Primjena logističke regresije u modeliranju vjernosti posjetitelja

Zovko, Mila

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

MILA ZOVKO

PRIMJENA LOGISTIČKE REGRESIJE U MODELIRANJU VJERNOSTI POSJETITELJA

DIPLOMSKI RAD

Split, rujan 2022.

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU

ODJEL ZA MATEMATIKU

PRIMJENA LOGISTIČKE REGRESIJE U MODELIRANJU VJERNOSTI POSJETITELJA

DIPLOMSKI RAD

Neposredna voditeljica:
dr. sc. Ana Perišić

Mentorica:
doc. dr. sc. Tea Martinić Bilać

Studentica:
Mila Zovko

Split, rujan 2022.

Uvod

Pojam regresije uveo je Sir Francis Galton uočivši da visina djece regradira prema prosjeku, odnosno da visoki roditelji imaju visoku djecu, no u prosjeku nižu od roditelja, dok niski roditelji imaju nisku djecu, u prosjeku nešto višu od roditelja. Ovaj efekt nazvao je regresija prema prosjeku. Regresijska analiza koristi se za donošenje zaključaka o slučajnoj varijabli koja ovisi o nezavisnoj varijabli ili više nezavisnih varijabli koje nazivamo prediktorima. Logistička regresija često je korištena u slučajevima kada je zavisna varijabla dihotomna. Kako je čest slučaj da je ishod zavisne varijable dihotoman, u posljednjem desetljeću logistička regresija postala je neizostavna metoda u biomedicini, biomatematici, kemiji, ekonomiji i općenito u statistici.

U ovom radu upravo ćemo proučavati logističku regresiju te njenu konkretnu primjenu. U prvom poglavlju upoznat ćemo se sa pojmom generaliziranih linearnih modela, zatim ćemo objasniti matematičku pozadinu modela jednostavne i složene logističke regresije, te ukratko objasniti ocjenu prilagodbe modela. U drugom poglavlju primjenit ćemo logističku regresiju u svrhu modeliranja vjernosti posjetitelja zaštićenih područja.

Sadržaj

Poglavlje 1

Logistička regresija

1.1 Osnovni pojmovi

Definicija 1.1 *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i neka je $\mathcal{P}$ familija vjerojatnosnih mjera na $(\Omega, \mathcal{F})$. Trojka $(\Omega, \mathcal{F}, \mathcal{P})$ zove se statistička struktura.*

Familija vjerojatnosti često je parametrizirana $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, gdje je θ parametarski prostor, odnosno skup vrijednosti parametara.

Definicija 1.2 *Neka je $X : \Omega \to \mathbb{R}^d$ slučajan vektor i $(\Omega, \mathcal{F}, \mathcal{P})$ statistička struktura, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Za $\theta \in \Theta$ označimo $F(x;\theta) = \mathbb{P}_\theta(X \leq x)$,$x \in \mathbb{R}^d$. Tada je $F(\cdot;\theta)$ funkcija distribucije od X uz vjerojatnost $\mathbb{P}_\theta \in \mathcal{P}$. Kažemo da X pripada statističkom modelu $\mathcal{P}' = \{F(\cdot;\theta) : \theta \in \Theta\}$.*

Definicija 1.3 *Slučajan uzorak duljine n ili n-dimenzionalni slučajni uzorak na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je konačan niz $X_1, X_2, \ldots, X_n$ slučajnih varijabli (vektora) na $(\Omega, \mathcal{F})$ tako da $\forall P \in \mathcal{P}$ su slučajne varijable $X_1, X_2, \ldots, X_n$ nezavisne i jednako distribuirane.*

Definicija 1.4 *Statistika na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je svaka slučajna varijabla (vektor) $T : \Omega \to \mathbb{R}^d$ takva da postoji $n \in \mathbf{N}$ i n-dimnezionalni*

slučajni uzorak $(X_1, X_2, \ldots, X_n)$ *na* $(\Omega, \mathcal{F}, \mathcal{P})$ *te izmjerivo preslikavanje* $t : \mathbb{R}^n \to \mathbb{R}^d$ *takvo da je* $T = t(X_1, X_2, \ldots, X_n)$.

1.2 Generalizirani linearni model

1.2.1 Uvod

Predmet mnogih istraživanja je utvrđivanje povezanosti između različitih varijabli. Nadalje, često je cilj ispitati ovisi li neka varijabla (zavisna varijabla) o drugim varijablama (nezavisne varijable ili kovarijati). Cilj je utvrditi i oblik veze. Analize se provode na temelju prikupljenih podataka.

Podaci će biti predstavljeni matricom podataka u kojem su redci indeksirani eksperimentalnim ili istraživačkim jedinicama. U ovom kontekstu, jedinice su objekti na kojima se provode opažanja, na primjer pacijenti u medicinskom istraživanju ili kliničkom ispitivanju. Stupci matrice podataka su varijable kao što su dob pacijenta, težina, spol i tako dalje. Neke od varijabli smatraju se varijablama odziva ili zavisnim varijablama, za čije se vrijednosti vjeruje da su pod utjecajem drugih eksplanatornih varijabli, zvane kovarijatima, odnosno nezavisnim varijablama. Kovarijati mogu biti kvantitativne ili kvalitativne varijable. Kvantitativne varijable poprimaju numeričke vrijednosti, a kvalitativne varijable poprimaju vrijednosti iz konačnog skupa nenumeričkih vrijednosti. Kvalitativne kovarijate nazivat ćemo faktorima. Zavisne varijable mogu biti numeričke, kontinuirane ili diskretne, ili mogu biti kvalitativne.

U matričnom zapisu skup opažanja zavisne varijable označen je vektorom stupca opažanja $Y = (y_1, y_2, \ldots, y_n)^T$. Skup kovarijata ili eksplanatornih varijabli predstavljen je kao $n \times p$ matrica X. Svaki redak matrice X odnosi se na jednu jedinicu opažanja, a svaki stupac na jednu kovarijatu.

1.2. Generalizirani linearni model

Najjednostavnija veza između varijable odziva i kovarijata je linearna veza. Osnovna pretpostavka linearnih modela je pretpostavka o postojanju linearne veze između očekivanja varijable odziva i kovarijata. Svakoj kovarijati pridružen je koeficijent ili parametar, obično nepoznat. Skup parametara je vektor dimenzije p, koji se obično označava s $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$.

Model logističke regresije pripada familiji generaliziranih linearnih modela koje dobijemo kao proširenja klasičnih linearnih modela. Opišimo prvo klasični linearni model. Neka je vektor opažanja $Y = (y_1, y_2, \ldots, y_n)^T$ realizacija slučajne varijable Y čije su komponente nezavisno distribuirane sa očekivanjem μ. Sustavni dio modela se sastoji od predstavljenog vektora μ izraženog preko (malog) broja nepoznatih parametara $\beta_1, \ldots, \beta_p$. U slučaju klasičnih linearnih modela, ova specifikacija ima oblik

$$\mu = \sum_{j=1}^{p} x_j \beta_j$$

gdje su β_j nepoznati parametri čije se vrijednosti procijenjuju iz danih podataka, a x_j vrijednosti j-te kovarijate. Neka je Y_i i-to opažanje, tada se sustavni dio modela može zapisati kao

$$\mathbb{E}(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij} \beta_j; \quad i = 1, \ldots, n,$$

gdje je x_{ij} vrijednost j-te kovarijate za opažanje i. Prethodni sustav možemo zapisati u matričnom obliku kao

$$\mu = X\boldsymbol{\beta},$$

gdje je X matrica modela dimenzija $n \times p$, a $\boldsymbol{\beta}$ vektor parametara dimenzija $p \times 1$.

Slučajni dio modela se sastoji od pretpostavke da su greške modela, definirane kao $\epsilon = Y - X\boldsymbol{\beta}$, nezavisne i varijanca konstantna. Ovo su jake pretpostavke i potrebno ih je provjeriti, koliko je to moguće, iz samih podataka. Nadalje, struktura sustavnog dijela pretpostavlja da znamo kovarijate koje utječu na srednju vrijednost i možemo ih učinkovito mjeriti bez pogreške, ovu pretpostavku također treba provjeriti, koliko je to moguće.

Daljnju specifikaciju modela dobijemo iz pretpostavke da greške zadovoljavaju Gaussovu ili normalnu distribuciju uz konstantnu varijancu σ^2.

Konačno, klasični linearni model možemo opisati kao model u kojem su komponente slučajne varijable Y nezavisne, normalno distribuirane sa konstantnom varijancom σ^2 te vrijedi

$$\mathbb{E}(Y) = \mu, \quad \text{gdje je} \quad \mu = X\beta, \tag{1.1}$$

odnosno

$$Y_i \sim N(\sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2).$$

1.2.2 Generalizacija

Kako bismo pojednostavili prijelaz sa klasičnih na generalizirane linearne modele, uvodimo podjelu komponenti na sljedeći način:

1. *Slučajna komponenta:* komponente od Y su nezavisne, normalno distribuirane sa $\mathbb{E}(Y) = \mu$ i konstantnom varijancom σ^2;

2. *Sustavna komponenta:* η je izražen preko kovarijata $x_1, x_2, \ldots, x_p$ s

$$\eta = \sum_{j=1}^{p} x_j \beta_j;$$

3. Treća komponenta daje vezu između slučajne i sustavne komponente:

$$\eta = \mu.$$

Ako označimo sa

$$\eta_i = g(\mu),$$

onda ćemo $g(\cdot)$ nazivati funkcijom poveznicom (link funkcija). Dakle, kod klasičnog linearnog modela pretpostavljamo da slučajna varijabla ima normalnu distribuciju dok za funkciju poveznicu biramo identitetu. Generalizirani linearni modeli dopuštaju da slučajna varijabla Y prati bilo koju distribuciju iz eksponencijalne familije, te dopuštaju bilo koju monotonu diferencijabilnu funkciju kao izbor za link funkciju.

1.2.3 Funkcije vjerodostojnosti za generalizirane linerne modele

Osnovna pretpostavka generaliziranih linearnih modela odnosi se na distribuciju komponenata slučajnih varijabli Y; pretpostavljamo da svaka komponenta slučajne varijable Y ima distribuciju iz eksponencijalne familije.

Definicija 1.5 *Distribucija slučajne varijable Y pripada nekoj eksponencijalnoj familiji ako joj gustoća ima oblik*

$$f_Y(y;\theta,\phi) = exp\{(y\theta - b(\theta))/a(\phi) + c(y,\phi)\} \tag{1.2}$$

za neke funkcije $a(\cdot), b(\cdot)$ i $c(\cdot)$, pri čemu je funckija b dva puta neprekidno diferencijabilna tako da je b' invertibilna.

Familija ima dva parametra: θ (prirodni parametar) i ϕ (parametar disperzije).Funkcija a parametra ϕ zove se funkcija disperzije te omogućuje dodatnu fleksibilnost u modelu, tako da ne moraju svi odzivi imati istu varijancu. Funkciju c tipično ignoriramo jer nema utjecaja u procesu procjene parametara generaliziranih linearnih modela.

Primjer 1.6 *Normalna distribucija pripada eksponencijalnoj familiji. Naime, njezina gustoća ima oblik*

$$\begin{aligned} f_Y(y;\theta,\phi) &= \frac{1}{\sqrt{(2\pi\sigma^2)}} exp\{-(y-\mu)^2/2\sigma^2\} \\ &= exp\{(y\mu - \mu^2/2)/\sigma^2 - \frac{1}{2}(y^2/\sigma^2 + log(2\pi\sigma^2))\}, \end{aligned}$$

iz čega slijedi da su $\theta = \mu$, $\phi = \sigma^2$, i

$$a(\phi) = \phi, \qquad b(\theta) = \theta^2/2, \qquad c(y,\phi) = -\frac{1}{2}\{y^2/\sigma^2 + log(2\pi\sigma^2)\}.$$

Promotrimo sada funkciju vjerodostojnosti $l(\theta,\phi;y) = \log f_Y(y;\theta,\phi)$ za dane θ, ϕ i y unutar neke ekponencijalne familije.

Iz (1.2) imamo

$$l(\theta;y) = \{y\theta - b(\theta)\}/a(\phi) + c(y,\phi),$$

Srednju vrijednost i varijancu od Y možemo dobiti iz relacija

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{1.3}$$

i

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right)^2 = 0. \tag{1.4}$$

Vrijedi

$$\frac{\partial l}{\partial \theta} = \{y - b'(\theta)\}/a(\phi) \tag{1.5}$$

i

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi). \tag{1.6}$$

Iz (1.3) i (1.5) imamo da je

$$0 = \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = \{\mu - b'(\theta)\}/a(\phi),$$

stoga,

$$\mathrm{E}(Y) = \mu = b'(\theta).$$

Konačno, iz (1.4), (1.5) i (1.6) imamo da je

$$0 = -\frac{b''(\theta)}{a(\phi)} + \frac{\mathrm{var}(Y)}{a^2(\phi)},$$

pa je

$$\mathrm{var}(Y) = b''(\theta)a(\phi).$$

Dakle, varijanca od Y je produkt dvije funkcije. Funkcija koja ovisi samo o parametru θ i naziva se funkcija varijance te ćemo je označiti s $b''(\theta)$, dok druga ne ovisi o θ, već ovisi samo o ϕ.

1.3 Model jednostavne logističke regresije

Kada zavisnu varijablu odziva želimo opisati pomoću jedne nezavisne varijable, koristimo model jednostavne logističke regresije. Cilj kod proučavanja ovog modela je pronaći najprikladniji i najjednostavniji razuman model koji opisuje odnos između zavisne varijable ishoda (odziva) i skupa nezavisnih (prediktorskih) varijabli. Najčešći primjer modeliranja je uobičajeni linearni regresijski model gdje se pretpostavlja da je varijabla ishoda kontinuirana. Za razliku od toga, u logističkom regresijskom modelu varijabla ishoda je binarna ili dihotomna, to jest ima dva moguća ishoda. Ova razlika između logističke i linearne regresije odražava se i u izboru parametarskog modela i u izboru pretpostavki. Nadalje, metode koje koristimo pri analizi pomoću logističke regresije slijede ista opća pravila koja koristimo u linearnoj regresiji. Stoga, tehnike koje koristimo u linearnoj regresijskoj analizi motiviraju naš pristup logističkoj regresiji. U daljnjem tekstu uvodimo dvije važne razlike između logističke i linearne regresije.

Prva razlika tiče se uvjetne distribucije varijable ishoda. U svakom regresijskom problemu ključna veličina je srednja vrijednost varijable ishoda, s obzirom na vrijednost nezavisne varijable. Ta se veličina naziva uvjetna srednja vrijednost i bit će označena s $E(Y|x)$, gdje Y označava varijablu ishoda, a x vrijednost nezavisne varijable. U modelu linearne regresije pretpostavljamo da se opažanje varijable ishoda može izraziti sa $y = E(Y|x) + \epsilon$. Veličina ϵ naziva se greška i izražava devijaciju opaženih vrijednosti od uvjetne srednje vrijednosti. Najčešće pretpostavljamo da ϵ slijedi normalnu distribuciju sa srednjom vrijednošću nula te s varijancom koja je konstanta za sve vrijednosti nezavisne varijable. Iz prethodnog slijedi da će uvjetna distribucija varijable ishoda, s obzirom na dani x, biti normalna sa srednjom vrijednošću $E(Y|x)$ i konstantnom varijancom. S druge strane,

u slučaju dihotomne varijable ishoda, možemo izraziti vrijednost varijable ishoda sa $y = \pi(x) + \epsilon$. U ovom slučaju pogreška ϵ može poprimiti jednu od dvije moguće vrijednosti. Ako je $y = 1$, onda je $\epsilon = 1 - \pi(x)$ s vjerojatnošću $\pi(x)$, a ako je $y = 0$, onda je $\epsilon = -\pi(x)$ s vjerojatnošću $1 - \pi(x)$. Dakle, ϵ ima distribuciju sa srednjom vrijednošću nula i varijancom $\pi(x)[1 - \pi(x)]$. Slijedi da je uvjetna distribucija varijable ishoda binomna distribucija s vjerojatnošću danom sa uvjetnom srednjom vrijednošću $\pi(x)$.

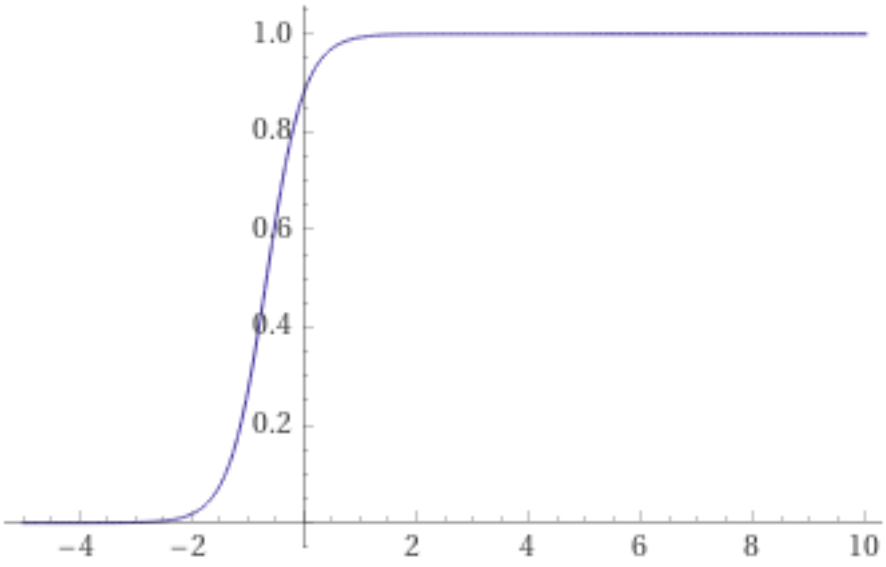
Druga važna razlika između modela linearne i logističke regresije se odnosi na prirodu odnosa između ishoda i nezavisnih varijabli. U linearnoj regresiji pretpostavljamo da se uvjetna srednja vrijednost varijable ishoda može izraziti kao linearna funckija u x, kao što je

$$E(Y|x) = \beta_0 + \beta_1 x.$$

Ovaj izraz implicira da je moguće da $E(Y|x)$ poprimi bilo koju vrijednost za $x \in < -\infty, +\infty >$. S druge strane, model koji koristimo logističkoj regresiji je oblika

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \ , \tag{1.7}$$

gdje smo uveli oznaku $\pi(x) = E(Y|x)$. Iz prethodnog izraza slijedi da je $0 \leq \pi(x) \leq 1$. To se može vidjeti na Slika 1.1. gdje je prikazan graf logističke funkcije.



Slika 1.1: Logistička funkcija

Promjena $\pi(x)$ po jedinici promjene x postaje progresivno manja kako se $\pi(x)$ približava nuli ili jedinici. Za takvu krivulju kažemo da ima S-oblik. Transformacija funkcije $\pi(x)$, koja je od velike važnosti u proučavanju logističke regresije, jest log transformacija te je definirana sa

$$\begin{aligned} g(x) &= \ln \frac{\pi(x)}{1-\pi(x)} \\ &= \beta_0 + \beta_1 x \ , \end{aligned}$$

izražena preko $\pi(x)$ definiranog kao u (1.7). Dakle, veza između slučajne i sistemske komponente, odnosno funckija poveznica, je logit $g(x)$. Važnost ove transformacije je u tome što $g(x)$ sadrži mnoga poželjna svojstva modela linearne regresije. Funkcija logit, $g(x)$, je linearna po svojim parametrima, može biti neprekidna te ovisno o vrijednostima varijable x može varirati od $-\infty$ do $+\infty$.

Konačno, možemo zaključiti da u regresijskoj analizi kada je varijabla ishoda dihotomna vrijedi sljedeće:

1. Uvjetna srednja vrijednost mora se formulirati tako da bude u granicama između nule i jedinice. Vidjeli smo da model logističke regresije $\pi(x)$ dan izrazom (1.7) zadovoljava ovo ograničenje.

2. Distribucija koju koristimo pri statističkoj analizi te opisujemo distribuciju pogrešaka je binomna distribucija.

3. Opća pravila na kojima temeljimo analizu pomoću linearne regresije, također koristimo pri analizi logističkom regresijom.

1.3.1 Procjena parametara modela

Pretpostavimo da imamo uzorak od n nezavisnih opažanja, (x_i, y_i), gdje y_i označava vrijednost dihotomne varijable ishoda, a x_i vrijednost nezavisne varijable za i-ti subjekt. Nadalje, pretpostavimo da je varijabla ishoda kodirana kao 0 ili 1, što predstavlja odsutnost, odnosno prisutnost, karakteristike. Da bismo logistički regresijski model u izrazu (1.7) prilagodili skupu podataka, potrebno je procijeniti vrijednosti nepoznatih parametara β_0 i β_1.

U linearnoj regresiji metoda koja se najčešće koristi za procjenu nepoznatih parametara je metoda najmanjih kvadrata. U metodi najmanjih kvadrata odabiremo one vrijednosti β_0 i β_1 koje minimiziraju zbroj kvadratnih odstupanja opaženih vrijednosti varijable Y od predviđenih vrijednosti dobivenih na temelju modela. Koristeći uobičajene pretpostavke za linearnu regresiju, metoda najmanjih kvadrata daje procjenitelje s nizom poželjnih statističkih svojstava. No, kada metodu najmanjih kvadrata primijenimo na model s dihotomnim ishodom, procjenitelji više nemaju ista svojstva.

1.3. Model jednostavne logističke regresije

Opća metoda procjene u modelu linearne regresije (kada su pogreške normalno distibuirane) koja dovodi do funkcije najmanjih kvadrata naziva se metoda maksimalne vjerodostojnosti. Na toj metodi temeljimo naš pristup u modelu logističke regresije. Općenito, metoda maksimalne vjerodostojnosti daje vrijednosti nepoznatih parametra koji maksimiziraju vjerojatnost dobivanja promatranog skupa podataka. Kako bismo primijenili ovu metodu na naš problem, prvo moramo konstruirati funkciju, koja se zove funkcija vjerodostojnosti. Ova funkcija izražava vjerojatnost opaženih podataka kao funkciju nepoznatih parametara. Procjenitelji maksimalne vjerodostojnosti parametara odabrani su tako da maksimiziraju ovu funkciju. Stoge se dobiveni procjenitelji najviše slažu s promatranim podacima. U daljenjem tekstu ćemo opisati kako pronaći te vrijednosti iz modela logističke regresije.

Izraz za $\pi(x)$ dan u jednadžbi (1.7) daje uvjetnu vjerojatnost da je Y jednako 1 za dani x. Ovo će biti označeno kao $P(Y = 1|x)$. Slijedi da veličina $(1 - \pi(x))$ daje uvjetnu vjerojatnost da je Y jednako nuli za dani x, $P(Y = 0|x)$. Dakle, za one parove (x_i, y_i), gdje je $y_i = 1$, doprinos funkciji vjerodostojnosti je $\pi(x_i)$, a za one parove gdje je $y_i = 0$, doprinos funkciji vjerodostojnosti je $(1 - \pi(x_i))$. Najprikladniji način za izraziti doprinos funkciji vjerodostojnosti za dani par (x_i, y_i) je dan sljedećim izrazom:

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{1.8}$$

Zbog pretpostavke da su opažanja nezavisna, funkcija vjerodostojnosti dobiva se kao produkt od n članova danih sa (1.8) na sljedeći način:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{1.9}$$

Načelo maksimalne vjerodostojnosti nam sugerira da za procjenu od $\boldsymbol{\beta}$ koristimo vrijednost koja maksimizira izraz u jednadžbi (1.9). Međutim, radi matematičke jednostavnosti proučavat ćemo logaritam od (1.9) kojeg nazivamo log vjerodostojnost te je dan

$$\boldsymbol{L}(\boldsymbol{\beta}) = \ln [l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln [\pi(x_i)] + (1 - y_i) \ln [1 - \pi(x_i)]\}. \tag{1.10}$$

Iz nužnih uvjeta ekstrema za $\boldsymbol{L}(\boldsymbol{\beta})$ slijede jednadžbe vjerodostojnosti:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{1.11}$$

i

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0 \tag{1.12}$$

Kod linearne regresije jednadžbe vjerodostojnosti su linearne u nepoznatim parametrima i stoga se lako rješavaju. S druge strane, za logističku regresiju izrazi u (1.11) i (1.12) su nelinearni u β_0 i β_1, te stoga zahtijevaju posebne metode za njihovo rješavanje. Takve metode su iterativne prirode i programirane su u dostupnom softveru za logističku regresiju.

Vrijednost procijenjenog parametra $\boldsymbol{\beta}$ dobivena rješavanjem jednadžbi (1.11) i (1.12) naziva se procjena maksimalne vjerodostojnosti i označava se kao $\widehat{\boldsymbol{\beta}}$. Nadalje, $\widehat{\pi}(x_i)$ je procjena maksimalne vjerodostojnosti za $\pi(x_i)$. Ova veličina daje procjenu uvjetne vjerojatnosti $E[Y = 1|x]$, te predstavlja

prilagođenu ili predviđenu vrijednost za model logističke regresije. Posljedica jednadžbe (1.11) je

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\pi}(x_i),$$

to jest, zbroj opaženih vrijednosti y jednak je zbroju predviđenih (očekivanih) vrijednosti.

1.3.2 Testiranje značajnosti parametara modela

Nakon procjene koeficijenata modela, od interesa je ispitati značajnost varijabli u modelu. Naš pristup testiranju značajnosti koeficijenta varijable je vezan uz pitanje da li nam model koji uključuje predmetnu varijablu govori više o varijabli ishoda od modela koji ne uključuje tu varijablu.

Primjena opće metode za procjenu značajnosti varijabli na linearnom regresijskom modelu nas motivira za pristup koji koristimo u logističkoj regresiji.

Procjeni značajnosti regresijskog koeficijenta u linearnoj regresiji pristupa se formiranjem i analizom tablice varijance. Ova tablica razdvaja ukupni zbroj kvadratnih odstupanja opaženih vrijednosti od njihove srednje vrijednosti na dva dijela: (1) zbroj kvadratnih odstupanja opažanja od regresijskog pravca SSE (ili rezidualni zbroj kvadrata), i (2) zbroj kvadrata predviđenih vrijednosti, temeljenih na regresijskom modelu, od srednje vrijednosti zavisne varijable SSR, (ili zbog regresijskog zbroja kvadrata). Ako y_i označava opaženu vrijednost, a $\widehat{y_i}$ označava predviđenu vrijednost za i-to opažanje u modelu, onda

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2.$$

1.3. Model jednostavne logističke regresije

U modelu koji ne sadrži nezavisnu varijablu, jedini parametar je β_0 i $\beta_0 = \overline{y}$ ($\overline{y}$ je srednja vrijednost varijable odgovora). U ovom slučaju, $\widehat{y_i} = \overline{y}$ i SSE je jednak varijanci varijable odgovora. Kada u model uključimo nezavisnu varijablu, svako smanjenje SSE bit će posljedica činjenice da je koeficijent nagiba β_1 različit od nula. Promjena vrijednosti SSE je nastala zbog regresijskog izvora varijabilnosti, označenog kao SSR,

$$\text{SSR} = \sum_{i=1}^{n}(y_i - \overline{y_i})^2 - \sum_{i=1}^{n}(y_i - \widehat{y_i})^2.$$

U linearnoj regresiji naš interes je usmjeren na veličinu SSR-a. Velika vrijednost SSR-a sugerira da je nezavisna varijabla važna, dok mala vrijednost sugerira da nezavisna varijabla nije od pomoći u predviđanju varijable odziva.

Isti princip koristimo i kod logističke regresije, to jest, uspoređujemo promatrane vrijednosti varijable odgovora kako bismo predvidjeli vrijednosti dobivene iz modela sa i bez promatrane varijable. U logističkoj regresiji, usporedba opaženih s predviđenim vrijednostima temelji se na logaritamskoj funkciji vjerodostojnosti definiranoj u jednadžbi (1.10). Za bolje razumijevanje ove usporedbe, korisno je razmišljati o opaženoj vrijednosti varijable odgovora kao o predviđenoj vrijednosti koja proizlazi iz saturiranog modela. Saturirani model je model koji sadrži točno onoliko parametara koliko ima opaženih vrijednosti.

Usporedba opaženih s predviđenim vrijednostima pomoću funkcije vjerodostojnosti dana je sa sljedećim izrazom:

$$D = -2\ln\left[\frac{\text{vjerodostojnost procijenjenog modela}}{\text{vjerodostojnost saturiranog modela}}\right]. \tag{1.13}$$

Veličina unutar zagrada u gornjem izrazu naziva se omjer vjerodostojnosti. Potrebno je umetnuti minus dva puta unutar logaritma za dobivanje veličine čija je distribucija poznata i može se koristiti u svrhu testiranja hipoteza. Takav test naziva se test omjera vjerodostojnosti. Korištenjem jednadžbe (1.10), jednadžba (1.13) postaje

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi}_i}{1-y_i}\right)\right], \tag{1.14}$$

gdje je $\widehat{\pi}_i = \widehat{\pi}(x_i)$.

Statistika D u jednadžbi (1.14) se naziva devijancom ili statistikom odstupanja i igra središnju ulogu u nekim pristupima ocjenjivanja usklađenosti modela. Devijanca za logističku regresiju ima istu ulogu kao i rezidualni zbroj kvadrata u linearnoj regresiji. Zapravo, devijanca dana sa (1.14), kada se izračuna za linearnu regresiju, identična je SSE-u.

Nadalje, u modelu gdje su vrijednosti varijable ishoda 0 ili 1, vjerodostojnost saturiranog modela je 1. Posebno, iz definicije saturiranog modela slijedi da je $\widehat{\pi}_i = y_i$, a vjerodostojnost je dana sa

$$l(\text{saturirani model}) = \prod_{i=1}^{n} y_i^{y_i} \times (1-y_i)^{1-y_i} = 1.$$

Nadalje, iz jednadžbe (1.13) slijedi da je devijanca

$$D = -2\ln(\text{vjerodostojnost procijenjenog modela}). \tag{1.15}$$

Za potrebe ocjene značajnosti nezavisne varijable promatramo promjenu u D kao posljedicu uključivanja nezavisne varijable u model. Spome-

nuta promjena dobiva se kao:

$$G = D(\text{model bez varijable}) - D(\text{model s varijablom}).$$

Ova statistika ima istu ulogu u logističkoj regresiji kao brojnik parcijalnog F testa u linearnoj regresiji. Budući da je vjerodostojnost saturiranog modela zajednička za obje vrijednosti D koje se razlikuju za izračun G, može se izraziti kao

$$G = -2\ln\left[\frac{\text{vjerodostojnost modela bez varijable}}{\text{vjerodostojnost modela s varijablom}}\right]. \tag{1.16}$$

Posebno, u slučaju jedne nezavisne varijable može se pokazati da kada varijabla nije u modelu, maksimalna procjena vjerodostojnosti β_0 je $\ln(n_1/n_0)$ gdje je $n_1 = \sum_{i=1}^{n} y_i$ i $n_0 = \sum_{i=1}^{n}(1-y_i)$, a predviđene vrijednosti su konstantne, n_1/n. Ovdje vrijednost od G je:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\widehat{\pi}_i^{\,y_i}(1-\widehat{\pi}_i)^{(1-y_i)}}\right] \tag{1.17}$$

to jest

$$G = 2\left\{\sum_{i=1}^{n}\Big[y_i\ln(\widehat{\pi}_i)+(1-y_i)\ln(1-\widehat{\pi}_i)\Big]-\Big[n_1\ln(n_1)+n_0\ln(n_0)-n\ln(n)\Big]\right\}. \tag{1.18}$$

Ako pretpostavimo istinitost nul hipoteze (to jest, $H_0 : \beta_1 = 0$), onda statistika G slijedi χ^2 distribuciju s jednim stupnjem slobode. Za velike vrijednosti statistike G odbacit ćemo nul hipotezu , a to nas upućuje na

zaključak da je varijabla značajna. Također, potrebne su i dodatne matematičke pretpostavke, ali, za gornji slučaj one su nerestriktivne i zahtijevaju veliku količinu uzorka, n.

Izračun logističke vjerodostojnosti i test omjera vjerodostojnosti standardni su dio svih računalnih alata za logističku regresiju. To nam olakšava provjeru važnosti dodavanja novih varijabli u model. U jednostavnom slučaju jedne nezavisne varijable, prvo promatramo model koji sadrži samo konstantni član. Zatim promatramo model koji sadrži nezavisnu varijablu zajedno s konstantom. To dovodi do nove log vjerodostojnosti. Test omjera vjerodostojnosti dobiva se množenjem razlike između ove dvije vrijednosti s -2.

Predložena su još dva slična, statistički ekvivalentna testa. To su Wald test i Score test. Pretpostavke koje su nam potrebne za ove testove iste su kao one za test omjera vjerodostojnosti u (1.17). Opširnije značajke o ovim testovima i njihovim pretpostavkama mogu se naći u Rao (1973).

Waldov test se temelji na usporedbi procjene maksimalne vjerodostojnosti parametra $\widehat{\beta_1}$ i procjene njegove standardne pogreške. Rezultirajući omjer, uz pretpostavku istinitosti nul hipoteze ($\beta_1 = 0$), pratit će standardnu normalnu distribuciju. Hauck i Donner (1977.) ispitali su izvedbu Waldovog testa i otkrili da često ne uspijeva odbacili nul hipotezu kada je koeficijent značajan te su preporučili korištenje testa omjera vjerodostojnosti.

Test značajnosti varijable koji ne zahtijeva izračunavanje procjene maksimalne vjerodostojnosti za β_1 je Score test. Zagovornici Score testa navode smanjeni računalni napor kao njegovu glavnu prednost. Korištenje testa ograničeno je time što se ne može dobiti iz nekih programskih paketa. U univarijantnom slučaju, ovaj se test temelji na uvjetnoj distribuciji derivacije u jednadžbi (1.12), s obzirom na derivaciju u jednadžbi (1.11). U tom slučaju

možemo zapisati izraz za Score test. Test koristi vrijednost jednadžbe (1.12), izračunatu pomoću $\beta_0 = \ln(n_1/n_0)$ i $\beta_1 = 0$. Kao što je ranije navedeno za te vrijednosti parametara $\widehat{\pi} = n_1/n = \overline{y}$. Stoga, lijeva strana jednadžbe (1.12) postaje $\sum_{i=1}^{n} x_i(y_i - \overline{y})$. Može se pokazati da je procijenjena varijanca jednaka $\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})$. Testna statistika za Score test (ST) je

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

1.3.3 Procjena pouzdanim intervalom

Važan dodatak testiranju značajnosti modela je izračun i interpretacija intervala pouzdanosti za parametre od interesa. Kao u slučaju linearne regresije, možemo ih odrediti za koeficijent nagiba, slobodni koeficijent i logit. U nekim slučajevima, moglo bi biti od interesa pružiti procjene intervala za predviđene vrijednosti.

Osnova za konstrukciju intervalnih procjenitelja je ista statistička teorija koju smo koristili za formuliranje testova značajnosti modela. Konkretno, procjenitelji intervala pouzdanosti za koeficijent nagiba i slobodnog koeficijenta povezani su s Waldovim testovima. Rubne točke $100(1-\alpha/2)\%$ pouzdanog intervala za koeficijent nagiba su

$$\widehat{\beta_1} \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta_1}), \tag{1.19}$$

a za slobodni koeficijent su

$$\widehat{\beta_0} \pm z_{1-\alpha/2}\widehat{SE}(\widehat{\beta_0}), \tag{1.20}$$

gdje je $z_{1-\alpha/2}$ $(1-\alpha/2)$ kvantil standardne normalne distribucije, to jest gornja rubna točka od $100(1-\alpha/2)\%$ pouzdanog intervala standardne normalne distribucije, dok $\widehat{SE}(\cdot)$ označava procjenitelje standardne pogreške procjenitelja parametra.

Logit je linearni dio modela logističke regresije. Procjenitelj logita je

$$\widehat{g}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x, \tag{1.21}$$

dok je procjenitelj varijance procijenjenog logita dan sa

$$\widehat{Var}\big[\widehat{g}(x)\big] = \widehat{Var}\big(\widehat{\beta}_0\big) + x^2\widehat{Var}\big(\widehat{\beta}_1\big) + 2x\widehat{Cov}\big(\widehat{\beta}_0, \widehat{\beta}_1\big). \tag{1.22}$$

Rubne točke $100(1-\alpha/2)\%$ intervala pouzdanosti dane su s

$$\widehat{g}(x) \pm z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)], \tag{1.23}$$

gdje je

$$\widehat{SE}[\widehat{g}(x)] = \sqrt{\widehat{Var}[\widehat{g}(x)]}.$$

Procjenitelj logita i pripadni interval pouzdanosti pružaju osnovu za procjenu prilagođene vrijednosti, u ovom slučaju logističke vjerojatnosti, i procjenu intervala pouzdanosti. Posebno, korištenjem (1.12), ako je x vrijednost nezavisne varijable onda je procijenjena logistička vjerojatnost

$$\widehat{\pi}(x) = \frac{e^{\widehat{g}(x)}}{1 + e^{\widehat{g}(x)}}, \tag{1.24}$$

a rubne točke $100(1-\alpha/2)\%$ intervala pouzdanosti dobivene su iz odgovarajućih rubnih točaka intervala pouzdanosti za logit. Krajnje točke intervala

pouzdanosti dane su s

$$\frac{e^{\widehat{g}(x)\pm z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}{1+e^{\widehat{g}(x)\pm z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}. \tag{1.25}$$

Prilagođena vrijednost izračunata u (1.24) analogna je odgovarajućoj točki na regresijskom pravcu dobivenoj linearnom regresijom. U linearnoj regresiji svaka točka na procijenjenom regresijskom pravcu daje procjenu srednje vrijednosti zavisne varijable u populaciji subjekata s vrijednošću kovarijata x. Stoga je vrijednost $\widehat{\pi}(x)$ u (1.24) procjena srednje vrijednosti (tj. udjela) subjekta s kovarijantnom vrijednošću x u uzorku populacije koja ima dihotomnu varijablu ishoda kodiranu s 1. Svaki pojedini ispitanik u uzorku kovarijantnom vrijednošću x ili ima dihotomnu varijablu ishoda kodiranu s 1 ili s 0. Interval pouzdanosti sugerira da se ta srednja vrijednost kreće između

$$\frac{e^{\widehat{g}(x)+z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}{1+e^{\widehat{g}(x)+z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}. \tag{1.26}$$

i

$$\frac{e^{\widehat{g}(x)-z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}{1+e^{\widehat{g}(x)-z_{1-\alpha/2}\widehat{SE}[\widehat{g}(x)]}}. \tag{1.27}$$

s pouzdanošću od $100(1-\alpha/2)\%$.

1.3.4 Interpretacija parametara modela

Tumačenje bilo kojeg modela zahtijeva da budemo u mogućnosti donijeti praktične zaključke iz procijenjenih koeficijenata u modelu. Pitanje kojim se bavimo je: *"Što nam procijenjeni koeficijenti u modelu govore o istraživačkim pitanjima koja su motivirala studiju?"* Procijenjeni koeficijenti za nezavisne varijable predstavljaju promjenu funkcije zavisne varijable po

jedinici promjene u nezavisnoj varijabli. Stoga, interpretacija uključuje dva problema: određivanje funkcionalnog odnosa između zavisne varijable i nezavisne varijable i odgovarajuće definiranje jedinice promjene za nezavisnu varijablu.

Prvi korak je odrediti koja funkcija zavisne varijable daje linearnu funkciju nezavisnih varijabli. To se zove link funkcija. U modelu logističke regresije link funkcija je logit transformacija

$$g(x) = \ln\{\pi(x)/[1-\pi(x)]\} = \beta_0 + \beta_1 x.$$

Koeficijent nagiba predstavlja promjenu u logit koja odgovara promjeni jedne jedinice u nezavisnoj varijabli (tj. $\beta_1 = g(x+1) - g(x)$). Ispravno tumačenje koeficijenta u logističkom regresijskom modelu ovisi o mogućnosti pripisivanja značenja razlici između dva logita. Tumačenje ove razlike drugačije je od slučaja do slučaja jer se odnosi izravno na definiciju i značenje promjene jedne jedinice u nezavisnoj varijabli.

Nadalje, razmotrit ćemo i interpretaciju koeficijenata logističke regresije u slučaju kad je nezavisna varijabla nominalne ljestvice i dihotomna. Ovaj slučaj predstavlja konceptualni temelj za sve ostale situacije.

Pretpostavljamo da je nezavisna varijabla x kodirana kao nula ili kao jedan. Razlika u logitu za subjekt $x = 1$ i $x = 0$ je

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

U ovom slučaju logit razlika jednaka je β_1. Kako bismo protumačili ovaj rezultat, moramo uvesti i raspraviti mjeru povezanosti koja se naziva omjer izgleda, te definirati izglednost događaja.

Tablica 1.1: Vrijednosti modela logističke regresije s dihotomnom nezavisnom varijablom

	nezavisna varijabla (X)	
varijabla odziva (Y)	x=1	x=0
y=1	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
y=0	1-$\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	1-$\pi(0) = \frac{1}{1+e^{\beta_0}}$

Definicija 1.7 *Neka je A događaj. Izglednost događaja A definiramo sa* $\frac{P(A)}{1-P(A)}$.

Moguće vrijednosti logističkih vjerojatnosti mogu se prikladno prikazati u tablici 2×2 kao što je prikazano na Tablici 1.1.

Izglednost da će ishod biti prisutan među pojedincima s $x = 1$ definiran je kao $\pi(1)/[1-\pi(1)]$. Slično tome, izglednost da će ishod biti prisutan među pojedincima za koje je $x = 0$ definiran je kao $\pi(0)/[1-\pi(0)]$. Omjer izgleda, označen s OR, definiran je kao omjer izglednosti za $x = 1$ i izglednosti za $x = 0$, a dan je sa

$$\text{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}. \tag{1.28}$$

Uvrštavajući izraz za logistički regresijski model prikazan u Tablici 1.1 u (1.28) dobivamo

$$\begin{aligned}\mathrm{OR} &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\frac{1}{1+e^{\beta_0}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{(\beta_0+\beta_1)-\beta_0} \\ &= e^{\beta_1}.\end{aligned}$$

Dakle, za logističku regresiju s dihotomnom nezavisnom varijablom kodiranom s 1 i 0, odnos između omjera izglednosti i regresijskih koeficijenata je

$$\mathrm{OR} = e^{\beta_1}. \tag{1.29}$$

Ovaj jednostavan odnos između koeficijenta i omjera izglednosti temeljni je razlog zašto se logistička regresija pokazala tako moćnim analitičkim istraživačkim alatom.

U slučaju kada je OR = 1 izglednost prisutnosti ishoda jednaka je za objekte za koje je $x = 1$ i $x = 0$. U suprotnom, kažemo da je izglednost prisutnosti ishoda veća OR puta (OR > 1) ili manja OR puta (OR < 1) u skupini objekata za koje je $x = 1$ u odnosu na skupinu za koju je $x = 0$.

U mnogim slučajevima OR aproksimira veličinu koja se naziva relativni rizik. Ovaj parametar jednak je omjeru $\pi(1)/\pi(0)$.

Zbog jednostavnog tumačenja, OR obično uzimamo za parametar od interesa u logističkoj regresiji. Međutim, njegova procjena, $\widehat{OR}$, ima distribuciju koja je asimetrična dok je u teoriji, za dovoljno velike uzorke, distribucija $\widehat{OR}$ normalna. Nažalost, ovaj zahtjev za veličinom uzorka obično premašuje

zahtjev većine studija. Stoga se zaključci obično temelje na distribuciji uzorka $\ln(\widehat{OR}) = \widehat{\beta}_1$, koja slijedi normalnu distribuciju za mnogo manje uzorke. Procjenu $100(1-\alpha/2)\%$ intervala pouzdanosti za omjer izgleda dobijemo tako da prvo izračunamo rubne točke intervala pouzdanosti za koeficijent β_1, a zatim a zatim ih eksponenciramo. Općenito, krajnje točke dane su izrazom

$$\exp\big[\widehat{\beta}_1 \pm z_{1-\alpha/2} \times \widehat{SE}(\widehat{\beta}_1)\big].$$

Prije zaključivanja slučaja dihotomne varijable, važno je uzeti u obzir učinak koji kodiranje varijable ima na izračun procijenjenog omjera izgleda. Već smo zaključili da je procjena omjera izgleda $\widehat{OR} = exp(\widehat{\beta}_1)$. Ovo je točno kada je nezavisna varijabla kodirana kao 0 ili 1. Drugačije kodiranje može zahtijevati da izračunamo vrijednost razlike logita za specifično kodiranje koje smo koristili, a zatim potenciramo tu razliku kako bismo procijenili omjer izgleda.

1.4 Model složene logističke regresije

U ovom poglavlju ćemo opisati kako generalizirati model logističke regresije u slučaju kada zavisnu varijablu odziva želimo opisati pomoću više nezavisnih varijabli. Promotrimo kolekciju od p nezavisnih varijabli označenih s vektorom $\boldsymbol{X} = \big(X_1, X_2, \ldots, X_p\big)$. Nadalje, pretpostavimo, za sada, da je svaka od ovih varijabli na intervalnoj skali. Uvjetnu vjerojatnost prisutnosti nekog ishoda označimo s $\pi(\boldsymbol{X}) = P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$, gdje je $\boldsymbol{x} = (x_1, x_2, \ldots, x_p\big)$. Logit multivarijatne logističke regresije dan je s

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \tag{1.30}$$

Tablica 1.2: Primjer dummy varijable koja ima tri različita levela

X_j	D_1	D_2
0	0	0
1	1	0
2	0	1

a model logističke regresije dan je s

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}. \tag{1.31}$$

Općenoti, u model možemo uključiti prediktore mjerene na različitoj skali. Ako su neke od nezavisnih varijabli diskretne nominalne skale , kao što su primjerice rasa ili spol, nije ih prikladno uključiti u model kao da su varijable intervalne ljestvice. Brojevi koje koristimo za reprezentaciju diskretne varijable na različitim nivoima samo su indifikatori i nemaju numerički značaj. U ovoj situaciji metoda izbora je korištenje dizajna varijabli (ili dummy varijable). Ako nominalno skalirana varijabla X_j poprima k_j različitih vrijednosti (odnosno, ima k_j kategorija), onda uvodimo $k_j - 1$ dummy varijabli. Naprimjer, u slučaju $k_j = 3$ uvodimo dvije varijable D_1 i D_2, dok jedna kategorija uvijek služi kao referentna. Dizajn u ovom slučaju možemo prikazati kao u Tablici 1.1. Stoga bi logit za model s p varijabli u slučaju kada je j-ta varijabla diskretna (uz pretpostavku da su ostale varijable numeričke) bio oblika

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{j-1} x_{j-1} + \sum_{l=1}^{k_j - 1} \beta_l D_l + \beta_{j+1} x_{j+1} + \ldots + \beta_p x_p.$$

1.4.1 Procjena parametara modela

Pretpostavimo da imamo uzorak od n nezavisnih opažanja $(\boldsymbol{x_i}, y_i) = (x_{i_1}, x_{i_2}, \ldots, x_{i_p}, y_i)$, $i = 1, 2, \ldots, n$. Kao i u univarijatnom slučaju, da bismo mogli kontruirati model za dana opažanja moramo prvo dobiti procijenjene parametre $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_p)$. Metoda procjene korištena u slučaju više varijabli bit će ista kao u modelu jednostavne logističke regresije, to jest koristit ćemo metodu maksimalne vjerodostojnosti. Funkcija vjerodostojnosti je gotovo identična funkciji danoj sa (1.9). Jedina razlika je što je sada $\pi(\boldsymbol{x})$ definiran sa izrazom (1.31). U ovom slučaju ćemo imati $p + 1$ jednadžbi vjerodostojnosti koje se dobivaju deriviranjem logaritamske funkcije vjerodostojnosti s obzirom na $p + 1$ koeficijenata. Rezultirajuće jednadžbe vjerodostojnosti možemo izraziti na sljedeći način:

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

i

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

za $j = 1, 2, \ldots, p$.

Kao i u modelu jednostavne logističke regresije, rješavanje jednadžbi vjerodostojnosti zahtijeva poseban softver koji je dostupan u većini, ako ne i u svim statističkim paketima. Označimo sa $\widehat{\boldsymbol{\beta}}$ rješenja gornjih jednadžbi. Stoga, vrijednosti koje računamo su $\widehat{\boldsymbol{\pi}}(\boldsymbol{x_i})$, to jest vrijednosti izraza (1.31) dobivene uvrštavanjem $\widehat{\boldsymbol{\beta}}$ i $\boldsymbol{x_i}$.

Sada kada smo model logističke regresije generalizirali za slučaj više varijabli, potrebno je detaljnije ramotriti procjenu standardne pogreške.

Metoda procjene varijanci i kovarijanci procijenjenih koeficijenata slijedi iz dobro razvijene teorije metode maksimalne vjerodostojnosti. Procjenitelji su dobiveni iz matrice drugih parcijalnih derivacija logaritamske

funkcije vjerodostojnosti. Ove parcijalne derivacije imaju opći oblik

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \tag{1.32}$$

i

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i) \tag{1.33}$$

za $j, l = 0, 1, 2, \ldots, p$, gdje π_i označava $\pi(\boldsymbol{x_i})$. Neka je s $\boldsymbol{I}(\boldsymbol{\beta})$ označena $(p+1) \times (p+1)$ matrica koja sadrži članove $-\frac{\partial^2 L(\beta)}{\partial \beta_j^2}$ i $-\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l}$, to jest $\big[\boldsymbol{I}(\boldsymbol{\beta})\big]_{ij} = -\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l}$. Matricu $\boldsymbol{I}(\boldsymbol{\beta})$ nazivamo opaženom informacijskom matricom. Varijance i kovarijance procijenjenih koeficijenata su dobivene iz inverza matrice $\boldsymbol{I}(\boldsymbol{\beta})$ kojeg označavamo s $\mathrm{Var}(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$. Elemente matrice $\boldsymbol{I}(\boldsymbol{\beta})$ nije moguće eksplicitno izraziti osim u malom broju posebnih slučajeva. Stoga, j-ti dijagonalni element matrice ćemo označiti sa $\mathrm{Var}(\beta_j)$, što je varijanca od $\widehat{\beta}_j$, a proizvoljni nedijagonalni element sa $\mathrm{Cov}(\beta_j, \beta_l)$, što je zapravo kovarijanca od $\widehat{\beta}_j$ i $\widehat{\beta}_l$. Procijenitelje varijanci i kovarijanci, koje ćemo označiti s $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})$, dobivamo računanjem $\mathrm{Var}(\boldsymbol{\beta})$ za $\widehat{\boldsymbol{\beta}}$. Dakle, za vrijednosti elemenata matrice koristit ćemo oznake $\widehat{\mathrm{Var}}(\widehat{\beta}_j)$ i $\widehat{\mathrm{Cov}}(\widehat{\beta}_j, \widehat{\beta}_l)$, $j, l = 0, 1, 2, \ldots, p$. U većini slučajeva ćemo koristiti samo procijenjene standardne pogreške procijenjenih koeficijenata, koje ćemo označiti sa

$$\widehat{\boldsymbol{SE}}\big(\widehat{\boldsymbol{\beta_j}}\big) = \sqrt{\Big[\widehat{\boldsymbol{Var}}\big(\widehat{\boldsymbol{\beta_j}}\big)\Big]} \tag{1.34}$$

za $j = 0, 1, 2, \ldots, p$. Danu notaciju ćemo koristiti i u razvoju metoda za testiranje koeficijenata i procjenu intervala pouzdanosti.

U daljnim razmatranjima bit će nam koristan sljedeći oblik informacijske matrice:

$$\widehat{\mathbf{I}}\big(\widehat{\boldsymbol{\beta}}\big) = \mathbf{X}^{\boldsymbol{T}}\mathbf{V}\mathbf{X},$$

gdje je $\mathbf{X}$ matrica dimenzije $n \times (p+1)$ koja sadrži podatke za svakog ispitavanika, a $\mathbf{V}$ je dijagonalna matrica dimenzije $n \times n$ sa dijagonalnim elementima $\widehat{\pi}_i(1-\widehat{\pi}_i)$. Odnosno, matrica $\mathbf{X}$ je

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

a matrica $\mathbf{V}$ je

$$\mathbf{V} = \begin{pmatrix} \widehat{\pi}_1(1-\widehat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \widehat{\pi_n}(1-\widehat{\pi_n}) \end{pmatrix}.$$

1.4.2 Testiranje značajnosti parametara modela

Test omjera vjerodostojnosti za značajnost modela s uključenih p parametara provodi se kao u univarijatnom slučaju. Osnova testa je ista: usporediti log vjerodostojnost procijenjenog modela i log vjerodostojnost modela koji sadrži samo β_0 (takozvani nul model). Test se temelji na statistici G navedenoj izrazom (1.16). Jedina je razlika što su izračunate vrijednosti $\widehat{\pi}$ za model bazirane na vektoru koji sadrži $p+1$ parametara $\widehat{\boldsymbol{\beta}}$. Uz pretpostavku istinitosti nul hipoteze, to jest da je p koeficijenata nagiba za kovarijate u modelu jednako nula ($\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$), G će imati χ^2 distribuciju s p stupnjeva slobode. P-vrijednost tada računamo kao $p = P(\chi^2(p) > g)$, gdje g označava opažene vrijednosti testne statistike na temelju našeg uzorka. Ako je p-vrijednost jako mala onda odbacujemo nul hipotezu i zaključujemo da je barem jedan koeficijent različit od nula, odnosno, model je značajan.

Test značajnosti pojedinih prediktora također provodimo na isti način kao u univarijatnom slučaju. Ako promatramo značajnost j-tog prediktora, onda nul i alternativna hipoteza glase:

$$\mathrm{H}_0 : \beta_j = 0,$$
$$\mathrm{H}_1 : \beta_j \neq 0.$$

Nadalje, računamo testnu statistiku $\mathrm{W} = \frac{\widehat{\beta}_j}{\widehat{\mathrm{SE}}(\widehat{\beta}_j)} \overset{\mathrm{H}_0}{\sim} \mathrm{N}(0,1)$ (Waldov test) i pripadnu p-vrijednost $p = P(\mathrm{Z} > |w|)$, gdje je w opažena vrijednost testne statistike na temelju našeg uzorka. Ako je p-vrijednost jako mala, odbacujemo nul hipotezu i zaključujemo da je j-ti prediktor značajan. Isto tako možemo testirati značajnost prediktora pomoću testne statistike

$$\mathrm{G} = \text{model bez varijable } - \text{model s varijablom}$$

Uz pretpostavku istinitosti nul hipoteze (varijabla nije značajna) G će imati χ^2 distribuciju s 1 stupnjem slobode.

Ako je naš cilj dobiti najprikladniji model uz minimiziranje broja parametara, onda je idući korak procijeniti jednostavniji model koji sadrži samo one varijable za koje se smatra da su značajne i usporediti ga sa punim modelom koji sadrži sve varijable. Usporedba modela također se provodi uspoređivanjem log vjerodostostojnosti modela sa i bez uključenim podskupom varijabli, kao i u (1.16). U ovom slučaju će sad G imati χ^2 distribuciju s onoliko stupnjeva slobode koliko smo isključili varijabli. Nadalje, uključivanje i isključivanje varijabli ne vrši isključivo na temelju rezultata testova. Odluka o uključivanju i isključivanju varijabli donosi se u skladu s teorijskim postavkama u ovisnosti o prirodi modela.

Ukoliko je kategorijalna varijabla uključena u model, tada svi njeni leveli moraju biti uključeni. Dakle, ako zadržimo varijablu zbog značajnosti

barem jednog levela, onda ostavljamo i sve druge. Inače, podrazumijeva se da smo rekodirali varijablu, a to (nekad) nema smisla.

Za univarijantni model smo opisali dva testa ekvivalentna testu omjera vjerojatnosti za procjenu značajnosti modela. To su Wald i Score testovi. Ukratko ćemo navesti multivarijatne verzije ovih testova. Ovi testovi dostupni su u mnogim programskim paketima.

Multivarijatni analog Waldovog testa dobiva se iz sljedećeg izračuna:

$$\begin{aligned}\mathbf{W} &= \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \left[\widehat{\mathbf{Var}}\left(\widehat{\boldsymbol{\beta}}\right)\right]^{-1} \widehat{\boldsymbol{\beta}} \\ &= \widehat{\boldsymbol{\beta}}^{\boldsymbol{T}} \left(\mathbf{X'VX}\right) \widehat{\boldsymbol{\beta}}.\end{aligned}$$

Uz pretpostavku istinitosti nul-hipoteze da je svaki od $p+1$ koeficijenata jednak nuli ($H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0$), $\mathbf{W}$ prati χ^2 distribuciju s $p+1$ stupnjeva slobode. Testovi za samo p koeficijenta koji se nalaze uz prediktorske varijable modela dobivaju se eliminacijom $\widehat{\beta_0}$ iz $\widehat{\boldsymbol{\beta}}$ i relevantnog retka (prvog ili zadnjeg) i stupca (prvog ili zadnjeg) iz $\mathbf{X^T V X}$.

Multivarijatni analog Score testa za značajnost modela temelji se na distribuciji derivacija funkcija vjerodostojnosti u odnosu na $\boldsymbol{\beta}$. Izračun ovog testa je jednako kompliciran kao i Waldov test jer je za njegovo detaljno definiranje potrebno uvođenje dodatnih notacija koje kasnije ne bismo ni koristili.

1.4.3 Procjena pouzdanim intervalom

Metode korištene za procjene pouzdanim intervalima u slučaju multivarijatne logističke regresije slične su kao u slučaju univarijatne regresije kako je prikazano u poglavlju 1.2.3.

Rubne točke $100(1-\alpha/2)\%$ intervala pouzdanosti za koeficijent dobivaju se iz (1.19) za koeficijente nagiba i iz (1.20) za slobodni koeficijent.

Njihova interpretacija bi glasila da s pouzdanošću od $100(1-\alpha/2)\%$ procijenjujemo kako se porastom vrijednosti nezavisne varijable x_j za jednu jedinicu logaritam omjera izglednosti mijenja između $\widehat{\beta}_j - z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_j)$ i $\widehat{\beta}_j + z_{1-\alpha/2}\widehat{SE}(\widehat{\beta}_j)$ uz fiksne vrijednosti ostalih varijabli.

Procjenitelj intervala pouzdanosti za logit je malo kompliciraniji za model s više varijabli od rezultata prikazanog u (1.23). Osnovna ideja je ista, samo što je sada uključeno više izraza. Iz (1.30) slijedi da je opći izraz za procjenitelj logit modela koji sadrži p kovarijata dan sa

$$\widehat{g}(\boldsymbol{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p. \tag{1.35}$$

Alternativni način da se izrazi procjenitelj logita u (1.35) je korištenjem vektorskog zapisa kao $\widehat{g}(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}$, gdje vektor $\widehat{\boldsymbol{\beta}}^T = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)$ označava procjenitelje $p+1$ koeficijenata i vektor $\boldsymbol{x}^T = (x_0, x_1, x_2, \ldots, x_p)$ koji predstavlja konstantu i skup vrijednosti p kovarijabli u modelu, gdje je $x_0 = 1$.

Iz (1.22) slijedi da je izraz za procjenitelja varijance procijenjenog logita iz (1.35) dan sa

$$\widehat{\mathrm{Var}}\big[\widehat{g}(\boldsymbol{x})\big] = \sum_{j=0}^{p} x_j^2 \widehat{\mathrm{Var}}\big(\widehat{\beta}_j\big) + \sum_{j=0}^{p}\sum_{k=j+1}^{p} 2x_j x_k \widehat{\mathrm{Cov}}\big(\widehat{\beta}_j, \widehat{\beta}_k\big). \tag{1.36}$$

Prethodni rezultat možemo izraziti u kraćoj formi korištenjem matričnog izraza za procjenitelj varijance procjenitelja koeficijenata. Iz izraza za opaženu informacijsku matricu imamo

$$\widehat{\mathrm{Var}}\big(\widehat{\beta}\big) = \big(\mathbf{X}^{\boldsymbol{T}}\mathbf{V}\mathbf{X}\big)^{-1}. \tag{1.37}$$

Sada iz (1.37) slijedi da je izraz za procjenitelja u (1.36) ekvivalentan izrazu

$$\begin{aligned}\widehat{\mathrm{Var}}\left[\left(\widehat{g}(\mathbf{x})\right)\right] &= \mathbf{x}^{\boldsymbol{T}}\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})\mathbf{x} \\ &= \mathbf{x}^{\boldsymbol{T}}(\mathbf{X'VX})^{-1}\mathbf{x}.\end{aligned} \tag{1.38}$$

1.5 Ocjena prilagodbe modela

Postoji više metoda ocjene prilagodbe modela. Ukratko prikazat ćemo ocjenu prilagodbe temeljem klasifikacijske tablice što uključuje pokazatelje osjetljivosti i specifičnosti, ROC krivulje, odnosno AUC mjere, AIC kriterija i HL testa budući da su iste korištene u poglavlju gdje je provedena primjena.

Klasifikacijska tablica jedan je od načina da prikažemo procijenjenu snagu dihotomnog regresijskog modela. U tablici se klasificira dihotomna varijabla ishoda s obzirom na predikcije. U slučajevima kada je zavisna varijabla kategorijalna sa više od dvije kategorije ili kontinuirana varijabla, potrebno je odrediti graničnu vrijednost na temelju koje će se vrijednost varijable ishoda klasificirati kao jedan, odnosno nula. Predviđena vrijednost varijable ishoda je nula (to jest $\widehat{Y} = 0$) kada je $\widehat{\pi}_i > \pi_0$, a predviđena vrijednost varijable ishoda je jedan (to jest $\widehat{Y} = 1$) kada je $\widehat{\pi}_i \leq \pi_0$ za neku određenu granicu π_0. Različite granične vrijednosti daju različite klasifikacijske tablice.

Definicija 1.8 *Osjetljivost (Sensitivity) modela je vjerojatnost da model predvidi pozitivan ishod opažanja kada je vrijednost varijable ishoda doista pozitivna, to jest predstavlja sposobnost modela da ispravno prepozna pozitivne ($Y = 1$) ishode.*

Definicija 1.9 *Specifičnost (Specificity) modela je vjerojatnost da model predviđa negativan ishod promatranja kada je vrijednost varijable ishoda doista nega-*

tivna, to jest predstavlja spospobnost modela da ispravno prepozna negativne ($Y = 0$) ishode.

U svrhu boljeg objašnjena promotrimo model u kojem širina i boja oklopa ženke raka predstavljaju nezavisne varijable, dok prisutnost mužjaka u njezinoj blizini predstavlja dihotomnu varijablu ishoda. Na temelju opisanih varijabli procijenjen je logistički model M_1. U tablici je prikazana klasifikacijska tablica koristeći granice $\pi_0 = 0.5$ i $\pi_0 = 0.642$.

Tablica 1.3: Klasifikacijska tablica za podatke o ženki raka

	$\pi_0 = 0.64$		$\pi_0 = 0.5$		
	$\widehat{Y} = 1$	$\widehat{Y} = 0$	$\widehat{Y} = 1$	$\widehat{Y} = 0$	Ukupno
$Y = 1$	74	37	94	17	111
$Y = 0$	20	42	37	25	62

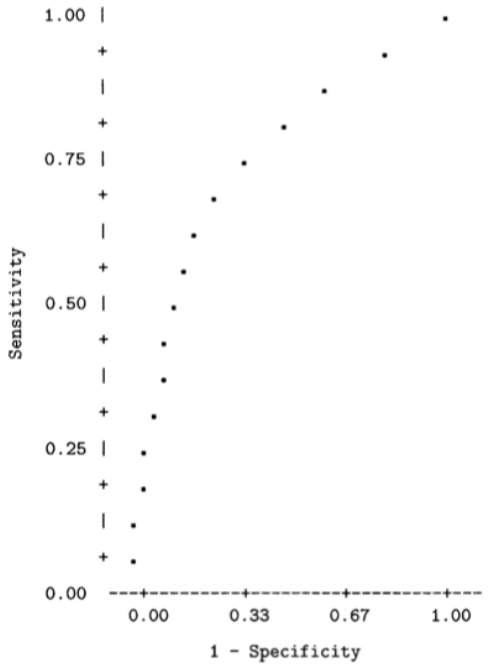
Iz Tablice 1.3 možemo vidjeti da različite granične vrijednosti daju i različite vrijednosti osjetljivosti i specificnosti. Naime, kada je $\pi_0 = 0.642$ procijenjena osjetljivost je jednaka $74/111 = 0,667$ i procijenjena specifičnost je jednaka $42/62 = 0,667$, a kada je $\pi_0 = 0.5$ procijenjena osjetljivost je jednaka $94/111 = 0,847$ i procijenjena specifičnost je jednaka $25/62 = 0,403$.

Jedan od načina vizualizacije osjetljivosti i specifičnosti modela je crtanje ROC krivulje.

Definicija 1.10 *ROC (Receiver operating characteristic) krivulja je grafički prikaz osjetljivosti kao funkcije (1 – specifičnost) za sve moguće granične vrijednosti π_0.*

ROC krivulja nam prikazuje više informacija od klasifikacijske tablice jer sadrži ocjenu prilagodbe modela za sve moguće granične vrijednosti π_0. Za

određenu specifičnost, veća ocjena prilagodbe modela odgovara većoj osjetljivosti. Dakle, što je veća snaga predviđanja, to je ROC krivulja viša. Slika prikazuje ROC krivulju za model M_1.



Slika 1.2: ROC krivulja za logistički regresijski model M_1

Definicija 1.11 *Površina ispod krivulje ("area under ROC curve") ili kraće AUC je mjera sposobnosti modela da uspješno klasificira podatke, to jest vjerojatnost da će model rangirati slučajno odabrani ishod koji je jednak jedinici ($Y = 1$) višim nego što će rangirati slučajno odabrani ishod koji je jednak nuli ($Y = 0$), uz pretpostavku da ishodi $Y = 1$ imaju veći rang.*

AUC može poprimiti vrijednosti između nula i jedan. Što je vrijednost *AUC* veća, to je prediktivna sposobnost modela veća. Uobičajno interpretiramo vrijednost *AUC* na sljedeći način:

$AUC = 0.5$ model je bezvrijedan,

$AUC \in [0.7, 0.8 >$ model je prihvatljiv,

$AUC \in [0.8, 0.9 >$ jako dobra ocjena modela,

$AUC \geq 0.9$ izvrsna ocjena modela.

Jedan od često korištenih informacijskih kriterija je AIC ("Akaike information criterion") koji ocjenjuje prilagodbu modela računanjem koliko su njegove prilagođene vrijednosti bliske pravim očekivanim vrijednostima. Optimalni model je onaj koji nastoji imati svoje prilagođene vrijednosti najbliže stvarnim vjerojatnostima ishoda, kažnjavajući model s većim brojem procijenjenih parametara, to jest model koji minimizira

$$AIC = -2(\log \text{vjerodostojnost} - \text{broj procijenjenih parametara u modelu}).$$

Hosmer- Lemeshow test (HL test) ispituje u kojoj mjeri stvarni podaci odgovaraju onima predviđenim modelom. Ukratko, HL test ispituje odgovaraju li opažene stope događaja očekivanim stopama događaja u podskupovima populacije u istraživanju. Potrebno je odrediti broj podskupina te rezultat testa može varirati s obzirom na odabrani broj grupa u smislu da nas dovodi do suprotnih zaključaka.

Poglavlje 2

Primjena

2.1 Opis problema

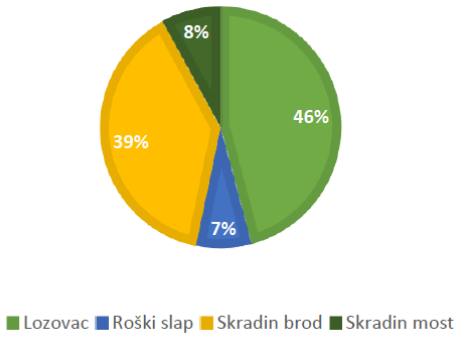
Modeliranje vjernosti potrošača aktivno je istraživačko područje. Vjernost potrošača usko je vezana sa zadovoljstvom potrošača. Logistička regresija jedna je od najpopularnijih metoda korištenih za analizu povezanosti vjernosti i zadovoljstva potrošača, te općenito izgradnju prediktivnih modela koji se bave pitanjem vjernosti potrošača. U ovom radu logistička regresija primijenjena je u svrhu modeliranja vjernosti posjetitelja zaštićenih područja. Vjernost je promatrana kroz namjeru ponovne posjete. Osim prikaza primjene logističke regresije, cilj analize bio je ispitati možemo li vjernost posjetitelja predvidjeti na temelju njihovog zadovoljstva. Također, u svrhu prikaza izgradnje modela sa različitim vrstama varijabli, odnosno prediktora, vjernost modeliramo i uz jednu kategorijalnu varijablu koja separira domaće i strane goste.

2.2 Uzorak

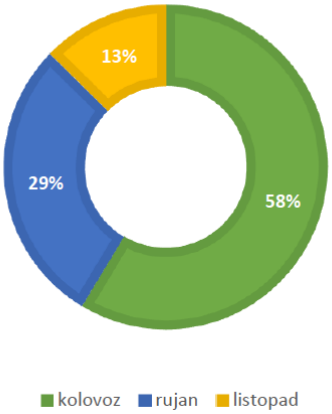
Populaciju u ovom istraživanju čine svi posjetitelji koji su tijekom kolovoza, rujna i listopada 2019. godine posjetili NP Krka. Ispitanik je osoba starija od osamnaest godina koja je u individualnom posjeti NP Krka bila pred krajem svog boravka. U dizajnu uzorka korišten je stratificirani slučajni uzorak prema:

1. Lokaciji izlaza,
2. Kalendarskom mjesecu posjete,
3. Vremenu posjete.

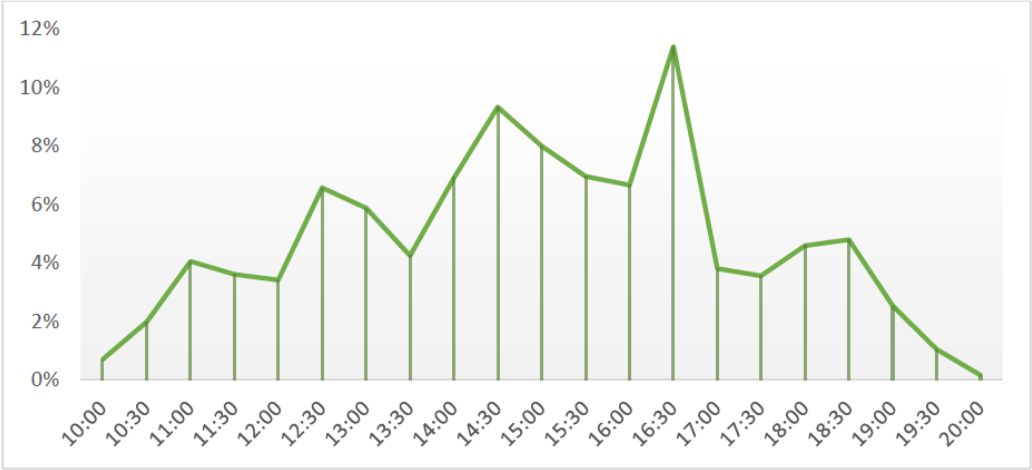
Stratifikacija prema lokaciji i prema mjesecu posjete provedena je na osnovu podataka u broju prodanih karata u razdoblju 2015- 2017. Obrađena su ukupno 2053 anketna upitnika. Anketiranje posjetitelja se provodilo na izlazima na 4 lokacije nacionalnog parka: izlaz Lozovac (lokacija gdje posjetitelji čekaju prijevoz do Lozovca), izlaz Skradin (lokacija gdje posjetitelji čekaju prijevoz brodom do Skradina), izlaz Skradin-most i izlaz Roški slap. Kvote ispitanike prema mjestu i vremenu prikupljanja bile su unaprijed određene. Struktura prikupljenih upitnika prema lokaciji prikupljanja prikazana je na slikama 2.1., dok je struktura prema mjesecu prikupljanja prikazana na slikama 2.2..



Slika 2.1: Lokacija prikupljanja



Slika 2.2: Mjesec prikupljanja



Slika 2.3: Vrijeme anketiranja

2.3 Opis varijabli

Kako je naglašeno u uvodu, u ovoj analizi ograničili smo skup prediktora na varijable vezane uz porijeklo i zadovoljstvo posjetitelja. Vrijabla Zemlja porijekla poslužila je kao primjer izgradnje modela sa kategorijalnom nezavisnom varijablom. Nezavisna varijabla Zemlje porijekla kodirana je s dvije kategorije, a to su "1" za domaće posjetitelje, te "0" ako je ispitanik stranog

porijekla.

Zadovoljstvo posjetitelja ponudom Parka analizirano je za 19 elemenata turističke ponude, a to su: Stručnost osoblja nacionalnog parka, Ljubaznost osoblja nacionalnog parka, Signalizacija do parka, Signalizacija u parku, Korisnost vodiča i karata za posjetitelje koji su dostupni u parku, Dostupnost informacija o biljkama i životinjama u parku, Informacije o lokalnim običajima i tradiciji, Širina spektra dostupnih aktivnosti, Rekreativni sadržaji, Edukativni sadržaji, Dostupnost lokalnih proizvoda, Ponuda suvenira, Kvaliteta i održavanost staza i šetnica, Sanitarni čvorovi, Visina cijene ulaznice, Vrijednost za novac posjeta parku, Ukupno zadovoljstvo parkom, Zadovoljenost motiva dolaska, te Gužva.

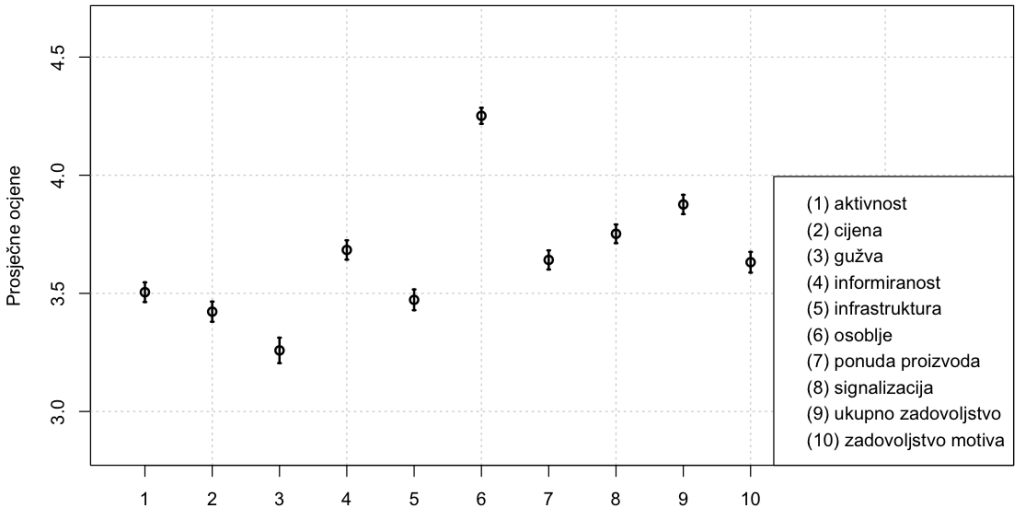
Ispitanici su svaki element ponude ocjenjivali ocjenama od 1 (izrazito nezadovoljan) do 5 (izrazito zadovoljan). Budući da modeli logističke regresije s visoko koreliranim nezavisnim varijablama mogu dovesti do netočnih rezultata, i budući da prvi i drugi element zadovoljstva zapravo govori o zadovoljstvu osobljem, za potencijalnu prediktorsku varijablu modela (zadovoljstvo osobljem) izračunata je aritmetička sredina ocijena prvog i drugog elementa. Zbog istih argumenata promatraju se aritmetičke sredine trećeg i četvrtog, petog, šestog i sedmog, osmog, devetog i desetog, jedanaestog i dvanaestog, trinaestog i četrnaestog i petnaestog i šestanestog elementa kao potencijalne prediktorske varijable signalizacije, informiranosti, aktivnosti, ponude proizvoda, infrastrukture i cijena, redom.

Koncept vjernosti operacionaliziran je varijablom koja mjeri namjeru ponovne posjete parku. Ispitanici su ocijenjivali namjeru ponovne posjete parku odgovorima "Da", "Ne" i "Ne znam". Varijabla odziva za naš model upravo je ova ocijena namjere ponovnog dolaska i kodirana je kao "0" ako je ispitanik odgovorio sa "Ne" ili "Ne znam" i "1" ako je odvorio sa "Da".

2.4 Deskriptivna varijabli

Vizualizacija podataka možda je najbrži i najkorisniji način da se sažmu podaci i saznamo više o njima.

Na Slika 2.4 prikazane su prosječne vrijednosti zadovoljstva zajedno sa 95% pouzdanim intervalima, dok su sve numeričke vrijednosti dane u tablici 2.1.



Slika 2.4: Usporedba 95% pouzdanih intervala prosječnih ocjena elementa zadovoljstva posjetitelja

Element osoblje ima najveću prosječnu ocjenu (4.25), zatim slijedi ukupno zadovoljstvo sa prosječnom ocjenom 3.88. U prosjeku najlošije ocijenjeni element je gužva sa prosječnom ocjenom 3.26. Najveću razinu raspršenosti bilježi element gužva (37%).

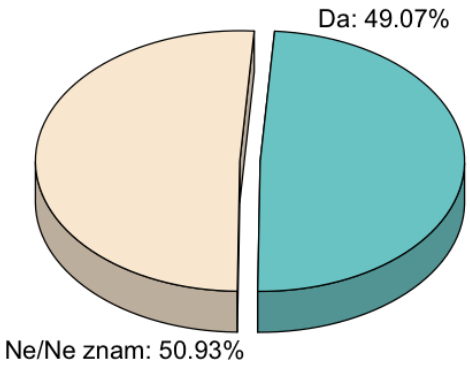
2.4. Deskriptivna varijabli

Tablica 2.1: Prosječne ocjene, koeficijent varijacije i granice 95% pouzdanih intervala

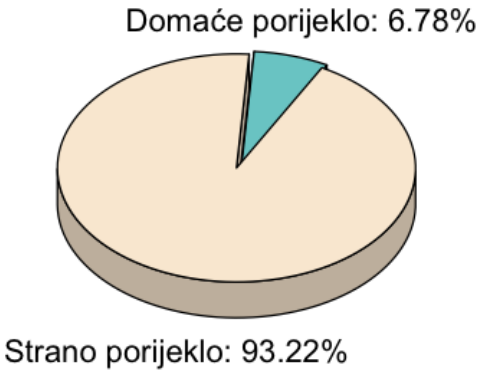
	prosječna ocjena	Koef. varijacije (%)	Donja granica	Gornja granica
Aktivnosti	3.50	26%	3.46	3.55
Cijena	3.42	28%	3.38	4.47
Gužva	3.26	37%	3.20	3.31
Informiranost	3.68	25%	3.64	3.72
Infrastruktura	3.43	28%	3.43	3.52
Osoblje	4.25	18%	4.22	4.29
Ponuda prizvoda	3.64	25%	3.60	3.68
Signalizacija	3.75	24%	3.71	3.79
Ukupno zadovoljstvo	3.88	24%	3.84	3.92
Zadovoljstvo motiva	3.63	27%	3.59	3.68

Na Slika 2.5 prikazani su stavovi o namjeri ponovne posjete posjetitelja Parka. Nešto više od polovine posjetitelja nije iskazalo namjeru ponovne posjete parku (50.93%), dok nešto manje od polovine posjetitelja ima namjeru ponovno posjetiti park.

Na Slika 2.6 prikazano je sociodemografsko obilježje zemlje porijekla posjetitelja Parka. Gotovo većina posjetitelja je stranog porijekla (93.22%), dok je nešto malo posjetitelja domaćeg porijekla (6.78%).



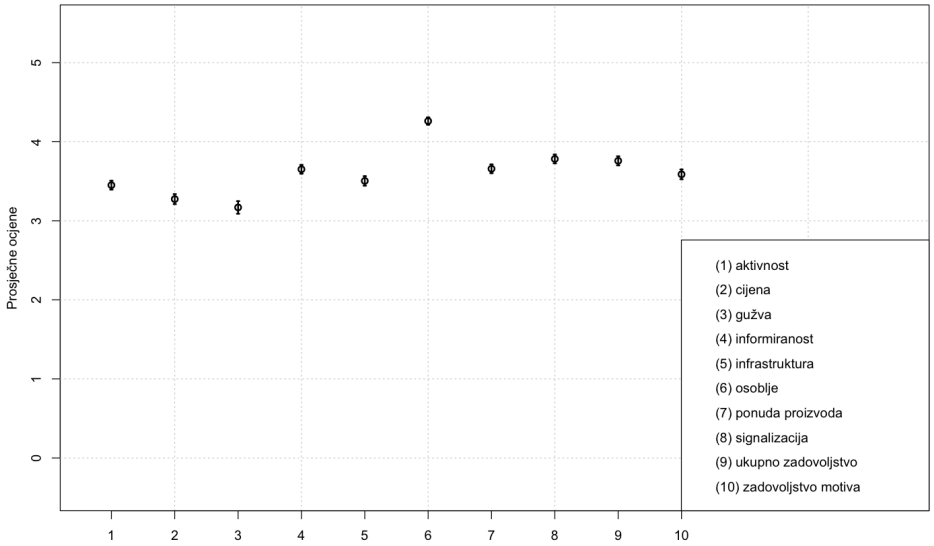
Slika 2.5: Namjera ponovne posjete



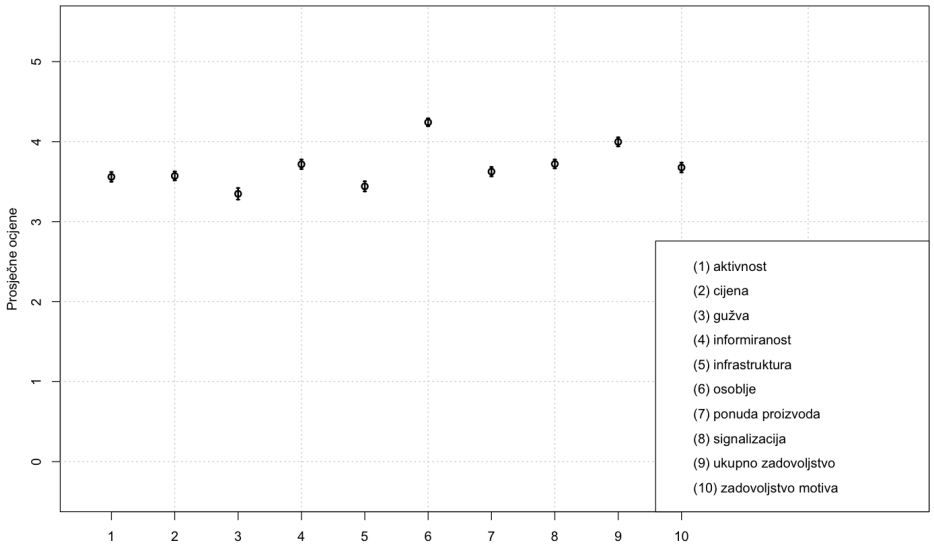
Slika 2.6: Zemlja porijekla posjetitelja

Na Slikama 2.7 i 2.8 prikazane su prosječne vrijednosti zadovoljstva zajedno sa 95% pouzdanim intervalima s obzirom na varijablu namjere ponovne posjete Parka.

2.4. Deskriptivna varijabli



Slika 2.7: Usporedba 95% pouzdanih intervala prosječnih ocjena elementa zadovoljstva posjetitelja koji nemaju namjeru ponovne posjete Parka



Slika 2.8: Usporedba 95% pouzdanih intervala prosječnih ocjena elementa zadovoljstva posjetitelja koji imaju namjeru ponovne posjete Parka

2.5 Modeli s kvantitavnim nezavisnim varijablama

U ovom poglavlju prikazat ćemo izgradnju logističkog modela s kvantitativnim varijablama. Generalno, univarijatna analiza koja prethodi multivarijantnoj analizi treba uključivati sve moguće prediktore bez obzira na vrstu varijable (kategorijalna/kvantitavina). No, radi lakše ilustracije izgradnje modela ovdje smo razdvojili početnu analizu s obzirom na vrstu varijable. Dodatno, sve numeričke varijable uključene u model vezane su uz koncept zadovoljstva posjetitelja Parka.

2.5.1 Univarijatni modeli

Za svaki ocijenjeni element zadovoljstva (x) izgrađen je univarijatan model logističke regresije, $g(x) = \beta_0 + \beta_1 x$, gdje je varijabla ishoda namjera ponovne posjete Parka, u svrhu odlučivanja koju varijablu uključiti u konačni najadekvatniji multivarijantni model.

U Tablici 2.2 prikazani su rezultati procjena parametara za svaki element zadovoljstva. Standardna je praksa koristiti p-vrijednosti za odlučivanje hoće li se varijable uključiti u konačni model. Niska p-vrijednost (obično $< 0,05$) znači da možemo odbaciti nultu hipotezu $H_0 : \beta_1 = 0$, odnosno da je varijabla značajna. Iz izračunatih p-vrijednosti prikazanih u Tablici 2.2 vidimo da su statistički značajne varijable Aktivnost,Cijena,Ukupno zadovoljstvo, Zadovoljstvo motiva i Gužva, na razini značajnosti od 5%.

Demonstracije radi, interpretirajmo peti univarijatni model gdje je nezavisna varijabla element Aktivnost, a zavisna, kao i u svakom univarijatnom modelu, namjera ponovne posjete Parka. Budući da koeficijenti logističke regresije daju promjenu u logaritamskoj izglednosti ishoda za povećanje

prediktorske varijable za jednu jedinicu interpretacija je sljedeća: Za svako povećanje jedne jedinice u prosječnoj ocjeni elementa Aktivnosti, izglednost ponovne posjete Parku (u odnosu da namjere ponovne posjete Parka nema) povećava se $e^{0.131} = 1.14$ puta.

Tablica 2.2: Univarijatna logistička regresija

Varijable		procijenjeni koef.	standardna pogreška	z-vrijednost	p-vrijednost
Infrastruktura	β_0	0.216	0.168	1.287	0.198
	β_1	-0.066	0.047	-1.415	0.157
Signalizacija	β_0	0.255	0.194	1.310	0.190
	β_1	-0.076	0.050	-1.502	0.133
Osoblje	β_0	0.108	0.255	0.421	0.674
	β_1	-0.034	0.059	-0.575	0.565
Informiranost	β_0	-0.315	0.194	-1.624	0.104
	β_1	0.080	0.051	1.571	0.116
Aktivnost	β_0	-0.456	0.185	-2.461	0.0139
	β_1	0.131	0.051	2.561	0.0104
Ponuda prizvoda	β_0	0.118	0.193	0.614	0.539
	β_1	-0.041	0.051	-0.802	0.423
Cijena	β_0	-1.174	0.179	-6.550	$5.77e^{-11}$
	β_1	0.341	0.0505	6.763	$1.35e^{-11}$
Ukupno zadovoljstvo	β_0	-1.147	0.203	-5.636	$1.74e^{-08}$
	β_1	0.290	0.051	5.681	$1.34e^{-08}$
Zadovoljstvo motiva	β_0	-0.342	0.175	-1.955	0.051
	β_1	0.093	0.047	1.994	0.0461
Gužva	β_0	-0.395	0.131	-3.017	0.003
	β_1	0.123	0.038	3.252	0.001

2.5.2 Multivarijatni modeli

Nakon utvrđivanja statistički značajnih varijabli analizom univarijatnih modela, provedena je Stepwise procedura. Stepwise procedura je postupak korišten pri identifikaciji najadekvatnijeg modela u kontekstu višestruke regresije. Općenito, razlikujemo Forward,Backward te kombinaciju Forward i Backward procedure.

Forward, odnosno dabir prema naprijed, počinje od nezavisne varijable koja daje najviše informacija o zavisnoj varijabli. Ova se varijabla zadržava u svim budućim modelima. U drugoj fazi postupak razmatra preostale varijable i određuje koja, u kombinaciji s prvom varijablom, pruža najviše dodatnih informacija o zavisnoj varijabli. Ovaj postupak se nastavlja sve dok nema daljnjih varijabli koje daju značajan dodatni doprinos prilagodbi modela. Uzastopni doprinosi se uspoređuju korištenjem različitih kriterija. Na primjer moguće je donijeti odluku koju varijablu uključiti na temelju Akaikeovog informacijskog kriterija (AIC) pri čemu biramo model s najmanjom vrijednosti AIC.

Backward, donosno eliminacija unatrag, počinje od modela koji sadrži sve varijable i uklanja neinformativne varijable, jednu po jednu. Neinformativnost varijable može se mjeriti različitim kriterijima. Kao i u slučaju Forward odabira, najčešće se oslanjamo na statističke testove ili različite informacijske kriterije.

Kombinacija Forward i Beckward procedure počinje od modela koji sadrži sve varijable te nakon izbacivanja neinformativne varijable u budućim koracima ponovno uključuje isključenu varijablu te onda računa vrijednost kriterija.

Stepwise procedura provedena je s početnim skupom prediktora koji uključuje statističke značajne varijable analizom univarijatnih modela, te

Tablica 2.3: Prvi korak stepwise procedure za dobivene statistički značajne varijable analizom univarijatnih modela

	Df	Deviance	AIC
- Aktivnost	1	2019.4	2029.4
- Gužva	1	2020.5	2030.5
- Motiv	1	2020.8	2030.8
- Ukupno zadovoljstvo	1	2025.5	2035.5
- Cijena	1	2032.1	2042.1

kao rezultat daje da je najbolji multivarijantan model onaj koji sadrži dvije nezavisne varijable,a to su element Cijena i Ukupno zadovoljstvo.

Promotrimo malo detaljnije korake stepwise procedure provedene za dobivene statistički značajne varijable analizom univarijatnih modela. Na početku, gledamo model u kojem su kao nezavisne varijable uključene sve varijable. Startna vrijednost AIC cijelog modela iznosi 2031.25. Zatim promatramo kolika će biti vrijednost AIC modela kada izostavimo pojedinu varijablu. Oznaka "-" u tablici predstavlja da je ta varijabla izostavljena iz modela. Vrijednosti prvog koraka procedure su prikazane u Tablica 2.3.

Iz tablice vidimo da ako isključimo varijablu koja predstavlja element Aktivnost model će imati najmanji AIC (2029.4), i taj AIC je manji od 2031.25, pa u prvom koraku izostavljamo tu varijablu iz modela.

U drugom koraku procedure vrijednosti su prikazane u Tablica 2.4.

Primjetimo, iako smo u prvom koraku isključili varijablu koja predstavlja element Aktivnost, svejedno se uzima u obzir i njeno ponovno uključenje u model označeno sa znakom "+". U drugom koraku vidimo da ako izostavimo varijablu koja predstavlja element Gužva model će imati najmanji AIC

Tablica 2.4: Drugi korak stepwise procedure za dobivene statistički značajne varijable analizom univarijatnih modela

	Df	Deviance	AIC
- Gužva	1	2020.6	2028.6
- Motiv	1	2021.2	2029.2
+ Aktivnost	1	2019.2	2031.2
- Ukupno zadovoljstvo	1	2025.6	2033.6
- Cijena	1	2032.1	2040.1

(2028.6) i taj AIC je manji od 2029.37, pa se ta varijabla izostavlja iz modela.

U trećem koraku procedure vrijednosti su prikazane u Tablica 2.5.

Sada će model imati najmanji AIC ako izostavimo varijablu koja predstavlja element Motiv i taj AIC (2027.7) je manji od 2028.57 pa se ta varijabla izostavlja iz modela.

U četvrtom koraku procedure vrijednosti su prikazane u Tablica 2.6.

Najmanju vrijednost AIC (2028.6) model će imati ako dodamo varijablu koja predstavlja element Motiv ,no ta vrijednost nije manja od 2027.66 pa procedura staje.

Promotrimo sada dobiveni model i označimo ga sa M_2. Model je sljedeći:

$$\ln \frac{\pi(Y=1)}{1-\pi(Y=1)} = -1.66728 + 0.27229 * x_1 + 0.17925 * x_2 \ ,$$

gdje $Y = 1$ označava da ispitanik ima namjeru ponovno posjetiti Park, x_1 označava vrijednosti varijable koja predstavlja element Cijena i x_2 označava vrijednost varijable koja predstavlja element Ukupno zadovoljstvo.

U Tablici 2.7 prikazani su rezultati procjena parametara, standardna pogreška, z-vrijednost i p-vrijednost modela M_2.

Tablica 2.5: Treći korak stepwise procedure za dobivene statistički značajne varijable analizom univarijatnih modela

	Df	Deviance	AIC
- Motiv	1	2021.7	2027.7
+ Gužva	1	2019.4	2029.4
+ Aktivnost	1	2020.5	2030.5
- Ukupno zadovoljstvo	1	2027.2	2033.2
- Cijena	1	2036.3	2042.3

Tablica 2.6: Četvrti korak stepwise procedure za dobivene statistički značajne varijable analizom univarijatnih modela

	Df	Deviance	AIC
+ Motiv	1	2020.6	2028.6
+ Gužva	1	2021.2	2029.2
+ Aktivnost	1	2021.5	2029.5
- Ukupno zadovoljstvo	1	2027.2	2031.2
- Cijena	1	2036.3	2040.3

Tablica 2.7: Multivarijantna logistička regresija - model M_2

Varijable		procijenjeni koef.	standardna pogreška	z-vrijednost	p-vrijednost
slobodni koefcijent	β_0	-1.667	0.259	-6.440	$1.19e^{-10}$
Cijena	β_1	0.272	0.072	3.799	0.000145
Ukupno zadovoljstvo	β_2	0.179	0.077	2.341	0.19257

Procijenjene parametre modela interpretiramo na sljedeći način:

β_0: Za svako povećanje ocjene elementa Cijene od jedne jedinice, izglednost namjere ponovne posjete Parka raste $e^{0.27229} = 1.313$ puta uz fiksnu ocjenu elementa Ukupno zadovoljstvo.

β_2: Za svako povećanje ocjene elementa Ukupno zadovoljstvo od jedne jedinice, izglednost namjere ponovne posjete Parka raste $e^{0.17925} = 1.196$ puta, uz fiksnu ocjenu elementa Cijena.

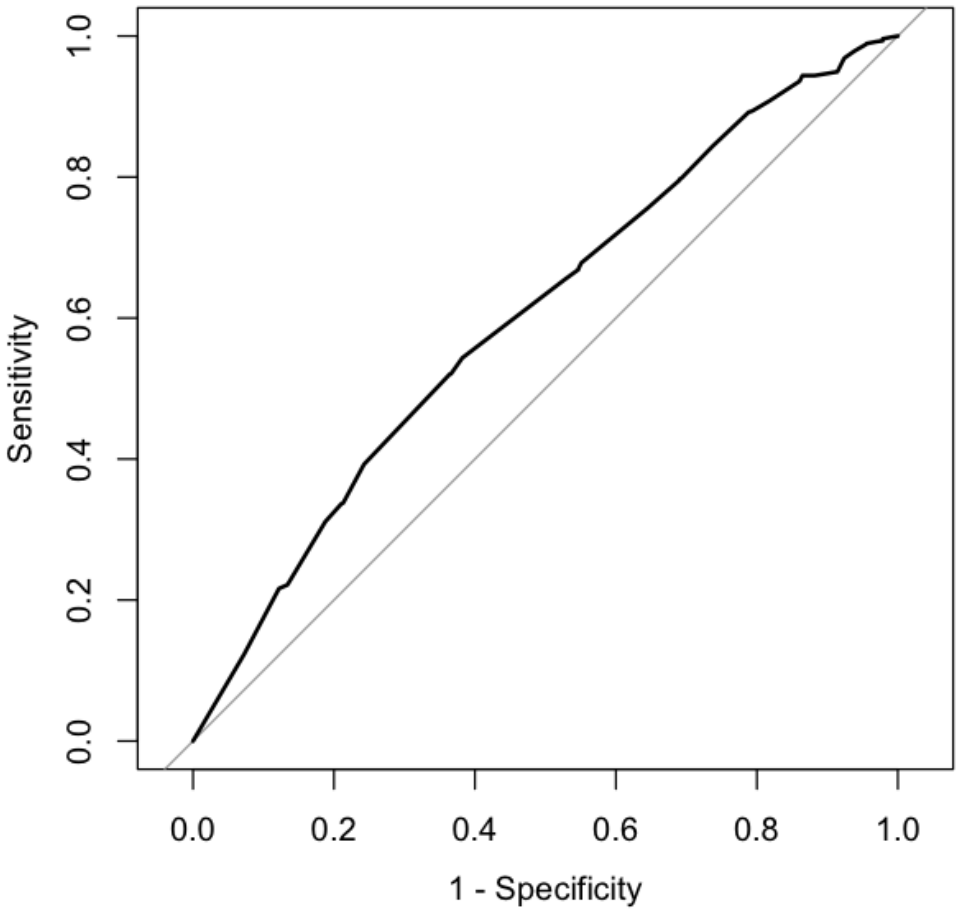
Prikažimo ocijene prilagodbe modela M_2 dobivene primjenom HL testa i na temelju ROC krivulje. Rezultat provedbe HL testa za model M_2 je prikazan u Tablica 2.8.

Tablica 2.8: Rezultat provedbe HL testa za model M_2

χ^2	7.321
df	8
p-vrijednost	0.5024

Iz izračunate p-vrijednosti prikazane u Tablica 2.8 na razini značajnosti od 5% zaključujemo da nema dovoljno dokaza da je prilagodba modela loša.

Na Slika 2.9 prikazana je ROC krivulja M_2 modela. Vrijednost AUC za model M_2 je 0.603 iz čega zaključujemo da vjernost posjetitelja Parka nije dovoljno modelirati samo sa ocjenama zadovoljstva.



Slika 2.9: Prikaz ROC krivulje modela M_2.

2.6 Modeli s kategorijalnom nezavisnom varijablom

U ovom poglavlju prikazat ćemo izgradnju logističkog modela s kategorijalnom varijablom. Budući da prethodno nije provedena posebna analiza kojom bismo reducirali skup prediktorskih varijabli vezanih uz zadovoljstvo multivarijatni logistički model M_2 izgrađen je prvo samo sa uključenim varijablama zadovoljstva. Provedimo prvo univarijatnu analizu, a zatim modelu M_2 nadodajmo kategorijalnu varijablu Zemlja porijekla.

2.6.1 Univarijatan model

Promotrimo univarijatan model logističke regresije $M_3 = g(x) = \beta_0 + \beta_1 x$ gdje je varijabla ishoda Namjera ponovne posjete Parka, a x je vrijednost elementa Zemlje porijekla. U Tablica 2.9 prikazani su rezultati procjena parametara, standardne pogreške, z-vrijednost i pripadne p-vrijednost.

Tablica 2.9: Univarijatna logistička regresija - model M_3

Varijable		procijenjeni koef.	standardna pogreška	z-vrijednost	p-vrijednost
slobodni koefcijent	β_0	-0.076	0.054	-1.404	0.16
Zemlja porijekla 1	β_1	1.038	0.209	4.974	6.57e^{-07}

U Tablica 2.9 varijabla imena "Zemlja porijekla 1" predstavlja kategoriju gdje je varijabla Zemlja porijekla jednaka 1, to jest predstavlja kategoriju domaćeg stanovništva. Iz izračunate p-vrijednosti prikazane u Tablici 2.9 ($6.57e^{-07}$) na razini značajnosti od 5% možemo zaključiti da je varijabla

Zemlja porijekla značajna. Interpretacija parametra β_1 modela M_3 glasi: Izglednost namjere ponovne posjete Parka domaćih posjetitelja je $e^{1.038} = 2.824$ puta veća od izlednosti namjere ponovne posjete Parka stranih posjetitelja.

2.6.2 Multivarijatni modeli

Nadodajmo sada multivarijantnom modelu M_2 kategorijalnu nezavisnu varijablu Zemlja porijekla uključujući i interakciju. Dakle, promatramo model

$$M_4 = \ln \frac{\pi(Y=1)}{1-\pi(Y=1)} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \beta_4 * x_1 * x_3 + \beta_5 * x_1 * x_3 \ ,$$

gdje $Y = 1$ označava da ispitanik ima namjeru ponovno posjetiti Park, x_1 označava vrijednosti varijable koja predstavlja element Cijena, x_2 označava vrijednost varijable koja predstavlja element Ukupno zadovoljstvo, x_3 označava vrijednost varijable koja predstavlja kategoriju domaćeg stanovništva.

Tablica 2.10: Multivarijantna logistička regresija - model M_4

Varijable		procijenjeni koef.	standardna pogreška	z-vrijednost	p-vrijednost
slobodni koefcijent	β_0	-1.887	0.275	-6.871	$6.37e^{-12}$
x_1	β_1	0.365	0.078	4.712	$2.46e^{-06}$
x_2	β_2	0.131	0.080	1.625	0.104
Zemlja porijekla 1	β_3	2.019	1.005	2.008	0.0446
x_1*Zemlja porijekla 1	β_4	-0.427	0.259	-1.652	1
x_2*Zemlja porijekla 1	β_5	0.123	0.329	0.374	0.708

Iz izračunatih p-vrijednosti (1 i 0.708) iz Tablica 2.10 na razini značajnosti od 5% možemo zaključiti da nema značajnih interakcija.

Promotrimo sada multivarijantan model s istim nezavisnim varijablama kao u modelu M_4 ali bez interakcija. Dakle, promatramo model

$$M_4 = \ln \frac{\pi(Y=1)}{1-\pi(Y=1)} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 \ ,$$

gdje su x_1, x_2 i x_3 označavaju vrijednosti istih varijabli kao u modelu M_4. U Tablica 2.6 prikazani su rezultati procjena parametara, standardna pogreška, z-vrijednost i p-vrijednost modela M_5.

Tablica 2.11: Multivarijantna logistička regresija - model M_5

Varijable		procijenjeni koef.	standardna pogreška	z-vrijednost	p-vrijednost
slobodni koefcijent	β_0	-1.777	0.262	-6.776	$1.24e^{-11}$
x_1	β_1	0.322	0.073	4.387	$1.15e^{-05}$
x_2	β_2	0.141	0.078	1.822	0.0685
Zemlja porijekla 1	β_3	1.133	0.215	5.268	$1.38e^{-07}$

Iz izračunatih p-vrijednosti prikazanih u Tablica 2.11 ($6.57e^{-07}$) možemo zaključiti da su varijable koje predstavljaju element Cijenu, Ukupno zadovoljstvo i Zemlju porijekla značajne. Interpretacija parametara modela je sljedeća:

β_1: Za povećanje ocjene elementa Cijena od jedne jedinice, izglednost Namjere ponovne posjete raste za $e^{0.322} = 1.38$ puta za ispitanike iste zemlje porijekla (domaće ili strano stanovništvo) uz fiksnu ocijenu elementa Ukupno zadovoljstvo.

β_2: Za povećanje ocjene elementa Ukupno zadovoljstvo od jedne jedinice, izglednost Namjere ponovne posjete raste za $e^{0.141} = 1.151$ puta

za ispitanike iste zemlje porijekla (domaće ili strano stanovništvo) uz fiksnu ocjenu elementa Cijena.

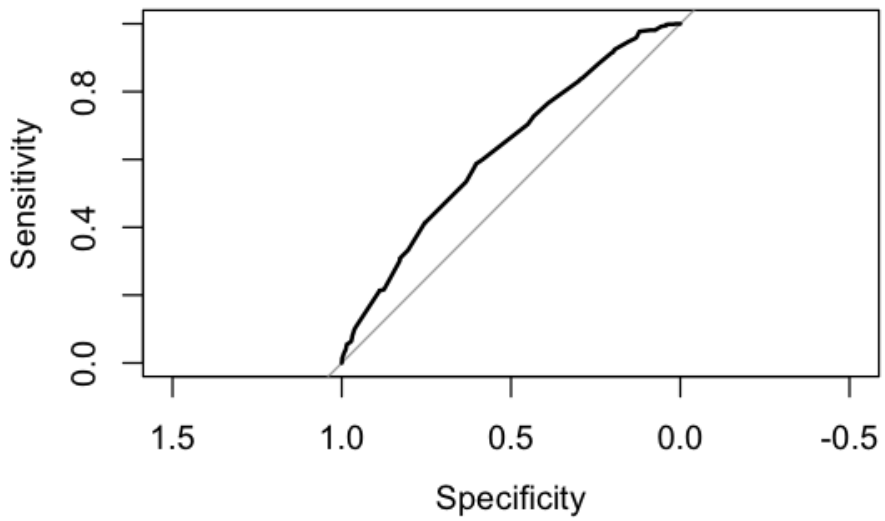
β_3: Izglednost namjere ponovne posjete za domaće stanovnike je $e^{1.133} = 3.105$ puta veća od izglednosti namjere ponovne posjete stranog stanovništva uz fiksne ocjene elemenata Cijena i Ukupno zadovoljstvo.

Prikažimo ocijene prilagodbe modela M_5 dobivene primjenom HL testa i na temelju ROC krivulje. Rezultat provedbe HL testa za model M_4 je sljedeći

Tablica 2.12: Rezultat provedbe HL testa za model M_4

χ^2	10.118
df	8
p-vrijednost	0.2569

Iz izračunate p-vrijednosti prikazane u Tablica 2.12 na razini značajnosti od 5% zaključujemo da nema dovoljno dokaza da je prilagodba modela loša.Na Slika 2.10 prikazana je ROC krivulja M_5 modela. Vrijednost AUC za model M_5 je 0.6256 što je malo bolje nego dobiveno u modelu M_2, no ponovno zaključujemo da vjernost posjetitelja Parka nije dovoljno modelirati samo sa ocjenama zadovoljstva i zemljom porijekla.



Slika 2.10: Prikaz ROC krivulje modela M_5.

Zaključak

Za modeliranje binarnih podataka najčešće se koristi logistički model upravo zbog svoje jednostavnosti i lake interpretacije. Logistička regresija koristi se u ispitivanju povezanosti dihotomne zavisne i jedne ili više eksplanatornih varijabli. Postupak je prilično sličan multivarijatnoj linearnoj regresiji s temljenom razlikom da je varijabla odziva binomna. Rezultat je utjecaj svake varijable na omjer izglednosti promatranog događaja od interesa. Najpopularnija metoda za izgradnju prediktivnih modela koji se bave vjernosti potrošača upravo je logistička regresija. U ovome radu prikazana je primjena logističke regresije u modeliranju vjernosti posjetitelja zaštićenih područja pri čemu su kao nezavisne varijable odabrane varijable zadovoljstva i varijabla vezana uz zemlju porijekla posjetitelja. No, ipak nakon provedene analize promatrajući vjernost kroz predviđanje namjere ponovne posjete Parku na temelju potrošačevog zadovoljstva i demografskim separiranjem, zaključili smo da vjernost posjetitelja ipak Parka nije dovoljno modelirati samo sa ocjenama zadovoljstva i zemljom porijekla.

Literatura

[1] David W. Hosmer, Stanley Lemeshow (2000) Applied Logistic Regression, John Wiley Sons, 2000.

[2] Rao, C.,R. (2009) Linear Statistical Inference and Its Application, Wiley Series in Probability and Statistics, ISSN 1940-6517

[3] Agresti, A. (2018) An Introduction to Categorical Data Analysis, Wiley Series in Probability and Statistics

[4] McCullagh, P., and Neider, J. A. (1989). Generalized Linear Models, Second Edition. Chapman Hall, London.

[5] Rao, C. R. (1973). Linear Statistical Inference and Its Application, Second Edition. Wiley, Inc., New York.

[6] Hauck, W. W., and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association, 72, 851-853.

[7] Perišić, A., Beljo, I., Livaić, T. (2019.) Struktura i stavovi posjetitelja Nacionalnog parka Krka, Veleučilište u Šibeniku

TEMELJNA DOKUMENTACIJSKA KARTICA

PRIRODOSLOVNO–MATEMATIČKI FAKULTET
SVEUČILIŠTA U SPLITU
ODJEL ZA MATEMATIKU

DIPLOMSKI RAD

PRIMJENA LOGISTIČKE REGRESIJE U MODELIRANJU VJERNOSTI POSJETITELJA

Mila Zovko

Sažetak:

Logička regresija analizira odnos izmedu dihotomne zavisne varijable i jedne ili više nezavisnih varijabli. Najčešće se koristi u prediktivnom modeliranju, gdje model procjenjuje vjerojatnost kategorija u klasifikacijskim problemima. U prvom dijelu rada koji se odnosi na teorijsku podlogu logističke regresije upoznajemo se sa familijom generaliziranih linearnih modela. Zatim, izvodimo pripadajući univarijatni i multivarijatni logistički model, metodom maksimalne vjerodostojnosti procijenjujemo njihove parametre, testiramo značajnost parametara te objašnjavamo interpretaciju modela. U drugom dijelu prikazujemo primjenu logističke regresije u modeliranju vjernosti posjetitelja gdje ispitujemo možemo li vjernost posjetitelja predvidjeti na temelju njihovog zadovoljstva, te u svrhu prikaza izgradnje modela sa različitim vrstama varijabli, vjernost modeliramo i uz jednu kategorijalnu varijablu koja se odnosi na zemlju porijekla posjetitelja.

Ključne riječi:

logistička regresija, vjernost posjetitelja, prediktivni model

Podatci o radu:

TEMELJNA DOKUMENTACIJSKA KARTICA

broj stranica 58, broj slika 12 i tablica 14, broj literaturnih navoda 7, jezik izvornik: hrvatski)

Mentorica: *doc. dr. sc., Tea Martinić Bilać*

Neposredna voditeljica: *dr. sc. Ana Perišić*

Članovi povjerenstva:

dr. sc. Ana Laštre, pred.

Povjerenstvo za diplomski rad je prihvatilo ovaj rad *23. rujna 2022.*

TEMELJNA DOKUMENTACIJSKA KARTICA

FACULTY OF SCIENCE, UNIVERSITY OF SPLIT
DEPARTMENT OF MATHEMATICS

MASTER'S THESIS

THE APPLICATION OF LOGISTIC REGRESSION IN MODELLING VISITOR LOYALTY

Mila Zovko

Abstract:

Logistic regression is used to examine the association of one dichotomous variable with one or more independent variables. It is extensively used in predictive modeling, where the model estimates probabilities for classification problems with two possible outcomes. The first part of the thesis refers to the theoretical foundation of logistic regression. In this part, generalised linear models are introduced, which is followed by the theoretical background of univariate and multivariate logistic regression modelling. We present the method of maximum likelihood parameter estimation, significance testing and explain the interpretation of the model. In the second part of the thesis, we present the application of logistic regression in visitor loyalty modeling. We perform logistic regression in order to examine whether visitor loyalty can be predicted by visitor satisfaction. Also, to present the model building with different types of variables, loyalty is modeled with one categorical variable referring to the country of origin.

Key words:

logistic regression, visitor loyalty, predictive model

Specifications:

TEMELJNA DOKUMENTACIJSKA KARTICA

58 pages, 12 figures, 14 tables, 7 references, original in: Croatian

Mentor: *assistant professor Tea Martinić Bilać*

Immediate mentor: *Ana Perišić, PhD*

Committee:

Ana Laštre, Phd, Lecturer

This thesis was approved by a Thesis commettee on *September 23, 2022.*