarstvo. Analizirao se uticaj šest različitih insekticida: Attracap, Force 20CS, Force 1.5G, Buteo Start, Lumiposa i Sonido. Attracap i Force 1.5G predstavljaju tretmane koji se primenjuju inkorporacijom, dok se ostali insekticidi koriste za tretman semena pre setve.

Prilikom sprovođena binarne logističke regresije u programskom paketu SPSS i korišćenjem teorijskog dela rada, ustanovilo se da je binarni logistički model pogodan u predviđanju verovatnoće oštećenja suncokreta zavisno od lokaliteta i tretmana. Unutar grupe insekticida koji se unose u zemljište, najveće smanjenje šanse za oštećenjem biljke u odnosu na šansu za oštećenjem netretiranih biljka ima Force1.5G, odnosno njegov učinak u zaštiti bilja od oštećenja prilikom najznačajnijih zemljišnih štetočina žičara je efikasniji nego primena Attracap granula. Među insekticidima kojima se seme tretira pre setve najveće smanjenje šanse za oštećenje biljaka u odnosu na šansu za oštećenjem biljaka unutar kontrolne grupe ima Force 20CS. Posmatranjem obe tehnike tretmana, najmanju uspešnost ima Buteo Start i najveću Force 1.5G.

Prilikom primene multinomne logističke regresije uvele su se kategorije koje predstavljaju nivo oštećenja biljke: 0 - nema oštećenja, 1 - jedva vidljivo oštećenje, 2 - vidljiva oštećenja koja ne utiču na vitalnost biljke, 3- biljka vidno oštećena, ali ima šanse da se oporavi, 4- biljka vidno oštećenja, uvenula bez mogućnosti oporavka ili je u potpunosti uništena. Tada se logistički model koristio u predviđanju verovatnoće za pojavom određenog nivo oštećenja prilikom fiksiranog lokaliteta i vrste tretmana. Putem Hosmer-Lemeshow testa ustanovljenoje da multinomna logistička regresija nema dobro slaganje sa podacima, te se binarna logistička regresija bolje pokazala u sprovođenju posmatrane analize. Unutar svakog nivoa oštećenja (sem nivoa 0) najveću uspešnost je takođe zabeležio Force 1.5G, u smislu najvećem doprinosu smanjenja šanse za pojavom fiksiranog nivoa oštećenja u odnosu na šansu za pojavom oštećenja prilikom korišćenja drugih insekticida.

Literatura

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

[2] Gorica Gvozdić, Primenjena logistička regresija, master rad, 2011.

[3] Zagorka Lozanov-Crvenković, Statistika, 2012

[4] Alan Agresti, Categorical Data Analysis, 2013.

[5] Ronald Christensen, Log-Linear Models and Logistic Regression. Springer, 1997.

[6] Hallett C. David, Goodness of Fit Test in Logistic Regression, 1999.

[7] John O. Rawlings, Sastry G. Pantual, David A. Dickey, Applied Regression Analysis, 1998.

[8] Joseph M. Hilbe, Logistic regression Models, 2009.

[9] Gvozdenac S, Ovuka J, Miklič V, Cvejić S, Tanasković S, Bursić V, Sedlar A, The effect of seed treatments on wireworm (Elateridae) performance, damages, and yield traits of sunflower (Helianthus annuus L), 2019.

[10] Poggi, S; Le Cointe, R; Lehmhus, J; Plantegenest, M; Furlan, L; Alternative Strategies for Controlling Wireworms in Field Crops: A Review, 2021.

[11] Muhammad Alamgir Islam, Mode od Transportation Choice in Bangladesh, naučni rad, 2020.

[12] Multinomial goodnes-of-fit test for logistic regresion model, Statistics in Medicine, 27, 2008.

[13] Menard Scott, Applied Logistic Regression Analysis - second edition, 2001.

[14] Morten W. Fagerland, David W. Hosmer, A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models, 2012.

[15] Scott A. Czepiel, Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation.

[16] Morten W. Fagerland, David W. Hosme, How to test for goodness of fit in ordinal logistic regression models, 2017.

[17] Danijela Rajter-Ćirić, Verovatnoća, 2009.

[18] Menard, S. Applied logistic regression analysis, 2nd Edition, 2002.

[19] Julie Pallant, SPSS priručnik za preživljavanje, prevod (Miljenko Šućur) treće izdanje, 2009.

[20] IBM SPSS Guide

[21] Daniel Arkkelin, Using SPSS to Understand Research and Data Analysis, 2014.