



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



Miloš Gavrilović

MULTINOMINALNI LOGIT REGRESIONI MODEL SA PRIMENOM U TURIZMU

-MASTER RAD-

Novi Sad, 2021.

Sadržaj

Uvod	2
1 Regresioni modeli	3
1.1 Promenljive	3
1.2 Ocenjivanje parametara.....	6
1.3 Jednostavni linearni regresioni model	8
1.3.1 Metod najmanjih kvadrata	9
1.3.2 Metoda maksimalne verodostojnosti	10
1.4. Višestruka linearna regresija.....	11
1.4.1 Ocenjivanje metodom najmanjih kvadrata.....	12
1.4.2 Ocenjivanje metodom maksimalne verodostojnosti	14
1.5 Odabir modela	15
2. Logistička (logit) regresija	18
2.1 Logit model	18
2.1.1 Ocenjivanje parametara.....	21
2.2. Multinomialna logit regresija.....	22
2.2.1 Ordinalna logistička regresija.....	22
2.2.2 Ocenjivanje parametara.....	23
2.2.3. Nominalna logistička regresija	24
2.3 Odabir modela	26
3 Primena multinomialne regresije u oblasti turizma	28
3.1. Mogućnost korišćenja multinomialnog logit modela za ispitivanje načina prevoza	28
3.2. Mogućnost korišćenja multinomialnog logit modela za izbor turističke destinacije.....	34
4 Zaključak.....	41
Literatura	42
Prilog: Kodovi u R-u za multinomialne regresione modele	44

Uvod

Korišćenje podataka i modeliranje predstavljaju preduslov za dobro zaključivanje i predviđanje. U gotovo svim oblastima delatnosti postoje uslovi za primenu ovakvog pristupa. U ovom radu akcenat je stavljen na primenu regresionih modela u jednoj izuzetno važnoj privrednoj grani, a to je turizam.

U prvom delu rada date su teorijske osnove statističke analize koje obuhvataju osnovne pojmove teorije ocenjivanja i prikaz linearnog regresionog modela za posebnim osvrtom na uključivanje kategorijalnih nezavisnih promenljivih.

U drugom delu predstavljani su generalizovani linearni modeli i to logistički (logit) model i multinominalni regresioni model. Osnovna karakteristika logit modela ogleda se u modeliranju šansi da se neki događaj realizuje, odnosno modelira se binarna zavisna promenljiva. Multinominalni regresioni model predstavlja dalje uopštavanje logit modela kada je zavisna promenljiva kategorijalna i kada skup njenih realizacija sadrži bar tri kategorije. Kod multinominalne logističke regresije postoje ordinalna i nominalna. Ordinalna logistička regresija je slučaj kada zavisna promenljiva ima određen redosled u dodeljivanju vrednosti, dok kod nominalne logističke regresije kategorije nemaju prirodni poredak.

Za razvoj turizma neophodno je dobro razumevanje potreba i navika turista. zato su istraživanja u ovoj oblasti postala neophodan uslov za donošenje strateških pravaca razvoja. U trećem delu predstavljena su dva primera primene multinominalnog modela u turizmu. U prvom primeru, odnosno modelu, kao zavisna promenljiva posmatran je odabir prevoznog sredstva, dok je u drugom primeru modelirano zadovoljstvo turista izabranom destinacijom kao zavisna promenljiva.

Što se tiče primene multinominalne logističke regresije u turizmu u Srbiji postoje određeni problemi. Dostupnost podataka je jedan od tih problema. Na sajtu Republičkog zavoda za statistiku (<https://www.stat.gov.rs/>) podaci koji su dostupni vezano za turizam su: Dolasci noćenja domaćih turista u aranžmanu domaćih i turističkih agencija, Dolasci i noćenja turista po regionima i oblastima - godišnji podaci, Dolasci i noćenja turista do decembra 2010. godine - mesečni podaci i Dolasci i noćenja turista po regionima. Na sajtovima turističkih organizacija (Turistička organizacija Srbije i Turistička organizacija Beograda) nema dostupnih podataka. Kao preporuka može se navesti da bi odgovarajuće institucije zadužene za turizam, počevši od Ministarstva trgovine, turizma i telekomunikacija, pa do pojedinačnih turističkih agencija, trebalo da ulože dodatni napor u korišćenju savremenih metoda u planiranju, kao što je primena složenih matematičkih modela zasnovanih na relevantnim podacima.

1 Regresioni modeli

Modeliranje realnih pojava je oduvek bilo neophodno za razumevanje samih pojava. Regresioni modeli su jedan podskup takvih modela i njihova osnovna ideja jeste da ponašanje, najčešće, jedne pojave (zavisne promenljive) objasne koristeći skup nezavisnih promenljivih.

Preciznije, regresiona analiza je skup statističkih procesa za procenu odnosa između *zavisne promenljive* (koja se često naziva „promenljiva ishoda“ ili „odgovora“, eng. 'outcome' or 'response') i jedne ili više *nezavisnih promenljivih* (često se nazivaju „prediktori“, „kovarijate“, „promenljive objašnjenja“ ili „karakteristike“, 'predictors', 'covariates', 'explanatory variables' or 'features'). Takođe se ravnopravno za pojam promenljive koriste i pojam obeležje i pojam varijable.

Regresiona analiza prvenstveno se koristi u dve različite svrhe. Prvo, regresiona analiza se široko koristi za predikciju. Drugo, u nekim se situacijama regresiona analiza može koristiti za zaključivanje uzročno -posledičnih veza između zavisne i nezavisnih promenljivih. Regresije same po sebi otkrivaju samo odnose između zavisne promenljive i skupa nezavisnih promenljivih u fiksnom skupu podataka. Da bi se regresije koristile za predikciju ili zaključivanje uzročno -posledičnih odnosa, potrebno je pažljivo obrazložiti zašto postojeći odnosi imaju moć predviđanja za novi kontekst ili zašto odnos između posmatranih promenljivih ima uzročno -posledično tumačenje.

Regresijski modeli uključuju sledeće komponente:

Nepoznate parametre (β),

Nezavisne promenljive (X_i),

Zavisnu promenljivu (Y_i),

Grešku, odnosno šum (ε_i),

i najčešće prikazuje zavisnu promenljivu u funkciji od nezavisne promenljive i parametra, gde ε_i predstavlja aditivnu grešku tj. opšti oblik regresionog modela je

$$Y_i = f(X_i, \beta) + \varepsilon_i.$$

Ključni deo modeliranja je odabir funkcionalne forme f i ocenjivanje parametara tako da dobijeni model ispuni svoju svrhu. Ovaj postupak je složen i zavisi od velikog broja faktora. Jedan od njih su karakteristike posmatranih promenljivih koji će ukratko biti prikazani u nastavku.

1.1 Promenljive

Osnovna i najjednostavnija podela promenljivih je na osnovu osobina skupa mogućih realizacija. Numeričke promenljive mogu da uzmu vrednosti iz nekog podskupa realnih brojeva, a atributivne promenljive vrednosti uzimaju iz skupa kategorija koje predstavljaju moguće realizacije. Numeričke promenljive se dalje dele na neprekidne i diskretne. Neprekidne promenljive su one promenljive koje mogu uzeti bilo koju vrednost iz nekog skupa realnih brojeva. Diskretna promenljiva može imati samo određene vrednosti na brojevnoj osi, odnosno samo vrednosti iz

konačnog ili prebrojivog skupa. Podela promenljivih na numeričke neprekidne, numeričke diskretne i atributivne, naziva se i podela prema sadržini.

Druga podela promenljivih vrši se na osnovu primenjene merne skale. Razlikovanje skale koja se koristi za merenje realizacije je jedan od najvažnijih zadataka u pravilnom odabiru analize za obradu.

Prilikom merenja najčešće se koriste četiri merne skale, a to su:

1. nominalna skala = skup kategorija (vrednosti, stanja) koje nisu uređene, odnosno rangirane. Kod nominalne skale koriste se imena za klasifikaciju. Ako se koriste brojevi, oni samo služe za kraći zapis, a nemaju svojstva brojeva kao što je uređenje
2. ordinalna skala = skup kategorija koje su uređene, odnosno rangirane, ali gde nije poznato rastojanje između kategorija i gde rastojanje između dve kategorije može biti različito od rastojanja drugog para kategorija. Ako se koriste brojevi oni označavaju rang, ali ne pokazuju veličinu njihovog odnosa.
3. intervalna skala = skup vrednosti je podskup skupa realnih brojeva i skala poseduje jedinicu mere, pa je rastojanje između svake dve susedne kategorije jednako.
4. skala odnosa je intervalna skala koja ima pravu (fizičku) nulu.

Promenljive čije realizacije su određene nominalnom ili ordinalnom skalom nazivaju se kvalitativne ili kategorijalne promenljive, dok se promenljive dobijene na intervalnoj skali ili skali odnosna, a koja mogu biti diskretna i neprekidna, nazivaju kvantitativne promenljive. Sa stanovišta modeliranja najvažnija razlika između kategorijalnih i kvantitativnih promenljivih jeste što kategorijalne uzimaju vrednosti koji nemaju pravo numeričko značenje i matematičke operacije nad njima su besmislene. Numeričke promenljive uzimaju vrednosti koji nastaju kao rezultat merenja ili prebrojavanja i matematičke operacije nad njima imaju smisla.

Jasno je da kategorijalne promenljive razdvajaju objekte posmatranja u jasno razgraničene grupe po određenoj karakteristici ili osobini. Na osnovu kardinalnosti skupa mogućih realizacija (kategorija) dele se na binarne i polinomne promenljive.

Binarne promenljive mogu da uzmu samo jednu od dve vrednosti, dok je kod polinomnih kardinalnost skupa kategorija tri ili više. Binarne promenljive se nazivaju i dihotomne. U oblasti regresione analize, zbog svog značenja, često se za binarne promenljive koristi pojam dummy (fiktivne, zamenske) promenljive. U regresionoj analizi polinomna promenljiva se prikazuje preko skupa binarnih/dummy promenljivih na sledeći način. Neka polinomna promenljiva C može da uzme m različitih realizacija tj. neka je skup kategorija $\{c_1, c_2, \dots, c_m\}$ tada definišemo $m - 1$ dummy promenljivu $\{X_1, X_2, \dots, X_{m-1}\}$ na sledeći način

$$\begin{aligned} X_1 &= \begin{cases} 1, & \text{ako je } C = c_2 \\ 0, & C \in \{c_1, \dots, c_m\} \setminus c_2 \end{cases} \\ X_2 &= \begin{cases} 1, & \text{ako je } C = c_3 \\ 0, & C \in \{c_1, \dots, c_m\} \setminus c_3 \end{cases} \\ &\vdots \end{aligned}$$

$$X_{m-1} = \begin{cases} 1, & \text{ako je } C = c_m \\ 0, & C \in \{c_1, \dots, c_{m-1}\} \end{cases}$$

Vrednost $C = c_1$ u ovom slučaju je izabrana za referentnu vrednost i njoj odgovaraju sledeće vrednosti uvedenih promenljivih $X_1 = X_2 = \dots = X_{m-1} = 0$.

Na ovaj način multinominalna promenljiva sa m različitih kategorija se uključuje u model koristeći $m - 1$ dummy promenljivih. Tabelarni zapis opisanog kodiranja je

kategorije promenljive C	dummy promenljive			
	X_1	X_2	...	X_{m-1}
c_1	0	0	...	0
c_2	1	0	...	0
c_3	0	1	...	0
...				
c_m	0	0	...	1

Tabela 1

U postupku modeliranja zbog razumevanja i tumačenja rezultata važan je odabir referentne vrednosti, odnosno odabir kategorije promenljive C za koju će sve pridružene dummy promenljive uzeti vrednosti 0.

Načini određivanje funkcionalne forme, kao i ocenjivanje parametara regresionih modela zavise od osobina, odnosno vrsta promenljivih koje ulaze u model. U tabeli 2 dat je prikaz kombinacija koje su od interesa u daljem radu.

zavisna promenljiva	nezavisna promenljiva	metod za analizu
neprekidna	neprekidna	regresija, korelacija
neprekidna	kategorijalna	regresija, ANOVA
binarna	kategorijalna	logit/probit , loglinear
binarna	neprekidna	logit/probit
multinominalna	kategorijalna	loglinear, multinomial logit
multinominalna	neprekidna	multinomial logit

Tabela 2

Pojave koje će biti modelirane u delu spadaju u grupu modela gde je zavisna promenljiva multinominalna, dok nezavisne promenljive mogu biti kategorijalne ili neprekidne. Iz tog razloga u nastavku će biti dat prikaz linearne regresije, a zatim njeno proširenje na logit i multinominalni logit model¹.

¹ Više detalja u Powers, D. A. , Xie, Y. (1999). *Statistical Methods for Categorical Data Analysis*, Academic Press

1.2 Ocenjivanje parametara

Kako je već napomenuto jedna od komponenti regresionih modela jesu nepoznati parametri. Iz tih razloga u prvom delu ovog poglavlja biće predestavljeni osnovni pojmovi teorije ocenjivanja parametara.

Prilikom ispitivanja obeležja Y , često su poznate neke informacije o raspodeli tog obeležja. Na primer, poznato je da obeležje ima normalnu raspodelu, ali nisu poznate vrednosti njegovog matematičkog očekivanja ili disperzije, koji su tada nepoznati parametri. Znači, poznato je da raspodela obeležja pripada nekoj familiji raspodela $\{f(y, \theta), y \in \mathbb{R}, \theta \in \Theta\}$, gde je Θ skup parametara, koji može biti jednodimenzionalan ili višedimenzionalan, kada je nepoznato nekoliko parametara. Ova familija raspodela se naziva dopustiva familija raspodela za obeležje Y , a skup Θ se naziva dopustiv skup parametara. Slično u opštem regresionom modelu nepoznati parametri određuju funkcionalnu formu između zavisne i nezavisne promenljive. U svakom slučaju postupak ocenjivanja nepoznatih parametara zasniva se na određivanju statistike koja obezbeđuje poželjne osobine i omogućuje izračunavanje nepoznatih parametara na osnovu statističkih podataka, odnosno uzorka (Lozanov-Crvenković, 2012).

Ako obeležje Y , ima dopustivu familiju raspodela $\{f(y, \theta), y \in \mathbb{R}, \theta \in \Theta\}$, i ako je $(Y_1, \dots, Y_n)$ prost slučajan uzorak obeležja Y , ocenjivanje parametara θ se vrši izborom neke statistike $\hat{\theta} = f(Y_1, \dots, Y_n)$, da bi se kao ocena nepoznatog parametara uzela realizovana vrednost ove statistike.

$\hat{\theta} = f(Y_1, \dots, Y_n)$ ćemo nazivati ocenjivač i on kao takav ima neku raspodelu. Izbor ocenjivača u velikoj meri zavisi od osobina koje obezbeđuje, a koje navodimo u nastavku.

Ocenjivač $\hat{\theta}$ je *centriran (nepristrasan)* ocenjivač parametara θ ako je

$$E(\hat{\theta}) = \theta.$$

Ocenjivač $\hat{\theta}$ je *asimptotski centriran (nepristrasan)* ocenjivač parametara θ ako je

$$\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta.$$

Ocenivač $\hat{\theta}$ je *necentriran (pristrasan)* ako nije $E(\hat{\theta}) = \theta$, a odstupanje od centriranosti se izražava kao $b(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta$. Veličina $b(\hat{\theta}, \theta)$ se naziva *mera pristrasnosti*.

Kvadratni koren iz varijanse $\sqrt{Var(\hat{\theta})}$ se naziva *standardna greška ocene*.

Neka su data dva centrirana ocenjivača $\hat{\theta}_1$ i $\hat{\theta}_2$ tada je $\hat{\theta}_1$ *efikasniji* ocenjivač od $\hat{\theta}_2$ ako važi

$$Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2).$$

Ocenjivač $\hat{\theta}$ je *postojan (stabilan, konzistentan)* ocenjivač parametara θ ako je za svako $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P\{|\hat{\theta} - \theta| \geq \varepsilon\} = 0.$$

Preciznost ocene se može meriti pomoću srednje kvadratne greške

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2.$$

Koristeći disperziju ocenjivača i njegovu pristrasnost, srednje kvadratno odstupanje se može prikazati na sledeći način

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + b^2(\hat{\theta}, \theta),$$

pa zaključujemo da je dovoljan uslov za konzistentnost kada postoji disperzija raspodele ocenjivača da važi $\lim_{n \rightarrow \infty} MSE(\hat{\theta}) = 0$, odnosno da se sa povećanjem uzorka smanjuje pristrasnost i varijansa i da obe teže 0.

U praksi se koristi nekoliko metoda za dobijanje ocenjivača. Najpoznatiji su metod najmanjih kvadrata i metod maksimalne verodostojnosti.

Metod najmanjih kvadrata je najjednostavniji postupak i zasniva se na problemu kvadratnog programiranja. Ocenjivač se dobija kao rešenje problema minimizacije $\min_{\theta} \sum_{i=1}^n (Y_i - Y_i(\theta))^2$. Tačnije, ideja ovog metoda zasniva se na određivanju ocene koja će minimizirati kvadratnu grešku. Ocenjivači dobijeni ovom metodom apriorno ne moraju da poseduju ni jednu od poželjnih osobina.

Sa druge strane, *metod maksimalne verodostojnosti* daje ocenjivač koji maksimizira funkciju verodostojnosti. Intuitivno, na ovaj način maksimiziramo saglasnost izabranog modela sa posmatranim podacima. Primene ovog metoda garantuje da je dobijeni ocenjivač asimptotski centriran i konzistentan. Međutim za primenu ovog metoda potrebno je poznavanje oblika raspodele, dok kod metoda najmanjih kvadrata to nije potrebno.

Da bismo formulisali metodu maksimalne verodostojnosti, potrebno je definisati funkciju verodostojnosti.

Definicija 1. Funkcija verodostojnosti za prost slučajni uzorak $(Y_1, \dots, Y_n)$ na osnovu realizovanog uzorka obima n je

$$L(\theta, Y_1, \dots, Y_n) = \begin{cases} p(Y_1, \theta) \dots p(Y_n, \theta) & \text{za diskretnu raspodelu} \\ \varphi(Y_1, \theta) \dots \varphi(Y_n, \theta) & \text{za neprekidnu raspodelu} \end{cases}$$

Definicija 2. Neka je

$$L(\theta) = L(\theta, Y_1, \dots, Y_n)$$

funkcija verodostojnosti za prost slučajni uzorak $(Y_1, \dots, Y_n)$ na osnovu realizovanog uzorka. Ako je vrednost $\hat{\theta} = u(Y_1, \dots, Y_n)$ u skupu θ i ako maksimizira vrednost funkcije verodostojnosti $L(\theta)$, odnosno ako je $\hat{\theta}$ rešenje problema

$$\max_{\theta \in \Theta} L(\theta)$$

tada je statistika $\hat{\theta} = u(Y_1, \dots, Y_n)$ ocenjivač nepoznatog parametra θ dobijena metodom maksimalne verodostojnosti.

U velikom broju slučajeva funkcija verodostojnosti zadovoljava uslove regularnosti, pa se njen maksimum može odrediti rešavanjem jednačine

$$\frac{\partial L(\theta)}{\partial \theta} = 0.$$

Takođe, kako je funkcija $L(\theta)$ pozitivna i kako je $\ln$ monotona transformacija, funkcije $L(\theta)$ i $\ln L(\theta)$ postižu maksimum za istu vrednost θ , često je. zbog oblika raspodele, lakše naći maksimum prirodnog logaritma funkcije verodostojnosti pa umesto prethodne jednačine rešavamo

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0.$$

1.3 Jednostavni linearni regresioni model

Jednostavni ili prost linearni regresioni model, kraće linearni regresioni model, koristi se za modeliranje linarnog odnosa između zavisne i nezavisne promenljive. To je najjednostavniji oblik stohastičkog odnosa između dve promenljive X i Y i model se izražava u obliku

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n \quad (1)$$

u kojem je Y zavisna promenljiva, X nezavisna promenljiva i ε šum, dok su β_0 i β_1 nepoznati parametri regresije. Parametar β_0 zovemo odsečak, dok je parametar β_1 nagib. Oznaka i se odnosi na i -to posmatranje, odnosno posmatramo uzorak

$$(Y_i, X_i), \quad i = 1, \dots, n.$$

Da bismo upotpunili specifikaciju regresijskog modela, dodajemo sledeće pretpostavke:

1. ε_i ima normalnu raspodelu (normalnost)
2. $E(\varepsilon_i) = 0$ (očekivanje)
3. $Var(\varepsilon_i) = \sigma^2$ (homoskedastičnost)
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ za $i \neq j$ (odsustvo autokorelacije)
5. X je nestohastička promenljiva sa fiksnim vrednostima takva da je, za bilo koji uzorak veličine n , $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \neq 0$. Kad $n \rightarrow \infty$ izraz, $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ je konačan, gde je $\bar{X} = \frac{\sum X_i}{n}$

Model (1) sa navedenih pet pretpostavki nazivamo **klasični normalni linearni regresioni model**.

Koristeći pretpostavke za šum, kao i odgovarajuće osobine za očekivanje, varijansu i normalnu raspodelu dobijamo sledeće karakteristike promenljive $Y_i, i = 1, \dots, n$

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

$$Var(Y_i) = E(Y_i - E(Y_i))^2 = \sigma^2$$

$$Cov(Y_i, Y_j) = 0 \text{ za } i \neq j$$

što daje

$$Y_i: \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2).$$

Klasični jednostavni linearni model sadrži tri nepoznata parametra β_0 , β_1 i σ^2 . Za njihovo ocenjivanje primenićemo dva najčešće korišćena metoda.

1.3.1 Metod najmanjih kvadrata

Počnimo prvo s izvođenjem ocenjivača parametara β_0 i β_1 metodom najmanjih kvadrata. Problem najmanjih kvadrata se svodi na $\min(\sum_{i=1}^n (Y_i - E(Y_i))^2)$. Neka je S funkcija koja je jednaka izrazu $\sum_{i=1}^n (Y_i - E(Y_i))^2$, tada u slučaju klasičnog linearnog modela dobijamo

$$\min_{\beta_0, \beta_1} S = \min_{\beta_0, \beta_1} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Potreban uslov za minimum daje sistem poznat pod nazivom sistem normalnih jednačina

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0 \quad (2)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) X_i = 0 \quad (3)$$

Rešavanjem prethodnog sistema dobijemo ocenjivače parametara $\hat{\beta}_0$ i $\hat{\beta}_1$

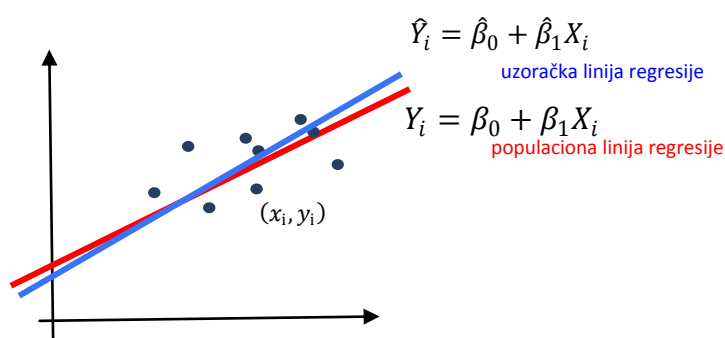
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad (4)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad (5)$$

i na ovaj način dobijamo regresionu liniju uzorka, slika 1,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (6)$$

koja prolazi kroz tačku $(\bar{X}, \bar{Y})$. Vrednost $\hat{\beta}_0$ meri odsečak, a vrednost $\hat{\beta}_1$ nagib regresijske linije.



Slika 1

U postupku najmanjih kvadrata iskorišćene su samo dve pretpostavke klasičnog modela i to pretpostavke 1. i 5. Sam postupak ne garantuje nijednu osobinu ocenjivača.

1.3.2 Metoda maksimalne verodostojnosti

Jedna od metoda ocenjivanja parametara je i metoda maksimalne verodostojnosti. Osnovna ideja ove metode je da se pomoću izabranog uzorka odaberu vrednosti nepoznatih parametara koje daju najveću verovatnoću da baš taj uzorak bude izabran.

U slučaju našeg regresijskog modela, uzorak sadrži posmatranja za n promenljivih $Y_1, \dots, Y_n$. Te promenljive su normalno distribuirane. One imaju sredine $(\alpha + \beta X_1), \dots, (\alpha + \beta X_n)$ i zajedničku varijansu σ^2 .

Pošto su Y_i normalno distribuirane sa sredinom $(\beta_0 + \beta_1 X_i)$ i varijansom σ^2 , na temelju formule za normalnu distribuciju imamo

$$\ln \varphi(\beta_0, \beta_1, \sigma^2; Y_i) = -\frac{1}{2} \ln 2\pi\sigma^2 - \frac{1}{2} \left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right)^2.$$

Sumiranjem po i dobijemo

$$\ln L(\beta_0, \beta_1, \sigma; Y_1, \dots, Y_n) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2.$$

U L se javljaju tri nepoznata parametara β_0, β_1, σ . Diferenciranjem funkcije L i izjednačavanjem sa nulom s obzirom na svaki parametar dobijamo sistem jednačina

$$\frac{\partial \ln L}{\partial \beta_0} = \frac{1}{2\sigma^2} \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 X_i) = 0,$$

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{1}{2\sigma^2} \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 X_i) X_i = 0,$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n 2(Y_i - \beta_0 - \beta_1 X_i)^2 = 0.$$

Iz prve dve jednačine dobijemo

$$\sum_{i=1}^n Y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^n X_i,$$

$$\sum_{i=1}^n X_i Y_i = \hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2.$$

što je sistem dat sa (2) i (3) u slučaju primene metoda najmanjih kvadrata. To povlači da su ocene parametara β_0 i β_1 u slučaju primene metoda maksimalne verodostojnosti jednake ocenama dobijenih metodom najmanjih kvadrata. Preostala treća jednačina daje ocenjivač σ^2 , koji glasi

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2,$$

Kako su ocenjivači $\hat{\beta}_0$ i $\hat{\beta}_1$ jednaki kao i u metodi najmanjih kvadrata, onda je $\hat{\sigma}^2$ jednak

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2 \quad (7)$$

gde su članovi e_i reziduali najmanjih kvadrata (odnosno $e_i = Y_i - \hat{Y}_i$).

1.4. Višestruka linearna regresija

Višestruki linearni regresioni model se koristi za proučavanje veza između zavisne promenljive i više nezavisnih promenljivih. Generička forma linearnog regresionog modela je

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK} + \varepsilon_i, \quad i = 1, \dots, n \quad (8)$$

u kojem je Y zavisna promenljiva, X_j za $j = 1, \dots, K$ nezavisne promenljive i ε šum tj. odstupanje dok su β_0 i β_j za $j = 1, \dots, K$ nepoznati parametri. Oznaka i se odnosi na i -to posmatranje. Broj nezavisnih promenljivih je K . Da bismo upotpunili specifikaciju regresijskog modela, dodajemo sledeće osnovne pretpostavke:

1. ε_i ima normalnu raspodelu
2. $E(\varepsilon_i) = 0$
3. $Var(\varepsilon_i) = \sigma^2$
4. $Cov(\varepsilon_i, \varepsilon_j) = 0$ za $i \neq j$
5. Nijedna nezavisna promenljiva nije slučajna. Svaka ima fiksne vrednosti za različite uzorke i takva je bez obzira na veličinu uzorka, $\frac{1}{n} \sum_{i=1}^n (X_{iK} - \bar{X}_K)^2 \neq 0$. Kad $n \rightarrow \infty$ izraz $\frac{1}{n} \sum_{i=1}^n (X_{iK} - \bar{X}_K)^2$ je konačan
6. $n > K$ tj. broj posmatranja je veći od broja koeficijenata koje treba oceniti
7. Nezavisne promenljive su linearno nezavisne.

Vidimo da su pretpostavke na šum, prve četiri pretpostavke, identične kao u slučaju jednostruke regresije, dok je peta pretpostavka proširena na svih K nezavisnih promenljivih. Šesta i sedma pretpostavka su nove u višestrukoj linearnoj regresiji. Kada kombinujemo (8) sa ovih sedam pretpostavki dobijemo klasični normalni višestruki linearni regresioni model.

Iz pretpostavki dobijamo da je očekivana vrednost promenljive Y_i

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK},$$

dok je njena varijansa

$$Var(Y_i) = \sigma^2,$$

što daje

$$Y_i: \mathcal{N}(\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}, \sigma^2).$$

Opis klasičnog normalnog višestrukog linearnog regresionog modela se obično prikazuje matričnom formom.

$$Y = X\beta + \varepsilon,$$

gde su

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad Y_i \in \mathbb{R}$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1K} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nK} \end{bmatrix}, \quad X_{ij} \in \mathbb{R}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}, \quad \beta_i \in \mathbb{R}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon_i \in \mathbb{R}$$

Značenje, odnosno tumačenje koeficijenata regresionog modela, zavisi od vrste nezavisne promenljive. Ako je nezavisna promenljiva X_j neprekidna, tada koeficijent β_j predstavlja vrednost za koliko će se promeniti zavisna promenljiva Y kada se X_j poveća za jednu jedinicu, a ostale nezavisne promenljive ostanu konstante, odnosno

$$\beta_j = \frac{\partial Y}{\partial X_j}.$$

Ako je X_j kategorijalna slučajna promenljiva, specijalno ako je binarna, tada koeficijent β_j predstavlja za koliko se razlikuje vrednost zavisne promenljive vrednost Y u dve posmatrane kategorije ($X_j = 0$, $X_j = 1$). Ako promenljiva X_j ima m kategorija, onda uvodimo $m - 1$ dummy (binarnu) promenljivu kao što je opisano u tabeli, i tumačenje koeficijenata se posmatra istovremeno za svih $m - 1$ promenljivih².

1.4.1 Ocenjivanje metodom najmanjih kvadrata

Postupak metode najmanjih kvadrata se sada svodi na određivanje minimuma funkcije po $K + 1$ promenljivoj. Tačnije, funkcija cilja je

$$S(\beta_0, \beta_1, \dots, \beta_K) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_K X_{iK})^2,$$

pa rešavamo problem

² Više detalja u McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. 2nd Ed. John Wiley & Sons, Inc

Greene, W. H. (2018), *Econometric Analysis*, 8th Edition, Pearson, New York kao i Hill, R. C., Griffiths, W. E., Lim., G. C. (2018), *Principles of econometrics*, 5th Edition, John Wiley & Sons

$$\min_{\beta_0, \beta_1, \dots, \beta_K} S(\beta_0, \beta_1, \dots, \beta_K).$$

Diferenciranjem S po nepoznatim parametrima $\beta_0, \beta_1, \dots, \beta_K$ i izjednačavanjem sa nulom dobijamo sistem linearnih jednačina

$$\begin{aligned}\frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^n Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_K X_{iK} = 0, \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_K X_{iK}) X_{i1} = 0, \\ &\vdots \\ \frac{\partial S}{\partial \beta_K} &= -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_K X_{iK}) X_{iK} = 0.\end{aligned}$$

Transformacijama prethodnog sistema dobijemo ekvivalentni sistem koji nazivamo sistem normalnih jednačina čiji je oblik

$$\begin{aligned}\sum_{i=1}^n Y_i &= \beta_0 n + \beta_1 \sum_{i=1}^n X_{i1} + \dots + \beta_K \sum_{i=1}^n X_{iK}, \\ \sum_{i=1}^n Y_i X_{i1} &= \beta_0 \sum_{i=1}^n X_{i1} + \beta_1 \sum_{i=1}^n X_{i1}^2 + \beta_2 \sum_{i=1}^n X_{i1} X_{i2} + \dots + \beta_K \sum_{i=1}^n X_{i1} X_{iK} \\ &\vdots \\ \sum_{i=1}^n Y_i X_{iK} &= \beta_0 \sum_{i=1}^n X_{iK} + \beta_1 \sum_{i=1}^n X_{i1} X_{iK} + \dots + \beta_K \sum_{i=1}^n X_{iK}^2\end{aligned}$$

odnosno dobijamo linearni sistem $K + 1$ jednačine sa $K + 1$ nepoznatom, čija rešenja predstavljaju ocenjivače nepoznatih parametara regresije.

Koristeći matrični zapis imamo

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n Y_i X_{i1} \\ \vdots \\ \sum_{i=1}^n Y_i X_{iK} \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \dots & \sum_{i=1}^n X_{iK} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \dots & \sum_{i=1}^n X_{i1}X_{iK} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{iK} & \sum_{i=1}^n X_{i1}X_{iK} & \dots & \sum_{i=1}^n X_{iK}^2 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.$$

što daje zapis za sistem normalnih jednačina.

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\boldsymbol{\beta},$$

Rešenje prethodnog sistema $\boldsymbol{\beta}$ sada jednostavno postaje

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}).$$

Poslednja sedma pretpostavka obezbeđuje da matrica $\mathbf{X}'\mathbf{X}$ bude regularna, odnosno da uvek postoji jedinstveno rešenje posmatranog sistema.

1.4.2 Ocenjivanje metodom maksimalne verodostojnosti

Kod metode maksimalne verodostojnosti nema suštinske razlike kada se radi bilo u jednostrukoj bilo u višestrukoj regresiji. Kada funkciju verodostojnosti izrazimo u logaritamskom obliku kao kod jednostruke regresije, i iskoristimo da je u ovom slučaju $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}$, dobijemo:

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) &= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_K X_{iK})^2 \\ &= -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0. \end{aligned}$$

Ocenjivači metode maksimalne verodostojnosti se dobiju diferenciranjem funkcije L s obzirom na $\beta_0, \beta_1, \dots, \beta_K$ i σ^2 i izjednačavanje svakog parcijalnog izvoda sa nulom. Lako se uočava da je prvih $K + 1$ jednačina u tako određenom sistemu u potpunosti jednako normalnim jednačinama metode najmanjih kvadrata. To znači da su maksimalno verodostojni ocenjivači koeficijenata regresije jednaki ocenjivačima metode najmanjih kvadrata. Maksimalno verodostojan ocenjivač od σ^2 je:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_K X_{iK})^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

ili

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2.$$

Ponovo zaključujemo da kada su ispunjene sve klasične pretpostavke ne postoji razlika u dobijenim ocenjivačima primenom ove dve metode. Međutim, u slučaju narušavanja pretpostavki, kao što je npr. heteroskedastičnost, metod najmanjih kvadrata više nije odgovarajući jer po svojoj konstrukciji ne uključuje pretpostavke. Iz tih razloga je korišćenje metoda maksimalne verodostojnosti daleko zastupljenije u postupcima ocenjivanja parametara u statističkim modelima (Kmenta, 1971).

1.5 Odabir modela

Kada su koeficijenti modela ocenjeni, postavlja se pitanje koliko je dati model dobar. To možemo da posmatramo iz više uglova.

Ako nas, pre svega, interesuje povezanost ili uticaj nezavisne promenljive β_j na zavisnu promenljivu Y onda je od interesa testiranje nulte hipoteze $H(\beta_j = 0)$ protiv jednostrane ili dvostrane alternativne hipoteze.

Sa druge strane, u nekim slučajevima nam je od primarne važnosti da imamo model koji će dobro predviđati zavisnu promenljivu pomoću skupa nezavisnih promenljivih. U ovom slučaju koristimo poređenje različitih modela i biramo onaj koji najbolje opisuje posmatranu pojavu. U nastavku ćemo prikazati postupke za izbor modela, tj. postupke ili kriteririjume koji porede različite modele za istu zavisnu promenljivu.

Svakako jedan od najzastupljenijih postupaka zasniva se na vrednosti prilagođenog koeficijenta determinacije $\bar{R}^2$ definisan sa

$$\bar{R}^2 = 1 - \frac{SSE/(n-K)}{SST/(n-1)},$$

gde je

$$SST = \sum (Y_i - \bar{Y})^2$$

ukupna suma kvadrata,

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

regresiona suma kvadrata, i

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

suma kvadrata greške, za koje važi

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2, \text{ odnosno } SST = SSR + SSE.$$

Posmatrajmo dva modela nad istim uzorkom. Neka je u prvom modelu uključeno K_1 nezavisnih promenljivih, a u drugom $K_1 + K_2$. Za oba modela možemo odrediti prilagođene koeficijent determinacije,

$$\overline{R_I^2} = 1 - \frac{SSE_I/(n - K_1)}{SST/(n - 1)} \quad \text{i} \quad \overline{R_{II}^2} = 1 - \frac{SSE_{II}/(n - K_1 - K_2)}{SST/(n - 1)}.$$

Kriterijum $\overline{R^2}$ tada dovodi do izbora *Modela I* ako je

$$\overline{R_I^2} \geq \overline{R_{II}^2};$$

u protivnom izbor je *Model II*.

Drugi pristup poređenja modela zasniva se na proceni greške predviđanja. Najpoznatiji kriterijum iz ovog skupa je Akaike informacioni kriterijum (AIC). AIC procenjuje kvalitet svakog modela koji može da se formira na osnovu datog skupa podataka u odnosu na svaki drugi model i računa se na sledeći način

$$AIC = 2 * K - 2 * \ln(\hat{L})$$

gde je K broj ocenjenih parametara, dok je $\hat{L}$ vrednost funkcije verodostojnosti za dati model. Kako bolji model ima veću vrednost funkcije verodostojnosti, to zaključujemo da vrednost AIC statistike treba da bude što manje. Primenom AIC kriterijuma biramo model sa najmanjom vrednošću statistike AIC.

Još jedan kriterijum iz ove grupe Bejzov informacioni kriterijum BIC (Bayesian Information Criteria) koji je dat sa

$$BIC = K * \ln(n) - 2 * \ln(\hat{L})$$

gde je K broj ocenjenih parametara, $\hat{L}$ je vrednost funkcija verodostojnosti, a n je obim uzoraka. Analogno kao kod AIC kriterijuma, prilikom odabira između nekoliko modela, preferira se onaj sa najnižim BIC -om. Često se u praksi posmatra razlika u vrednosti između dva modela koji označavamo sa ΔBIC . Ako je ΔBIC između 0 i 2, onda je razlika između dva modela mala, ako je ΔBIC između 2 i 6 onda postoji značajna razlika između dva posmatrana modela, ako je ΔBIC između 6 i 10 onda je razlika jak, ako je ΔBIC veći od 10, onda je razlika veoma jaka.

Poslednji kriterijum koji ćemo predstaviti, a koji se koristi prilikom izbora modela zasniva se na razlici funkcije verodostojnosti kada je model pun, tj. kada su uključene sve nezavisne promenljive, i funkcije verodostojnosti za redukovani model, odnosno kada je uključen deo nezavisnih promenljivih. Kada je razlika ove dve vrednosti relativno mala smatramo da je redukovani sistem dovoljno dobar. Ovaj test naziva se Likelihood ratio test tj. test odnosa verodostojnosti.

Vrednost testa se računa na sledeći način

$$-2(\ln L_{reduced} - \ln L_{full})$$

gde je $\ln L_{reduced}$ maksimalna vrednost funkcije verodostojnosti za redukovan model i $\ln L_{full}$ maksimum funkcije verodostojnosti za potpun model ³.

³ Detaljnije videti na Hosmer, D.W. & Lemeshow, S. (1989), *Applied Logistic Regression*, John Wiley & Sons

2. Logistička (logit) regresija

U ovom delu biće predstavljeno jedno proširenje lineranih regresionih modela na slučaj kada je zavisna promenljiva binarna ili multinominlna. Razlozi za popularnost logističkog modela su veliki. jer njegova primena omogućuje modeliranje šansi da se neki događaj realizuje, što je često predmet posmatranja i analize u medicini, inženjerstvu i drugim oblastima.

2.1 Logit model

Često je u centru istraživanja, odnosno modeliranja, neka pojava čiju verovatnoću želimo da procenimo. Na primer, želimo da procenimo uspešnost lečenja novim tretmanom, pa nam je promenljiva od interesa da li je došlo do izlečenja ili možemo da posmatramo uspešnost studiranja, i tada nam je promenljiva od interesa da li je student uspešno ili ne završio studije. Veliki je broj primera gde je promenljiva od interesa binarna, tj. gde ima samo dve vrednosti uspeh i neuspeh, dok skup nezavisnih promenljivih, odnosno promenljivih koje objašnjavaju uspeh ili neuspeh, mogu biti različite. U nastavku dajemo definiciju logističkog, ili skraćeno logit, modela ⁴.

Da bismo definisali logit model, polazimo od logističke funkcije koja glasi

$$f(z) = \frac{1}{1+e^{-z}}, \quad z \in \mathbb{R} \quad (9)$$

koja ima nekoliko važnih osobina:

$$\lim_{z \rightarrow +\infty} f(z) = 1,$$

$$\lim_{z \rightarrow -\infty} f(z) = 0,$$

$$f(z) \in [0,1]$$

$$f'(z) = \frac{e^{-z}}{(1+e^{-z})^2} > 0, \text{ tj. funkcija je striktno rastuća}$$

Iz navedenih osobina proizilazi da je logistička funkcija podobna da opiše odnosno modelira verovatnoću nekog događaja.

Neka je Y binarna slučajna promenljiva. Želimo da procenimo uticaj K nezavisnih promenljivih na verovatnoću da se događaj realizuje odnosno događaja $Y = 1$. Tačnije modeliramo sledeću uslovnu verovatnoću

$$P(Y_i = 1 | X_{i1}, \dots, X_{iK}) \quad (10)$$

Ako nezavisne promenljive posmatramo kao lineranu kombinaciju dobijamo

$$z = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK} \quad (11)$$

Kombinujući (9) i (1), i koristeći da je modeliramo uslovnu verovatnoću (10) dobijamo logit regresioni model

⁴ Više detalja u Kleinbaum, D. G., Klein, M. (2010). *Logistic Regression A Self-Learning Text*, Springer

$$P(Y_i = 1|X_{i1}, \dots, X_{iK}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^K \beta_j X_{ij})}} \quad (12)$$

Parametri $\beta_0, \beta_1, \dots, \beta_K$ su nepoznati u modelu i treba ih oceniti na osnovu uzorka. Matrica uzorka za nezavisne promenljive je X i ista kao u slučaju višestruke linearne regresije, dok vektor realizacija zavisne promenljive Y sadrži 0 i 1, u zavisnosti da li se događaj realizovao ili ne.

Radi lakše notacije verovatnoću $P(Y_i = 1|X_{i1}, \dots, X_{iK})$ ćemo označiti sa p_i . Ako iz modela

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^K \beta_j X_{ij})}}$$

izrazimo $\beta_0 + \sum_{j=1}^K \beta_j X_{ij}$ dobijamo

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^K \beta_j X_{ij} \quad (13)$$

Uvodimo oznaku

$$\text{logit } p_i = \ln\left(\frac{p_i}{1-p_i}\right),$$

pa dobijamo zapis logističkog modela

$$\text{logit } p_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK} \quad (14)$$

Da bismo mogli da izvršimo ocenjivanje parametara modela (14) neophodno je da na osnovu uzorka izračunamo realizovane vrednosti verovatnoće p_i . Ovu verovatnoću dobijamo primenom ocenjivača za nepoznatu proporciju kod binomne raspodele. Neka u uzorku M elemanata gde su vrednosti nezavisne promenljive identične i iznose $x_{i1}, \dots, x_{iK}$. Neka je kod tih M elemenata uzorka vrednost odgovarajuće zavisne promenljive bila jednaka 1 u $m < M$ elemanata, tada je realizovana verovatnoća $p_i = \frac{m}{M}$.

Pretpostavke za logit model su:

1. $Y_1, \dots, Y_n$ su nezavisni.
2. Greške moraju biti nezavisne, ali ne i normalno distribuirane.

Logit model je izuzetno prikladan kada treba da se opišu kvote/šanse za uspeh ili drugi suštinski ishod, ili šanse za uspeh sa kojima se suočava jedna grupa u odnosu na drugu. Šanse su definisane kao količnik verovatnoće jednog ishoda u odnosu na drugi. Za primer, neka je p_i verovatnoća uspeha i $1 - p_i$ verovatnoća neuspeha. Kvota uspeha je

$$\omega = \frac{p_i}{1-p_i}.$$

Dalje ω ćemo nazivati *odds*.

Na primer, ako je verovatnoća uspeha jednaka $p_i = 0.25$, onda je $1 - p_i = 0.75$. Pa $odds = 0.25/0.75$. To znači da su šanse za uspeh 1 prema 3 ili 3 prema 1 da neće uspeti. U stvari kada je $odds > 1$, znači da su veće šanse za uspeh, dok za $odds < 1$, veće su šanse za neuspeh. Ako se dve grupe suočavaju sa istim šansama za uspeh, onda je odnos kvota jednak 1.

Ako iz relacije

$$\frac{p_i}{1 - p_i} = odds$$

izrazimo p_i dobijamo

$$p_i = \frac{odds}{odds + 1}. \quad (15)$$

Iz relacije (13) dobijamo

$$\ln(odds) = \beta_0 + \sum_{j=1}^K \beta_j X_{ij},$$

odnosno

$$odds = e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}},$$

pa konačno iz (15) dobijamo

$$p_i = \frac{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}}}.$$

Što se tiče interpretacije, e^{β_0} su šanse da je karakteristika prisutna, odnosno da se događaj realizovao, kada su svi $X_{ij} = 0$, odnosno kada su sve nezavisne promenljive uzele vrednost 0. U tom slučaju p_i jeste $e^{\beta_0} / (1 + e^{\beta_0})$.

Kada je X_{ik} numerička promenljiva i onda X_{ik} povećamo za jedinicu, množe se izgledi da je karakteristika prisutna, ovo je slično linearnoj regresiji, ali umesto adaptivne promene to je multiplikativna promena stope. Uzmimo da je $X_{ik} = const$, gde je X_{ik} k -ta promenljiva. sada treba da vidimo kakve su šanse grupe kojoj je došlo do povećanja jedinice u odnosu na onu gde nije došlo do povećanja jedinice.

$$\frac{odds_1}{odds_2} = \frac{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij} + \beta_k(const+1)}}{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij} + \beta_k const}} = e^{\beta_k}$$

e^{β_k} su šanse da ishod vezan za prvu grupu bude uspešan ako joj se poveća jedinica u odnosu na drugu grupu u kojoj nije došlo do povećanje te nezavisne promenljive.

Kada je nezavisna promenljiva binarna (odnosno uzima 0 ili 1), tumačenje se obavlja na sledeći način. Uzmimo da je u prvoj grupi $X_{ik} = 1$, dok u drugoj grupi $X_{ik} = 0$.

$$\frac{odds_1}{odds_2} = \frac{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij} + \beta_k}}{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}}} = e^{\beta_k}$$

e^{β_k} su šanse da prva grupa uspe ako je $X_{ik} = 1$ u odnosu na drugu grupu u kojoj je $X_{ik} = 0$. Ako je $e^{\beta_k} > 1$ tj. $\beta_k > 0$, tada je su u grupi $X_{ik} = 1$ veće šanse za $(e^{\beta_k} - 1) \cdot 100\%$. Obrnuto, ako je $e^{\beta_k} < 1$ tj. $\beta_k < 0$ šanse u prvoj grupi su manje za $(1 - e^{\beta_k}) \cdot 100\%$.

Ako je nezavisna promenljiva kategorijalna (dummy), onda odaberemo jednu kategoriju koja je referentna među m kategorija i deninišemo $m - 1$ dummy promenljivu kao u tabeli 1. Za referentnu vrednost, kao što je to ranije navedeno, svih $m - 1$ dummy promenljivih uzima vrednost 0. Dobijeni koeficijenti kod $m - 1$ novodefinisane binarne promenljive predstavljaju odnos šansi posmatrane kategorije u odnosu na referentnu.

2.1.1 Ocenjivanje parametara

Kod logit modela funkcija verodostojnosti je

$$L(\boldsymbol{\beta}, \mathbf{X}) = \prod_i c_i p_i^{Y_i} (1 - p_i)^{n_i - Y_i}. \quad (16)$$

gde n_i predstavlja broj elemenata uzorka kod kojih su sve vrednosti nezavisnih promenljivih jednake, Y_i predstavlja broj realizacija posmatranog događaja, odnosno broj uspeha, u posmatranih n_i slučajeva i c_i je konstanta.

Koristeći da je

$$p_i = \frac{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}}} \text{ i } 1 - p_i = \frac{1}{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}} + 1}$$

iz (16) dobijamo

$$L(\boldsymbol{\beta}, \mathbf{X}) = \prod_i c_i \left(\frac{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}}}{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}} + 1} \right)^{Y_i} \left(\frac{1}{e^{\beta_0 + \sum_{j=1}^K \beta_j X_{ij}} + 1} \right)^{n_i - Y_i},$$

odnosno funkcija verodostojnosti postaje

$$L(\boldsymbol{\beta}, \mathbf{X}) = \prod_i c_i \frac{e^{Y_i(\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK})}}{(1 + e^{(\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK})})^{n_i}}. \quad (17)$$

Kako je logaritmovanje monotona transformacija, maksimum funkcije verodostojnosti se neće promeniti, pa za određivanje ekstremne vrednosti posmatramo funkciju

$$\ln L(\boldsymbol{\beta}, \mathbf{X}) = \sum_i (\ln c_i + Y_i \ln(\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}) - (n_i) \ln(1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}})). \quad (18)$$

Pošto treba da odredimo maksimum funkcije po parametrima β_k , $k \in 0, 1, \dots, K$, određujemo stacionarne tačke kao rešenje sistema $\frac{\partial \ln L(\boldsymbol{\beta}, \mathbf{X})}{\partial \beta_k} = 0$, $k = 0, 1, 2, \dots, K$ odnosno dobijamo sistem

$$\begin{aligned} \frac{\partial \ln L(\boldsymbol{\beta}, \mathbf{X})}{\partial \beta_0} &= \sum_i \frac{Y_i}{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}} - n_i \frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}}} = 0 \\ \frac{\partial \ln L(\boldsymbol{\beta}, \mathbf{X})}{\partial \beta_k} &= \sum_i \frac{Y_i X_{ik}}{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}} - n_i \frac{e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}} X_{ik}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \dots + \beta_K X_{iK}}} = 0, k = 1, 2, \dots, K \end{aligned}$$

Prethodni sistem sa $(K + 1)$ nepoznatom je sistem nelinearnih jednačina pa se za njegovo rešenje koristi neki od iterativnih postupaka. Najčešće se primenjuje Njutn-Rafsonov postupak ili iterativno ponderisani najmanji kvadrati (Pregibon, 1981).

2.2. Multinomialna logit regresija

Do sada smo razmatrali samo one situacije koje su uključivale samo dva moguća ishoda tako da je zavisna promenljiva bila binarna. Sada ćemo razmatrati situacije koje uključuju više od dva moguća ishoda, odnosno kada zavisna promenljiva Y može da uzme J različitih vrednosti. regresioni modeli koji služe za modeliranje ovakve zavisne promenljive nazivaju se multinomialni logistički regresioni modeli ili multinomialni logit regresioni modeli, ili skraćeno multinomialni regresioni modeli.

U zavisnosti od karakteristika skupa J mogućih vrednosti (kategorija) promenljive Y razlikujemo dva slučaja. Prvi slučaj podrazumeva da zavisna promenljiva ima 3 ili više kategorija koje imaju prirodni poredak i tada odgovarajući model nazivamo ordinalna logistička regresija, dok u drugom slučaju, kada među kategorijama ne postoji prirodni poredak, model nazivamo nominalni logistički (logit) model.

Očigledno je da je razlika između logit modela i multinomialnog logit model u tome što je zavisna promenljiva kod logit modela je binarna (ima vrednosti 0 ili 1), dok kod multinomialnog logit modela zavisna promenljiva ima više mogućih kategorija. Ipak, multinomialni logit modeli se svode na logit model, što će biti prikazano u nastavku.

2.2.1 Ordinalna logistička regresija

Neka zavisna promenljiva Y ima J kategorija ($Y = 0, 1, \dots, J - 1$), onda su tu $J - 1$ načina da dihotomizujemo kategorije: ($Y \geq 1$ prema $Y < 1$; $Y \geq 2$ prema $Y < 2$, ..., $Y \geq J - 1$ prema $Y < J - 1$). Posmatrajmo dva događaja $Y \geq j$ i $Y < j$, tada je važi

$$P(Y < j) + P(Y \geq j) = 1,$$

odnosno za događaj $Y \geq j$ je događaj $Y < j$ suprotan, pa problem svodimo na posmatranje jedne nove binarne promenljive. Sada možemo da primenimo teoriju logističke regresije i da posmatrani odnos šansi za ova dva događaja,

$$\text{odds}(Y \geq j) = \frac{P(Y \geq j)}{P(Y < j)}, j = 0, 1, \dots, J - 1.$$

Dalje neka je

$$P(Y_{ij} \geq j | X_{i1}, \dots, X_{iK}),$$

za svako $i \in \{1, \dots, n\}$, $j \in \{0, 1, \dots, J - 1\}$ gde je n obim uzorka.

sada za svaku moguću realizaciju j zavisne promenljive Y možemo definisati odgovarajući logistički model

$$P(Y_{ij} \geq j | X_{i1}, \dots, X_{iK}) = \frac{1}{1 + e^{-(\beta_{j,0} + \sum_{k=1}^K \beta_{jk} X_{ik})}},$$

gde je $i \in \{1, \dots, n\}$, a $j \in \{0, \dots, J-1\}$

Verovatnoća $p_{ij} = P(Y_{ij} \geq j | X_{i1}, \dots, X_{iK})$ znači da je verovatnoća da i -ti element uzorka bira j -tu mogućnost koja je veća ili jednaka j . Pa samim tim verovatnoća da i -ti element uzorka bira j -tu mogućnost koja je manja od j zapravo $1 - p_{ij}$, ili

$$P(Y_{ij} < j | X_{i1}, \dots, X_{iK}) = 1 - P(Y_{ij} \geq j | X_{i1}, \dots, X_{iK})$$

Što se tiče šansi, dobijemo da je

$$odds = \frac{P(Y_{ij} \geq j | X_{i1}, \dots, X_{iK})}{P(Y_{ij} < j | X_{i1}, \dots, X_{iK})} = \frac{1}{\frac{1 + e^{-(\beta_{j,0} + \sum_{k=1}^K \beta_k X_{ik})}}{e^{-(\beta_{j,0} + \sum_{k=1}^K \beta_k X_{ik})}}} = e^{\beta_{j,0} + \sum_{k=1}^K \beta_k X_{ik}}$$

Ako je X_{ik} binarna, onda su šanse jednake e^{β_k} .

Kakva je interpretacija koeficijenata $\beta_{j,0}$. Koeficijenti $\beta_{j,0}$ je kad sve nezavisne promenljive imaju vrednost jednaka nuli ($X_{i1} = \dots = X_{iK} = 0$) i $\beta_{j,0}$ su ln izgledi za $Y \geq j$. Štaviše, ln izgledi za $Y \geq j$ veći nego log izgledi za $Y \geq (j+1)$. Pa samim tim to znači da je $\beta_{0,0} > \beta_{1,0} > \dots > \beta_{J-1,0}$.

Tumačenje koeficijenata je isto kao i za logit model.

Primer 2. Neka je data zavisna promenljiva Y_{ij} gde postoje četiri mogućnosti izbora. I neka su date dve nezavisne promenljive X_{i1} i X_{i2} , gde je i broj za i -ti element uzorka.

Multinomialni logit model ovako izgleda

$$P(Y_{ij} \geq j | X_{i1}, X_{i2}) = \frac{1}{1 + e^{-(\beta_{j,0} + \beta_1 X_{i1} + \beta_2 X_{i2})}}$$

gde je $i \in \{1, \dots, n\}$, a $j \in \{0, 1, 2, 3\}$

Pošto postoje četiri mogućnosti da i -ti element uzorka odabere j -tu mogućnost, koeficijente koje treba oceniti su $\beta_{0,0}, \beta_{1,0}, \beta_{2,0}$ koji su slobodni članovi i β_1 i β_2 su koeficijent vezani za nezavisne promenljive. U ovom slučaju ocenjeni koeficijenti su $\hat{\beta}_{0,0}, \hat{\beta}_{1,0}, \hat{\beta}_{2,0}, \hat{\beta}_1, \hat{\beta}_2$.

Ovaj primer pokazuje da u opštem slučaju broj koeficijenata koje treba oceniti jeste $J - 1 + K$.

Odnosno ocenjeni koeficijenti su $\hat{\beta}_{0,0}, \dots, \hat{\beta}_{J-1,0}, \hat{\beta}_1, \dots, \hat{\beta}_K$ za parametre $\beta_{0,0}, \dots, \beta_{J-1,0}, \beta_1, \dots, \beta_K$.

2.2.2 Ocenivanje parametara

Da bismo definisali funkciju verodostojnosti, potrebna nam je verovatnoća posmatranog ishoda za svako posmatranje.

U ordinalnoj logističkoj regresiji, modeliraćemo verovatnoću za $Y \geq j$. Da biste dobili izraz za $P(Y = j)$ možemo koristiti vezu da je $P(Y = j) = P(Y \geq j) - P(Y \geq j+1)$. Na ovaj način možemo

koristiti model da bi se dobio izraz za verovatnoću da je pojedinac u određenoj kategoriji ishoda za date promenljive ($X_1, \dots, X_K$).

Funkcija verodostojnosti je

$$L(\boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^n \prod_{j=0}^{J-1} P(Y_{ij} = j | X_{i1}, \dots, X_{iK})^{Y_{ij}},$$

gde je

$$Y_{ij} = \begin{cases} 1, & \text{ako je } i - \text{ti element uzorka i } Y = j \\ 0, & \text{inače} \end{cases}.$$

Ova funkcija može se napisati i u logaritamskom obliku

$$\ln L(\boldsymbol{\beta}, \mathbf{X}) = \sum_{i=1}^n \sum_{j=0}^{J-1} Y_{ij} \ln P(Y_{ij} = j | X_{i1}, \dots, X_{iK})$$

Ta funkcija verodostojnosti u logaritamskom obliku globalno je konkavna pa njeno maksimiziranje nije teško sve dokle nema mnogo nepoznatih parametara.

Multinomni model ima potencijalnu slabost povezanu s činjenicom da se pretpostavlja da su izbori između bilo koje dve mogućnosti nezavisni o preostalim mogućnostima. To je poznato kao „nezavisnost nevažnih mogućnosti”. Kakve to probleme može stvoriti najbolje ilustruje sledeći čuveni primer „crveno-plavog autobusa”. Pretpostavite da postoje tri načina prevoza između kojih pojedinac bira: automobil, crveni autobus i treća mogućnost. Sada se u skladu s multinomijalnim logit modelom prednosti izbora automobila umesto crvenog autobusa specificiraju na isti način, bez obzira je li treća mogućnost voz ili plavi autobus (Kmenta, 1971). To je očito neumesno. Taj problem se javlja kad god su neke dostupne mogućnosti bliski supstituti. On se može izbeći integriranjem mogućnosti kod Y , što je katkada osetljivo. Međutim, ako supstituti nisu dovoljno bliski da bi bili lagano spojivi, već tako bliski da bi njihovo zanemarivanje bilo sumnjivo, multinomijalni logit model može biti neprikladan. U takvoj situaciji može se upotrebiti test hipoteze o „nezavisnosti nevažnih mogućnosti”⁵.

2.2.3. Nominalna logistička regresija

Kao što je ranije rečeno, kod ove regresije zavisna promenljiva ima 3 ili više kategorija koje nisu u prirodnom poretku.

Sa nominalnom logističkom regresijom, jedna od određenih kategorija promenljivih kao referentna kategorija i svaki od ostalih kategorija se upoređuju sa referentnom kategorijom. Izbor referentne kategorije može biti proizvoljan i po nahođenju istraživača. Promena referentne kategorije ne menja oblik modela, ali menja tumačenje ocenjenih parametara u modelu.

⁵ Vidite Hausman, J. i McFadden, D. „Specification Tests for Multinomial Logit Model”, *Econometrica*, 52 (septembar 1984), str. 1219-1240.

Ako neku zavisnu promenljivu sa tri izbora kodiramo sa 0 ili 1 ili 2, onda biramo neku od njih da poredimo sa preostale dve. Ovi izrazi su kao i u ordinalnoj regresiji takođe u obliku odnosa verovatnoće.

Kod nominalne logističke regresije sa tri nivoa definisaćemo naš model koristeći dva izraza za količnik odnosa šansi. Prvi je odnos verovatnoće da je rezultat kategorije 1 podeljen sa verovatnoćom da je ishod u kategoriji jednak 0; drugi je odnos verovatnoće da je rezultat kategorije 2 podeljen sa verovatnoćom da je ishod u kategoriji jednak 0. Odnosno $\frac{P(Y=1)}{P(Y=0)}$ i $\frac{P(Y=2)}{P(Y=0)}$.

Kada postoje tri kategorije ishoda, zbir verovatnoća za tri kategorije ishoda moraju biti jednaki 1 ($P(Y = 0) + P(Y = 1) + P(Y = 2) = 1$). Svako poređenje razmatra samo dve verovatnoće.

Model se proširuje na više od tri nivoa, neka ishod ima J mogućnosti $(0, 1, \dots, J - 1)$. Sada postoje $J - 1$ mogućih poređenja sa referentnom kategorijom.

Ako je referentna kategorija 0, možemo da definišemo model.

$$\frac{P(Y_i = j | X_{i1}, \dots, X_{iK})}{P(Y_i = 0 | X_{i1}, \dots, X_{iK})} = e^{\beta_{j,0} + \sum_{k=1}^K \beta_{j,k} X_{ik}}$$

za $j \in \{1, \dots, J - 1\}$

Primer 3. Neka je data zavisna promenljiva Y_{ij} gde ima tri mogućnosti izbora. I neka su date dve nezavisne promenljive X_{i1} i X_{i2} , gde je i broj za i -ti element uzorka.

Multinomialni logit model ovako izgleda

$$\frac{P(Y_i = j | X_{i1}, \dots, X_{iK})}{P(Y_i = 0 | X_{i1}, \dots, X_{iK})} = e^{\beta_{j,0} + \sum_{k=1}^K \beta_{j,k} X_{ik}}$$

gde je $i \in \{1, \dots, n\}$, a $j \in \{1, 2\}$

Pošto postoje tri mogućnosti da i -ti element uzorka odabere j -tu mogućnost, koeficijente koje treba oceniti su $\beta_{1,0}, \beta_{2,0}$ koji su slobodni članovi i $\beta_{1,1}, \beta_{2,1}, \beta_{1,2}, \beta_{2,2}$ su koeficijent vezani za nezavisne promenljive. U ovom slučaju ocenjeni koeficijenti su $\hat{\beta}_{1,0}, \hat{\beta}_{2,0}, \hat{\beta}_{1,1}, \hat{\beta}_{2,1}, \hat{\beta}_{1,2}, \hat{\beta}_{2,2}$.

Ovaj primer pokazuje da u opštem slučaju broj koeficijenata koje treba oceniti jeste $(J - 1) * (K + 1)$. Odnosno ocenjeni koeficijenti su $\hat{\beta}_{1,0}, \dots, \hat{\beta}_{J-1,0}, \hat{\beta}_{1,1}, \dots, \hat{\beta}_{J-1,1}, \hat{\beta}_{1,2}, \dots, \hat{\beta}_{J-1,2}, \dots, \hat{\beta}_{1,K}, \dots, \hat{\beta}_{J-1,K}$ za parametre $\beta_{1,0}, \dots, \beta_{J-1,0}, \beta_{1,1}, \dots, \beta_{J-1,1}, \beta_{1,2}, \dots, \beta_{J-1,2}, \dots, \beta_{1,K}, \dots, \beta_{J-1,K}$.

Koja je interpretacija koeficijenata $\beta_{1,0}, \dots, \beta_{J-1,0}$. Oni predstavljaju izgleda gde su nezavisne promenljive jednake nuli. β_{jK} može se tumačiti kao: povećanje šansi za upadne u kategoriju j u odnosu na druge kategorije koje je povezano sa povećanje prediktora za jednu jedinicu, držeći ostale pojmove konstantnim.

Model će se jednako dobro uklopiti, postići istu verovatnoću i proizvesti iste prilagođene vrednosti. Promeniće se samo vrednosti i tumačenje koeficijenata. Ako je odgovor redni, obično se bira najviša ili najniža kategorija na rednoj skali.

Odnos šansi se računa u zavisnosti da li je neka promenljiva binarna ili neprekidna.

Ako je neka promenljiva binarna, onda se šanse ovako računaju

$$\begin{aligned} odds_j &= \frac{(P(Y_i = j|X_{i1}, \dots, X_{is} = 1, \dots, X_{iK})/P(Y_i = 0|X_{i1}, \dots, X_{is} = 1, \dots, X_{iK}))}{(P(Y_i = j|X_{i1}, \dots, X_{is} = 0, \dots, X_{iK})/P(Y_i = 0|X_{i1}, \dots, X_{is} = 0, \dots, X_{iK}))} \\ &= \frac{e^{\beta_{j,0} + \sum_{k=1}^K \beta_{j,k} X_{ik} + \beta_{j,s} * 1}}{e^{\beta_{j,0} + \sum_{k=1}^K \beta_{j,k} X_{ik} + \beta_{j,s} * 0}} = e^{\beta_{j,s}} \quad j \in \{1, \dots, J-1\} \end{aligned}$$

gde je X_{is} određena promenljiva $s \in \{1, \dots, K\}$.

Ako je neka promenljiva neprekidna, onda šanse ovako računamo

$$\begin{aligned} odds_j &= \frac{(P(Y_i = j|X_{i1}, \dots, X_{is} = const_1, \dots, X_{iK})/P(Y_i = 0|X_{i1}, \dots, X_{is} = const_1, \dots, X_{iK}))}{(P(Y_i = j|X_{i1}, \dots, X_{is} = const_2, \dots, X_{iK})/P(Y_i = 0|X_{i1}, \dots, X_{is} = const_2, \dots, X_{iK}))} \\ &= \frac{e^{\beta_{j,0} + \sum_{k=1}^K \beta_{j,k} X_{ik} + \beta_{j,s} * const_1}}{e^{\beta_{j,0} + \sum_{k=1}^K \beta_{j,k} X_{ik} + \beta_{j,s} * const_2}} = e^{\beta_{j,s} * (const_1 - const_2)} \quad j \in \{1, \dots, J-1\} \end{aligned}$$

Što se tiče definisanje funkcije verodostojnosti, ona je slično definisana kao i u ordinalnoj regresiji.

$$L = \prod_{i=1}^n \prod_{j=0}^{J-1} P(Y_{ij} = j|X_{i1}, \dots, X_{iK})^{Y_{ij}}$$

gde

$$Y_{ij} = \begin{cases} 1, & \text{ako je } i - \text{ti element uzorka i } Y = j \\ 0, & \text{inače} \end{cases}$$

Ova funkcija može se napisati i u logaritamskom obliku

$$\ln L = \sum_{i=1}^n \sum_{j=0}^{J-1} Y_{ij} \ln P(Y_{ij} = j|X_{i1}, \dots, X_{iK}).$$

2.3 Odabir modela

Pored kriterijuma koji su predstavljeni u delu linearnih modela, za odabir logističkih modela postoje dodatni kriterijumi koji uzimaju u obzir prirodu logističkih modela.

Tako su najzastupljenija proširenje adjugovanog koeficijenta determinacije, McFadden-ov, Cox-Snell-ov i Nagelkerke-ov koeficijent determinacije. Najjednostavnije je McFadden-ov R^2 koji se izračunava na sledeći način

$$R_{McF}^2 = 1 - \frac{\ln L(\hat{\theta})}{\ln L(0)}$$

dok je Cox-Snell-ov R^2 definisan sa

$$R_{C\&S}^2 = 1 - \left(\frac{L(0)}{L(\hat{\theta})} \right)^{\frac{2}{n}},$$

gde $L(0)$ predstavlja vrednost funkcije verodostojnosti kada u model nije uključena nijedna nezavisna promenljiva, odnosno postoji samo slobodan član, dok $L(\hat{\theta})$ predstavlja vrednost funkcije verodostojnosti za posmatrani model. Tumačenje koeficijenta je isto kao i u slučaju adjugovanog koeficijenta determinacije, odnosno veća vrednost koeficijenta ukazuje na bolji model. Nagelkerke-ov koeficijent determinacije predstavlja modifikaciju Cox-Snell-ov koeficijenta. Kako je najveća moguća vrednost za Cox-Snell-ov koeficijent $R_{max}^2 = 1 - (L(0))^{\frac{2}{n}}$, Nagelkerke-ov koeficijent dobija se kao

$$R_{Neg}^2 = \frac{R^2}{R_{max}^2}.$$

Kako je Cox-Snell-ov R^2 zasnovan na oceni maksimalne verodostojnosti ($L(\hat{\theta})$) za model u poređenju ocene maksimalne verodostojnosti ($L(0)$) za osnovni (prazan) model njegova teoretski maksimalna vrednost manja je od 1, čak i za „savršeni“ model. Nagelkerkeov R^2 je prilagođena verzija Cox-Snell-ov R^2 koji prilagođava skalu statistike tako da pokriva čitav opseg od 0 do 1. MekFadenov R^2 je još jedna verzija, zasnovana na jezgrima ocene maksimalne verodostojnosti ($\ln L$) za model sa konstantom i celog modela.

3 Primena multinominalne regresije u oblasti turizma

Postoje velike mogućnosti u primeni različitih regresionih modela u realnim pojavama. U ovom delu biće prikazana primena multinominalne logističke regresije (MLR) u modeliranje dve pojave u oblasti turizma.

Prvi model se odnosi na ispitivanje načina prevoza koji turisti preferiraju, tj. koriste u zavisnosti od više faktora, odnosno nezavisnih promenljivih, Istraživanje je sprovedeno u Norveškoj, dok je drugi model razmatra koje socio-demografske karakteristike utiču na nivo zadovoljstva turista u Crnoj Gori.

3.1. Mogućnost korišćenja multinominalnog logit modela za ispitivanje načina prevoza

Krister Tran (Christer Thrane, 2015) je predstavio model koji je korišćen za ispitivanje načina prevoza u slučaju dužih putovanja. Opcije koje turista može da koristi predstavljaju zavisnu promenljivu (transportation mode choices =TMC) i ima tri kategorije: privatni prevoz, avio prevoz i javni prevoz. Putem ankete prikupljeni su podaci o većem broju faktora, odnosno nezavisnih promenljivih koje su podelje u tri veće grupe: promenljive vezane za udaljenost (Travel distance variables=TD), promenljive o karakteristikama putovanja (Trip-related characteristics=TRC) i socio-demografske promenljive (Socio-demographic variables=SDV). Model koji leži u osnovi ove studije može se zapisati u sledećoj funkcionalnoj formi

$$TMC = f(TD, TRC, SDV).$$

Istraživanje putem upitnika, koje je finansirano od strane Veća za istraživanje Norveške u okviru projekta TOURIMPACT čiji je cilj unapređenje i širenje statistike u Norveškoj, je sprovedeno u toku marta i aprila 2009. godine i obuhvatilo je 2139 tursita. Istraživanje je ograničeno na osobe starosti od 18 do 80 godina koji su prešli najmanje 100 kilometara (jedan smer).

Zavisna promenljiva Vrsta prevoza je, kako je to već napomenuto, nominalna promenljiva sa tri kategorije. Kako tri posmatrane kategorije nemaju prirodni redosled, izvršeno je označavanje na sledeći način. Za referentnu vrednost je izabran prevoz privatnim automobilom, dok su druge dve karakteristike avio prevoz i javni prevoz. Oko 65% norveških turista koristilo je privatni automobil, svaki peti, 22%, je putovao avionom dok je ostatak koristio javni transport (Podaci su u tabeli 3).

Tabela 3: zavisna promenljiva

Promenljiva	kod	proporcija (%)
<i>Vrsta prevoza</i>		
Privatni automobil	0	65%
Avio prevoz	1	22%
Javni prevoz	2	13%

Cilj istraživanja je bio da se ustanovi koji turisti biraju koju vrstu prevoza, tj. na osnovu kojih karakteristika se može predvideti koja vrsta prevoza će biti odabrana.

Turisti su najčešće putovali između 200 i 400 kilometara i slično tome dužina puta je najčešće bila između 2,5 i 4 sata. Prosečna dužina boravka najčešće je bila malo duže od 3,5 dana, dok je najčešće

putovale 4 osobe (kao grupa, porodica). U proseku su turisti posetili 1,5 okruga. Dve najčešće namene putovanja su bile posete prijateljima kao i drugim domovima, oko 50% turista je imalo jednog pratioca. Prosečan turista je imao 41 godinu i prijavljen je godišnji prihod od 686 500 norveških kruna.

Izdvojeno je deset promenljivih koje su uključene u model. Tri promenljive su kategorične promenljive: Razlog putovanja, Saputnici i Pol. Kod ove tri promenljive ne postoji prirodan redosled, pa je izbor referentne vrednosti bio na istraživačima. Dve numeričke promenljive *dužina putovanja (km)* i *dužina putovanja (sati)* prevedene u kategorijalne gde je zadržan prirodni redosled, pa je za referentnu vrednost uzeta najniža kategorija. U tabeli 4 date su deskriptivni pokazatelji (procenat turista sa posmatranom kategorijom u uzorku) za pet kategorijalnih promenljivih koji ulaze u model i označene su referentne vrednosti.

Tabela 4: Nezavisne promenljive kategoričnog tipa

	Promenljiva	proporcija (%)
TD	<i>Dužina putovanja (km)</i>	
	Manje od 150	18%
	Od 150 do 199	14%
	Od 200 do 399	35%
	Od 400 do 599	16%
	Preko 599	17%
	<i>Dužina putovanja (sati)</i>	
	Manje od 2	10%
	Od 2 do 2.49	15%
	Od 2.5 do 3.99	31%
TRC	<i>Razlog putovanja</i>	
	Poseta prijateljima	37%
	Drugi dom	27%
	Skijanje	15%
	Događaj	10%
	Drugi razlog	11%
	<i>Saputnici</i>	
	Sam	16%
	Dve osobe	49%
	Tri osobe	26%
SDV	<i>Pol</i>	
	Ženski	47%
	Muški	53%

Numeričke promenljive koje pripadaju grupi TRC promenljivih su: Dužina boravka (dani), Veličina putničke grupe, Broj posećenih okruga, Urbano-ruralne destinacije dok grupi SDV promenljivih pripadaju: godine turista i godišnji prihod. Dodatno je u modelu uključena i promenljiva koja predstavlja godine turista na kvadrat.

Tabela 5: Nezavisne promenljive neprekidnog tipa

	Promenljiva	prosek	stand. ods.
TRC	<i>Dužina boravka (dani)</i>	3.66	2.82
	<i>Veličina putničke grupe</i>	3.99	5.44
	<i>Broj posećenih okruga</i>	1.55	1.11
	<i>Urban-rural destinacija</i>	3.09	1.52
SDV	<i>Godine</i>	40.99	13.74
	<i>Godine na kvadrat</i>	1868.54	1212.92
	<i>Godišnj prihod</i>	686500.00	313316.00

Naredni korak u modeliranju je prevođenje kategorijalnih promenljivih (dummy promenljivih) sa više od dve kategorije u skup binarnih promenljivih. Na taj način je svaka promenljiva sa pet kategorija (Dužina putovanja (km), Dužina putovanja (sati), prevedena u četiri binarne promenljive. Kodiranje je dato u tabelama 6-8, dok su promenljivoj Saputnici pridružene tri binarne, tabela 9. Na ovaj način je definisano 19 binarnih promenljivih koje predstavljaju nezavisne promenljive u modelu.

Tabela 6: Prevođenje promenljive *Dužina putovanja (km)* u binarne promenljive

	X_1	X_2	X_3	X_4
Manje od 150	0	0	0	0
Od 150 do 199	1	0	0	0
Od 200 do 399	0	1	0	0
Od 400 do 599	0	0	1	0
Preko 599	0	0	0	1

Tabela 7: Prevođenje promenljive *Dužina putovanja (sati)* u binarne promenljive

	X_5	X_6	X_7	X_8
Manje od 2	0	0	0	0
Od 2 do 2.49	1	0	0	0
Od 2.5 do 3.99	0	1	0	0
Od 4 do 5.49	0	0	1	0
Preko 5.49	0	0	0	1

Tabela 8: Prevođenje promenljive *Razlog putovanja* u binarne promenljive

	X_{13}	X_{14}	X_{15}	X_{16}
Poseta prijateljima	0	0	0	0
Drugi dom	1	0	0	0
Skijanje	0	1	0	0
Događaj	0	0	1	0
Drugi razlog	0	0	0	1

Tabela 9: Prevođenje promenljive *Saputnici* u binarne promenljive

	X_{17}	X_{18}	X_{19}
Sam	0	0	0
Dve osobe	1	0	0
Tri osobe	0	1	0
Četiri osobe	0	0	1

Što se tiče karakteristike putovanja kao i socio-demografske karakteristike, te promenljive su neprekidne promenljive. *Dužina boravka* (izražena u danima), *Veličina putničke grupe* (broj osoba), *Broj posećenih okruga* i *Urbano-ruralne destinacije* (1-5 pri čemu je 1 za veliki grad, dok je 5 za divljinu) su nezavisne promenljive označene sa X_9, X_{10}, X_{11} i X_{12} .

Kod socio-demografskih promenljivih starost (18-80), godine na kvadrat, godišnji prihod su neprekidne promenljive i označene su sa X_{20}, X_{21} i X_{22} . Pol je jedina binarna promenljiva koja je posmatrana kao karakteristika turista i ona je iznačena sa X_{23} .

Pošto u ovom primeru ima tri izbora i 23 promenljive, treba oceniti 48 koeficijenata $((J - 1) * (K + 1))$.

Jednačina za multinominalni logit model kada poredimo izbor avio prevoza u odnosu na putnički automobil (referentna kategorija) je

$$\frac{P(Y_i = 1|X_{i1}, \dots, X_{i23})}{P(Y_i = 0|X_{i1}, \dots, X_{i23})} = e^{\beta_{1,0} + \sum_{k=1}^{23} \beta_{1,k} X_{ik}}. \quad (19)$$

Jednačina za multinominalni logit model kada poredimo izbor javnog prevoza u odnosu na putnički automobil je

$$\frac{P(Y_i = 2|X_{i1}, \dots, X_{i23})}{P(Y_i = 0|X_{i1}, \dots, X_{i23})} = e^{\beta_{2,0} + \sum_{k=1}^{23} \beta_{2,k} X_{ik}}. \quad (20)$$

U tabeli 10 su prikazani koeficijenti za multinominalni logit model.

Tabela 10: Ocenjeni koeficijenti za modele (19) i (20)

Nezavisne promenljive	prom.	mod. (19)	Avio prevoz	mod. (20)	Javni prevoz
Dužina putovanja (km)					
Od 150-199	X_1	$\beta_{1,1}$	-0.158 (0.468)	$\beta_{2,1}$	0.44 (0.315)
Od 200-399	X_2	$\beta_{1,2}$	0.33 (0.418)	$\beta_{2,2}$	-0.197 (0.304)
Od 400 do 599	X_3	$\beta_{1,3}$	3.051 (0.406) ***	$\beta_{2,3}$	-0.154 (0.329)
Više od 600	X_4	$\beta_{1,4}$	4.487 (0.434) ***	$\beta_{2,4}$	0.448 (0.364)
Dužina putovanja (sati)					
Od 2 do 2.49	X_5	$\beta_{1,5}$	-0.207 (0.288)	$\beta_{2,5}$	-0.288 (0.354)
Od 2.5 do 3.99	X_6	$\beta_{1,6}$	-0.586 (0.316)	$\beta_{2,6}$	-0.032 (0.356)
Od 4 do 5.49	X_7	$\beta_{1,7}$	-0.721 (0.363) *	$\beta_{2,7}$	0.167 (0.383)
Više od 5.5	X_8	$\beta_{1,8}$	-2.624 (0.369) ***	$\beta_{2,8}$	0.71 (0.374)
Dužina boravka (danima)	X_9	$\beta_{1,9}$	0.189 (0.036) ***	$\beta_{2,9}$	0.092 (0.031) **
Veličina provoda	X_{10}	$\beta_{1,10}$	0.047 (0.015) **	$\beta_{2,10}$	0.043 (0.013) ***
Broj posećenih okruga	X_{11}	$\beta_{1,11}$	-0.368 (0.084) ***	$\beta_{2,11}$	-0.042 (0.059)
Urbano-ruralne destinacije	X_{12}	$\beta_{1,12}$	-0.541 (0.086) ***	$\beta_{2,12}$	-0.337 (0.062) ***
Razlog putovanja					
Druga kuća	X_{13}	$\beta_{1,13}$	-1.825 (0.415) ***	$\beta_{2,13}$	-0.883 (0.257) ***
Skijanje	X_{14}	$\beta_{1,14}$	-1.466 (0.422) ***	$\beta_{2,14}$	-0.492 (0.282)
Događaj	X_{15}	$\beta_{1,15}$	0.101 (0.236)	$\beta_{2,15}$	0.633 (0.238) **
Drugi razlog	X_{16}	$\beta_{1,16}$	0.47 (0.246)	$\beta_{2,16}$	0.581 (0.236) *
Saputnici					
Dve osobe	X_{17}	$\beta_{1,17}$	-1.325 (0.226) ***	$\beta_{2,17}$	-1.391 (0.196) ***
Tri osobe	X_{18}	$\beta_{1,18}$	-2.366 (0.3) ***	$\beta_{2,18}$	-1.974 (0.249) ***
Četiri i više osoba	X_{19}	$\beta_{1,19}$	-3.517 (0.69) ***	$\beta_{2,19}$	-2.997 (0.496) ***
Godine (18-80)	X_{20}	$\beta_{1,20}$	-0.063 (0.04)	$\beta_{2,20}$	-0.115 (0.034) ***
Godine na kvadrat (x 1000)	X_{21}	$\beta_{1,21}$	0.67 (0.453)	$\beta_{2,21}$	1.171 (0.381) ***
Godišnji prihod (x 1000)	X_{22}	$\beta_{1,22}$	0.612 (0.299) *	$\beta_{2,22}$	-0.925 (0.304) **
Pol	X_{23}	$\beta_{1,23}$	0.122 (0.166)	$\beta_{2,23}$	-0.004 (0.157)
Konstanta		$\beta_{1,0}$	2.621	$\beta_{2,0}$	3.611

Rezultat u zagradi je standardna greška, a zvezdice označavaju nivo značajnosti (* - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.001$).

Ako vidimo koeficijente iz tabele 10, vidimo da veći značaj ima odabir avio prevoza u odnosu na privatni automobil za distance od 400 do 599 kilometara i za udaljenosti više od 599 kilometara. Ti koeficijenti su: 3.051 i 4.487.

Što se tiče dužine putovanja u satima najveći značaj u odabir avio prevoza u odnosu na privatni automobil je kada turista putuje više od 5 i po sati. Taj koeficijent iznosi -2.624.

Značajni su i karakteristika vezana za putovanja, osim na broj posećenih okruga na odluku da putnik odabere javni prevoz u odnosu na automobil (Koeficijent vezan za broj posećenih okruga iznosi -0.042, dok za dužinu boravka, veličinu putničke grupe i urbano-ruralne destinacije iznose: 0.092, 0.043 i -0.337 redom).

Drugi dom svakako ima najveći uticaj u odabiru avio i javnog prevoza umesto privatni automobil. Koeficijenti vezani za drugi dom su -1.825 (avio prevoz) i -0.883 (javni prevoz).

Broj saputnika je takođe značajan u odabiru prevoznog sredstva za turiste. Koeficijenti za broj saputnika koji su vezani za odabir avio prevoza su: -1.325 (dve osobe), -2.366 (tri osobe) i -3.517 (četiri i više osoba). Koeficijenti za broj saputnika koji su vezani za odabir putničkog prevoza su: -1.391 (dve osobe), -1.974 (tri osobe) i -2.997 (četiri i više osoba).

Čini se da su socio-demografske karakteristike od manjeg značaja kada je u pitanju odabir prevoza. Jedino marginalno značajniji efekat je prihod domaćinstava, tačnije veći prihod domaćinstava utiče na to da će putnik pre izabrati avio prevozu u odnosu na javni prevoz i automobil.

Kako se povećava broj okruga koji se posećuju tokom putovanja, tako smanjuje verovatnoća da će putnik putovati avio prevozom umesto automobilom. To se vidi iz $e^{-0.368} = 0.692$. Odnosno za povećanje broja okruga za jednu jedinicu šanse se smanjuju za $1 - 0.692 = 0.308$. Što je određite ruralnije, veća je verovatnoća da turisti preferiraju automobil u odnosu na vazdušni i javni prevoz. To se vidi iz toga da $e^{-0.541} = 0.582$ i $e^{-0.337} = 0.714$ tj. verovatnoće se za promenu jednu jedinicu smanjuju za 0.418 odnosno 0.286 da turisti preferiraju vazdušni odnosno javni prevoz u odnosu na automobil.

Predviđene verovatnoće mogu da se izračunaju za svaki od nivoa ishoda. Predviđene verovatnoće su bazirane na koeficijentima iz tabele 10. Predviđene verovatnoće su računane preko funkcije *fitted* u R-u.

U tabeli 11 ćete videti predviđene verovatnoće odabira vrste prevoza u zavisnosti broju dana boravka turista.

Tabela 11

Dužina boravka	Privatni automobil	Avio prevoz	Javni prevoz
2 dana	0.79	0.12	0.09
4 dana	0.73	0.17	0.1
6 dana	0.67	0.22	0.11
8 dana	0.6	0.29	0.12

Odnosno što se tiče dužine putovanja (km) uzimamo vrednosti koje su uzete iz tabele 4. To su: 0.18, 0.14, 0.35, 0.16 i 0.17. Kod dužine putovanja (sati) uzete su vrednosti koje iznose: 0.1, 0.15, 0.31, 0.23 i 0.21. Kod karakteristika putovanja uzimamo vrednosti koje su uzete iz tabele 5. To su: 3.99, 1.55 i 3.09, dok kod dužine boravka u danima su uzete vrednosti od 2 do 8, a ne 3.66.

Ono što je vidljivo u tom slučaju jeste da kako se broj dana povećava tako je veća verovatnoća da putnik uzme avio prevoz, analogno tome je da je veća verovatnoća da putnik uzme javni prevoz (vidljivo u procentima 12, 17, 22 i 29 za avio prevoz i 9, 10, 11 i 12 za javni prevoz). A samim tim opada verovatnoća da putnik uzme privatni automobil (verovatnoće su 79, 73, 67 i 60).

Verovatnoće se računaju koristeći

$$P(Y_i = j) = \frac{e^{\beta_{j,0} + \beta_{j,1}X_{i1} + \dots + \beta_{j,K}X_{i23}}}{1 + \sum_{s=1}^2 e^{\beta_{s,0} + \beta_{s,1}X_{i1} + \dots + \beta_{s,K}X_{i23}}}, j \in \{1,2\}$$

i za $j = 0$

$$P(Y_i = 0) = \frac{1}{1 + \sum_{s=1}^2 e^{\beta_{s,0} + \beta_{s,1}X_{i1} + \dots + \beta_{s,K}X_{i23}}}$$

Ovom modelu su dati koeficijenti za Nagelkerke-ov koeficijent determinacije, McFadden-ov koeficijent determinacije i Cox-Snell-ov koeficijent determinacije. I oni iznose: 0.65, 0.438 i 0.539 respektabilno, dok koeficijenti za $L(0)$ i $L(\hat{\theta})$ iznose: -1064.564 i -1892.726.

3.2. Mogućnost korišćenja multinominalnog logit modela za izbor turističke destinacije

U ovom delu će biti prikazani primer multinominalnog logit modela gde se koriste pri izboru turističke destinacije. Neki od tih radova su: Nurlaela (2018), Frangos, Karapistolis, Stalidis, Fragkos, Sotiropoulos, Manolopoulos (2014), Jušten i Hezinger (2020), Ju, Jun i Park (2018) i Perović, Stanovčić, Morić i Peković (2012).

Kod Nurlaele (2018), rađena je analiza multinominalnog logit modela za izbor destinacije. Izbori koji su ponuđeni su: „Tropical North Queensland”, „California Beach”, „the Cornwall Beach” i da nema putovanja. Multinominalni logit model je bio od koristi za identifikovanje najboljeg scenarija koji je „Tropical North Queensland” doneo vođstvo na tržištu u skladu sa dobrim budžetskim performansama. Razvijen je scenario testiranjem promenljivih cena u kombinaciji sa dummy promenljivama za povećanje porodičnih preduzeća; zatim testiranje scenarija na osnovi kombinacije spremnosti za plaćanje i analiza elastičnosti. Dalje, analiza elastičnosti je sprovedena radi razumevanja uticaja promena politike na rezultate modela (Nurlaela, 2018).

U radu „Odanost turista je najvažnija za cene, kulturu i sunce: Multinominalna logistička regresija turista u poseti Atini” cilj je identifikovati prediktivne faktore lojalnosti turista turističkoj destinaciji glavnog grada Atine. Uzorkovanje je sprovedeno među članovima slučajnog uzorka od 461 turista u Atini. Upitnik je sadržao odeljke o zadovoljstvu iz Atine, imidž Atine, imidž Grčke, lojalnost Atini kao odredište, usmena objava za putovanje u Atinu, demografski podaci i pitanja o namerama za ponovljene turističke posete u Atini. Metode analize podataka bile su Analiza istraživačkog faktora (EFA) slike grada Atine i slika Grčke, dobijene tokom posete turista. EFA o slici Atine je otkrio tri faktora: Prirodno okruženje i ponašanje gradskog stanovništva prema turistima. Istorijski spomenici, vreme i Mediteranska kuhinja i kvalitet gradskog smeštaja i ostalih usluga. Logistička regresija za zavisnom promenljivom: Odanost turističkog posetioca Atini, koja je otkrila sledeće statistički značajne nezavisne faktore koji pozitivno utiču na lojalnost destinaciji: zadovoljstvo cenama putovanja, čistoća Atine, cene proizvoda i usluga, Istorijski spomenici i prirodna lepota. Interesantna je druga Logistička regresija sa zavisnom promenljivom: Usmena reč ponovljenog posetioca Atine zbog posete gradu, koja je statistički otkrila dve značajne nezavisne promenljive: Zadovoljstvo cenom putovanja i sunčanim prirodnim okruženjem Atine i Grčke (Frangos, Karapistolis, G., Fragkos, Sotiropoulos, & Manolopoulos, 2014).

U radu iz 2020 godine pod nazivom: „Izvan grada - ali kako i gde? Destinacija kao model izbora za urbano-seoski turistička putovanja u Austriju” govori o tome da u seoskom turizmu dominiraju putovanja automobilom. Da bi privukli turiste potrebno je veće razumevanje o faktorima koji pokreću turističke odluke. Ovaj rad ispituje odredište i izbor način prevoza kao kombinovani izbor u kontekstu urbano-ruralnog turizma u Austriji. Da bi to učinili, ovaj članak istražuje dve različite

strukture modela, na kraju koristeći multinominalni logit model, koji je ukorenjen u teoriji slučajnih korisnosti. Analizirani podaci se zasnivaju na velikom turizmu, sa naznačenim dodatnim karakteristikama putovanja i odredišta kako bi se omogućio očekivani fokus na faktore na strani ponude. Rezultati pokazuju da su pod jedan, odabrani odredište i način prevoza isprepletene odluke, pod dva, putnici vide različito automobil i javni prevoz vreme putovanja i udaljenost, pod tri, visokokvalitetno prisustvo na mreži je najjači atraktivni cilj, pod četiri, prohodnost olakšava atraktivnost odredišta i javnog prevoza i pod pet dnevna i turistička mobilnost je povezana kroz osnovne kulture mobilnosti (Juschten & Hössinger, 2020).

Što se tiče rada: „Motivacija turizma: integralni pristup do izbora odredišta”, njegova svrha jeste da se raspravlja o prevladavajućim socio-psihološkim modelima koji ispituju kako potrebe i motivacije turista utiču na njihov izbor odredišta kolektivnim razmatranjem (Plog 1974), psihografski profili, turistička tipologija (Koen 1979) i lestvice putničkih karijera (Pis 1988). Trenutna studija tvrdi da nijedan pojedinačni mode ne može adekvatno objasniti postupak odabira destinacije turista kao i putnička ponašanja i uvodi novu integrisanu perspektivu postojećih psiholoških modela. Anketirana su 202 stanovnika Hong Konga koji su putnici i ova studija deli psihografiju ispitanika na tri vrste. Što se tiče multinominalnog logit modela korišćenjem te analize, studija otkriva da motivacija za putovanje turista i postavke odredišta mogu se razlikovati prema njihovim psihografskim tipovima (Yoo, Yoon, & Park, 2018).

Što se tiče Crne Gore, u radu koji se zove: „Koje socio-demografske karakteristike utiču na nivo zadovoljstva turista u Crnoj Gori? Empirijska analiza”, koji je vezan za turizam u Crnoj Gori. Da bi se u budućnosti postigli strateški ciljevi održivog razvoja turizma, glavni fokus treba dati analizi glavnih odrednica zadovoljstva turista. Koristeći bazu podataka koja pruža informacije o ponašanju i zadovoljstvu turističkih putovanja tokom boravka u Crnoj Gori, sprovedena je empirijska analiza kako bi se razumelo da li su socio-demografske karakteristike povezane sa nivoom zadovoljstva turista. Učinak socio-demografskih karakteristika se meri kroz pet dimenzija: pol, starost, država prebivališta, zanimanje i zarada. Koristeći multinominalni logit model analiziramo gore pomenute socio-demografske karakteristike i njihov uticaj na nivo zadovoljstva turista u Crnoj Gori (Perović, Stanovčić, Morić, & Sanja, 2012).

Anketirano je 1244 turista. U tabeli 12 su date prikazane frekvencije, 388 (31%) turista je izjavilo da je veoma zadovoljno, 783 turista je bilo zadovoljno (63%), dok ostatak od 73 (6%) turista nije bilo zadovoljno. Promenljiva Satisfakcija turista je kodirana tako da 0 uzima turista koji nije zadovoljan, 1 ako turista veoma zadovoljan i 2 ako je turista zadovoljan. Neprekidne promenljive su realnog tipa pa je tu najmanji problem, ostale promenljive se kodiraju sa 0 ili 1. Kod promenljive tip smeštaja, referentna kategorija jeste druge mogućnosti, što se tiče vrste prevoza referentna kategorija je voz, zemlje bivše Jugoslavije su rerefentna kategorija za zemlje odakle dolaze turisti, kod zaposlenosti referentna kategorija je zaposleni u privatnoj kompaniji,a kod plate to je manje od 900 Evra.

Tabela 12: zavisna promenljiva

Promenljiva	kod	proporcija (%)
<i>Satisfakcija turista</i>		
Nije zadovoljan	0	6%
Veoma zadovoljan	1	31%
Zadovoljan	2	63%

U tabeli 13 su dati deskriptivni pokazatelji (procenat turista sa posmatranom kategorijom u uzorku) sa sedam kategorijalnim promenljivama koje ulaze u model i označene su referentne vrednosti.

Tabela 13: Nezavisne promenljive kategoričnog tipa

Promenljiva	proporcija (%)
<i>Ranije iskustvo</i>	
Nema	32%
Ima	68%
<i>Tip smeštaja</i>	
Druge mogućnosti	19%
Noćenje	42%
Noćenje i doručak	9%
Polupansion	17%
Pansion	9%
Ol inkluziv	4%
<i>Vrsta prevoza</i>	
Voz	10%
Avion	33%
Autobus	23%
Automobil	34%
<i>Pol</i>	
Ženski	48%
Muški	52%
<i>Zemlje iz kojih dolaze turisti</i>	
Bivša Jugoslavija	61%
Evropska Unija	25%
Bivša SSSR-a	11%
Ostatak sveta	3%
<i>Zanimanje turista</i>	
Samozaposleni	18%
Zaposleni u privatnoj kompaniji	42%
Državni funkcioner	6%
Penzioner	7%
Student	23%
Nezaposlen	4%
<i>Plata turista (Evro)</i>	
Manje od 900	15%
Od 900 do 3000	50%
Više od 3000	35%

U tabeli 14 su date nezavisne promenljive koje su numeričke. To su: Broj noćenja, Broj odmora i Uзраст.

Tabela 14: Nezavisne promenljive neprekidnog tipa

Promenljiva	prosek	stand. ods.
<i>Broj noćenja</i>	3.00	1.51
<i>Broj odmora</i>	1.45	0.93
<i>Uзраст</i>	34.59	12.78

Naredni korak u modeliranju je prevođenje kategorijalnih promenljivih (dummy promenljivih) sa više od dve kategorije u skup binarnih promenljivih. Na taj način je svaka promenljiva sa šest kategorija (*Tip smeštaja* i *Zanimanje turista*) prevedena u pet binarnih promenljivih. Kodiranje je dato u tabelama 15 i 16. Promenljivama *Vrsta prevoza* i *Zemlje iz koje dolaze turisti* su pridružene tri promenljive (jer postoje četiri kategorije), tabela 17 i 18. Promenljivoj *Plata turista* su pridružene dve promenljive (jer postoje tri kategorije), tabela 19. Na ovaj način je kodirano 23 binarnih promenljivih koje predstavljaju nezavisne promenljive u modelu.

Tabela 15: Prevođenje promenljive *Tip smeštaja* u binarne promenljive

	X_4	X_5	X_6	X_7	X_8
Druge mogućnosti	0	0	0	0	0
Noćenje	1	0	0	0	0
Noćenje i doručak	0	1	0	0	0
Polupansion	0	0	1	0	0
Pansion	0	0	0	1	0
Ol inkluziv	0	0	0	0	1

Tabela 16: Prevođenje promenljive *Zanimanje turista* u binarne promenljive

	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}
Samozaposleni	0	0	0	0	0
Zaposleni u privatnoj kompaniji	1	0	0	0	0
Državni funkcioner	0	1	0	0	0
Penzioner	0	0	1	0	0
Student	0	0	0	1	0
Nezaposlen	0	0	0	0	1

Tabela 17: Prevođenje promenljive *Vrsta prevoza* u binarne promenljive

	X_9	X_{10}	X_{11}
Voz	0	0	0
Avion	1	0	0
Autobus	0	1	0
Automobil	0	0	1

Tabela 18: Prevođenje promenljive *Zemlje iz koje dolaze turisti* u binarne promenljive

	X_{14}	X_{15}	X_{16}
Bivša Jugoslavija	0	0	0
Evropska Unija	1	0	0
Bivšeg SSSR-a	0	1	0
Ostatak sveta	0	0	1

Tabela 19: Prevođenje promenljive *Plata turista* u binarne promenljive

	X_{22}	X_{23}
Manje od 900	0	0
Od 900 do 3000	1	0
Više od 3000	0	1

Što se tiče promenljivih kao što su: *Broj noćenja*, *Broj odmora* i *Uzrast*, one su neprekidne promenljive. Te promenljive se označavaju za X_1, X_3 i X_{13} . Dok binarne promenljive kao što su *Ranije iskustvo* i *Pol* one se označavaju sa X_2 i X_{12} .

Pošto i u ovom primeru ima tri izbora i 23 promenljive, treba oceniti 48 koeficijenata $((J - 1) * (K + 1))$. Koeficijenti se mogu oceniti na više načina.

Jednačina za multinominalni logit model kada poredimo kategoriju “veoma zadovoljan” u odnosu na “nezadovoljan” data je sa (19), dok je jednačina za multinominalni logit model kada poredimo kategoriju “zadovoljan” u odnosu na “nezadovoljan” data sa (20). Objedinjeno ove dve jednačine zapisujemo na sledeći način

$$\frac{P(Y_i = j | X_{i1}, \dots, X_{iK})}{P(Y_i = 0 | X_{i1}, \dots, X_{iK})} = e^{\beta_{j,0} + \sum_{k=1}^{23} \beta_{j,k} X_{ik}}, \quad j \in \{1, 2\}.$$

U tabeli 20 su prikazani ocenjeni koeficijenti za multinominalni logit model.

Tabela 20: Koeficijenti multinomilanog regresionog modela

Nezavisne promenljive	prom.	mod. (19)	Veoma zadovoljan	mod. (20)	Zadovoljan
Broj noćenja	X_1	$\beta_{1,1}$	-0.18 (0.09) +	$\beta_{2,1}$	-0.25 (0.08) **
Ranije iskustvo	X_2	$\beta_{1,2}$	0.72 (0.36) +	$\beta_{2,2}$	0.55 (0.34)
Broj odmora	X_3	$\beta_{1,3}$	-0.34 (0.11) **	$\beta_{2,3}$	-0.44 (0.11) **
Tip smeštaja					
Noćenje	X_4	$\beta_{1,4}$	-0.2 (0.4)	$\beta_{2,4}$	0.06 (0.38)
Noćenje i doručak	X_5	$\beta_{1,5}$	-1.01 (0.56) +	$\beta_{2,5}$	-0.14 (0.51)
Polupansion	X_6	$\beta_{1,6}$	-0.73 (0.46)	$\beta_{2,6}$	-0.8 (0.44) +
Pansion	X_7	$\beta_{1,7}$	-0.28 (0.52)	$\beta_{2,7}$	-0.62 (0.51)
Ol inkluziv	X_8	$\beta_{1,8}$	13.48 (575.7)	$\beta_{2,8}$	13.3 (575.7)
Vrsta prevoza					
Avion	X_9	$\beta_{1,9}$	0.47 (0.53)	$\beta_{2,9}$	-0.54 (0.48)
Autobus	X_{10}	$\beta_{1,10}$	1.18 (0.52) *	$\beta_{2,10}$	0.39 (0.47)
Automobil	X_{11}	$\beta_{1,11}$	0.7 (0.48)	$\beta_{2,11}$	-0.09 (0.43)
Pol	X_{12}	$\beta_{1,12}$	-0.29 (0.28)	$\beta_{2,12}$	-0.25 (0.27)
Uzrast	X_{13}	$\beta_{1,13}$	0.02 (0.02)	$\beta_{2,13}$	0.02 (0.02)
Zemlje iz koje dolaze turisti					
Evropska Unija	X_{14}	$\beta_{1,14}$	1.18 (0.4) **	$\beta_{2,14}$	0.78 (0.38) **
Zemlje bivšeg SSSR-a	X_{15}	$\beta_{1,15}$	2.31 (0.69) **	$\beta_{2,15}$	1.55 (0.68) **
Ostatak sveta	X_{16}	$\beta_{1,16}$	14.96 (650.1)	$\beta_{2,16}$	14.32 (650.1)
Zanimanje turista					
Zaposleni u privatnoj kompaniji	X_{17}	$\beta_{1,17}$	0.54 (0.37)	$\beta_{2,17}$	0.07 (0.35)
Državni funkcioneri	X_{18}	$\beta_{1,18}$	-0.41 (0.39)	$\beta_{2,18}$	-0.67 (0.46)
Penzioneri	X_{19}	$\beta_{1,19}$	1.7 (1.13)	$\beta_{2,19}$	1.14 (1.12)
Studenti	X_{20}	$\beta_{1,20}$	0.8 (0.44) +	$\beta_{2,20}$	0.7 (0.41) +
Nezaposleni	X_{21}	$\beta_{1,21}$	-0.29 (0.73) **	$\beta_{2,21}$	0.04 (0.67)
Plata turista					
Od 900 do 3000	X_{22}	$\beta_{1,22}$	1.19 (0.35) **	$\beta_{2,22}$	1.11 (0.33) **
Preko 3000	X_{23}	$\beta_{1,23}$	0.95 (0.4) **	$\beta_{2,23}$	1.32 (0.38) **
Konstanta		$\beta_{1,0}$	-0.22	$\beta_{2,0}$	1.85

Rezultat u zagradi je standardna greška, a zvezdice označavaju nivo značajnosti (+- $p < 0.1$, * - $p < 0.05$, ** - $p < 0.01$).

U analizi ovih rezultata vidimo da broj noćenja i broj godišnjih odmora značajan, ali negativan uticaj na nivo zadovoljstva turista (Koeficijenti su -0.18 i -0.34 za veoma zadovoljnog turistu -0.25 i -0.44 za zadovoljnog turistu). Rezultati takođe pokazuju da je promenljiva Prethodno iskustvo ima značajan i pozitivan efekat da je turista veoma zadovoljan (Koeficijent je 0.72), ali je beznačajan da turista bude zadovoljan (Koeficijent je 0.55).

Među vrstama smeštaja, noćenje sa doručkom i polupansion imaju značajan, ali negativan uticaj na zadovoljstvo turista (koeficijenti za veoma zadovoljnog turistu su: -1.01 i 0.73 , dok koeficijenti za zadovoljnog turistu su: -0.14 i -0.8). Što se tiče prevoza, turisti koji su putovali autobusom izražavaju veći nivo zadovoljstva. Ostali prevozi nemaju uticaja na zadovoljstvo (Koeficijenti su: 1.18 i 0.39).

Ovo istraživanje ukazuje na to da pol i godine ne utiču na nivo zadovoljstva jer te promenljive nisu značajne (Koeficijenti su: -0.29 i 0.02 za veoma zadovoljnog turistu, odnosno -0.25 i 0.02 za zadovoljnog turistu). Što se tiče zanimanja, kod studenata se predviđa zadovoljstvo (Koeficijent su: 0.8 i 0.7). Štaviše, to što turista nije zaposlen ima negativan uticaj na to da bude veoma zadovoljan (Koeficijenti su: -0.29 i 0.04). Što se tiče zemlje prebivališta, to što je turista iz Evrope ima pozitivan i značajan uticaj na to da turista bude zadovoljan i veoma zadovoljan (Koeficijenti su: 1.18 i 0.78). Zanimljivo je da turisti iz bivšeg Sovjetskog Saveza izražavaju veći nivo zadovoljstva (Koeficijenti su: 2.31 i 1.55). Iz procenjenih rezultata možemo zaključiti da plata ima pozitivan i značajan uticaj na zadovoljstvo (Koeficijenti su: 1.19 i 1.11 za platu od 900 do 3000 Evra i 0.95 i 1.32 za platu preko 3000 Evra.).

Koristeći bazu podataka koja pruža informacije o ponašanju i zadovoljstvu turističkih putovanja tokom njihovog boravka u Crnoj Gori, sprovedena je empirijska analiza kako bi bilo shvaćeno da li su socio-demografske karakteristike povezane sa nivoom zadovoljstva turista. Učinak socio-demografskih karakteristika meri se kroz 10 dimenzija klasifikovanih u dve kategorije : karakteristike odmora (broj noćenja, prethodno iskustvo, broj odmora, vrsta smeštaja i vrsta prevoza) i turističke socio-demografske karakteristike (pol, starost država prebivališta, zanimanje i zarada). Otkrivamo da je nivo zarade pozitivno i značajno povezan sa zadovoljstvom turista. Takođe, turisti iz Evrope i bivšeg Sovjetskog Saveza veoma su zadovoljni boravkom u Crnoj Gori. S druge strane, nalazimo da broj noćenja i broj odmora imaju značajan efekat, ali sa negativnim predznakom, što dovodi do toga da turisti koji duže borave u Crnoj Gori i putuju više tokom godine nisu zadovoljni svojim boravkom u Crnoj Gori. Međutim, turisti koji su već putovali u Crnu Goru izražavaju veliku ko zadovoljstvo svojim boravkom (Perović, Stanovčić, Morić, & Sanja, 2012).

Jedna od implikacija ovih rezultata je da kreatori politike moraju da ih uključe u razvoj buduće marketinške strategije destinacija i stvaranje superiornog nivoa iskustva za različite segmente tržišta. Rezultati procene nedvosmisleno pokazuju da su turisti sa višom platom i oni koji dolaze iz Evrope i okruga bivšeg Sovjetskog Saveza veoma zadovoljni boravkom u Crnoj Gori. Međutim, ta studija je samo prvi korak ka potpunoj proceni ponašanja i zadovoljstva turističkih putovanja. Buduća istraživanja treba da razviju dopunske indikatore kako bi se procenili efekti različitih socio-ekonomskih i demografskih karakteristika na zadovoljstvo turista. Ovakva istraga mogla bi baciti

ново svetlo na pitanja koja bi mogla poboljšati ukupnu turističku strategiju Crne Gore (Perović, Stanovčić, Morić, & Sanja, 2012).

4 Zaključak

Rad sa podacima predstavlja osnovu za donošenje odluka kako ličnih, tako i poslovnih. Primena statističkih modela kao naprednih statističkih metoda pruža velike mogućnosti u razumevanju posmatranih pojava kao i predviđanja.

Regresiona analiza je jedna od najširih statističkih oblasti. U ovom radu akcenat je stavljen na logističku kao i multinominalnu regresiju koja predstavlja specijalne slučajeve generalizovanog linearnog modela.

U ovom radu korišćeni su linearni regresioni model, logit model, ordinalni logit model i nominalni logit model. Linearni regresioni model je služio za proučavanje veza između zavisne i nezavisne promenljive, gde je zavisna promenljiva neprekidnog tipa. U slučaju kada je zavisna promenljiva binarna, odnosno kada ima dva moguća ishoda, koristi se logistička regresija. Kada je zavisna promenljiva kategorijalna, ali ima više od dva moguća ishoda koji imaju određeni redosled primenjuje se ordinalni model, U slučaju kada ishodi nemaju određen redosled koristi se nominalna regresija.

Nominalni model je korišćen u primeni u turizmu. Postoji nekoliko načina primene nominalnog modela. Neki od primena su odabir prevoza sredstva kao i odabir turističke destinacije.

Turizam je važna privredna grana koja može značajno da utiče na BDP. Planiranje razvoja turizma zahteva donošenje odluka koje bi morale da budu donete na osnovu relevantnih podataka. To znači da je rad sa podacima izuzetno važan i u ovom segmentu. U master radu je dat prikaz korišćenja statističkih modela u turizmu. U prvom primeru gde se multinominalni logit model koristio jeste odabir prevoznog sredstva putnika u Norveškoj. Putnici su u anketi kad su odgovarali na pitanja birali kojim prevoznim sredstvom su putovali. Prevozna sredstva su bili privatni automobil, avio prevoz i javni transport. Multinominalni logit model je pomogao da se prouče koji faktori su uticali na to koje prevozno sredstvo će putnik da odabere. Faktori koji imaju veći uticaj na izbor prevoznog sredstva kada putnici putuju više od 400 kilometara i kada putuju više od 4 sata. Na izbor utiču i karakteristike vezane za putovanja. Putovanja do drugog doma i odlazak na skijanje imaju značaj u odabiru prevoza. Broj saputnika takođe ima uticaja na izbor. U drugom primeru gde se multinominalni logit model koristio jeste ispitivanje koji od faktora utiču na raspoloženje turista na letovanju u Crnoj Gori. Turisti su se izjašnjavali da su nezadovoljni, veoma zadovoljni i zadovoljni. Faktori koji su najviše uticali na raspoloženje turista jesu i broj odmora, da li turista ima noćenje i doručak, da li dolazi iz EU, SSSR-a. Nezaposlenost turista kao i visina plate su imali uticaj na zadovoljstvo turista.

Kao što je ranije rečeno multinominalna logistička regresija ima široku primenu, ali da bi moglo da se sprovede istraživanje potrebno je da bude dovoljno dostupnih podataka da bi istraživanje bilo kvalitetno.

Literatura

Frangos, C.C. , Karapistolis, D. , Stalidis, G. , Fragkos, C. , Sotiropoulos I. and Manolopoulos, I. (2014). Tourist Loyalty is All About Prices, Culture and the Sun: A Multinomial Logistic Regression of Tourists Visiting Athens. *Procedia - Social and Behavioral Sciences*, 175, 32-38.

Greene, W. H. (2018), *Econometric Analysis*, 8th Edition, Pearson, New York

Hausman, J., McFadden, D. „Specification Tests for Multinomial Logit Model”, *Econometrica*, 52 (septembar 1984), str. 1219-1240.

Hill, R. C., Griffiths, W. E., Lim., G. C. (2018), *Principles of econometrics*, 5th Edition, John Wiley & Sons

Hosmer, D.W. & Lemeshow, S. (1989), *Applied Logistic Regression*, John Wiley & Sons

Juschten M. and Hössinger Reinhard. (2020). Out of the city - but how and where? A mode-destination choice model for urban-rural tourism trips in Austria, *Current Issues in Tourism* <https://doi.org/10.1080/13683500.2020.1783645>

Kleinbaum, D. G., Klein, M. (2010). *Logistic Regression A Self-Learning Text*, Springer

Kmenta, J. (1971). *Elements of econometrics*, Taylor & Francis

Lozanov-Crvenković, Z. (2012). *Statistika*, Prirodno-matematički fakultet, Novi Sad.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. 2nd Ed. John Wiley & Sons, Inc

Nurlaela, S. (2018). An application of Multinomial Logit Model (MNL) on tourist destination choices. *IOP Conf. Series: Earth and Environmental Science*, 202, 012021, doi:10.1088/1755-1315/202/1/012021.

PennState Eberly College of Science, *Baseline-Category Logit Model*, The Pennsylvania State University © 2021, viewed 21 September 2021, <<https://online.stat.psu.edu/stat504/lesson/8/8.2>>

PennState Eberly College of Science, *Polytomous (Multinomial) Logistic Regression*, The Pennsylvania State University © 2021, viewed 21 September 2021, <<https://online.stat.psu.edu/stat504/lesson/8/8.1>>

PennState Eberly College of Science, *The Proportional Log-Odds Cumulative Logit Model*, The Pennsylvania State University © 2021, viewed 21 September 2021, <<https://online.stat.psu.edu/stat504/lesson/8/8.4>>

Perović, Đ., Stanovčić, T., Morić, I., Peković, S. (2012), What socio-demographic characteristics do influence the level of tourist's satisfaction in Montenegro? Empirical analysis. *Journal of tourism - studies and research in tourism*, vol. 14(14), pages 5-10

Powers, D. A. , Xie, Y. (1999). *Statistical Methods for Categorical Data Analysis*, Academic Press

Pregibon, D. (1981). Logistic Regression Diagnostics. *Annals of Statistics*, 9, 705-724

Republički zavod za statistiku, Dolasci i noćenja domaćih turista u aranžmanu domaćih turističkih agencija, © 2021 - Republički zavod za statistiku, viewed 20 September 2021, <<https://data.stat.gov.rs/Home/Result/220204?languageCode=sr-Cyrl>>

Republički zavod za statistiku, Dolasci i noćenja domaćih turista u aranžmanu domaćih turističkih agencija, © 2021 - Republički zavod za statistiku, viewed 20 September 2021, <<https://data.stat.gov.rs/Home/Result/220201?languageCode=sr-Cyrl>>

Republički zavod za statistiku, Dolasci i noćenja turista do decembra 2010. godine - mesečni podaci, © 2021 - Republički zavod za statistiku, viewed 20 September 2021, <<https://data.stat.gov.rs/Home/Result/220202?languageCode=sr-Cyrl>>

Republički zavod za statistiku, Dolasci i noćenja turista po regionima - mesečni podaci, © 2021 - Republički zavod za statistiku, viewed 20 September 2021, <<https://data.stat.gov.rs/Home/Result/220203?languageCode=sr-Cyrl>>

Thrane, C. (2015). Examining tourists' long-distance transportation mode choices using a Multinomial Logit regression model. *Tourism Management Perspectives*, 15, 115–121.

Turistička organizacija Beograda, #GoBelgrade, © 2021. Copyright by TOB.rs | SEO Optimizacija by Studio 77 | TOB Belgrade, viewed 22 September 2021, < <https://www.tob.rs/sr> >

Turistička Organizacija Srbije, Srbija, © 2021, Turistička organizacija Srbije, viewed 22 September 2021, <<https://www.srbija.travel/>>

Yoo, C., Yoon, Donghwan and Park, E.. (2018). Tourist motivation: an integral approach to destination choices. vol. 73 no. 2, pp. 169-185, © Emerald Publishing Limited, ISSN 1660-5373.

Prilog: Kodovi u R-u za multinominalne regresione modele

Na nekoliko načina možemo vektore da definišemo. Jedan način je ako podaci nisu dovoljno veliki, možemo ručno da uvedemo vrednosti.

$Y \leftarrow c(y_1, \dots, y_n)$ - Učitavanje vektora Y , dok su $y_1, \dots, y_n$ vrednosti vektora.

$X1 \leftarrow c(x_{11}, \dots, x_{1n})$ - Učitavanje vektora $X1$, dok su $x_{11}, \dots, x_{1n}$ vrednosti vektora.

$X2 \leftarrow c(x_{21}, \dots, x_{2n})$ - Učitavanje vektora $X2$, dok su $x_{21}, \dots, x_{2n}$ vrednosti vektora.

Drugi način za unos vektora jeste ako imate sačuvane podatke u fajlu.

```
library(readxl)
```

```
imefajla <- read_excel("imefajla.xlsx")
```

```
View(imefajla)
```

U ovom primeru će biti prikazane naredbe u R-u kako se mogu izračunati i prikazati koeficijenti modela. Postoji nekoliko načina za učitavanje podataka. Prvo treba kodirati neke promenljive koje su kategorične. Neke promenljive mogu da se kodiraju tako što se dodeljuju brojevi od 0 do $n - 1$ (ili od 1 do n) gde je n broj mogućnosti, ako su promenljive dummy, onda te promenljive se kodiraju tako da za svaku takvu promenljivu svaku kategoriju definišemo, ako osoba pripada toj kategoriji 1, ako ne onda je 0. Prvi korak je učitavanje potrebnih paketa što se postiže korišćenjem naredbe. `install.packages(„naziv paketa“)`, ili za starije verzije, `update.packages()`. Nakon toga je potrebno učitati bazu sa podacima.

```
library(MASS)
```

```
model1 <- polr(Y ~ X1 + ... + XK, data = imefajla, Hess = TRUE)
```

```
summary(model1)
```

Ovde će biti date naredbe za nominalnu logističku regresiju. Ovde će biti korišćena funkcija `multinom` iz paketa `nnet` da bi se ocenio nominalni model. U drugim R paketima postoje i druge funkcije sposobne za nominalnu regresiju. Odabrana funkcija je `multinom` zato što ne zahteva preoblikovanje podataka (kao što to čini `mlogit`). Prvo treba da se učitavaju podaci, zatim kodirati promenljive, pa onda treba da odaberemo referentnu kategoriju i da navedemo u funkciji `relevel`.

```
imefajla$Y <- relevel(imefajla$zavisna_promenljiva, ref = „referentna_kategorija“)
```

```
model2 <- multinom(Y ~ X1 + ... + XK, data = imefajla)
```

```
summary(model2)
```

Postoji funkcija koja računa predviđene verovatnoće za multinominalni logit model za sve nivoe.

```
model3 <- multinom(Y ~ X1 + ... + XK, data = imefajla)
```

```
pp <- fitted(model3)
```

Biografija

Miloš Gavrilović je rođen 11. juna 1995. godine. Završio je osnovnu školu „Petar Kočić” u Temerinu 2011. godine. Te godine je upisao gimnaziju „Jovan Jovanović Zmaj” u Novom Sadu, smer obdareni učenici u matematičkoj gimnaziji, koju je završio 2015. godine.

Osnovne akademske studije upisao je na Prirodno-matematičkom fakultetu u Novom Sadu, na Departmanu za matematiku, školske 2015/2016. godine, smer Matematika (M3)- modul Matematika finansija i završio ih je 2018. Iste godine je upisao master akademske studije na smeru Primenjena matematika (MB). Poslednji ispit je položio jula 2020. godine.

**UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA**

Redni broj:

RBR

Identifikacioni broj:

IBR

Tip dokumentacije: Monografska dokumentacija

TD

Tip zapisa: Tekstualni štampani materijal

TZ

Vrsta rada: Master rad

VR

Autor: Miloš Gavrilović

AU

Mentor: dr Zorana Lužanin

MN

Naslov rada: Multinominalni regresioni model sa primenom u turizmu

NR

Jezik publikacije: srpski (latinica)

JP

Jezik izvoda: srpski i engleski

JI

Zemlja publikovanja: Srbija

ZP

Uže geografsko područje: Vojvodina

UGP

Godina: 2021

GO

Izdavac: Autorski reprint

IZ

Mesto i adresa: Departman za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, Trg Dositeja Obradovica 4, Novi Sad

MA

Fizički opis rada: 4/45/25/20/1/0/0

(broj poglavlja/strana/lit.citata/tabela/slika/grafika/priloga)

FO

Naučna oblast: Matematika

NO

Naučna disciplina: Metodika matematike

ND

Predmetna odrednica/Ključne reči: linearni regresioni model, logit model, multinominalni logit model, turizam

PO

UDK:

Čuva se: Biblioteka Departmana za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

ČU

Važna napomena:

VN

Izvod: Tema ovog master rada je istraživanje primena multinominalnog logit modela u turizmu. Glavni deo rada se sastoji iz tri dela. U prvom delu prikazana teorijska osnova ekonometrijskih modela koji će biti iskorišćeni za modeliranje pojava u turizmu. Posebno će biti prikazan postupak ocenjivanja parametara i odgovarajućih testova u slučaju uopštenih linearnih modela. Logistička regresija (logit modeli) se koriste u slučaju kada je zavisna promenljiva binarna (ima samo dva ishoda, odnosno posmatrani događaj se realizovao ili ne). U drugom delu je dat teorijski deo multinominalne logističke regresije. U poslednjem delu je dat prikaz mrinene multinominalne logističke regresije na primerima u oblasti turizma. Prvi primer se odnosi na modeliranje izbora prevoza na duže relacije, gde nezavisne promenljive uključuju udaljenost putovanja u kilometrima i satima i određene socio-demografske promenljive utiču na izbor načina prevoza. Takođe će biti predstavljena mogućnost korišćenja multinominalne regresionog modela pri izboru turističke destinacije.

IZ

Datum prihvatanja teme od strane NN veća: 2.09.2021.

DP

Datum odbrane:

DO

Članovi komisije:

KO

Predsednik: dr Andreja Tepavčević, redovni profesor,
Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Mentor: dr Zorana Lužanin, redovni profesor,
Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Sanja Rapajić, redovni profesor,
Prirodno-matematički fakultet, Univerzitet u Novom Sadu

**UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
KEY WORDS DOCUMENTATION**

Accession number:

ANO

Identification number:

INO

Document type: Monograph type

DT

Type of record: Printed text

TR

Contents Code: Master's thesis

CC

Author: Miloš Gavrilović

AU

Mentor: Zorana Lužanin, Ph.D.

MN

Title: Multinomial logit regression model with application in tourism

TI

Language of text: Serbian

LT

Language of abstract: Serbian and English

LA

Country of publication: Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2021

PY

Publisher: Author's reprint

PU

Publ. place: Novi Sad, Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4

PP

Physical description: 4/45/25/20/1/0/0

(chapters/pages/literature/tables/pictures/graphics/appendices)

PD

Scientific field: Mathematics

SF

Scientific discipline: Teaching of mathematics

SD

Subject/Key words: linear regression model, logit model, multinomial logit model, tourism

SKW

UC:

Holding data: The Library of the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad

HD

Note:

N

Abstract: The topic of this master's thesis is the research of the application of the multinomial logit model in tourism. The main part of the paper consists of three parts. The first part presents the theoretical basis of econometric models that will be used to model phenomena in tourism. In particular, the procedure for estimating parameters and corresponding tests in the case of generalized linear models will be presented. Logistic regression (logit models) are used in the case when the dependent variable is binary (there are only two outcomes, ie the observed event was realized or not). The second part gives the theoretical part of multinomial logistic regression. In the last part, the presented multinomial logistic regression is presented on examples in the field of tourism. The first example relates to modeling long-distance transportation choices, where independent variables include distance traveled in kilometers and hours and certain socio-demographic variables influence the choice of transportation mode. The possibility of using a multinomial regression model when choosing a tourist destination will also be presented.

Accepted by the Scientific Board on: 2.09.2021.

ASB

Defended:

DE

Thesis defend board:

DB

President: Andreja Tepavčević, Ph.D., Full Professor,
Faculty of Sciences, University of Novi Sad
Mentor: Zorana Lužanin, Ph.D., Full Professor,
Faculty of Sciences, University of Novi Sad
Member: Sanja Rapajić, Ph.D., Full Professor,
Faculty of Sciences, University of Novi Sad