doi:10.14720/aas.2024.120.3.19353 Original research article / izvirni znanstveni članek

Logistična regresija in problem ločenosti

Hana ŠINKOVEC[1,2], Damijana KASTELEC[1], Luka BITEŽNIK[1]

Received July 22, 2024; accepted September 08, 2024.
Delo je prispelo 22. julij 2024, sprejeto 08. september 2024

Logistična regresija in problem ločenosti

Izvleček: Logistična regresija se uporablja za preučevanje povezanosti med binarno odzivno spremenljivko (nek dogodek se zgodi ali pa ne) in množico neodvisnih spremenljivk. Z modelom lahko napovemo verjetnosti dogodka za nove enote, pogojno na vrednosti neodvisnih spremenljivk. Poleg tega ocene parametrov modela, ki jih dobimo z metodo največjega verjetja, lahko interpretiramo kot logaritem razmerja obetov. Kadar so vzorci majhni ali dogodki, v kateri od skupin, ki določajo neodvisno spremenljivko, redki, se lahko zgodi, da algoritem po metodi največjega verjetja ne konvergira, ocene parametrov modela pa so posledično nesmiselne. Pojav se v statistiki imenuje »ločenost«. Ker programska oprema problema pogosto ne identificira, raziskovalci ločenost v praksi lahko prezrejo. Dobljeni rezultati raziskovalce lahko begajo ali pa jih napačno interpretirajo. S člankom zato želimo: motivirati uporabo logistične regresije za preučevanje povezanosti binarne odzivne spremenljivke z množico neodvisnih spremenljivk; bralcem predstaviti problem ločenosti na konkretnem primeru; pokazati, kako problem ločenosti premostiti.

Klučne besede: logistična regresija, metoda največjega verjetja, majhni vzorci, redki dogodki, ločenost, Firthov tip penalizacije

Logistic regression and the problem of separation

Abstract: Logistic regression is used to study the relationship between a binary outcome variable (one event may occur or not) and a set of covariates. Individualized prognosis can be obtained by estimating the probability of an event given the covariates. Moreover, regression coefficients, usually estimated by the method of maximum likelihood, can be interpreted as the log odds ratios. In situations where the data are small or sparse, the likelihood maximization algorithm may fail to converge, leading to implausible parameter estimates. In statistics, this situation is known as 'separation'. In practice, separation may go unnoticed due to software limitations in identifying the problem. The results obtained from such analyses can be puzzling and may be misinterpreted. Therefore, in this manuscript, we aim to: motivate the use of logistic regression to study the relationship between a binary outcome and a set of covariates; demonstrate the problem of separation with a real-data example; and show how to overcome separation.

Key words: logistic regression, maximum likelihood, small datasets, sparse datasets, separation, Firth's penalized likelihood

1 Univerza v Ljubljani, Biotehniška fakulteta, Oddelek za agronomijo, Ljubljana, Slovenija
2 korenspondenčni avtor: hana.sinkovec@bf.uni-lj.si

1 UVOD

Raziskovalci se v svojih raziskavah pogosto osredotočajo na preučevanje binarnih odzivnih spremenljivk, ki lahko zavzamejo le dve možni vrednosti – nek dogodek se zgodi ali pa ne. Tipični primeri iz biotehnike so: okuženost rastline z neko boleznijo (rastlina je okužena ali ne); zdravstveno stanje živine (prisotnost ali odsotnost bolezni); kalivost semen na določenih tleh (seme kali ali ne); učinkovitost posega, npr. gnojenja, pesticidov, s katerim dosežemo nek rezultat ali ne; prisotnost ali odsotnost določene fenotipske značilnosti; dovzetnost za zdravljenje (se odziva na zdravljenje ali ne); status preživetja (preživetje ali smrt).

V tovrstnih raziskavah nas pogosto zanima povezanost med binarno odzivno spremenljivko ter eno ali več neodvisnimi spremenljivkami. V primeru, ko je neodvisna spremenljivka opisna, podatke lahko preprosto analiziramo z računanjem deležev dogodka v vsaki od dveh ali več skupin, ki določajo neodvisno spremenljivko. Tak pristop postane problematičen, ko je neodvisna spremenljivka številska ali ko želimo opisati povezanost med odzivno spremenljivko ter več neodvisnimi spremenljivkami hkrati (Harrell, 2015), kar je tipično na primer v genetiki, kjer so neodvisne spremenljivke lahko številne genetske različice (npr. polimorfizem posameznega nukleotida – SNP). Raziskovalci tako pogosto številske spremenljivke kategorizirajo, kar privede do nepotrebne izgube informacije in pristranskosti, povezanost med odzivno in neodvisnimi spremenljivkami pa ovrednotijo za vsako neodvisno spremenljivko posebej, kar ne upošteva kombinacije hkrati prisotnih genetskih in okoljskih vplivov. To vodi do izgube moči pri odkrivanju pomembnih zvez med odzivno in neodvisnimi spremenljivkami.

Za analizo binarnih odzivnih spremenljivk je tako mnogo bolj primerna uporaba modela logistične regresije, v katerem je upoštevana informacija o vseh neodvisnih spremenljivkah hkrati, ki so lahko tako opisne kot tudi številske. Ocene parametrov modela, ki jih interpretiramo kot logaritem razmerja obetov, lahko dobimo z metodo največjega verjetja. V aplikacijah logistične regresije, predvsem v analizah majhnih vzorcev ali kadar so dogodki (v kateri od skupin, ki določajo neodvisno spremenljivko) redki (sparse data), pa se lahko zgodi, da algoritem po metodi največjega verjetja ne konvergira, ocene parametrov modela pa so posledično nesmiselne. Ta problem je v statistiki poznan kot »ločenost« (separation). Navadno nastane kot posledica (pre)majhnega števila enot v vzorcu, čeprav se lahko pojavi tudi pri analizah večjih vzorcev (van Smeden in sod., 2016). Ker programska oprema problema pogosto ne identificira, raziskovalci ločenost v praksi lahko prezrejo, in posledično rezultate analize povsem napačno interpretirajo. S člankom tako želimo: 1.) motivirati uporabo logistične regresije za preučevanje povezanosti binarne odzivne spremenljivke z množico neodvisnih spremenljivk; 2.) bralcem predstaviti problem ločenosti na konkretnem primeru; 3) pokazati, kako problem ločenosti premostiti.

2 LOGISTIČNA REGRESIJA

V praksi bi pogosto radi ocenili povezanost med naborom neodvisnih spremenljivk $X = \{X_1, X_2, \ldots, X_K\}$ z binarno odzivno spremenljivko Y. Vrednosti, ki jih lahko zavzame Y lahko zapišemo kot 1 ali 0, pri čemer bomo z $Y = 1$ označili pojav dogodka, ki nas zanima. V najpreprostejšem primeru je v podatkih le ena neodvisna spremenljivka (preučevani dejavnik), ki prav tako zavzema le dve vrednosti. Takšne podatke lahko strnemo v kontingenčno tabelo dimenzije 2 × 2:

		X	
		0	1
Y	0	f_{00}	f_{01}
	1	f_{10}	f_{11}

(1)

pri čemer so f_{00}, f_{01}, f_{10}, f_{11} absolutne frekvence različnih kombinacij vrednosti spremenljivk X in Y. Tovrstne podatke bi lahko analizirali z računanjem deležev dogodkov (relativnih frekvenc) v obeh skupinah, ki določata X, in tako ocenili verjetnost proučevanega dogodka pogojno na X: verjetnost dogodka v skupini $X = 0$, $P^0 = P(Y = 1|X = 0)$, bi ocenili kot $f_{10}/(f_{00} + f_{00})$, verjetnost dogodka v skupini $X = 1$, $P^1 = P(Y = 1|X = 1)$, pa kot $f_{11}/(f_{01} + f_{11})$. Vpliv X na Y bi potem lahko kvantificirali kot razliko verjetnosti $P^1 - P^0$ ali pa kot razmerje verjetnosti P^1/P^0 (Agresti, 1990).

Tovrstna analiza postane problematična, kadar so neodvisne spremenljivke številske ali pa ko želimo opisati povezanost Y z več neodvisnimi spremenljivkami X hkrati. Zato se za analizo binarnih odzivnih spremenljivk raje uporablja model logistične regresije, ki omogoča vključitev večjega števila neodvisnih spremenljivk ne glede na tip spremenljivk. Model je formuliran v smislu verjetnosti proučevanega dogodka pogojno na vrednosti neodvisnih spremenljivk:

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_K)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_K)},$$

(2)

pri čemer je β_0 presečišče, $\beta_1 \beta_{2, \ldots,} \beta_K$ pa so regresijski parametri za neodvisne spremenljivke $X_1, X_2, \ldots, X_k$. Model lahko zapišemo tudi v smislu logit transformacije verjetnosti $P(Y = 1|X) = P$, ki je logistični regresiji dala

ime. V tem primeru definiramo t. i. obete kot razmerje verjetnosti $P / (1 - P)$. Model za logaritem obetov je linearna kombinacija neodvisnih spremenljivk:

$$\text{logit}(P(Y = 1|X)) = \text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K. \quad (3)$$

Regresijski parameter β_k lahko interpretiramo kot razliko logaritma obetov ob povečanju X_k za eno enoto ter ob upoštevanju ostalih spremenljivk v modelu. Laže pa parametre interpretiramo v smislu $\exp(\beta_k)$, ki predstavlja razmerje obetov dveh enot, ki se razlikujeta le za eno enoto dane neodvisne spremenljivke X_k, medtem ko imata enake vrednosti ostalih neodvisnih spremenljivk. Kot tak je model logistične regresije primeren za dosego katerega koli od treh statističnih ciljev modeliranja: za opisovanje povezanosti nabora neodvisnih spremenljivk z odzivno spremenljivko Y; za ocenjevanje vpliva preučevanega dejavnika na odzivno spremenljivko Y ob upoštevanju ostalih spremenljivk v modelu; ter za napovedovanje verjetnosti dogodka za nove enote, pogojno na vrednosti neodvisnih spremenljivk (Shmueli, 2010).

V primeru kontingenčne tabele dimenzije 2 × 2 (1) lahko model logistične regresije zapišemo kot

$$\text{logit}(P(Y = 1|X = 0)) = \text{logit}(P^0) = \log\left(\frac{P^0}{1-P^0}\right) = \beta_0$$

$$\text{logit}(P(Y = 1|X = 1)) = \text{logit}(P^1) = \beta_0 + \beta_1.$$

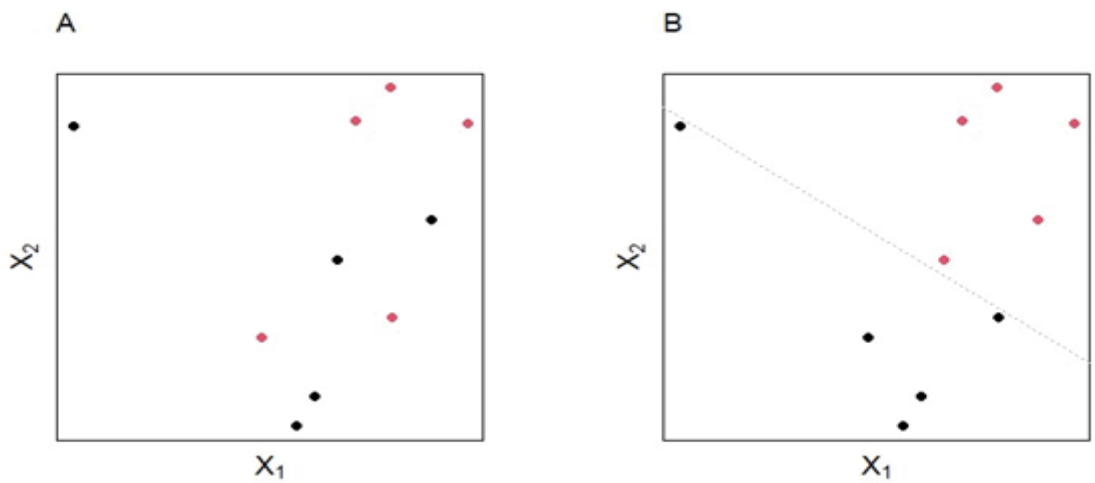
V danem primeru je β_0 logaritem obetov za Y= 1, ko X= 0, in β_1 razlika logaritma obetov, ko X= 1 v primerjavi z X= 0. Torej,

$$\beta_1 = \text{logit}(P^1) - \text{logit}(P^0) = \log\left(\frac{P^1}{1-P^1}\right) - \log\left(\frac{P^0}{1-P^0}\right) = \log\left(\frac{\left(\frac{P^1}{1-P^1}\right)}{\left(\frac{P^0}{1-P^0}\right)}\right).$$

Parametre modela logistične regresije po navadi ocenjujemo z metodo največjega verjetja. Za primer kontingenčne tabele dimenzije 2 × 2 (1) je ocena za regresijski parameter β_1 enaka

$$\hat{\beta}_1 = \log\left(\frac{f_{00} f_{11}}{f_{01} f_{10}}\right). \quad (4)$$

V večini primerov pa enačba zaprte oblike za ocenjevanje parametrov ne obstaja, ocene pa dobimo z uporabo iterativnih algoritmov; v praksi se pogosto upora-



Slika 1: Grafična ponazoritev situacije brez ločenosti (A) ter popolne ločenosti (B), ki nastane kot posledica linearne kombinacije dveh neodvisnih številskih spremenljivk X_1 in X_2. Črne in rdeče točke predstavljajo različne vrednosti odzivne spremenljivke Y, $y_i \in \{0,1\}$.
Figure 1: Graphical representation of situation without separation (A) and complete separation (B) by variables X_1 and X_2. Black and red circles represent different levels of the outcome Y, $y_i \in \{0,1\}$.

blja Newton-Raphsonova metoda (Newtonova metoda, 2023).

3 PROBLEM LOČENOSTI

Privlačnim lastnostim logistične regresije navkljub, ko je vzorec zadosti velik in naš model pravilno specificiran, pa ocene, dobljene z metodo največjega verjetja, lahko postanejo vprašljive. V primeru, ko je eden od dogodkov v kateri od skupin, ki določajo neodvisno spremenljivko, redek, kadar so vzorci majhni ali ko obstajajo močne povezave med odzivno spremenljivko Y ter neodvisnimi spremenljivkami X, se namreč lahko zgodi, da ocene parametrov po metodi največjega verjetja ne obstajajo, saj funkcija logaritma verjetja narašča v neskončnost. Tovrstno situacijo imenujemo ločenost (separation), saj sta na podlagi ene neodvisne spremenljivke ali linearne kombinacije več neodvisnih spremenljivk vrednosti odzivne spremenljivke Y ločeni. Čeprav problem pogosteje nastane, ko so neodvisne spremenljivke opisne, pa to ni izključno, temveč ločenost lahko povzroči tudi neodvisna številska spremenljivka oz. linearna kombinacija le-teh, kot je prikazano na Sliki 1B. Geometrijsko gledano ločenost torej nastane, ko v prostoru obstaja hiperravnina, ki vrednosti odzivne spremenljivke Y bodisi popolnoma loči (popolna ločenost, complete separation) (Slika 1B) ali pa jih popolnoma loči z izjemo nekaterih točk, ki ležijo točno na ravnini (kvazi-popolna ločenost, quasi-complete separation). To pomeni, da dobimo na podlagi neodvisnih spremenljivk v modelu popolne napovedi bodisi za vse enote v vzorcu bodisi za del enot v vzorcu (Albert in Anderson, 1984).

4 PRIMER: USPELOST PRIDOBIVANJA PODLAG ZA CEPLJENJE KONOPLJE

Globalni trg konoplje se širi (Rasera in sod., 2021), kar prinaša zahteve po izboljšanju agronomske prakse za povečanje učinkovitosti proizvodnje (García-Tejero in sod., 2019). Ena od novejših metod za pridobivanje sadik konoplje je cepljenje, starodavna tehnika vegetativnega razmnoževanja. Ta vključuje združitev vsaj dveh delov rastline – koreninskega sistema (podlage) in nadzemnega dela (cepiča), ki skupaj tvorita cepljeno rastlino (Salehi-Mohammadi in sod., 2009). Biteznik in sod. (2024) so za konopljo prilagodili standardno dvo-stopenjsko metodo cepljenja, uporabljeno pri vrtnarskih kulturah. Študija je preučevala vpliv različnih podlag na stopnjo preživetja, morfološke parametre in biokemijsko sestavo ženskih socvetij konoplje.

V članku se bomo osredotočili na uspelost pridobivanja podlag za cepljenje konoplje (*Cannabis sativa* L.) (Biteznik in sod., 2024). Odzivna spremenljivka Y (Uspelost podlage) zavzema dve vrednosti: podlaga se ukorenini, kar bomo označili z $Y = 1$, ali pa ne, $Y = 0$. V raziskavi so obravnavali naslednje podlage: sejance dvodomne sorte industrijske konoplje ‚Carmagnola' (CAR) in ‚Tiborszallasi' (TIB) ter sejance slovenske populacije dvodomne konoplje ‚Gorička Simba' (SIM). V poskusu je bilo 20 ponovitev za vsako podlago, skupno torej 60 podlag, pridobljenih iz sejancev. Obravnavali so tudi s potaknjenci razmnožene genotipe sort CAR, SIM in TIB, pri čemer je bil vsak genotip zastopan v 50 ponovitvah.

Za ilustracijo bomo primer sprva nekoliko poenostavili ter zanemarili razlike med sortami. Recimo, da nas zanimajo le razlike med uspelostjo pridobivanja podlag s potaknjenci razmnoženih rastlin v primerjavi s sejanci. Imamo torej eno neodvisno spremenljivko X (Sejanec), pri čemer bo $X = 1$ podlaga, pridobljena iz sejancev, ter $X = 0$ podlaga, pridobljena iz potaknjencev. Podatki so prikazani v Tabeli 1.

Tabela 1: Podatki o uspelosti pridobivanja podlag glede na potaknjence in sejance v študiji Biteznika in sod., 2014.
Table 1: Data on the rootstock rooting success using seedling or stem cutting, as reported in Biteznik et al., 2014.

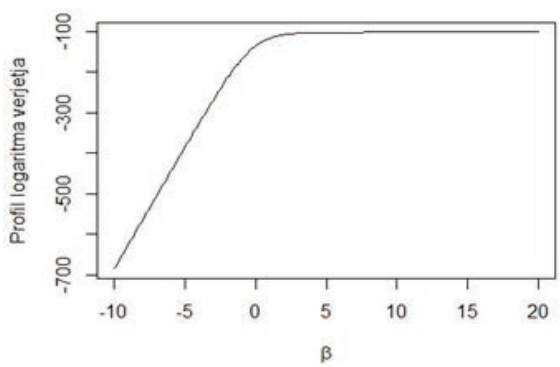
		Sejanec (X)	
		Ne ($X = 0$)	Da ($X = 1$)
Uspelost podlage (Y)	Ne ($Y = 0$)	64	0
	Da ($Y = 1$)	86	60

Podatki pričajo o močni povezanosti med Y in X, kar povzroči, da je ena od štirih celic kontingenčne tabele (1) prazna, $f_{01} = 0$: uspešne so bile vse podlage, ki so bile pridobljene iz sejancev. Ocena za verjetnost uspelosti podlage za sejance po metodi največjega verjetja je enaka P^1=60/60=1. Vseeno pa nas v analizah pogosto zanima tudi inferenca – radi bi kvantificirali in interpretirali vpliv neodvisne spremenljivke X na odzivno spremenljivko Y. Kot je razvidno iz enačbe (4), ocena regresijskega parametra $\hat{\beta}$=log((64·60)/(0·86)) ni definirana. Da torej lahko ocenimo parametre modela logistične regresije, je potreben dodaten pogoj – ocenjene verjetnosti morajo biti na intervalu (0, 1), izključujoč 0 oz. 1. Ta pogoj je smiseln, saj v praksi nikoli ne predpostavljamo, da lahko na podlagi ene ali več neodvisnih spremenljivk popolnoma napovemo nek dogodek v populaciji.

V primeru ločenosti podatkov se lahko zgodi, da iterativni algoritmi za ocene parametrov logistične regresije, implementirani v različni programski opremi, dajo različne ocene parametrov ter njihovih standardnih napak. Analiza z R-ovo (R Core Team, 2022) funkcijo glm da $\hat{\beta} = 18{,}27$ z ogromno standardno napako $s_{\beta} = 842{,}07$. Bega to, da program pri tem ne javi nobene napake in

izgleda, kot da je algoritem po Newton-Raphsonovi metodi konvergiral. Neizkušen raziskovalec bi tako lahko poročal nesmiselne ocene razmerij obetov za ukoreninjenje podlage iz sejanca glede na podlago iz potaknjenca, na primer exp (18,27) > 999,999 z zelo širokim 95 % Waldovim intervalom zaupanja exp ($\hat{\beta} \pm 1{,}96 \cdot s_{\hat{\beta}}$) = (0,001; > 999,999). IBM-ov SPSS (verzija 27) da oceno $\hat{\beta}$= 20,91 s standardno napako 5188,89. V kolikor izpis natančneje preučimo, SPSS za razliko od R-a raziskovalca opozori, da algoritem ni našel končne rešitve, saj se je ustavil preden je konvergiral, ko je bilo doseženo maksimalno število dvajsetih iteracij (`Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.`).

Čeprav oba programa vrneta oceno za parameter β, ki naj bi maksimirala logaritem verjetja, razlike v ocenah nastanejo zaradi drugačnih definicij konvergence; konvergenca je načeloma dosežena takrat, ko je sprememba ocen parametrov v naslednji iteraciji »minimalna«. V praksi spremembe ocen parametrov iz iteracije v iteracijo skoraj nikoli ne bodo natanko nič, temveč se v vsakem naslednjem koraku logaritem verjetja veča, čeprav po možnosti le malenkostno. Zato algoritem potrebuje definicijo minimalne spremembe, ki mora biti po absolutni vrednosti manjša od ε, ki pa je v različnih programih različno definiran (npr. 10^{-5}, 10^{-8}, 10^{-10}). Razlike v definicijah so po navadi irelevantne, v primeru ločenosti pa velike spremembe v ocenah parametrov privedejo do le majhne spremembe logaritma verjetja. Situacija je ponazorjena na Sliki 2, kjer logaritem verjetja sicer monotono narašča, a je skoraj raven za znaten obseg možnih vrednosti za β; tako na primer sprememba β s 5 na 15 ne izboljša prileganja modela v smislu logaritma verjetja, ravnost funkcije pa povzroči, da je standardna napaka ogromna in Waldov interval zaupanja za $\hat{\beta}$ izjemno širok, neinformativen in za prakso neuporaben, tako rekoč raztezajoč se od minus do plus neskončno (Mansournia in sod., 2018).



Slika 2: Funkcija profila logaritma verjetja za logaritem razmerja obetov β spremenljivke X (Sejanec), dobljena s fiksiranjem vrednosti za presečišče, ki je v primeru ločenosti skoraj ravna za znaten obseg možnih vrednosti za β.
Figure 2: Profile log likelihood function for the log odds ratio β of X (Seedling), obtained by fixing the intercept, which is almost flat over a vast range of possible values for β.

5 REŠITEV – FIRTHOVA PENALIZACIJA

Pri reševanju problema ločenosti je bistveno, da raziskovalec problem prepozna. Včasih statistični program raziskovalca na težavo opozori. V primeru podatkov o konoplji funkcija `glm` v `R`-u ni dala nobenega opozorila, lahko pa se pojavi sporočilo, da so ocenjene verjetnosti enake 0 ali 1 (`glm.fit: fitted probabilities numerically 0 or 1 occurred`). IBM-ov SPSS nas je po drugi strani opozoril, da je bilo doseženo maksimalno število iteracij, preden je funkcija konvergirala. Bistveno je, da ločenosti ne moremo odpraviti z večanjem števila iteracij. V splošnem velja, da ima ločenost za posledico velike ocene parametrov ter ogromne standardne napake bodisi za presečišče bodisi za neodvisne spremenljivke, ki ločenost povzročajo. Če je katera od teh spremenljivk opisna spremenljivka, kot priročno diagnostično orodje služi tabelni pregled podatkov: frekvenca nič v kateri koli celici kontingenčne tabele pomeni ločenost. Načeloma lahko problem sami prepoznamo tudi na podlagi napovedanih verjetnosti, ki bodo v primeru ločenosti vsaj za nekatere enote zelo blizu (ali enake) 0 ali 1.

Včasih lahko problem ločenosti razrešimo s smiselno revizijo podatkov. Tipični strategiji modeliranja, ki pogosto privedeta do ločenosti, sta kategorizacija številskih spremenljivk ali klasifikacija nominalnih spremenljivk v preštevilne razrede. Kadar je to smiselno, lahko kategorije za nominalne spremenljivke združimo.

V situacijah, ko ločenosti ne moremo preprosto odpraviti z revizijo podatkov, kot v primeru o konoplji, solidno rešitev problema ponujajo metode, ki modificirajo funkcijo verjetja. V kolikor logaritem verjetja, ki ga maksimiramo, rahlo modificiramo tako, da mu dodamo penalizacijski člen, lahko ocene parametrov skrčimo proti 0 in tako preprečimo, da gredo v smeri proti $\pm\infty$. Motivacija za frekventistične metode penalizacije je bodisi zmanjševanje pristranskosti bodisi zmanjševanje srednje kvadratne napake ocen. Te metode pa lahko motiviramo tudi s perspektive Bayesove statistike, ki obstoječe podatke združi z neko apriorno informacijo o porazdel-

itvi parametrov, ki po navadi ne predvideva ekstremnih vrednosti parametrov (Greenland in Mansournia, 2015).

V tem članku bomo predstavili Firthov tip penalizacije, ki je bil sprva motiviran z vidika zmanjševanja pristranskosti ocen parametrov v posplošenih linearnih modelih (Firth, 1993). Heinze in Schemper (2002) pa sta pokazala, da ponuja tudi dobro rešitev za problem ločenosti v logistični regresiji tako, da napovedane verjetnosti potegne stran od skrajnosti 0 in 1 proti 0,5. Firthov tip logistične regresije je implementiran v vseh glavnih statističnih programih: R, SAS, Stata, Statistica. V `R`-u model Firthove logistične regresije lahko ocenimo z uporabo funkcije `logistf` iz paketa `logistf` (Heinze in sod. 2022). Funkcija `STATS_FIRTHLOG` v SPSS-u prav tako bazira na `R`-ovi funkciji `logistf`. Na splošno je v primeru, ko lahko podatke strnemo v kontingenčno tabelo dimenzije 2×2 (1) ocena za β po Firthu enaka oceni po metodi največjega verjetja po tem, ko smo vsaki frekvenci v štirih celicah tabele dodali 0,5. V našem ilustrativnem primeru je ocena za $\hat{\beta}$ = log ((64,5·60,5)/(0,5·86,5))= 4,5 s standardno napako 1,43. Za namen statističnega sklepanja Heinze in Schemper (2002) predlagata uporabo intervalov zaupanja, ki temeljijo na profilu funkcije verjetja (profile likelihood confidence intervals) ter zaradi asimetričnosti omogočajo boljšo pokritost ocen v primerjavi z Waldovimi intervali zaupanja, ki so simetrični. V našem primeru je 95 % interval zaupanja za (2,53; 9,35), kar izključuje vrednost 0, torej je vpliv Sejanca na Uspelost podlage močno statistično značilen. Z modelom tako ocenjujemo, da ima podlaga, pridobljena iz sejanca, exp($\hat{\beta}$) = 90-krat večje obete, da se ukorenini, kot pa podlaga, pridobljena iz potaknjenca; pripadajoči 95 % interval zaupanja za razmerje obetov je (exp(2,53); exp(9,35)) = 12,6; 11498,8) kar nedvomno priča o tem, da bo podlaga, pridobljena iz sejancev, po vsej verjetnosti zares uspela.

V kolikor pa nas poleg ocen parametrov zanimajo tudi napovedane verjetnosti, je pri uporabi Firthove logistične regresije potrebna previdnost, saj so napovedane verjetnosti pristranske napram verjetnosti 0,5. Firthova logistična regresija namreč penalizira tudi oceno za presečišče. Za premostitev problema so Puhr in sod. (2017) predlagali metodo FLIC, ki popravi oceno za presečišče tako, da napovedane verjetnosti postanejo

Tabela 2: Rezultati analize podatkov o uspelosti pridobivanja podlag za cepljenje konoplje (Biteznik in sod., 2024) na podlagi modela logistične regresije (5) z neodvisnima spremenljivkama Sejanec in Sorta ter z njuno interakcijo, ki jih dobimo po metodi največjega verjetja ter s Firthovim tipom penalizacije.

Table 2: Analysis results on the rootstock rooting success (Biteznik et al., 2024) based on the logistic regression model (5), including seedling, variety and their interaction as covariates, which were obtained by the method of maximum likelihood and Firth's penalized likelihood, respectively.

Metoda	Spremenljivka	$\hat{\beta}$	Standardna napaka	95 % interval zaupanja	*p*-vrednost	Razmerje obetov
Metoda največjega verjetja	Presečišče	−0,08	0.28	(−0,63; 0,47)	0.777	
	Potaknjenec CAR – referenca					1,00
	Potaknjenec SIM	0,93	0,42	(0,11, 1,75)	0,03	2,53
	Potaknjenec TIB	0,24	0,40	(−0,55, −1,03)	0,55	1,27
	Sejanec CAR	18,65	1458,51	(−2840, −2877)	0,99	125286193
	Interakcija SIM	−0,93	2062,64	(−4044, −4042)	1	0,40
	Interakcija TIB	−0,24	2062,64	(−4044, −4042)	1	0,79
Firthov tip penalizacije	Presečišče	−0,08	0.28	(−0,63; 0,47)	0.779	
	Potaknjenec CAR - referenca					1,00
	Potaknjenec SIM	0,91	0,41	(0,11, 1,74)	0,026	2,48
	Potaknjenec TIB	0,24	0,40	(−0,54, 1,02)	0,523	1,27
	Sejanec CAR	3,79	1,46	(1,71, 8,66)	<0,001	44,35
	Interakcija SIM	−0,91	2,07	(−6,19, 4,38)	0,665	0,40
	Interakcija TIB	−0,24	2,06	(−5,52, 5,04)	0,909	0,79

nepristranske, medtem ko so ostale ocene parametrov enake kot pri Firthovi logistični regresiji.

Zdaj poglejmo rezultate analize podatkov o konoplji, pri čemer poleg spremenljivke Sejanec upoštevamo tudi spremenljivko Sorta s tremi kategorijami (CAR, SIM in TIB). Ker ima spremenljivka Sorta tri ravni, bo model vključeval dve umetni spremenljivki (dummy variables), ki sta dihotomni: vrednosti bodo torej enake 0 razen za sorto SIM pri X_1 oz. za sorto TIB pri X_2 bodo 1. Spremenljivka $X_3 = 1$ za sejance in $X_3 = 0$ za potaknjence. Poleg glavnih vplivov nas bo zanimalo tudi, ali obstaja interakcija med Sorto in Sejancem (torej, ali je vpliv sejanca drugačen glede na sorto). Naš model, na podlagi katerega lahko izračunamo razmerje obetov za 4 kombinacije vrednosti Sejanec in Sorta, zdaj zapišemo kot

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3, \quad (5)$$

kjer je potaknjenec CAR referenčna kategorija, β_1 in β_2 sta logaritma razmerja obetov za potaknjenca SIM oz. TIB glede na potaknjenec CAR, β_3 za sejanec CAR glede na potaknjenec CAR, β_4 in β_5 pa sta razliki logaritma razmerja obetov med sejancem SIM oz. TIB ter sejancem CAR.

Ocene parametrov modela (5), ki vključuje dva dejavnika ter njuno interakcijo, po metodi največjega verjetja ter z uporabo Firthovega tipa penalizacije so podane v Tabeli 2. Med metodama ni večjih razlik pri ocenah logaritma razmerij obetov za potaknjenca SIM in TIB. Zaradi ločenosti, saj je bila uspelost pridobivanja podlage za cepljenje konoplje za vse tri sejance 100 %, ocena logaritma razmerja obetov za sejanec CAR po metodi največjega verjetja ne obstaja. Statistični program R nam vrne neznačilne ocene razmerij obetov, ki pa se raztezajo tako rekoč od ±∞. Prav tako so tudi standardne napake za ocene interakcij ogromne. Po drugi strani Firthov tip penalizacije potrjuje, da ima spremenljivka Sejanec pomemben vpliv na uspelost pridobivanja podlage, in vrne močno značilno ($p < 0{,}001$) razmerje obetov za sejance sorte CAR v primerjavi s potaknjenci sorte CAR s precej asimetričnimi 95 % intervali zaupanja. Interakciji v modelu nista statistično značilni, kar nakazuje na to, da je vpliv Sejanca podoben ne glede na Sorto.

Pri načrtovanih poskusih modela navadno ne bi poenostavljali. V opazovalnih študijah pa bi se za voljo interpretabilnosti rezultatov raje odločili za bolj parsimoničen model, če sta ta dva enakovredna pri pojasnjevanju povezanosti med neodvisnimi spremenljivkami in odzivno spremenljivko. S testom razmerja verjetij z dvema stopinjama prostosti tako na primer lahko preverimo, da interakcija v modelu ni potrebna (p=0,17), ni torej pomembnih razlik med sejanci glede na sorto: vpliv sejanca je za vse sorte podoben. Model (5) tako lahko poenostavimo:

$$\text{logit}(P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \quad (6)$$

kjer je potaknjenec CAR referenčna kategorija, in sta logaritma razmerja obetov za potaknjenca SIM oz. TIB, pa kvantificira vpliv sejanca ob upoštevanju sorte. Rezultati modela (6), ocenjenega po Firthovi metodi, so podani v Tabeli 3.

Iz analize torej lahko sklenemo, da imajo potaknjenci sorte SIM 2,48-krat večje obete za uspelost podlage kot potaknjenci sorte CAR, pripadajoči 95 % interval zaupanja je (1,11; 5,67), medtem ko med potaknjenci sorte TIB in potaknjenci sorte CAR ni statistično značilnih razlik (p = 0,523). Primerjavo med sortama SIM in TIB lahko dobimo tako, da v modelu sorto SIM vzamemo

Tabela 3: Rezultati analize podatkov o uspelosti pridobivanja podlag za cepljenje konoplje (Biteznik in sod., 2024) na podlagi modela logistične regresije (6) z neodvisnima spremenljivkama Sejanec in Sorta, ki jih dobimo po metodi največjega verjetja ter s Firthovim tipom penalizacije.

Table 3: Analysis results on the rootstock rooting success (Biteznik et al., 2024) based on the logistic regression model (5), including Seedling and Variety as covariates, which were obtained by the method of maximum likelohood and Firth's penalized likelihood, respectively.

Metoda	Spremenljivka		Standardna napaka	95 % interval zaupanja	*p*-vrednost	Razmerje obetov
Firthov tip penalizacije	Presečišče		0.28	(,63; 0,47)	0.78	
	Potaknjenec CAR - referenca					1,00
	Potaknjenec SIM	0,91	0,41	(0,11, 1,74)	0,026	2,48
	Potaknjenec TIB	0,24	0,39	(0,54, 1,02)	0,523	1,27
	Sejanec	4,54	1,41	(2,56, 9,38)	<0,001	93,38

za referenčno kategorijo. Izkaže se, da med potaknjenci sorte TIB in potaknjenci sorte SIM ni značilnih razlik (p = 0,102). Za podlage iste sorte velja, da imajo sejanci 93-krat večje obete za uspelost podlage kot potaknjenci. Pripadajoči 95 % interval zaupanja je (13; 11870), močno asimetričen, a izraža pomemben vpliv sejancev v primerjavi s potaknjenci na uspelost pridobivanja podlag za cepljenje konoplje.

6 IMPLEMENTACIJA V R-U

V nadaljevanju dodajamo kodo za implementacijo logistične regresije po metodi največjega verjetja ter s Firthovim tipom penalizacije v R-u. Za ponovljivost najprej generiramo podatke, ki so prikazani v Tabeli 1.

```
R> podatki <- data.frame(sejanec = c(1,1,0,0), usp_podlage = c(1,0,1,0))
R> podatki <- podatki[c(rep(1,60), rep(2,0), rep(3, 86), rep(4, 64)), ]
```

Ocene za model logistične regresije po metodi največjega verjetja dobimo s funkcijo glm.

```
R> model <- glm(usp_podlage ~ sejanec, podatki, family = „binomial“)
```

Izpis povzetka modela vrne ukaz `summary(model)`:

```
Call:
glm(formula = usp_podlage ~ sejanec, family = „binomial“,
    data = podatki)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2955     0.1651   1.790   0.0735 .
sejanec      18.2706   842.0690   0.022   0.9827
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
      Null deviance: 258.23  on 209 degrees of freedom
Residual deviance: 204.71  on 208 degrees of freedom
AIC: 208.71
Number of Fisher Scoring iterations: 17
```

Napovedane verjetnosti za enote v podatkovnem okviru izračuna funkcija `predict(model, type = „response“)`.

Po tem, ko smo z ukazom `library(logistf)` naložili ustrezni paket, lahko ocene za Firthov model logistične regresije dobimo s funkcijo `logistf`.

```
R> model_firth <- logistf(usp_podlage ~ sejanec, podatki)
```

Izpis povzetka modela prikažemo s `summary (model_firth)`, v katerem so zapisani tudi 95 % intervali zaupanja za parametre modela, ki temeljijo na profilu funkcije verjetja, s pripadajočimi p-vrednostmi:

```
logistf(formula = usp_podlage ~ sejanec, data = podatki)
Model fitted by Penalized ML
Coefficients: coef  se(coef)  lower 0.95 upper 0.95     Chisq            p method
(Intercept)     0.2934792     0.1645131 -0.02713328  0.6191539  3.216735 7.288887e-02      2
sejanec           4.5023114  1.4295431 2.53155543  9.3466073 49.421415 2.064793e-12      2

Method: 1-Wald, 2-Profile penalized log-likelihood, 3-None

Likelihood ratio test = 49.42143 on 1 df, p = 2.064793e-12, n=210
Wald test = 14.58793 on 1 df, p = 0.0001337684
```

V kolikor nas poleg ocen parametrov zanimajo tudi napovedane verjetnosti, je priporočeno v funkcijo `logistf` dodati argument `flic = TRUE`.

```
R> model_flic <- logistf(usp_podlage ~ sejanec, podatki, flic = TRUE)
```

Iz izpisa povzetka modela `summary(model_flic)` lahko vidimo, da se model `model_flic` razlikuje od modela `model_firth` zgolj v oceni za presečišče. Napovedane verjetnosti dobimo z ukazom `model_flic$predict`.

Podatkom dodamo še spremenljivko sorta:

```
R> podatki$sorta <- c(rep(c(‚CAR‘, ‚SIM‘, ‚TIB‘), each = 20),
                      rep(‚CAR‘, 24), rep(‚SIM‘, 35), rep(‚TIB‘, 27),
                      rep(‚CAR‘, 26), rep(‚SIM‘, 15), rep(‚TIB‘, 23))
R> podatki$sorta <- factor (podatki$sorta)
```

Model z obema neodvisnima spremenljivkama ter njuno interakcijo naredimo z naslednjim ukazom:

```
R> model_firth_int <- logistf(usp_podlage ~ sorta*sejanec, podatki)
```

Model lahko poenostavimo tako, da iz modela izpustimo interakcijo, če ta ni potrebna:

```
R> model_firth_2 <- logistf(usp_podlage ~ sorta+sejanec, podatki)
```

Ali je interakcija v modelu potrebna ali ne, lahko

preverimo s testom razmerja verjetij z dvema stopinjama prostosti:

```
R> 1 - pchisq((-2)*(model_firth_int$loglik-model_firth_2$loglik)[1], df = 2)
```

Referenčno kategorijo iz sorte CAR v sorto SIM v modelu lahko zamenjamo z ukazom:

```
R> podatki$sorta <- relevel
(podatki$sorta, ref = ‚SIM‘)
```

7 SKLEPI

Pri analizah podatkov z modelom logistične regresije, ki omogoča preučevanje zveze med binarno odzivno spremenljivko ter eno ali več neodvisnimi spremenljivkami, se lahko zgodi, da ocene parametrov po metodi največjega verjetja ne obstajajo. Raziskovalce dobljene ocene razmerij obetov, ki jih ponujajo statistični programi in se praktično raztezajo od ±∞, lahko begajo, saj v praksi ne domnevamo, da je vpliv neke neodvisne spremenljivke na odzivno spremenljivko zares »neskončen«. Ločenost je tako prej posledica »smole pri vzorčenju« in jo lahko odpravimo z večanjem števila enot v vzorcu (Šinkovec in sod., 2019). Nenavadno se lahko zdi tudi to, da kljub očitni močni povezanosti med odzivno ter neodvisno spremenljivko zveza med njima ni statistično značilna. To je posledica ogromnih standardnih napak, ki vodijo do povsem neinformativnih Waldovih intervalov zaupanja. V članku smo ločenost predstavili na konkretnem primeru podatkov o konoplji, opisali njene posledice ter kot metodo, ki se je skozi čas že uveljavila kot dobra rešitev problema, predlagali Firthov tip penalizacije, ki prepreči, da ocene parametrov divergirajo.

Za konec naj še poudarimo, da je ločenost le skrajni primer, ko metoda logistične regresije odpove. Močno pristranske ocene parametrov (v smeri stran od 0) z velikimi standardnimi napakami lahko dobimo tudi pri analizah podatkov, ki niso nujno povsem ločeni (Greenland in sod., 2016). Podobno kot pri ločenosti težave (t. i. sparse data bias) navadno nastanejo pri analizah vzorcev, kjer so dogodki (v kateri od skupin, ki določajo neodvisno spremenljivko) redki, ali v analizah majhnih vzorcev. Podobno kot pri ločenosti lahko te težave omilimo oz. odpravimo z uporabo Firthove penalizacije. Predvsem je pomembno to, da se raziskovalci problemov zavedajo in jih lahko prepoznajo – le tako se bodo izognili poročanju nesmiselnih rezultatov svojih analiz.

8 VIRI

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons (etc.).

Albert, A. in Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*(1), 1–10. https://doi.org/10.1093/biomet/71.1.1

Biteznik, L., Štukelj, R., Flajšman, M. (2024). The efficiency of CBD production using grafted *Cannabis sativa* L. plants is highly dependent on the tzype of rootstock: A study. *Plants*, *13*(8), 1117. https://doi.org/10.3390/plants13081117

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. https://doi.org/10.2307/2336755

García-Tejero, I., Zuazo, V., Sánchez-Carnenero, C., Hernández, A., Ferreiro-Vera, C. in Casano, S. (2019). Seeking suitable agronomical practices for industrial hemp (*Cannabis sativa* L.) cultivation for biomedical applications. *Industrial Crops and Products, 139*. https://doi.org/10.1016/J.INDCROP.2019.111524

Greenland, S., in Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, *34*, 3133–3143. 10.1002/sim.6537

Greenland, S., Mansournia, M.A. in Altman, D.G. (2016). Sparse data bias: a problem hiding in plain sight. *BMJ*, *352*, i1981. https://doi.org/10.1136/bmj.i1981

Harrell, F. E., Jr. (2016). *Regression modeling strategies*. Springer International Publishing.

Heinze, G. in Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*, 2409–2419. https://doi.org/10.1002/sim.1047

Heinze, G., Ploner, M. in Jiricka L. (2022). *logistf: Firth's bias reduced logistic regression. R package version 1.24.1*. https://CRAN.R-project.org/package=logistf

IBM Corp. (2020). *IBM SPSS Statistics for Windows (Version 27.0)*. IBM Corp.

Mansournia, M.A., Geroldinger, A., Greenland, S. in Heinze, G. (2018). Separation in logistic regression: Causes, consequences, and control. *American Journal of Epidemiology*, *187*(4), 864–870. https://doi.org/10.1093/aje/kwx299

Newtonova metoda. (10. april 2024). Wikipedija. Pridobljeno s https://sl.wikipedia.org/w/index.php?title=Newtonova_metoda&oldid=5963185.

Puhr, R., Heinze, G., Nold, M., Lusa, L. in Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions? *Statistics in Medicine*, *36*, 2302–2317. 10.1002/sim.7273

R Core Team. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. Pridobljeno s https://www.R-project.org/

Rasera, G., Ohara, A. in Castro, R. (2021). Innovative and emerging applications of cannabis in food and beverage products: From an illicit drug to a potential ingredient for health promotion. *Trends in Food Science and Technology*, *115*, 31–41. https://doi.org/10.1016/J.TIFS.2021.06.035

Salehi-Mohammadi, R., Khasi, A., Lee, S.G., Huh, Y.C., Lee, J.M. in Delshad, M. (2009). Assessing survival and growth performance of 713 Iranian melon to grafting onto cucurbita rootstocks. *Korean Journal of Horticultural Science & Technology*, *27*(1), 1–6.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Šinkovec, H., Geroldinger, A. in Heinze, G. (2019). Bring more data!—A good advice? Removing separation in logistic regression by increasing sample size. *International Journal of Environmental Research and Public Health*, *16*(23), 4658. https://doi.org/10.3390/ijerph16234658

van Smeden, M., de Groot, J.A., Moons, K.G., Collins G.S., Altman D.G., Eijkemans M.J. in Reitsma J.B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, *16*, 163. https://doi.org/10.1186/s12874-016-0267-3