Logistička regresija u analizi smrtnosti

Bistrović, Ivana

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ivana Bistrović

Logistička regresija u analizi smrtnosti

Završni rad

Zagreb, 2017.

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO-MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Poslijediplomski specijalistički studij aktuarske matematike

Ivana Bistrović

Logistička regresija u analizi smrtnosti

Završni rad

Voditelj završnog rada: prof. dr. sc. Miljenko Huzak

Zagreb, 2017.

SADRŽAJ

1. UVOD

Regresija se upotrebljava u svrhu opisivanja i predviđanja odzivne ili zavisne varijable na temelju skupa nezavisnih varijabli. Ako je zavisna varijabla binarna, u tom slučaju logistička regresija pruža bolje rezultate od linearne pa se više i koristi. Kako se binarne varijable pojavljuju u mnogim segmentima života, logistička regresija ima široku primjenu. Otkako se pojavila pa sve do danas našla je primjene u medicini, biologiji, strojnom kodiranju, marketingu i raznim drugim poljima istraživanja. Na temelju nje se može proučiti učinkovitost lijeka, koji su vanjski utjecaji na bolesti, jesu li ispitanici zainteresirani za kupnju proizvoda i slično.

U praksi se često događa da je zavisna varijabla dihotomna, odnosno da može uzeti samo dvije vrijednosti, na primjer: muški ili ženski spol, kupovina je obavljena ili nije obavljena, osoba je živa ili nije živa. Logistička regresija je prikladna za rješavanje problema kada su u pitanju demografske varijable jer su one uglavnom kategorijske. Ona je posebno uspješna ako je kategorijska varijabla kao zavisna varijabla s jakom asimetrijom ili ako ima nelinearnu relaciju s ostalim varijablama.

Problemi ove vrste se mogu riješiti i preko višestruke linearne regresije tako što bi dvije vrijednosti varijable obilježili s dva cijela broja, obično s 0 i 1. Dobili bismo regresijski model koji bi mogao predvidjeti vrijednost zavisne varijable, zajedno s regresijskim koeficijentima koji bi pokazivali relativni utjecaj svake nezavisne varijable. Međutim, logistička regresija je bolje rešenje. S pozicije predviđanja, želimo znati u kojoj od dvije moguće skupine spada svaki ispitanik, odnosno jedinica promatranja. Preko višestruke linearne regresije dobit ćemo rješenje u kojem će zavisna varijabla imati vrijednost, negdje između 0 i 1. Predviđena vrijednost izgledat će kao vjerojatnost da će jedinica promatranja pripasti jednoj ili drugoj skupini.

Na primjer, ako je s (0) obilježen slučaj kupovine robe A, a s (1) kupovina robe B, a vrijednost zavisne varijable (kupovine) iznosi 0.65 za određenog kupca, onda je vjerojatnost veća da će kupac kupiti robu B jer je vrijednost bliža jedinici. Pretpostavka je da se kod višestruke linearne regresije dobivena vrijednost zavisne varijable u takvim slučajevima može promatrati kao vjerojatnost.

Ipak, problem koji se često javlja jest da se preko višestruke linearne regresije dobiju vrijednosti zavisne varijable koje su manje od nule ili veće od jedinice (na primjer, -0.2 ili 1.3). S obzirom da se ovakve vrijednosti ne mogu tumačiti kao vjerojatnosti, postaje jasno da predmetni model nije dobro rješenje. Potrebno je izvršiti određenu vrstu matematičke transformacije zavisne varijable u linearnom regresijskom modelu kako bi se dobio logistički regresijski model. Transformacija se izvodi tako što vrijednosti zavisne varijable postaju prirodni logaritam izgleda (eng. *natural logarithm of the odds ratio*). U pitanju je vrlo složena transformacija, ali na sreću za razumijevanje logističke regresije i njeno izvođenje uz pomoć statističkog računalnog programa njeno poznavanje nije nužno.

U ovom radu se na sažet i koncizan način teorijski i kroz praktični primjer analize smrtnosti, obradila suština logističke regresije. Rad je, uz uvod i zaključak, podijeljen u tri cjeline. U prvoj cjelini se terminološki određuju generalizirani linearni modeli i definiraju istraživački dizajn i pretpostavke logističke regresije. U drugoj cjelini se opisuje proces modeliranja i procjene koeficijenata, određivanje statističke značajnosti preko vrijednosti c-statistike i pseudo R^2 mjera te tumačenje i provjera modela logističke regresije.

U trećoj cjelini se uvodi logistički pristup modeliranja temeljen na regresiji za analizu stope smrtnosti osiguranika u Sjedinjenim Američkim Državama, uključujući jako staru dob osiguranika gdje su na raspolaganju manje vjerodostojni podaci. Kada je u pitanju provjera, na temelju modela procjene kreira se tablica iskustva koja se uspoređuje sa standardnim tablicama smrtnosti koje radi američko aktuarsko društvo (eng. *Society of Acturaries* - SOA).

2. TERMINOLOŠKO ODREĐENJE I TEORIJSKA PODLOGA

2.1. Generalizirani linearni modeli

Model logističke regresije pripada familiji generaliziranih linearnih modela (eng. *Generalized Linear Models* – GLM) koju su popularizirali McCullagh i Nelder u istoimenoj knjizi objavljenoj 1982. godine.

Generalizirani linearni model je definiran u terminima skupa međusobno neovisnih slučajnih varijabli $Y_1, Y_2, \ldots Y_n$ čije distribucije ne moraju biti jednake, ali distribucija svake od varijabli pripada eksponencijalnoj familiji i vrijede sljedeća svojstva (Barnett, Dobson, 2008.):

- distribucija od Y_i ima kanonsku formu i ovisi o samo konačno mnogo nezavisnih varijabli pri čemu je veza Y_i s pojedinom vrijednosti x_i nezavisne varijable X preko jedinstvenog parametra β i
- distribucije svih Y_i-eva su istog tipa.

Na temelju navedenih svojstava, možemo reći da su generalizirani linearni modeli definirani kroz tri komponente (Agresti, A. 2007.):

1. Slučajne komponente: ovisne varijable $Y_1, Y_2, \ldots Y_n$ koje dijele isti tip vjerojatnosti iz distribucije eksponencijalne familije,
2. Sustavne komponente: skup parametara i neovisne varijable čije mjerene vrijednosti kreiraju matricu dizajna X i
3. Funkcije poveznice ili poveznice (eng. *link function*): koja je monotona i derivabilna, a specificira vezu između slučajne i sustavne komponente, odnosno opisuje kako očekivanje ovisi o linearnom prediktoru neovisne varijable,

uz sljedeće pretpostavke:

- podaci $Y_1, Y_2, \ldots Y_n$ su neovisne slučajne varijable,
- ovisna varijabla Y_i ne mora biti normalno distribuirana, ali obično pretpostavlja distribuciju iz eksponencijalne obitelji,

- GLM ne preuzima linearni odnos između zavisne varijable i nezavisnih varijabli, ali pretpostavlja linearni odnos između transformiranog odgovora u smislu funkcije veze i objašnjavajućih varijabli,
- objašnjene varijable mogu biti čak i neke druge nelinearne transformacije izvornih nezavisnih varijabli,
- homogenost varijance ne mora biti zadovoljena (u mnogim slučajevima nije ni moguća),
- pogreške moraju biti neovisne, ali ne i normalno distribuirane,
- koristi metodu maksimalne vjerodostojnosti (eng. *maximum likelihood estimation* – MLE) umjesto metode najmanjih kvadrata (eng. *ordinary least squares* – OLS) za procjenu parametara i time se oslanja na aproksimacije velikih uzoraka i
- dobra prilagođenost modela se oslanja na dovoljno velike uzorke.

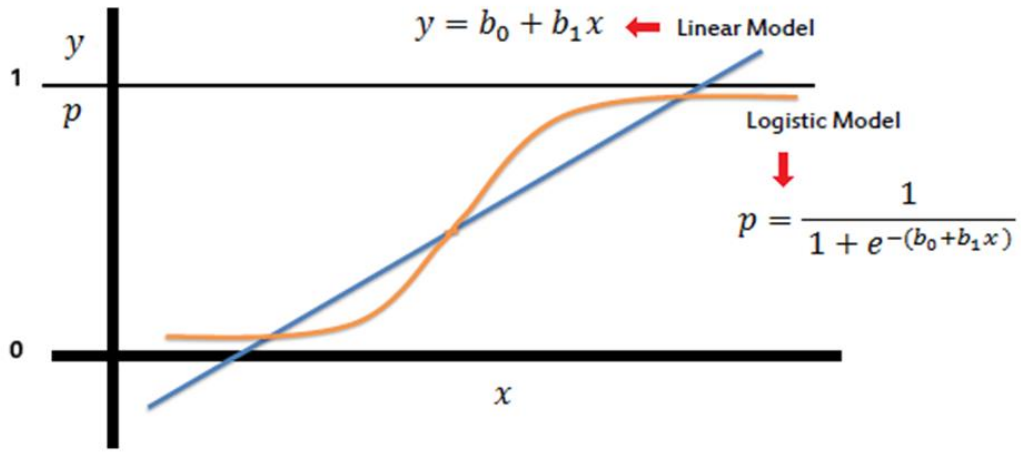
Prednosti generaliziranih linearnih modela u odnosu na tradicionalnu (OLS) regresiju su:
- ne trebamo transformirati *Y* da bi imao normalnu distribuciju,
- izbor poveznice je odvojen od odabira slučajne komponente pa imamo veću fleksibilnost u modeliranju,
- ako poveznica proizvodi aditivne učinke, tada ne trebamo konstantnu varijancu,
- parametri se procjenjuju pomoću metode maksimalne vjerodostojnosti čime se omogućuju optimalna svojstva procjenitelja,
- sve metode provjere modela za logističku regresiju vrijede i za druge generalizirane linearne modele (npr. Waldov i test omjera vjerodostojnosti, pouzdani intervali i sl.) i
- dostupni statistički računalni programi pretežito imaju samo jednu funkciju kojom se obuhvaća analiza naprijed navedenoga (npr. *PROC GENMOD* u SAS-u ili *glm* () u R-u, itd.) s mogućnosti variranja sve tri komponente.

Ali, postoje i neka ograničenja generaliziranih linearnih modela kao što su da linearna funkcija na primjer može imati samo linearni prediktor u sustavnoj komponenti te da odgovori moraju biti neovisni.

2.2. Logistička regresija

Logistička regresija odnosno logistički model se koristi za predviđanje vjerojatnosti događaja pomoću prilagođavanja podataka logističkoj krivulji prepoznatljivoj po svojemu S-obliku. Poveznica kojom je određen model logističke regresije je logit funkcija kojom se djeluje na vjerojatnost uspjeha u binomnoj distribuciji.

Slika 1. Usporedba linearnog i logističkog modela



Logistička regresija predstavlja vrstu regresijske analize u kojoj je zavisna (odzivna) varijabla dihotomna, odnosno binarna i kodira se s 0 ili 1 te postoji najmanje jedna nezavisna odnosno prediktorska varijabla. Navedeno u stvarnosti predstavlja modeliranje bilo kojeg problema kod kojeg se ciljni događaj može prevesti u kategorijsku varijablu (da/ne).

Primarna obilježja logističke regresije su sljedeća:

- logistički model se koristi kada je zavisna varijabla dihotomna,
- očekivanje zavisne varijable se transformira logit transformacijom te kod predviđanja logistički model izražava vjerojatnost da će neka jedinica opažanja ući u jednu dihotomnu skupinu umjesto u drugu te
- nezavisne varijable mogu biti kontinuirane, kategorijske, odnosno *dummy* varijable i u međusobnoj linearnoj zavisnosti, kao i kod obične regresije.

Regresijski koeficijenti kod logističke regresije se interpretiraju kao kod obične regresije – što veći koeficijent to je veći utjecaj nezavisne varijable na zavisnu, pod pretpostavkom male i nikakve kolinearnosti.

Obilježja logističke regresije možemo prikazati na primjeru logističkog modela za smrt jednog osiguranika, q = vjerojatnost štete po polici:

$$\ln\left(\frac{q}{1-q}\right) = \alpha + \beta_1 * dob + \beta_2 * spol \qquad (1)$$

s dvije nezavisne varijable: dob osiguranika kao kontinuiranu i spol osiguranika kao binarnu varijablu koja ima dvije vrijednosti: muške (1) i ženske (0) kategorije.

Najčešće korištena shema kodiranja je referentno kodiranje: kodiranje jedne kategorije kao 1, a druge kao 0 i nazvati kategoriju 0 referentnom kategorijom (npr. 1 za žena i 0 za muškarac i zatim je muškarac referentna kategorija).

Referentno kodiranje je korisno kada je primarni cilj studije usporedba smrtnosti između dvaju segmenata police. U okviru ovog kodiranja, možemo izračunati razliku od logaritama izgleda šteta između žena i muškaraca za istu dob (kontroliranje dobi, β_2)

$$ln\left(\frac{q_{ženki}}{1-q_{ženski}}\right) - ln\left(\frac{q_{muški}}{1-q_{muški}}\right) = \beta_2 \qquad (2)$$

ili koji je omjer izgleda smrti između žena i muškaraca

$$e^{\beta_2} = \left(\frac{q_{ženki}}{1-q_{ženski}}\right) / \left(\frac{q_{muški}}{1-q_{muški}}\right) \qquad (3)$$

Za kontinuirano promjenjivu dob, ako uzmemo razliku od log-izgleda između bilo koje dobi *x* i *x+1* za isti spol (kontroliranje spola, β_1), možemo izračunati omjer izgleda smrti kada dob poraste za jednu jedinicu

$$e^{\beta_1} = \left(\frac{q_{dob=x+1}}{1-q_{dob=x+1}}\right) / \left(\frac{q_{dob=x}}{1-q_{dob=x}}\right) \qquad (4)$$

Ako postavimo dob = 0 i spol = 0 (ili muški) i ovo uzmemo u obzir kao ukupnu referentnu skupinu, imamo sljedeće:

$$e^{\alpha} = \frac{q_{muški,dob=0}}{1 - q_{muški,dob=0}} \tag{5}$$

Općenito, funkcije pod (3), (4) i (5) pokazuju kako se koeficijenti logističkog modela mogu interpretirati kao omjeri izgleda referentnim kodiranjem (Schlotzhauer, 1993.):

- eksponencija koeficijenta binarne varijable predstavlja omjer izgleda nereferentne kategorije naspram referentne kategorije,
- eksponencija koeficijenta kontinuirane varijable predstavlja omjer izgleda kada se vrijednost varijable poveća za 1 jedinicu,
- eksponencija slobodnog člana predstavlja omjer izgleda referentnog slučaja (ovdje, muškarci u dobi 0) i
- rekodiranjem i transformacijom varijabli bilo koja kategorija može biti referenca.

U općenitijoj situaciji, ako kategorijska varijabla ima *k* kategorija vrijednosti $k > 2$, može se zamijeniti skupom $k - 1$ binarnih varijabli, tzv. *dummy* varijabli.

Na primjer, ako u funkciji (1) spol ima tri vrijednosti: muški, ženski i nepoznati, spol se može zamijeniti s $3 - 1 = 2$ binarnim varijablama x_1 i x_2, a ove tri kategorije mogu biti predstavljene uparenim x_1, x_2 kao:

	x_1	x_2
Ženski	1	0
Muški	0	1
Nepoznati	0	0

Ovo znači da x_1 služi kao ženski indikator, a x_2 kao muški indikator, a par (0,0) kao referenca. Model pod (1) se stoga reformira kao:

$$\ln\left(\frac{q}{1-q}\right) = \alpha + \beta_1 * dob + \beta_2 * x_1 + \beta_3 * x_2 \,. \tag{6}$$

Uzmimo u obzir logistički model za vjerojatnost štete q:

$$q = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots)}} . \quad (7)$$

Neka y bude indikator smrti, s vrijednošću 1 za smrt i 0 za biti živ, x označava vektor nezavisnih varijabli $x = (x_1, x_2, \ldots, x_k)$, β vektor koeficijenata $\beta = (\alpha, \beta_1, \beta_2, \ldots, \beta_k)$, a $q = Prob\,(y = 1|x)$ je funkcija od x. Pretpostavimo da imamo uzorak od n nezavisnih opažanja parova (x_i, y_i), i = 1, …, n. S obzirom da je vjerojatnost opažanog y_i s danim x_i jednak

$$q_i^{y_i}(1-q_i)^{1-y_i} , \quad (8)$$

zajednička vjerojatnost svih n opažanja je produkt ovih vjerojatnosti

$$l(\beta) = \prod_{i=l}^{n} q_i^{y_i}(1-q_i)^{1-y_i} . \quad (9)$$

Funkciju $l(\beta)$ zovemo vjerodostojnost od β.

Budući da je točka maksimuma log vjerodostojnosti identična točki maksimuma vjerodostojnosti, umjesto da maksimiziramo vjerodostojnost maksimizirat ćemo njen logaritam zbog jednostavnosti računa. Stoga, kako bismo našli β koji maksimizira vjerodostojnost funkcije tražimo β koji povećava log vjerodostojnost.

$$L(\beta) = \ln\big(l(\beta)\big) = \sum\{y_i \ln(q_i) + (1-y_i)\ln(1-q_i)\} \quad (10)$$

Rješenja dobivenih jednadžbi nije moguće dobiti u zatvorenoj formi već se rješavaju numerički, najčešće Newton-ovim iterativnim postupkom.

2.3. Istraživački dizajn logističke regresije

Upotreba logističke regresije se može promatrati uz šestorazinsku modelno-razvojnu perspektivu. Kao i kod svih multivarijatnih primjena, postavljanje ciljeva je prvi korak u procesu analize. Istraživač mora naznačiti specifične značajke dizajna i potvrditi zadane pretpostavke.

Analiza se nastavlja procjenom mogućnosti pojavljivanja u svakoj od skupina pomoću upotrebe logističke krivulje kao nezaobilaznog odnosa. Binarna mjera je prevedena u mogućnosti pojavljivanja, a zatim u logaritamsku vrijednost koja se ponaša kao zavisna mjera. Na kraju, model se potvrđuje putem zadržanog (eng. *holdout*) uzorka, odnosno dijela uzorka koji nije korišten u procesu modeliranja (Harrell, Frank, 2001.).

Logistička regresija posjeduje nekoliko obilježja koja utječu na istraživački dizajn. U nastavku se opisuju neka od tih obilježja.

- Binarna zavisna varijabla

Logistička regresija predstavlja dvije skupine interesa kao binarne varijable s vrijednostima 0 i 1. Nije bitno kojoj skupini je dodijeljena vrijednost 0 ili 1, ali zadatak mora biti naznačen u svrhu interpretacije koeficijenata (Trusty, 2000.).

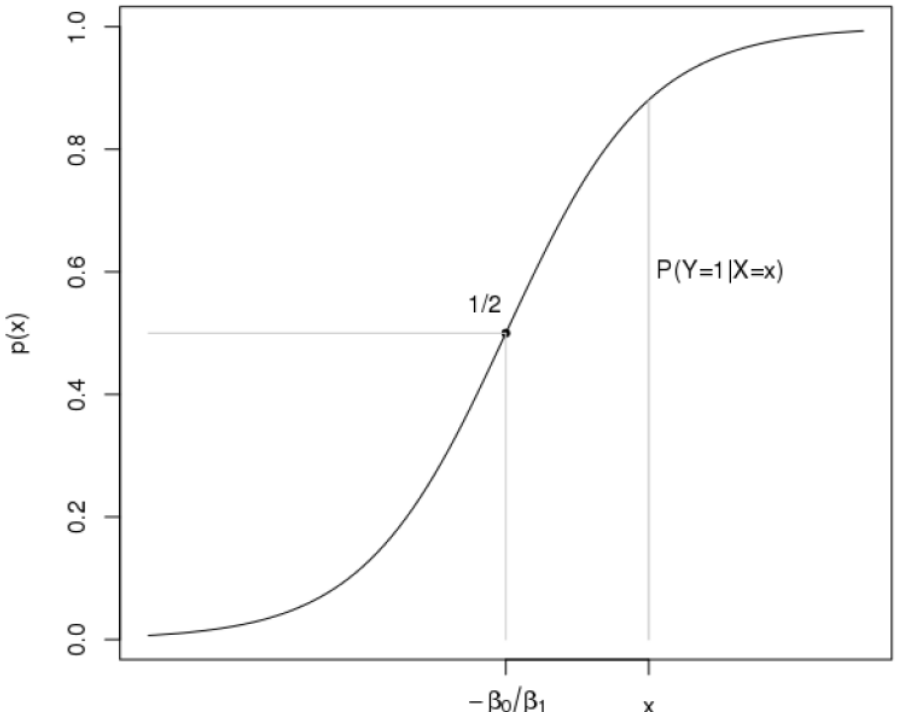
Ako skupine predstavljaju značajke (npr. spol) svakoj skupini se može dodijeliti vrijednost 1 (npr. ženama) i drugoj skupini vrijednost 0 (muškarci). U takvoj prilici koeficijenti bi odavali utjecaj nezavisnih varijabli na vjerojatnosti osobe koja je ženskog spola (skupina kodirana brojem 1).

Ako skupine prezentiraju događaje (npr. uspjeh ili neuspjeh), zadatak skupnih kodova također utječe na interpretaciju. Pretpostavimo da je skupina s uspjehom kodirana kao 1, a s neuspjehom s 0, tada koeficijenti predstavljaju utjecaje na vjerojatnosti uspjeha.

➢ Upotreba logističke krivulje

Logistička regresija koristi logističku krivulju za definiranje odnosa između nezavisnih i zavisnih varijabli. Uz pretpostavku pozitivne koreliranosti uspjeha s nezavisnim varijablama, na vrlo niskim razinama vrijednosti nezavisne varijable vjerojatnost doseže do 0, ali je nikada ne dostigne. Slično, s porastom vrijednosti nezavisne varijable vjerojatnost podiže krivulju, ali onda nagib počinje opadati tako da će uz povećanje razine nezavisne varijable vjerojatnost težiti do 1, ali nikada je ne dostignuti.

Slika 2. Logistička krivulja



➢ Osobina zavisne varijable

Zavisna varijabla (0 ili 1) prati Bernulijevu umjesto normalne distribucije i varijanca dihotomne varijable nije konstantna. Logistička regresija je razvijena specifično za modeliranje i rješavanje ovakvih situacija. Njezin jedinstveni odnos između zavisnih i nezavisnih varijabli zahtijeva drugačiji pristup u procjenjivanju varijable i interpretiranju koeficijenata.

- Veličina uzorka

Idealno je imati što više podataka po mjerenoj veličini, odnosno što veći uzorak. U stvarnosti su podaci često nedostupni pa istraživač treba razmotriti koji bi najmanje mogući uzorak bio dostatan za kvalitetno modeliranje.

Logistička regresija koristi metodu maksimalne vjerodostojnosti za procjenu parametara i time se oslanja na aproksimacije velikih uzoraka pa se i dobra prilagođenost modela oslanja na dovoljno velike uzorke slijedom čega će logistička regresija zahtijevati veću veličinu ukupnog uzorka, npr. veličine uzorka veće od 400 (Hosmer, Stanley, Rodney, 2013.).

Također treba uzeti u obzir da se ukupan uzorak dijeli na dva dijela: dio za prilagodbu modela i dio za testiranje modela (zadržani uzorak). U kreiranju ovog razdvajanja uzorka, potrebe veličine uzorka i dalje stoje za oba dijela zasebno, čime se udvostručava ukupna potrebna veličina uzorka zasnovana na specifikaciji modela (Trusty, 2000.).

2.4. Pretpostavke logističke regresije

Prednosti logističke regresije u usporedbi s drugim modelima je malobrojnost pretpostavki potrebnih za modeliranje i analizu. Logistička regresija ne zahtijeva nikakav poseban distribucijski oblik nezavisnih varijabli i ne zahtijeva linearne odnose između nezavisnih i zavisnih varijabli. Ona se odnosi na nelinearne učinke čak i kada eksponencijalni i polinomijalni termini nisu eksplicitno dodani kao dopunske nezavisne varijable, zbog logističkog odnosa.

Jedna od posebnih značajki logističke regresije je upotreba ranije opisanih logističkih veza kod obje procjene, procjene logističkog modela i procjene povezanosti između zavisnih i nezavisnih varijabli (Vaupel, 2014.).

3. MODELIRANJE I TESTIRANJE LOGISTIČKE REGRESIJE

3.1. Logistička transformacija

Kada je u pitanju višestruka linearna regresija predviđa se metrički zavisna varijabla, a isto se događa i kod logističke regresije, no u njenom slučaju vjerojatnost varijable ograničena je na interval između 0 i 1. Ali, kako se može sa sigurnošću tvrditi da procijenjena vrijednost ne pada izvan vrijednosti ovog intervala?

Postupak logističke transformacije se provodi kroz dva koraka (Thatcher, 1999.).

Prvi korak: iskazivanje vjerojatnosti kao izgleda (eng. *odds*).

U njihovom originalnom obliku, vjerojatnosti su ograničene na vrijednost između 0 i 1. Međutim, što ako preformuliramo vjerojatnost na način gdje nova varijabla neće uvijek biti u intervalu između 0 i 1? Može se preformulirati iskazivanjem vjerojatnosti kao izgleda: omjer vjerojatnosti dva međusobno suprotna ishoda ili događaja, *Probi*/(1- *Probi*). U ovom obliku, bilo koja vrijednost vjerojatnosti je svedena na kontinuiranu pozitivnu vrijednost koju možemo direktno procijeniti. Svaka vrijednost izgleda može biti ponovo pretvorena u vjerojatnost koja se nalazi između 0 i 1.

U ovom slučaju se koristi primjer vjerojatnosti uspjeha ili neuspjeha kako bi se prikazalo izračunavanje koeficijenta. Ako je vjerojatnost uspjeha 0.80, onda je poznato da je i vjerojatnost suprotnog događaja/neuspjeha 0.20 (1-0.80). Ova vjerojatnost znači da su izgledi uspjeha 4.0 (0.80/0.20) odnosno da je uspjeh 4 puta vjerojatniji od neuspjeha. U suprotnom slučaju, izgledi neuspjeha su 0.25 (0.20/0.80).

Možemo zaključiti da vjerojatnost od 0.50 rezultira izgledima od 1.0 (oba događaja imaju iste izglede da se dogode). Izgledi manji od 1.0 predstavljaju vjerojatnosti manju od 0.50, a izgledi veći od 1.0 odgovaraju vjerojatnostima većima od 0.50. Dakle, javlja se kontinuirana numerička varijabla koja uvijek može biti konvertirana u vjerojatnost unutar intervala 0 i 1.

Drugi korak: računanje log-vrijednosti (eng. *log odds values*).

Vrijednosti izgleda rješavaju problem procjene vjerojatnosti između 0 i 1, ali sada se javlja drugi problem. Kako sačuvati izglede od padanja ispod 0, što je najniža granica kod izgleda? Rješenje je izračunati log-vrijednost, koja se računa uzimanjem logaritma izgleda. Takvu transformaciju vjerojatnosti nazivamo logit od te vjerojatnosti.

Izgledi manji od 1.0 će imati negativnu log-vrijednost, izgledi veći od 1.0 će imati pozitivnu log-vrijednost, a izgledi od 1.0 imaju log-vrijednost od 0. Štoviše, bez obzira na to koliko su niske negativne vrijednosti i dalje mogu biti transformirane uzimanjem antilogaritma na vrijednost izgleda većeg od 0.

Tablica 1. Tipične vrijednosti vjerojatnosti, izgleda i log-vrijednosti

Vjerojatnosti	**Izgledi**	**Log-vrijednosti**
0.00	0.00	NI
0.10	0.111	-2.197
0.30	0.428	-0.847
0.50	**1.000**	**0.000**
0.70	2.333	0.847
0.90	9.000	2.197
1.00	NI	NI

NI = nemoguće izračunati

Pomoću log-vrijednosti se dobiva varijabla koja može posjedovati pozitivne i negativne vrijednosti i koja se uvijek može pretvoriti u vrijednosti vjerojatnosti u rasponu od 0 do 1. Ali, logit vrijednost nije moguće primijeniti na vjerojatnost od točno 0 ili 1. Ova vrijednost predstavlja zavisnu varijablu modela logističke regresije, a interpretira se kao logit-transformacija vjerojatnosti uspjeha.

3.2. Procjena modela

Postupak procjene logističkih koeficijenata je sličan kao kod linearne regresije, osim što se u ovom slučaju koriste samo dvije vrijednosti za zavisnu varijablu (0 i 1). Umjesto korištenja metode najmanjih kvadrata kao sredstva procjene modela, koristi se metoda maksimalne vjerodostojnosti.

Korištenjem procijenjenih koeficijenta za zadane vrijednosti nezavisne varijable možemo procijeniti odgovarajuću logit vrijednost

$$Logit_i = log\left(\frac{prob_{event}}{1 - prob_{event}}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n \tag{11}$$

ili procijeniti izglede uspjeha

$$Odds_i = \left(\frac{prob_{event}}{1 - prob_{event}}\right) = e^{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}. \tag{12}$$

Jednadžbe za oba modela su ekvivalentne, ali izbor između ove dvije jednadžbe odražava se i na procjenu koeficijenta. Mnogi statistički računalni programi omogućavaju izračunavanje logističkih koeficijenata pomoću obje jednadžbe. Ovaj proces može smjestiti jednu ili više nezavisnih varijabli i nezavisna varijabla može biti kontinuirana numerička ili ne (*dummy*).

3.3. Ocjena dobrote prilagodbe

Dobrota prilagodbe (eng. *goodness of fit*) logističkog modela može biti ocijenjena na nekoliko načina prema različitim pogledima, ali sa sličnim rješenjima. Jedan način je ispitivanje točnosti predviđanja pomoću statističkih testova, a drugi način je pomoću mjera snage predviđanja, odnosno statistike koja mjeri koliko dobro se može procijeniti zavisna na temelju nezavisne/-ih varijabli.

Osnovna mjera koja pokazuje koliko dobro procjena maksimalne vjerodostojnosti odgovara opaženim vrijednostima zavisne varijable odgovara dvostrukoj vrijednosti logaritamske vjerodostojnosti *-2LL* (eng. *two times log likelihood*). Minimalna vrijednost za *-2LL* je 0, što odgovara savršenom prilagođavanju (*likelihood* = 1 i *–2LL* = 0). Što je niža vrijednost *-2LL*, model je bolje prilagođen. Vrijednost *-2LL* se može koristiti za uspoređivanje jednadžbi ugniježđenih modela (eng. *nested models*) ili izračunavanje mjera usporedivih s R^2 mjerama u višestrukoj linearnoj regresiji.

Prilagodbe se mogu uspoređivati pomoću razlike log-vjerodostojnosti nultog i prilagođenog modela. Osnovni pristup prati tri koraka (Schlotzhauer, 1993.):

1. Ocjena nultog modela.

Prvi korak je ocijeniti nulti model koji predstavlja osnovu za usporedbu poboljšanja u prilagođenom modelu. Najčešće korišteni nulti model je model bez nezavisnih varijabli, koji je analogan računanju zbroja kvadrata reziduala korištenjem prosjeka kod višestruke linearne regresije. Logika korištenja ovog oblika modela je da predstavlja osnovu s kojom se bilo koji ugniježđeni model koji sadrži nezavisne varijable može usporediti.

2. Ocjena predloženog modela.

Ovaj model sadrži nezavisne varijable uključene u model logističke regresije. Model će se poboljšati u formi od nultog modela i rezultirati u nižoj *-2LL* vrijednosti. Bilo koji broj predloženih modela se može procijeniti (modeli s jednom, dvije i tri nezavisne varijable mogu se uzeti kao odvojeni predloženi modeli).

3. Procjena *-2LL* razlike.

Zadnji korak je procjena statističke značajnosti razlike *-2LL* vrijednosti između dva modela (nultog i predloženog modela) pri čemu je nul-hipoteza (H_0) da je nulti model točan u odnosu na alternativnu hipotezu (H_1) da je točan barem predloženi model. Ako statistički test rezultira odbacivanjem H_0 u korist H_1 na zadanoj razini značajnosti, tada se može reći da je skup nezavisnih varijabli u predloženom modelu značajan za poboljšanje procjene modela.

Na sličan način mogu se uspoređivati bilo koja dva predložena ugniježđena modela. U tom slučaju, značajnost razlika njihovih log-vjerodostojnosti ukazuje na potrebu korištenja složenijeg modela. Na primjer, model s dvije nezavisne varijable može se uspoređivati s modelom s tri nezavisne varijable za procjenu značajnosti poboljšanja dobivenih dodavanjem jedne nezavisne varijable. U ovakvim slučajevima, jednostavniji model je označen kao nulti model i uspoređuje se s drugim, složenijim modelom (Field, 2005.).

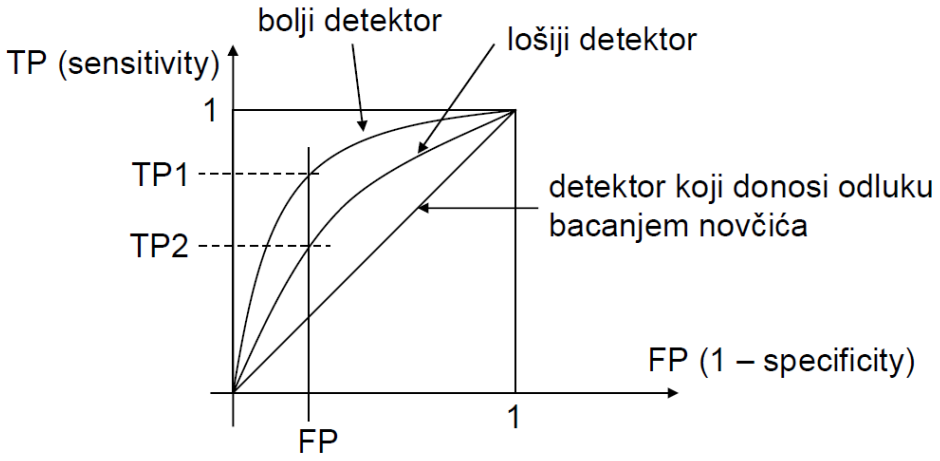
3.4. C-statistika i pseudo R^2 mjere

Jedna od standardnih mjera za ocjenu dobrote prilagodbe je c-statistika ili AUC (eng. *concordance statistic* ili *area under curve*) koja predstavlja površinu ispod ROC krivulje (eng. *Receiver Operating Characteristic curve*), a čije se vrijednosti kreću u rasponu od 0.5 do 1. Za modele s vrijednosti c-statistike manjom do 0.5 reći ćemo da su loše prilagođeni. Vrijednost c-statistike od 0.5 govori da model ne predviđa bolje od slučajnog izbora npr. bacanjem simetričnog novčića. Vrijednost c-statistike veća od 0.7 pretpostavlja dobro prilagođen model, dok vrijednost c-statistike veću od 0.8 imaju izvrsno prilagođeni modeli. Vrijednost c-statistike od 1 znači savršenu prilagodbu modela (Hosmer, Lemeshow 1980.).

ROC krivulja kombinira osjetljivost i specifičnost detektora za ocjenu dobrote prilagodbe modela. Osjetljivost detektora (eng. *sensitivity*) je jednaka vjerojatnosti korektno odabrane hipoteze H_1 (engl. *true positive* - TP), a specifičnost detektora (eng. *specificity*) je jednaka vjerojatnosti korektno odabrane hipoteze H_0 (eng. *true negative* - TN) pri čemu je hipoteza (H_1) da je dobar predloženi model (M_1) u odnosu na nul-hipotezu da M_1 nije dobar model.

Za neki detektor, u našem slučaju test značajnosti modela, ROC krivulja pokazuje vjerojatnost korektno odabrane hipoteze H_1 (TP) u ovisnosti o vjerojatnosti pogrešno odabrane hipoteze H_1 (engl. *false positive* - FP). Dakle kod ROC krivulje nanosimo na apscisu: FP (1-specifičnost), a na ordinatu: TP (osjetljivost), a bolji je onaj detektor koji za neki fiksni FP ima veći TP, odnosno što je veća površina ispod ROC krivulje.

Slika 3. Prikaz ROC krivulje na različitim razinama dobrote prilagodbe modela



Također, razvijeno je nekoliko mjera sličnih R^2 mjerama koje su dostupne u raznim statističkim računalnim programima. Ove pseudo R^2 mjere se tumače na način sličan koeficijentima determinacije u višestrukoj linearnoj regresiji. Pseudo R^2 za logistički model se računa na sljedeći način

$$R^2_{LOGIT} = \frac{-2LL_{nul} - (-2LL_{model})}{-2LL_{nul}}, \qquad (13)$$

gdje je $-2LL_{nul}$ log-vjerodostojnost nultog modela.

Kao i kod višestruke linearne regresije, logistička R^2 vrijednost se kreće u intervalu od 0.0 do 1.0. Kada predloženi model povećava prilagođenost podacima, vrijednost *-2LL* se smanjuje. Savršenu prilagodbu ima *–2LL* vrijednost od 0.0 i R^2_{LOGIT} od 1.0.

U literaturi (vidi Field, 2005.) spominje se Coxova i Snellova R^2 mjera koje se interpretiraju na isti način: vrijednosti bliže 1.0 tih statistika prikazuju bolje prilagođen model. Obje mjere odražavaju takav iznos varijacija procijenjenih logističkih modela, gdje bi vrijednost 1.0 predstavljala savršeno prilagođeni model.

Hosmer i Lemeshow su razvili klasifikacijski test gdje su slučajevi najprije podijeljeni na približno deset jednakih kategorija (vidjeti u McCullagh, Nelder, 1989.). Broj opaženih i predviđenih događaja uspoređuje se u svakoj kategoriji s hi-kvadrat statistikom. Ovaj test se ne temelji na vrijednosti vjerodostojnosti, već na usporedbi opaženih i predviđenih vrijednosti zavisne varijable. Prikladno korištenje ovog testa zahtijeva uzorak veličine najmanje 50 opažanja da bi bili sigurni kako svaka kategorija ima najmanje 5 opažanja, ali dopušta se da opažena frekvencija kategorija na repu/repovima bude barem 1.

Radi usporedbe, uz mjere za ocjenu modela logističke regresije, u tablici 2. su navedene slične (analogne) mjere za ocjenu modela višestruke linearne regresije.

Tablica 2. Usporedba primarnih elemenata prilagođenog modela

Višestruka linearna regresija	**Logistička regresija**
Ukupan zbroj kvadrata	*-2LL* nulti model
Zbroj kvadrata reziduala	*-2LL* predloženi model
Zbroj kvadrata zbog djelovanja regresije	Razlika *-2LL* nultog i predloženog modela
F-test prilagođenosti modela	Hi-kvadrat test *-2LL* razlika
Koeficijent determinacije (R)	Pseudo R^2 mjera

3.5. Tumačenje modela

Statističko modeliranje binarnih varijabli odziva podrazumijeva mjerenje izbora koje za svaki subjekt može biti uspješno ili neuspješno. Za binarnu varijablu Y i kvantitativnu nezavisnu varijablu X, neka $\pi(x)$ predstavlja vjerojatnost uspjeha, tj. vjerojatnost ishoda $Y = 1$, kada X ima vrijednost x. Ova vjerojatnost je parametar za binomnu distribuciju. Model logističke regresije ima linearni oblik za logit ove vjerojatnosti

$$\log it[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x. \qquad (14)$$

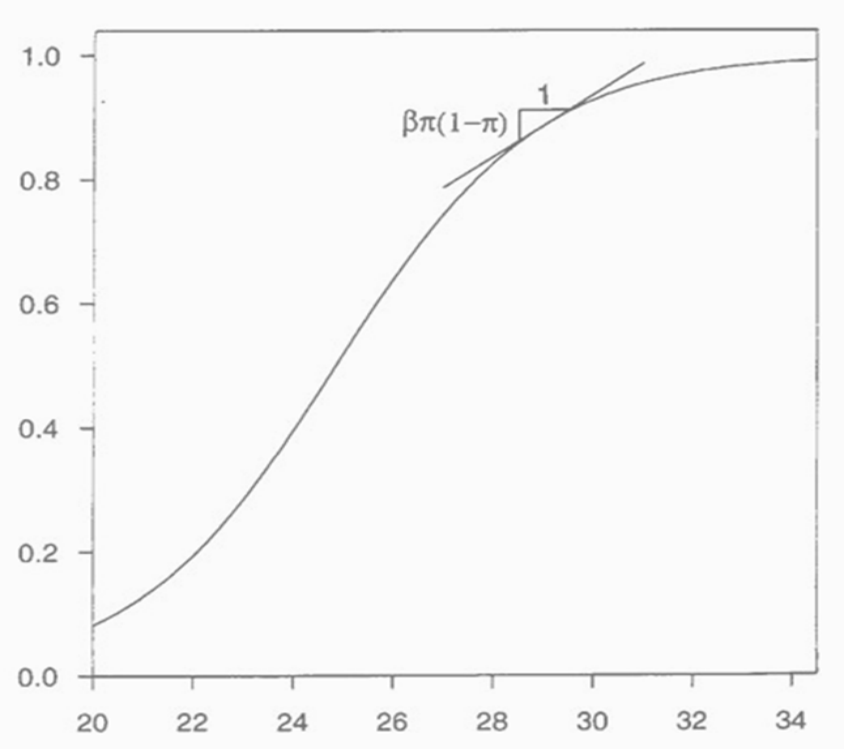
Ova jednadžba prikazuje kako $\pi(x)$ raste ili opada kao funkcija od x. S obzirom da model logističke regresije ima karakterističan „S“ oblik, što je prikazano na slici 4., $\pi(x)$ se još naziva i S-krivulja.

Druga jednadžba za logističku regresiju odnosi se izravno na vjerojatnost uspjeha. Ova jednadžba koristi eksponencijalnu funkciju $\exp(x) = e^x$ u obliku

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \qquad (15)$$

Parametar β određuje stopu rasta ili opadanja S-krivulje. Oznaka β ukazuje na to je li krivulja padajuća ili rastuća, kao i na stopu rasta promjene kako $|\beta|$ raste. Kada model ima vrijednost $\beta = 0$, desna strana jednadžbe pojednostavljuje se u konstantu, $\pi(x)$ je identičan za sve x te krivulja prelazi u horizontalnu ravnu liniju. Vjerojatnost uspjeha ($Y = 1$) postaje pritom konstanta neovisna o X (Trusty, 2000.).

Slika 4. Linearna aproksimacija logističke regresijske krivulje



Slika 4. prikazuje S-oblik modela logističke regresije za $\pi(x)$. Budući da ova funkcija ima zakrivljeni, a ne pravolinijski izgled, zaključuje se kako stopa promjene u $\pi(x)$ po jedinici promjene u x varira. Pravac koji predstavlja tangentu na krivulji za danu vrijednost x prikazuje stopu promjene u toj točki. Za istu vrijednost parametra β logističke regresije, taj pravac ima jednaki nagib za isti $\pi(x)$.

Na primjer, pravac tangente na krivulju za vrijednost x kod koje je $\pi(x) = 0.5$ ima nagib $\beta(0.5)(0.5) = 0.25\beta$. S druge strane, kada je $\pi(x) = 0.9$ ili 0.1, nagib iznosi 0.09β. Nagib se približava vrijednosti 0.0 kako se vjerojatnost približava vrijednosti 1.0 ili 0.0.

Najoštriji nagib krivulje se događa za vrijednost x kada je $\pi(x) = 0.5$. Tada je vrijednost $x = -\alpha/\beta$ što se ponekad naziva srednjom razinom učinkovitosti i označava se s EL50. Njime se prikazuje razina kod koje uspjeh ima vjerojatnost 50% (Vaupel, 2014).

Testiranje značajnosti koeficijenata

Logistička regresija testira hipoteze o pojedinačnim koeficijentima kao što je slučaj i kod višestruke linearne regresije. U višestrukoj linearnoj regresiji, statistički test se provodi kako bi se utvrdilo je li koeficijent različit od 0, jer ako je 0 onda nema utjecaja na zavisnu varijablu. U logističkoj regresiji također se koriste statistički testovi kako bi ustanovilo je li logistički koeficijent β različit od 0.

Međutim, u logističkoj regresiji korištenje logit kao zavisne mjere, vrijednost od 0 odgovara izgledu od 1.0 ili vjerojatnosti od 0.50 koja pokazuje da je vjerojatnost ista za obje grupe (Schlotzhauer, 1993.).

U višestrukoj linearnoj regresiji se koristi odgovarajuća studentova testna statistika (vrijednost t) za ocjenu značajnosti svakog koeficijenta. Logistička regresija koristi drugačije mjere, točnije Waldov test. Ovaj test omogućava testiranje statističke značajnosti za svaki koeficijent. Iz ranijeg procesa procjene se zna da su koeficijenti (β_1, β_2 , …, β_k) ustvari mjere promjene u intervalu vjerojatnosti. Međutim, logističke koeficijente teško je tumačiti u njihovom originalnom obliku jer su izraženi u obliku logaritma onda kada koristimo logit kao zavisnu mjeru.

Većina statističkih računalnih programa omogućava izračunavanje eksponencijalnog logističkog koeficijenta koji se koristi kao transformacija originalnog logističkog koeficijenta. Na taj način se za tumačenje može koristiti ili originalni ili eksponencijalni logistički koeficijent. Dva tipa logističkih koeficijenata se razlikuju u tome što odražavaju povezanost nezavisne varijable s dva oblika vjerojatnosti uspjeha zavisne varijable (Olshansky, 1998.).

Logistički koeficijent	**Odražava promjenu u:**
Izvorni	Logit (log izgledi)
Eksponencijalni	Izgledi

Pouzdani intervali za parametar β

Neka je β' procjenitelj (procjena) parametra β dobiven metodom maksimalne vjerodostojnosti. Pouzdani interval za parametar β u modelu logističke regresije, $\text{logit}[\pi(x)] = \alpha + \beta x$, je

$$\beta' \pm z_{\alpha/2}(ASE) \tag{16}$$

pri čemu ASE predstavlja asimptotičku standardnu pogrešku od procjenitelja (β'_{MLE}).

Kod modela logističke regresije, nulta hipoteza $H_0 : \beta = 0$ znači da je vjerojatnost uspjeha nezavisna od X. Kod većih uzoraka, testna statistika

$$z = \frac{\beta'}{ASE} \tag{17}$$

ima asimptotski standardnu normalnu distribuciju kada je $\beta = 0$.

Iako Waldov test dobro funkcionira kod velikih uzoraka, test omjera vjerodostojnosti je učinkovitiji i pouzdaniji za veličine uzorka koje koristimo u praksi.

Testna statistika uspoređuje maksimalni $-2LL_0$ log-funkcije vjerodostojnosti kada je $\beta = 0$ (to jest, kada $\pi(x)$ mora biti identična neovisno o vrijednostima x) do maksimalne $-2LL_1$ log-funkcije vjerodostojnosti za nerestriktivnu β. Testna statistika $-2LL_0 - (-2LL_1)$, što se kraće može zapisati kao $-2(L_0 - L_1)$, ima asimptotsku hi-kvadratnu distribuciju s jednim stupnjem slobode ako je točna hipoteza $\beta=0$.

Većina statističkih računalnih programa računa statistike maksimalne log-vjerodostojnosti $-2LL_0$ i $-2LL_1$, a statistika omjera vjerodostojnosti se dobiva iz ovih maksimuma (McCullagh, Nelder, 1989.).

Procijenjena vjerojatnost za $Y = 1$ pri fiksnoj vrijednosti nezavisne varijable x iznosi

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \qquad (18)$$

Većina statističkih računalnih programa prikazuje procjene kao i pouzdane intervale za prave vjerojatnosti.

Pouzdani intervali za vjerojatnost se mogu izvesti korištenjem matrice kovarijance procijenjenih parametara. Neka su α' i β' procjenitelji parametara α i β dobiveni metodom maksimalne vjerodostojnosti. Statistika $\alpha' + \beta' x$ u eksponentima jednadžbe predviđanja je procijenjeni linearni prediktor u logit transformaciji $\pi(x)$. Ta statistika ima asimptotsku standardnu pogrešku izračunatu iz asimptotske standardne pogreške za α i β te njihove asimptotske kovarijance izračunate kao

$$ASE(\alpha' + \beta' x) = \sqrt{Var(\alpha') + x^2 Var(\beta') + 2x Cov(\alpha', \beta')} \qquad (19)$$

Interval 95%-tne pouzdanosti za pravi logit je $(\alpha' + \beta' x) \pm 1.96ASE$ gdje je $ASE = ASE(\alpha' + \beta' x)$.

3.6. Provjera modela

Modeli logističke regresije se koriste za predviđanje vjerojatnosti da je $Y = 1$ za zadanu vrijednost x regresijske varijable X. Za svaku vrijednost x_i od X predviđena vjerojatnost ($\pi'(x_i)$) pomnoži se s brojem subjekata kako bi se dobila predviđena ili očekivana frekvencija uspjeha (realizacija događaja $Y = 1$). Slično tome, može se dobiti predviđena frekvencija za $Y = 0$ za svaki x_i. Test nulte hipoteze da je predloženi model točan uspoređuje očekivane i opažene frekvencije uz pomoć Pearsonovog X^2 testa ili testa omjera vjerodostojnosti G^2.

Bez obzira na to koliko podataka ima po kategorijama opisanih s raznim x_i od X, opažene i očekivane vrijednosti se mogu podijeliti prema predviđenim vjerojatnostima, odnosno formirati u skupine približno jednake veličine. Ako bi se formiralo 10 skupina, na primjer, jedan par opaženih i očekivanih vrijednosti se odnosi na $n/10$ opažanja koje imaju predviđenu vjerojatnost u istom intervalu vrijednosti, sljedeći par se odnosi na $n/10$ opažanja koji imaju predviđenu vjerojatnost u drugom intervalu vrijednosti i tako dalje.

Ovaj postupak je osnova testa prema Hosmeru i Lemeshowu. Njihova statistika zapravo nema hi-kvadratnu distribuciju, ali su simulacije pokazale kako je njihova distribucija približna hi-kvadratnoj distribuciji sa stupnjevima slobode $df = g - 2$, gdje g označava broj skupina (Hosmer, Stanley, Rodney, 2013.).

Statistika omjera vjerodostojnosti $-2(L_0 - L_1)$ se koristi za testiranje jesu li određeni parametri u nekom modelu jednaki 0. Ovim testom se usporedi maksimalna log-vjerodostojnost $(-2LL_1)$ za složeniji model s maksimalnom log-vjerodostojnosti $(-2LL_0)$ za jednostavniji model kojim se brišu neki parametri iz prvog modela. S M_1 može se označiti prilagođeni složeniji model, a s M_0 nulti model za koji su neki parametri 0 (Long, 1997.).

Statistika kvalitete prilagodbe $G^2(M)$ za testiranje prilagodbe modela logističke regresije M je specijalni slučaj statistike omjera vjerodostojnosti po kojem je $M_0 = M$, a M_1 je najsloženiji mogući model. Ovaj složeni model ima odvojeni parametar za svaki logit i pruža savršenu prilagodbu za logit uzorke. Takav model se naziva zasićeni model.

Neka *Ls* označava maksimalnu log-vjerodostojnost zasićenog modela, a L_M modela *M* tada je $G^2(M) = -2(L_M - L_S)$. Tako, na primjer, odstupanja za modele M_0 i M_1 iznose $G^2(M_0) = -2(L_0 - L_S)$ i $G^2(M_1) = -2(L_1 - L_S)$.

S $G^2(M_0 \mid M_1)$ možemo označiti statistiku omjera vjerodostojnosti za testiranje modela M_0 pod uvjetom da je M_1 prošireni funkcionalni model. Stoga, testna statistika za usporedbu ova dva modela iznosi

$$G^2\left(M_0|M_1\right) = -2\left(L_0 - L_1\right) = -2\left(L_0 - L_S\right) - \left[-2\left(L_1 - L_S\right)\right] = G^2\left(M_0\right) - G^2\left(M_1\right) \qquad (20)$$

kao razlika G^2 statistike kvalitete prilagodbe za ova dva modela. Drugim riječima, statistika omjera vjerodostojnosti za usporedbu ova dva modela je jednostavno razlika u odstupanjima ova dva modela. Ova statistika je velika kada je M_0 lošije prilagođen u usporedbi s M_1. Radi se o statistici s asimptotskom hi-kvadratnom distribucijom, sa stupnjevima slobode jednakima razlici između ostatka stupnjeva slobode vrijednosti za ova dva modela, a sve za slučaj istinitosti nultog modela.

4. PRIMJENA LOGISTIČKE REGRESIJE U ANALIZI SMRTNOSTI

Osiguravajućim društvima je za potrebe vrednovanja očekivanih obveza i održavanje potrebne solventnosti kapitala važna dobra procjena smrtnosti osiguranika. U praksi, postupak procjene smrtnosti osiguranika se susreće sa širokim rasponom podataka i analitičkim izazovima.

Za analizu smrtnosti osiguranika životnih osiguranja jednako su važna tri aspekta:

1. Trend smrtnosti: kako se smrtnost povećava ili smanjuje tijekom vremena,
2. Nagib krivulje smrtnosti: kako se smrtnost povećava po dobi ili trajanju i
3. Različitosti u smrtnosti: kako smrtnost, trend smrtnosti i/ili nagib krivulje smrtnosti variraju između segmenata osiguranika kao što su npr. muškarci i žene.

Aspekti 1. i 2. su povezani, ali se ne preklapaju. Trend smrtnosti predstavlja broj ukupne smrti po godini na svakih 100.000 živućih, a nagib krivulje smrtnosti predstavlja kretanje, odnosno povećanje smrtnosti u starijoj životnoj dobi i može se mjeriti parametrom k koji su definirali Horiuchi i Coale 1990. godine kao

$$k(x) = \frac{\frac{d\mu(x)}{dx}}{\mu(x)}. \tag{20}$$

Iako industrija životnog osiguranja prikuplja ogromnu količinu podataka, prikupljeni podaci obično ne pokrivaju dovoljno dugačko vremensko razdoblje za vjerodostojnu analizu trenda smrtnosti osiguranika. Trendovi smrtnosti osiguranika često se aproksimiraju na temelju studija smrtnosti opće populacije.

Ova ovisnost o studijama smrtnosti opće populacije za razumijevanje smrtnosti osiguranika vjerojatno će trajati sve dok industrija životnog osiguranja ne skupi i uspostavi odgovarajuću bazu podataka o osiguranicima, sličnu bazi podataka smrtnosti opće populacije (eng. *Human Mortality Database* - HMD), koja bi podržavala sveobuhvatnu studiju smrtnosti osiguranika na temelju stvarnih iskustvenih podataka.

Što se tiče razumijevanja nagiba krivulje smrtnosti i različitosti u smrtnosti, postoje specifični izazovi kao što su dinamična segmentacija osiguranika i velike razlike u smrtnost između segmenata, a koji se vjerojatno ne mogu riješiti korištenjem podataka opće populacije.

Prvo, u usporedbi s općom populacijom, struktura osiguranika je vrlo nestabilna. Osiguravajuća društva konstantno potiču klasificiranje rizika preko strožih pravila o preuzimanju rizika, prilagođavaju cijene preuzimanja rizika, šire tržišta i stvaraju nove proizvode, što utječe na antiselektivno ponašanje postojećih osiguranika (otkaz ili promjena postojeće police). Novi rizici i antiselektivno ponašanje mijenjaju osnovne karakteristike određenog segmenta osiguranika.

Segmenti osiguranika mogu biti grupirani osim po dobi i prema statusu pušenja, klasi pokrivenih rizika, vrsti proizvoda, individualnom ili grupnom osiguranju ili bilo kojoj drugoj karakteristici osiguranika koja je bitna osiguravajućem društvu za vrednovanje cijene police. Niti jedna od navedenih karakteristika nije bitna niti se analizira u podacima o smrtnosti opće populacije.

Drugo, trend i nagib krivulje smrtnosti se značajno razlikuju između segmenta osiguranika. Osiguravajuća društva se često na tržištu natječu cijenama police (premije) koje su određene na temelju analize po segmentima.

Treće, segmenti osiguranika često imaju relativno male veličine, kratke povijesti i višedimenzionalne karakteristike (npr. muškarac pušač s doživotnom policom).

Procjena trenda i različitosti u smrtnosti, odnosno prave vrijednosti svake karakteristike određenog segmenta postaje analiza smrtnosti osiguranika od neprocjenjive vrijednosti u tržišnoj utakmici i pozicioniranju osiguravajućih društava.

4.1. Definiranje obuhvata istraživanja

Podaci su preuzeti iz studije *Logistic Regression for Insured Mortality Experience Studies (*Zhu, Z., Li, Z., 2014.) koja koristi dva izvora podataka. Prvi izvor podataka je *Human Mortality Database* što pokriva razdoblje od 1933. do 2010. godine, a koristi se za procjenu stope smrtnosti opće populacije u Sjedinjenim Američkim Državama. Ti se podaci uglavnom koriste u svrhu usporedbe, a ne za aproksimaciju smrtnosti osiguranika.

Drugi izvor podataka su osiguravajuća društva i globalni reosiguratelji s iskustvenim podacima više od 60 osiguravatelja, policama zaključenima od 1912. godine i razdobljem izloženosti riziku od 2000. do 2009. godine. Ukupno je za analizu dostupno 174 milijuna godina izloženosti i 1,6 milijuna smrti (šteta). Ti se podaci koriste u svrhu procjene smrtnosti osiguranika u Sjedinjenim Američkim Državama.

Aktuari i analitičari za potrebe ocjene izdanih polica ili formiranje cijene novih polica tradicionalno koriste što više empirijskih podataka za analizu stope smrtnosti osiguranika. Ova jednostavna strategija može biti mač s dvije oštrice. Dok povećana količina podataka može biti korisna za vjerodostojnost analize, to također može smanjiti točnost ako su uključeni manje relevantni podaci.

Istraživanja (npr. Vaupel 2014.) su utvrdila da se u nekim razvijenim zemljama, životni vijek povećava za oko 2 i pol godine po desetljeću, što podrazumijeva da korištenje polica osiguranja izdanih prije 50 godina za procjenu smrtnosti budućih nositelja polica može rezultirati značajnom pristranošću. Stoga, da bi izbjegli pristrane nalaze analize u istraživanju se koristi manje podataka, ali relevantnijih za modeliranje i ekstrapoliranje procjena. To postiže zadovoljavajuću izvedbu modela.

Za potrebe određivanja cijene budućih polica životnog osiguranja koje pokrivaju prosječne rizike (standardna klasa rizika) primijenjeni su sljedeći kriteriji odabira podataka:

- Police izdane od 1950. i
- Iznos police osiguranja ≥ 50.000 $.

Dok bi za potrebe npr. ocjene izdanih polica životnog osiguranja bilo potrebno primijeniti neke druge kriterije koji bi bili relevantni za te potrebe.

Tablica 3. Sažetak podataka osiguranika

		Ukupni podaci				Odabrani podaci			
Spol	***Dostignuta dob***	***Broj šteta***	***Izloženost***	***q***	$\left(\frac{q}{1-q}\right)$	***Broj šteta***	***Izloženost***	***q***	$\left(\frac{q}{1-q}\right)$
Žene	**00–22**	1.371	6.619.606	0.00023	0.00023	295	1.758.271	0.00016	0.00016
	23–27	1.696	3.499.033	0.00034	0.00034	301	1.425.257	0.00020	0.00020
	28–32	2.320	5.993.700	0.00035	0.00035	606	3.133.315	0.00019	0.00019
	33–37	3.841	8.819.014	0.00041	0.00041	1.302	5.186.770	0.00024	0.00024
	38–42	6.636	10.403.254	0.00064	0.00064	2.495	6.266.131	0.00039	0.00039
	43–47	12.072	11.503.957	0.00103	0.00103	3.899	6.323.438	0.00061	0.00062
	48–52	18.630	10.672.826	0.00168	0.00168	5.219	5.405.578	0.00096	0.00096
	53–57	25.578	9.473.005	0.00275	0.00276	5.947	3.975.759	0.00150	0.00150
	58–62	32.781	7.007.003	0.00475	0.00477	5.541	2.408.077	0.00230	0.00231
	63–67	40.562	5.073.084	0.00836	0.00843	4.668	1.204.946	0.00387	0.00389
	68–72	51.493	3.891.708	0.01433	0.01454	4.099	642.219	0.00638	0.00642
	73–77	75.665	3.616.263	0.02402	0.02462	4.552	413.902	0.01100	0.01112
	78–više	300.442	5.387.950	0.06130	0.06531	14.515	482.748	0.03007	0.03100
Muškarci	**00–22**	3.924	6.803.996	0.00056	0.00056	781	1.827.197	0.00042	0.00042
	23–27	3.705	3.804.721	0.00094	0.00094	778	1.428.309	0.00054	0.00054
	28–32	4.676	6.164.348	0.00070	0.00070	1.354	3.463.059	0.00039	0.00039
	33–37	7.609	10.666.726	0.00071	0.00071	2.712	6.587.593	0.00041	0.00041
	38–42	14.318	14.284.777	0.00094	0.00095	5.144	8.933.857	0.00058	0.00058
	43–47	23.248	16.460.843	0.00143	0.00143	8.353	9.912.149	0.00084	0.00084
	48–52	37.557	16.712.735	0.00228	0.00229	12.003	9.305.648	0.00129	0.00129
	53–57	55.030	15.019.344	0.00369	0.00370	14.836	7.668.238	0.00193	0.00194
	58–62	73.922	12.672.379	0.00610	0.00614	16.423	5.396.172	0.00304	0.00305
	63–67	90.683	8.950.030	0.01065	0.01076	14.859	3.033.903	0.00489	0.00492
	68–72	109.899	6.293.104	0.01825	0.01859	12.866	1.584.710	0.00812	0.00819
	73–77	147.536	5.040.212	0.03158	0.03261	11.648	840.363	0.01386	0.01406
	78–više	491.936	6.939.737	0.07502	0.08111	19.907	586.764	0.03391	0.03510

4.2. Logistički modeli i modeliranje smrtnosti

Istraživači odavno koriste statističke modele za proučavanje opće smrtnosti stanovništva. Od uvođenja Gompertzovog zakona smrtnosti (1825.), proces modeliranja krivulje smrtnosti se poboljšao. Thatcher je 1999. godine odlično opisao i usporedio dostupne modele smrtnosti po dobi. Kod nekih pojednostavljenja kada je u pitanju smanjenje broja parametara i uporabom intenziteta smrtnosti kao zavisne varijable, koriste se četiri modela:

1. Gompertzov (1825.): $\mu \approx \alpha * \exp(\beta * x)$,
2. Weibullov (1951.): $\mu = \alpha * x^{\beta}$,
3. Heligmanov i Pollardov (1980.): $\mu \approx \alpha - \frac{1}{2}\beta + \beta * x$,
4. Kannistov (1994.): $\mu = \frac{\alpha * \exp(\beta * x)}{1 + \alpha * \exp(\beta * x)}$.

Od četiri modela, samo Kannistov model pretpostavlja da intenzitet smrtnosti ima konačnu asimptotu. Thatcherov zaključak je bio sljedeći: "Kada su ova četiri modela prilagođena sa stvarnim podacima (opće populacije), svi su relativno slični podacima o dobi u kojoj je većina smrtnih slučajeva koncentrirana, a time i relativno slični jedan drugome." Nije iznenađujuće kako je on također potvrdio korištenjem raznih podataka o stanovništvu da je Kannistov model i najbolji model (Thatcher, Kannisto i Vaupel 1998., Thatcher 1999.).

Kannistov model možemo koristiti na dva načina za studiju smrtnosti: korištenjem stope smrtnosti q umjesto intenziteta smrtnosti μ kao zavisne varijable i korištenjem ne samo dobi, već i mnogih drugih nezavisnih varijabli osiguranika.

Logistički model smrtnosti (model za q) ima opći oblik:

$$q = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{1.2} x_1 * x_2+\beta_{1,3} x_1 * x_3+\cdots)}}{1 + e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_{1.2} x_1 * x_2+\beta_{1,3} x_1 * x_3+\cdots)}} \tag{21}$$

ili

$$\ln\left(\frac{q}{1-q}\right) = \alpha + \sum_i \beta_i * x_i + \sum_{i,j} \beta_{i,j} * x_i * x_j + \cdots \tag{22}$$

Gdje su:

q – vjerojatnost smrti osiguranika do dobi (x+1), ako je doživio dob x,

x_i – su neovisne varijable (npr. dob, spol, trajanje police, proizvod),

α – konstanta, parametar modela i

β_t – koeficijenti zavisnih varijabli, parametri modela.

Logistički model za q ima mnoge prednosti za analizu smrtnosti osiguranika:

- modelira stopu smrtnosti q koji se izravno može koristiti u određivanju cijene police i upravljanju rizicima,
- može se lako prilagoditi za potrebe procjenjivanja razine, nagiba krivulje i razlika u smrtnosti (ključni pokazatelji koje se koriste u poslovnoj praksi) i
- može se primjenjivati široko dostupnima statističkim računalnim programima kao što su SAS, SPSS i R (u citiranoj analizi se koristio SPSS).

Za potencijalne nezavisne varijable modela je odabrano devet varijabli:

1. Spol osiguranika: muško i žensko;
2. Trajanje police osiguranja (od ugovaranja): kao kontinuirane varijable;
3. Pristupna dob (dob osiguranika pri ugovaranju police osiguranja na zadnji rođendan): od 1 do 99 kao kontinuirane varijable;
4. Status pušenja osiguranika: pušač, nepušač, nepoznato;
5. Vrsta proizvoda: doživotna polica, polica s određenim trajanjem;
6. Klasa pokrivenih rizika: preferirana (najmanji rizik), standard (prosječan rizik), agregirana (svi rizici);
7. Razdoblje izloženosti: od 2000. do 2009. godine kao kontinuirana varijabla;
8. Razdoblje rizika: četiri razdoblja definirana na temelju godine izdavanja;
9. Vrijednost police osiguranja ($): 50.000 – 99.000, 100.000 – 499.000, 500.000+.

Navedene varijable su odabrane zato što im najmanje nedostaju podaci i najčešće se koriste za donošenje odluka o cijenama, pokrivenim rizicima i marketinškim strategijama. Za razliku od općih studija smrtnosti stanovništva, za varijable koji bi predstavljale dob i vrijeme su odabrane pristupna dob osiguranika i trajanje police umjesto dosegnute dobi i kalendarske godine. Odabrani par ima bolji prikaz karakteristika osiguranika i određuje dimenziju tablica smrtnosti osiguranika.

Ako u općem obliku logističkog modela (21) i (22) za q izostavimo interakciju između nezavisnih varijabli, model postaje

$$q = \frac{e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots)}}{1+e^{(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots)}} \tag{23}$$

ili

$$\ln\left(\frac{q}{1-q}\right) = \alpha + \sum_i \beta_i * x_i \tag{24}$$

Iako uključivanje interakcije između nezavisnih modela ima potencijala za poboljšanje prilagodbe modela, testovi su pokazali da navedeno poboljšanje nije dovoljno značajno da se primjeni složeniji model. Također, u složenijemu modelu je i interpretacija koeficijenata puno složenija.

Za bolju usporedbu sa studijom Aktuarskog društva Sjedinjenih Američkih Država (*2013 Mortality Study, SOA*), odabrani podaci su podijeljeni u četiri podskupine i svaki podskup se analizira vlastitim modelom (24). Četiri podskupa su muški pušači, muški nepušači, ženski pušači i ženski nepušači.

Ovaj poseban dizajn modela omogućuje da svaki koeficijent modela bude procijenjen neovisno od ostala tri modela, što znači da svaka od četiri skupine osiguranika može imati vlastitu razinu i nagib krivulje smrtnosti te razlikovnih čimbenika bez toga da su ograničene utjecajem ostale tri skupine.

Od devet nezavisnih varijabli, spol i status pušenja osiguranika se koriste za osnovnu podjelu podataka studije, a sedam varijabli je ostavljeno kako bi bile uključene u ta četiri modela kao nezavisne varijable. Tablica 4. sumira p-vrijednosti testa značajnosti sedam varijabli za objašnjenje na svakom od četiri seta podataka.

Tablica 4. Analiza utjecaja

Pr > ChiSq (p-vrijednost)	Stupanj slobode	Žene		Muškarci	
		Nepušači	Pušači	Nepušači	Pušači
Trajanje police	1	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Pristupna dob	1	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Izloženost	1	0.1714	0.4597	0.1719	0.0017
Vrijednost police	2	0.0051	0.0040	< 0.0001	< 0.0001
Vrsta proizvoda	1	0.0157	0.9533	0.1363	< 0.0001
Razdoblje rizika	3	< 0.0001	0.0003	< 0.0001	< 0.0001
Klasa rizika	2	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Kao što se očekivalo, statistički, smrtnost osiguranika u sve četiri skupine značajno varira po trajanju police osiguranja, pristupnoj dobi osiguranika, vrijednosti police, razdoblju i klasi pokrivenih rizika. To potvrđuje da su ove varijable među najpouzdanijim pokazateljima stope smrtnosti.

Izloženost je uključena kao rezervno mjesto za poboljšanje stope smrtnosti u razdoblju od 10 godina koje su obuhvaćene podacima studije. Odgovarajuće p-vrijednosti od četiri modela znače da, nakon izdvajanja svega što je objašnjeno pomoću ostalih osam nezavisnih varijabli (uključujući spol i status pušenja), varijacija smrtnosti objašnjenja unutar godine izloženosti je statistički značajna na razini značajnosti $\alpha = 0.05$ samo za muškarce pušače. To može podrazumijevati da je više muškaraca pušača prestalo pušiti što je rezultiralo smanjenjem smrtnosti tijekom promatranog razdoblja.

Razlika stope smrtnosti prema vrsti proizvoda između doživotnih polica i polica s određenim trajanjem je samo statistički značajna za žene nepušače i muškarce pušače, nakon kontroliranja drugih osam nezavisnih varijabli.

Na razini značajnosti od 5%, svih sedam testiranih varijabli statistički su značajni u barem jednoj od četiri skupine polica pa su uključene u sva četiri logistička modela za q. Vinsonhaler i sur. (vidi Vinsonhaler, Nalini, Jeyaraj, Guy 2001.) su analizirali stopu smrtnosti osiguranika na temelju stvarnih iskustvenih podataka jednog privatnog mirovinskog društva koristeći sličan logistički model za q te su na istoj razini značajnosti pronašli samo jednu varijablu sa statističkom značajnošću u objašnjavanju varijacija u smrtnosti.

Podaci preuzeti iz studije (Zhu, Z., Li, Z., 2014.) u svrhu ilustracije interpretacije rezultata u tablici 5. umjesto kategorijalne varijable razdoblje rizika prikazuju status liječničkog pregleda osiguranika: pregled obavljen, pregled nije obavljen.

Tablica 5. Procjena izgleda

	Muškarci nepušači			**Muškarci pušači**		
Utjecaj	$\left(\frac{q}{1-q}\right)$	**Interval 95%-tne pouzdanosti**		$\left(\frac{q}{1-q}\right)$	**Interval 95%-tne pouzdanosti**	
Trajanje police	***1.141***	1.139	1.143	**1.118**	1.114	1.122
Pristupna dob	***1.101***	1.100	1.102	**1.093**	1.092	1.094
Godine izloženosti	***0.998***	0.995	1.001	**1.009**	1.003	1.014
Vrijednost police $100k-490k vs. $500k	***1.115***	1.096	1.135	**1.203**	1.143	1.265
Vrijednost police $50k-99k vs. $500k	***1.284***	1.258	1.311	**1.407**	1.335	1.484
Liječnički pregled vs. bez pregleda	***0.920***	0.902	0.939	**1.018**	0.986	1.050
Doživotna vs. polica s određenim trajanjem	***1.013***	0.996	1.030	**0.923**	0.890	0.958
Agregirani vs. standard rizici	***1.042***	1.027	1.057	**0.930**	0.893	0.967
Preferirani vs. standard rizici	***0.730***	0.719	0.741	**0.748**	0.717	0.781
	Žene nepušači			**Žene pušači**		
Utjecaj	$\left(\frac{q}{1-q}\right)$	**Interval 95%-tne pouzdanosti**		$\left(\frac{q}{1-q}\right)$	**Interval 95%-tne pouzdanosti**	
Trajanje police	**1.157**	1.153	1.160	**1.133**	1.126	1.139
Pristupna dob	**1.105**	1.104	1.105	**1.098**	1.096	1.099
Godine izloženosti	**0.997**	0.992	1.001	**1.004**	0.994	1.013
Vrijednost police $100k-490k vs. $500k	**1.000**	0.971	1.030	**0.926**	0.855	1.002
Vrijednost police $50k-99k vs. $500k	**1.037**	1.003	1.071	**0.988**	0.911	1.072
Liječnički pregled vs. bez pregleda	**0.950**	0.921	0.981	**1.044**	0.992	1.099
Doživotna vs. polica s određenim trajanjem	**1.033**	1.006	1.060	**0.998**	0.939	1.061
Agregirani vs. standard rizici	**1.038**	1.014	1.062	**0.938**	0.881	0.999
Preferirani vs. standard rizici	**0.740**	0.722	0.758	**0.767**	0.715	0.823

Na temelju podataka iz tablice 5. možemo interpretirati smrtnost muškaraca nepušača:

- *Trajanje police i pristupna dob:* ako je sve ostalo jednako, u prosjeku, smrtnost se povećava oko 14% po trajanju police i oko 10% po pristupnoj dobi (broju godina). 10%-tno povećanje smrtnosti za dob pri izdavanju police osiguranja također vrijedi i za opću populaciju (Thatcher 1999.).

- *Godine izloženosti:* ako je sve ostalo jednako, postoji statistički neznačajna niža smrtnost od 0.2% na godišnjoj razini smrtnosti.

- *Vrijednost police:* ako je sve ostalo jednako, u odnosu na velike police s nominalnim iznosom od najmanje 500.000 dolara, police s iznosom između 50.000 i 99.000 dolara će imati 28%, a police sa iznosom između 100.000 i 499.000 dolara 12% veću smrtnost.

- *Liječnički pregled:* ako je sve ostalo jednako, smrtnost osiguranika koji su imali liječnički pregled je 8% niža nego u onih bez pregleda.

- *Vrsta proizvoda:* ako je sve ostalo jednako, smrtnost osiguranika koji imaju doživotnu policu će biti statistički neznačajno viša (1,3%) od osiguranika koji imaju policu s određenim trajanjem.

- *Klasa pokrivenih rizika:* ako je sve ostalo jednako, smrtnost osiguranika preferirane klase je oko 27% niža od osiguranika standardne klase dok je smrtnost osiguranika svi-u-jednoj klase oko 4% veća od osiguranika standardne klase.

Na sličan način interpretiramo rezultate analize za ostale tri populacije.

Za razliku od potraživanja od zdravstvenog ili imovinskog osiguranja, potraživanja (štete) u životnom osiguranju pojavljuju se na puno nižoj frekvenciji i s mnogo stabilnijim uzorkom. Relativno mali broj šteta, ali vrlo sličnog obrasca dovodi do potrebe korištenja svih iskustvenih podataka u svrhu izrade modela bez mogućnosti odvajanja podataka u zadržani uzorak za provjeru modela.

Ocjena prilagodbe modela

Jedna od često korištenih mjernih statistika prilagodbe modela je c-statistika, odnosno površina ispod ROC krivulje. Sljedeća tablica prikazuje vrijednosti c-statistike za četiri modela.

Tablica 6. Podobnost modela

Povezanost predviđenih vjerojatnosti i opaženih odgovora	**Žene**		**Muškarci**	
	Nepušači	**Pušači**	**Nepušači**	**Pušači**
vrijednost c-statistike	0.682	0.753	0.679	0.747

Vinsonhaler i sur. (vidi Vinsonhaler, Nalini, Jeyaraj, Guy 2001.) su analizirali stopu smrtnosti osiguranika na temelju stvarnih iskustvenih podataka jednog privatnog mirovinskog društva koristeći sličan, ali jednostavniji logistički model za q sa samo jednom nezavisnom varijablom. Njihov model imao je vrijednost c-statistike u rasponu 0.51 – 0.59 za većinu dobnih skupina. Iako nije potrebno mjeriti vrijednost c-statistike prema dobnoj skupini, usporedba i dalje daje osjećaj da ova četiri modela imaju prilično visoku vrijednost c-statistike (oko 0,7) što predstavlja dobro prilagođen model.

Zanimljivo zapažanje je da dva modela nepušača imaju nižu vrijednost c-statistike od dva modela pušača. Ako se koristi vrijednost c-statistike kao mjera predvidljivosti, predvidivost po istom skupom nezavisnih varijabli za pušače iznosi oko 10% više nego za nepušače (u slučaju kada je poznat (ne)pušački status osiguranika).

Uz ranije spomenute prednosti, prilagodba statističkog modela za studije smrtnosti osiguranika donosi novi problem: cenzuru vremena smrti u podacima radi prestanka važenja police.

Zamislimo skupinu od 100 sadašnjih osiguranika. Ako je 10 umrlo u narednih 12 mjeseci, ali postoji samo pet generiranih šteta i ostalih pet je umrlo nakon prestanka pokrića osiguranja, stopa smrtnosti u skupini će biti 10%, ali stopa šteta će biti samo 5%. Stopa smrtnosti osiguranika ili stopa šteta je uvjetovana razdobljem rizika, odnosno trajanjem pokrića police osiguranja. To nije isto što i mjerenje smrtnosti za opću populaciju.

Kada se koriste modeli logističke regresije za procjenu stope smrtnosti osiguranika ili stopu šteta, oni ne uzimaju u obzir važenje police osiguranja pa imaju tendenciju precijeniti stopu šteta. Ako se model koristi za ekstrapolaciju smrtnosti za neku individualnu ili određenu grupu polica osiguranja, stopa šteta može biti značajno precijenjena. Stoga je potrebno uzeti u obzir i trajanje pokrića polica te sukladno tome modificirati model.

Prilagođavanja na temelju cenzuriranja

Recimo da je q stopa smrtnosti i pretpostavimo da svaka polica osiguranja ima tri vidljiva statusa na kraju svake godine:

- za istek važenja (q_l); uključuje sve osiguranike koji su doživjeli kraj godine, ali više nisu pod rizikom generiranja štete jer se prestaju pratiti zbog isteka police osiguranja,
- štete (q_c); uključuje sve osiguranike koji su tijekom promatrane godine generirali štetu (nisu doživjeli kraj godine) i
- važeća (q_i); uključuje sve osiguranike koji su doživjeli kraj godine i dalje su pod rizikom generiranja štete jer je polica osiguranja još na snazi,

tako da je $q_l + q_c + q_i = 100\%$.

S istim nezavisnim varijablama x_i možemo koristiti multinominalan logistički model (Hosmer i sur. 2013.) kako bi se modelirale tri vjerojatnosti kako slijedi

$$\begin{cases} q_c = \dfrac{e^{(\alpha_c+\beta_{c1}x_{c1}+\beta_{c2}x_2+\cdots)}}{1+e^{(\alpha_l+\beta_{l1}x_1+\beta_{l2}x_2+\cdots)}+e^{(\alpha_c+\beta_{c1}x_{c1}+\beta_{c2}x_2+\cdots)}} \\ q_l = \dfrac{e^{(\alpha_l+\beta_{l1}x_1+\beta_{l2}x_2+\cdots)}}{1+e^{(\alpha_l+\beta_{l1}x_1+\beta_{l2}x_2+\cdots)}+e^{(\alpha_c+\beta_{c1}x_{c1}+\beta_{c2}x_2+\cdots)}} \\ q_i = \dfrac{1}{1+e^{(\alpha_l+\beta_{l1}x_1+\beta_{l2}x_2+\cdots)}+e^{(\alpha_c+\beta_{c1}x_{c1}+\beta_{c2}x_2+\cdots)}} \end{cases} \quad (25)$$

Nazovimo ovaj model logistički q_c model kako bismo naglasili procjenu stope šteta. Dodana komponenta isteka važenja police osiguranja q_l igra ulogu pri procjeni dijela izlaganja pri isteku važenja police te ih isključuje iz procjene da ne bi doprinosili stopi šteta u slučaju smrti.

Asimptotički, usporedbom prijašnjih modela imamo

$$\lim_{trajanje\to\infty} q = \lim_{trajanje\to\infty} (q_c + q_l) = 1 \tag{26}$$

što ukazuje kako (26) rastavlja ukupnu stopu smrti na dio šteta i na dio isteka važenja police osiguranja te stoga imamo

$$\lim_{trajanje\to\infty} q_c = \lim_{trajanje\to\infty} \frac{1}{1 + e^{(\alpha_l - \alpha_c) + (\beta_{l1} - \beta_{c1}) * trajanje}} = \begin{cases} 1 & \beta_{l1} < \beta_{c1} \\ \frac{1}{1 + e^{(\alpha_{l1} - \alpha_{c1})}} & \beta_{l1} = \beta_{c1} \\ 0 & \beta_{l1} > \beta_{c1} \, . \end{cases} \tag{27}$$

4. 3. Ograničenja i moguća poboljšanja modela

Između ostaloga, u studiji stope smrtnosti osiguranika se mogu pojaviti tri vrste pristranosti:

1. pristranost procjene parametra,
   je sustavna pristranost koja odražava tehnička ograničenja odabranog modela i ovisi o metodi procjene (na primjer, korištenjem linearnog modela za prilagođavanje podataka s nesimetričnom populacijskom razdiobom);

2. pristranost zbog uzorkovanja
   događa se kada se koristi zamjenski skup podataka za predstavljanje ciljne populacije pri čemu zamjenski skup nema iste karakteristike populacije (npr. korištenje prijašnjeg iskustva kao što je smrtnost u 1950. godini za projiciranje smrtnosti u 2020. godini) i

3. pristranost podataka
   je raskorak između podataka i stvarnosti (npr. pogrešno prijavljena starost umrlih ili neevidentirani prestanak važenja police osiguranja).

Slijedi nekoliko ograničenja korištenja logističke regresije za procjenu smrtnosti osiguranika:

- Model nije prilagođen smrtnosti novorođenčadi i adolescenata zato što te skupine imaju visoku stopu smrtnosti (smrtnost novorođenčadi, odnosno stopa samoubojstava ili prometnih nesreća adolescenata).

- Model nije prilagođen za velike poremećaje kao što su istek velikog broja polica u istom razdoblju ili povećanje smrtnosti (šteta) radi pandemije neke bolesti.

- Kada nema puno podataka na raspolaganju, kao što su vrlo stara dob pri izdavanju police ili duga trajanja (pristranosti podataka), logistička funkcija će biti primarni alat za projekciju q ili q_c. Radi točnije projekcije, obično su potrebne konzultacije s aktuarima koji imaju iskustvo s takvim policama.

- Sadašnji nedostatak dosljedno prikupljene dugoročne baze podataka o osiguranicima ograničava optimizaciju svih analiza u osiguravajućim društvima, uključujući i logističke modele smrtnosti (uzorkovanje i podatkovna pristranost). Na primjer, nisu sva društva niti sve informacije o proizvodima dosljedne ili razmjerno prikazane. Posebna briga je potrebna pri interpretaciji ishoda modela koji implicitno pretpostavljaju konzistenciju. Kako tehnologija obrade podataka i analitičke metodologije napreduju, industrija će s vremenom moći uspostaviti mehanizam za stalno prikupljanje sveobuhvatnih podataka o osiguranicima za detaljne, dubinske analize.

5. ZAKLJUČAK

Logistička regresija predstavlja statističku tehniku prikladnu u slučaju kada problem istraživanja obuhvaća jednu binarnu kategorijsku zavisnu varijablu i jednu ili više kontinuiranih, kategorijskih ili *dummy* nezavisnih varijabli. Njena relativna snaga dolazi iz sposobnosti fleksibilnosti u više istraživačkih pretpostavki, a njena robusnost potječe od minimalnog skupa temeljnih pretpostavki i sličnosti s višestrukom linearnom regresijom za potrebe tumačenja. Rezultat je širok raspon primjene kako u znanosti tako i u industriji.

U industriji osiguranja logistička regresija se može primijeniti u analizi smrtnosti osiguranika jer dobro projektirani logistički modeli mogu učinkovitije iskoristiti raspoložive podatke za više različitih potreba. Za analizu smrtnosti osiguranika u Sjedinjenim Američkim Državama, a na temelju preuzetih podataka iz studije (Zhu, Z., Li, Z., 2014.) se primjenjuje pristup modeliranja primjenom logističke regresije. Prilagodba pristupa modeliranju više varijabli, dostupnost velike količine podataka o osiguranicima i korištenje moderne računalne tehnologije pružaju mnoge prednosti nad konvencionalnim metodama studija smrtnost opće populacije.

Navedene prednosti uključuju: model temeljen na empirijskim podacima s više nezavisnih varijabli, projekciju smrtnosti dugogodišnjih osiguranika i osiguranika u dobi 65+ kombiniranjem prošlog iskustva i modela ekstrapolacije, izglađivanje razlika između smrtnosti dugogodišnjih i novih osiguranika funkcijom poveznicom modela, provjeru pouzdanosti studije i s testnom statistikom, a ne sama na temelju broja šteta, izradu tablice smrtnosti osiguranika za različite segmente na temelju modela i provedivost sa široko dostupnim statističkim računalnim programima.

Kao i u bilo kojoj drugoj velikoj bazi podataka, podaci o osiguranicima imaju dosta propusta kao što je nedostatak podataka i nekonzistentno kodiranje podataka između osiguravajućih društava, koje mogu kompromitirati kvalitetu predviđanja logističkog modela pa je potrebno uz konzultacije s aktuarima napraviti određene korekcije. Poseban izazov u prilagodbi statističkog modela za studije smrtnosti osiguranika donosi cenzura vremena smrti u podacima radi prestanka važenja police.

LITERATURA

1. Agresti, A. (2007). Categorical Data Analysis (Second Edition). New York: John Wiley.
2. Alexander, K. L., Dauber, S. L., & Entwisle, D. R. (1996). Children in motion: School transfers and elementary school performance. The Journal of Educational Research, 90(1), 3–11.
3. Barnett, A. G., Dobson, A. J. (2008), An introduction to generalized linear models, CRC Press, Boca Raton.
4. Christensen, R. (1997) Log-linear models and logistic regression 2nd ed. New York: Springer-Verlag.
5. Draper, N.R. and Smith, H. (1981). Applied Regression Analysis. 2nd ed. New York: John Wiley.
6. Field, A. (2005). Discovering Statistics Using SPSS. London: Sage.
7. Harrell, Frank E. Jr. (2001). Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer-Verlag.
8. Heligman, L, and John H. P. (1980). "The Age Pattern of Mortality." Journal of the Institute of Actuaries 107 (01): 49–80.
9. Hosmer D.W. and S. Lemeshow (1980) "A goodness-of-fit test for the multiple logistic regression model." Communications in Statistics A10:1043-1069.
10. Hosmer, D. W., Stanley L., and Rodney X. S. (2013). Applied Logistic Regression. 3rd ed. Hoboken, N.J.: John Wiley & Sons Inc.
11. Human Mortality Database. University of California, Berkeley. Dostupno na: http://www.mortality.org.
12. Kannisto, V. (1994). Development of Oldest-Old Mortality, 1950-1990: Evidence from 28 Developed Countries. Odense Monographs on Population Aging 1. Odense University.
13. Long, J.S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage.
14. McCullagh P and Nelder J.A, (1989). Generalized Linear Models (Second Edition). Chapman and Hall.
15. Olshansky, S. Jay. (1998). "On the Biodemography of Aging: A Review Essay." Population and Development Review 24 (2): 381–93.

16. Schlotzhauer, D.C (1993). “Some issues in using PROC LOGISTIC for binary logistic regression”. Observations: The Technical Journal for SAS Software Users. Vol. 2, No. 4.
17. Society of Actuaries. 2008 Valuation Basic Table (VBT) Report & Tables.
18. Strauss, D. (1992). “The many faces of logistic regression.” The American Statistician, Vol. 46, No. 4, pp. 321-326.
19. Thatcher, A. R., Vaino K., and James W. V. (1998). “The Force of Mortality at Ages 80 to 120.” Odense Monographs on Population Aging 5, Odense University Press, Odense, Denmark.
20. Thatcher, A. R. (1999). “The Long-Term Pattern of Adult Mortality and the Highest Attained Age.” Journal of the Royal Statistical Society: Series A 162 (1): 5–43.
21. Trusty, J. (2000). High educational expectations and low achievement: Stability of educational goals across adolescence. The Journal of Educational Research, 93(6), 356–365.
22. Vaupel, J. (2014). “The Advancing Frontier of Human Survival.” General Session 1 presentation, Living to 100 Symposium, Jan. 8, Orlando.
23. Vinsonhaler, Charles, Nalini Ravishanker, Jeyaraj Vadiveloo, and Guy Rasoanaivo. 2001. “Multivariate analysis of Pension Plan Mortality Data.” North American Actuarial Journal, Vol 5, Issue 2, 126-135.
24. Zhu, Z. and Li, Z. (2014) Logistic Regression for Insured Mortality Experience Studies [online] Presented at the Living to 100 Symposium, Orlando, Florida. January 8 – 10, 2014. Dostupno na: https://www.scor.com/images/stories /pdf/library/scor-inform/scor_inform_us_dec2013.pdf

POPIS TABLICA

POPIS ILUSTRACIJA

SAŽETAK

Osiguravajućim društvima je za potrebe vrednovanja očekivanih obveza i održavanje potrebne solventnosti kapitala važna dobra procjena smrtnosti osiguranika. U praksi, postupak procjene smrtnosti osiguranika se susreće sa širokim rasponom podataka i analitičkim izazovima. U ovom radu, za analizu smrtnosti osiguranika u Sjedinjenim Američkim Državama primjenjuje se pristup modeliranja primjenom logističke regresije koji omogućuje korištenje manje podataka, ali relevantnijih za modeliranje i ekstrapoliranje procjena.

Logistička regresija predstavlja vrstu regresijske analize u kojoj je zavisna (odzivna) varijabla dihotomna, odnosno binarna i kodira se s 0 ili 1 te postoji najmanje jedna nezavisna odnosno prediktorska varijabla koja može biti kategorijska ili kontinuirana. Navedeno u stvarnosti predstavlja modeliranje bilo kojeg problema kod kojeg se ciljni događaj može prevesti u kategorijsku varijablu (da/ne, živ/mrtav, zdrav/bolestan).

Logistički model za q ima mnoge prednosti za analizu smrtnosti osiguranika: modelira stopu smrtnosti q koji se izravno može koristiti u određivanju cijene police i upravljanju rizicima, može se lako prilagoditi za potrebe procjenjivanja razine, nagiba krivulje i razlika u smrtnosti (ključni pokazatelji koje se koriste u poslovnoj praksi) i može se primjenjivati široko dostupnima statističkim računalnim programima.

Za potencijalne nezavisne varijable modela odabrano je devet varijabli zato što im najmanje nedostaju podaci i najčešće se koriste za donošenje odluka o cijenama, pokrivenim rizicima i marketinškim strategijama. Od devet nezavisnih varijabli, spol i status pušenja osiguranika se koriste za osnovnu podjelu podataka studije, a sedam varijabli je ostavljeno kako bi bile uključene u ta četiri modela kao nezavisne varijable.

Kao što se očekivalo, statistički, smrtnost osiguranika u sve četiri skupine značajno varira po trajanju police osiguranja, pristupnoj dobi osiguranika, vrijednosti police, razdoblju i klasi pokrivenih rizika. To potvrđuje da su ove varijable među najpouzdanijim pokazateljima stope smrtnosti. Na razini značajnosti od 95%, svih sedam testiranih varijabli statistički su značajni u barem jednoj od četiri skupine polica pa su uključene u sva četiri logistička modela za q.

SUMMARY

Insured population mortality estimation is essential to insurers' developing liability expectations and maintaining required solvency capital. In practice, insured mortality measurement needs to deal with a broad range of data and analytical challenges. In this paper, we introduce a logistic regression-based modeling approach for analyzing the U.S. insured mortality experience. This approach allows use of less but more relevant data to address multiple challenges in quantifying insured mortality.

Logistic regression is a regression model where the dependent variable is categorical, binary variable where the output can take only two values, 0 and 1, and is related to at least one of explanatory variables, which can be discrete and/or continuous. In practice, this is modeling of any problem type which answer represent outcomes such as win/lose, alive/dead or healthy/sick (discrete variable).

A logistic q model has many advantages for insured experience studies: it models mortality q that is directly used in business operation and risk management; it can be flexibly configured for estimating mortality levels, slopes and differentials that are key metrics used in business practices; and it can be developed with widely available commercial software systems.

Nine observable explanatory variables are selected as potential independent variables for model development. These variables are selected because they have the least missing values and are the most frequently used for pricing decisions, underwriting adjustments and marketing strategies. Of the nine explanatory variables, gender and smoking status are used to split the study data and seven are left to be included in the models.

As expected, insured mortality varies significantly statistically by duration, issue age, underwriting class, underwriting era and face band for all four subgroups. This confirms that these variables are among the most reliable mortality predictors. At 95 percent confidence level, all seven tested variables have statistical significance in explaining mortality variation in at least one of the four policy groups so we included them in all four logistic *q* models.

BIOGRAFIJA

Ivana Bistrović je rođena u Zagrebu. Nakon završetka osnovne škole i prirodoslovno-matematičke gimnazije zapošljava se u Gradu Zagrebu, Gradskom uredu za gospodarstvo, rad i poduzetništvo na radno mjesto stručne referentice za dokumentaciju i polaže državni stručni ispit.

Uz rad se stručno usavršava na području jezika, informatike i menadžmenta uredskog poslovanja te napreduje na radno mjesto administrativne tajnice pročelnika. Nastavlja formalno obrazovanje na Ekonomskom fakultetu Sveučilišta u Zagrebu i završava Stručni studij poslovne ekonomije, Preddiplomski sveučilišni studij poslovne ekonomije i Diplomski sveučilišni studij, smjer Turizam te stječe akademski naziv magistra ekonomije s najvećim pohvalama (mag. oec. *summa cum laude*).

Tijekom studiranja, dva puta je izabrana za sudjelovanje na specijalnom programu međunarodne edukacije studenata turizma ITHAS (*International Tourism and Hospitality Academy at Sea*): *Intensive study module on Creation of Tourism Product* i *Intensive study module on Small Scale Tourism Development* te je dobitnica Dekanove nagrade za diplomski rad na temu: Značaj razvoja kongresnog turizma za Grad Zagreb.

U Gradskom uredu za gospodarstvo, rad i poduzetništvo napreduje od radnog mjesta više referentice za turizam na radno mjesto više stručne suradnice za praćenje stanja u gospodarstvu te odlučuje nastaviti formalno obrazovanje na Prirodoslovno-matematičkom fakultetu, Odsjeku za matematiku Sveučilišta u Zagrebu na Specijalističkom poslijediplomskom studiju Aktuarske matematike.

Predstavlja Grad Zagreb u udruzi *EUROCITIES* na *Economic Development Forumu* i u radu radnih skupina Atraktivnost i brendiranje gradova, Međunarodni ekonomski odnosi, Malo i srednje poduzetništvo i Inovacije te podnosi uspješnu prijavu Grada Zagreba za dobivanje domaćinstva za održavanje jednog od *Economic Development Foruma*.

Natječe se na javnom međunarodnom natječaju koji je objavilo Veleposlanstvo Sjedinjenih Američkih Država u Republici Hrvatskoj u suradnji sa *US Department of State* iz Washington DCa i izabrana je za sudjelovanje na *Professional Fellowship Programu* (program razmjene stručnjaka financiran od *US Department of State*) na temu promocije i razvoja tehnoloških inovacija, poduzetništva i obrazovanja. Za instituciju domaćina je izabran *City of Chicago, Department of Innovation and Technology*. Nakon završenog programa sudjeluje na *Professional Fellows Congressu* što se održao u Washington DC-u, Sjedinjene Američke Države.

U Gradu Zagrebu napreduje na radno mjesto stručne savjetnice za plan i analizu u Gradskom uredu za prostorno uređenje, izgradnju Grada, graditeljstvo, komunalne poslove i promet. Poslovi radnog mjesta uključuju izradu: prijedloga za osiguranje sredstava u proračunu Grada Zagreba, financijskog plana i izvješća Ureda, programa i planova radova po područjima nadležnosti, izvješća o izvršenju programa i planova radova na području kapitalnih ulaganja u objekte za društvene djelatnosti, analiza, izvješća i drugih stručnih materijala iz područja nadležnosti Ureda te pružanje podataka, informacija i drugih stručnih podloga za rad ureda/zavoda/službi gradske uprave Grada Zagreba.