UNIVERZITET U NOVOM SADU
PRIRODNO – MATEMATICKI FAKULTET
DEPARTMAN
ZA MATEMATIKU I INFORMATIKU

Albert Koložvari

Primena skoring modela u izračunavanju premije kasko osiguranja

-master rad-

Mentor:
Prof. dr Zorana Lužanin

Novi Sad, 2022

Sadržaj

1 Uvod

Razvojem računara i internet mreže u privatnoj i poslovnoj sferi društva krajem 20. veka započeto je digitalno (informacijsko) doba. Informacija predstavlja jedan od najvažnijih činilaca današnjeg sveta i gotovo je nezaobilazan faktor kako čoveka, tako i svakog činioca poslovnog sveta.
Danas se mnoge osiguravajuće kuće međusobno takmiče u povećanju prodaje njihovih proizvoda i uvođenju novih tarifa. Matematički analitičari (aktuari) koji rade za „takmičare“ u toj tržišnoj utakmici, imaju zadatak da pomognu menadžmentu osiguravajuće kuće u donošenju brze i pouzdane poslovne odluke. Jedan od načina je analiza istorijskih podataka i konstrukcija modela u svrhu uočavanja pravilnosti na tržištu, te pravljenja plana za buduće poslovanje.
Skoring modeli (modeli bodovanja) su statistički modeli pomoću kojih se, na osnovu istorijskih podataka može utvrditi verovatnoća nekog dogadjaja, tj. predvidja se ponašanje pojedinca ili grupe na osnovu istorijskih podataka. Takodje, ovi modeli mogu imati ulogu da opišu veze izmedju nekih parametara tržišta.
Do sada, najpoznatiji skoring modeli su kreditni skoring modeli koji se koriste u bankarskom sektoru i imaju ulogu u kreditnoj analizi i proceni svakog klijenta, na način da će neki klijenti dobijati kredite po povoljnijim uslovima, dok neki neće imati priliku da dobiju kredit od banke. Sa druge strane, primena skoring sistema u industriji osiguranja nije u tolikoj meri razvijena kao u bankarskom sektoru. Medjutim, postoji sve više naučnih radova u kojima se daju primeri u računanju premija, predvidjanju šteta, itd.
Skoring modeli se baziraju na istorijskim podacima. Razvojem računara i softvera, kompanije sada sadrže ogromne baze podataka, koji su uglavnom nestruktuirani. Zbog toga, konstrukcija skoring modela sadrži nekoliko etapa koje ćemo navesti u nastavku.

1.1 Koraci u konstrukciji modela

Svaki projekat koji uključuje modeliranje ima različite ciljeve, pa je zato važno znati i razumeti sve korake u izgradnji i evaluaciji modela. Bez obzira na različite ciljeve u istraživanju, svaka konstrukcija modela treba da sadrži sledeće korake:

- Postavljanje ciljeva istraživanja
- Prikupljanje i obrada podataka za analizu
- Sprovodjenje istraživačke analize podataka
- Odredjivanje oblika modela
- Procena rezultata modela
- Validacija modela
- Implementacija modela
- Održavanje modela
- Obnova modela

Postavljanje ciljeva istraživanja
Pre konstrukcije bilo kakvog modela, potrebno je utvrditi ciljeve istraživanja. Ciljevi istraživanja se utvrdjuju u komunikaciji sa sektorima koji su usko povezani sa konstrukcijom modela, npr. IT sektor, menadžment kompanije, itd. U zavisnosti od ciljeva istraživanja, varira i izgled i obim podataka koji će se koristiti u istraživanju, zatim, varira i vremenski period u kojem se projekat može završiti i na kraju, što je možda i najvažnije, varira i cena izrade modela.

Prikupljanje i obrada podataka za analizu
Prikupljanje i obrada podataka često oduzima najviše vremena u konstrukciji modela. Podaci koji se koriste su obično „neuredni“ („sirovi podaci“), pa se velika količina vremena ulaže u njihovo sredjivanje. Posle sredjivanja podataka, obično se podaci podele u dva podskupa, jedan koji se koristi za konstrukciju modela (eng. training set), dok se drugi koristi za testiranje modela (eng. test set).

Sprovodjenje istraživačke analize podataka
Istraživačka analiza podataka se nadovezuje na obradu podataka. Istraživačka analiza ima ulogu u prikazu odnosa zavisnih i nezavisnih promenljivih u modelu. Obično se analiza sprovodi grafički, tj. crtaju se grafici na kojima je prikazan odnos zavisne i nezavisne promenljive. Jedan od takvih prikaza je i matrični grafički prikaz (eng. scatter plot matricies) odnosa promenljivih u modelu.

Odredjivanje oblika modela
U ovom koraku konstrukcije modela, potrebno je odrediti model koji najbolje opisuje cilj projekta i podatke koje imamo. Neki od oblika modela koji se često koriste su stabla odlučivanja, neuronske mreže, uopšteni linearni modeli, itd. Kod uopštenih linearnih modela se u ovom koraku bira i link funkcija koja najbolje povezuje zavisnu i nezavisne promenljive.

Procena rezultata modela
Sa dobijanjem prvih preliminarnih rezultata modela, potrebno je analizirati model u cilju njegovog poboljšanja. Analiza modela uključuje procenu statističke značajnosti svakog parametra modela. Krajnji rezultat ovog koraka je model koji najbolje opisuje date podatke. Jedan od najčešćih statističkih metoda koji se koriste u ovom koraku kod uopštenih linearnih modela je stepenasta statistička metoda.

Validacija modela
Validacija modela podrazumeva testiranje modela na podskupu podataka koji je namenjen za testiranje modela. Postoji nekoliko načina testiranja dobijenog modela, kao što su grafičko predstavljanje predvidjenih i stvarnih vrednosti zavisne promenljive, kvantilni grafici (namenjeni za poredjenje dva modela), površina ispod ROC krive (za logističke modele), itd.

Implementacija modela
Glavni cilj većine modela je da se model pretvori u neku vrstu proizvoda. Na primer, u industriji osiguranja, taj proizvod može biti skoring sistem za odredjene klase osiguranika, odnosno pomoću modela se prave bodovne skale, koje se primenjuju u pravljenju nove tarife.

Održavanje modela
Vremenom, moć konstruisanog modela opada najčešće zbog promena na tržištu. Zbog toga je potrebno „održavati“ model. Održavanje modela se sprovodi periodično, kako bi se moć modela održala. Takodje, bitno je pri osvežavanju modela, osvežiti i podatke, tj. koristiti novije podatke za konstrukciju modela.

Obnova modela
Obnova modela podrazumeva ponovnu konstrukciju modela. To se radi kako zbog korišćenja novog seta podataka, tako i ako želimo dodati novu promenljivu u model. Na primer, osiguravajuća kuća je dobila podatke o istoriji šteta postojećih osiguranika koje pri ranijim konstrukcijama modela nije imala i menadžment želi da poveća buduće premije za osiguranike koji imaju zahteve za odštetu.

2 Osnovni pojmovi osiguranja

2.1 Uvod

Istorija osiguranja seže daleko u prošlost, u doba Vavilonaca koji su pre 4 000 godina primenjivali oblik osiguranja koji se sprovodio tako što se nadoknadjivala šteta vlasniku broda koji je pretrpeo štetu na trgovačkom putovanju, a zauzvrat svaki od trgovaca (vlasnika brodova) je morao isplatiti jedan deo svoje dobiti ako njegov brod nije pretrpeo nikakvu štetu. Prva sačuvana osigurana polisa potiče iz Lombardije 1182. godine. U to vreme pomorsko osiguranje je bilo veoma razvijeno, što je rezultat velikog broja trgovačkih putovanja Mediteranom, a i šire.
Suština postojanja osiguranja se nije mnogo promenila ni danas. Ono ima ulogu, pre svega, u materijalnoj zaštiti ljudi (osiguranih lica) od štetnih dogadjaja. Osnova osiguranja je rizik, tj. rizik da će doći do oštećenja ili gubitka imovine usled nekog neočekivanog dogadjaja ili rizik da će doći do narušavanja zdravlja ili čak gubitka života takodje usled nekog neočekivanog dogadjaja, ... **Rizik** možemo definisati kao neizvesnost ishoda budućeg dogadjaja.
Na taj način možemo definisati **osiguranje** kao nauku koja se bavi proučavanjem delovanja ostvarenja rizika, ekonomskim posledicama ostvarenog rizika, te izučavanjem načina upravljanjem rizikom kako bi se umanjile i eventualno sprečile mogućnosti nastanka rizika.[1]
Kada dodje do ostvarenja rizika, uvek imamo za posledicu materijalni (gubitak imovine) ili moralni gubitak (gubitak života ili zdravlja). Ukoliko osiguranik kupi osiguranje (polisu osiguranja) , taj rizik na sebe preuzima osiguravajuća kuća. **Polisom osiguranja** osiguravajuća kuća se obavezuje da će, usled ostvarenja rizika, osiguraniku isplatiti odredjenu osiguranu sumu novca u zavisnosti od veličine nastale štete. Cenu polise osiguranja nazivamo **premijom osiguranja**.
Naučna osnova modernog osiguranja omogućava da ono funkcioniše na principima ekonomski racionalnog poslovanja, koje se postiže korišćenjem matematike osiguranja čiju osnovu predstavlja aktuarska matematika[1]. Teškoće u predvidjanju dogadjaja su problemi koje aktuarska matematika uspešno rešava koristeći se zakonom velikih brojeva i računom verovatnoće. Zakon velikih brojeva podrazumeva da pri objedinjavanju velike mase rizika u jednu zajednicu rizika, slučajnost kao karakteristika pojedinih rizika se sve manje ističe.

2.2 Vrste osiguranja

Osiguranje možemo podeliti po raznim kreiterijumima i predmetima osiguranja. Neka osnovna podela osiguranja u većini zemalja, pa tako i kod nas je podela na **životna** i **neživotna** osiguranja koja se kasnije dele prema riziku koji pokriva osiguranje. Zakonom o osiguranju u Republici Srbiji , izvršena je podela životnih osiguranja na

1. Osiguranje života
2. Osiguranje za slučaj venčanja i rodjenja
3. Rentno osiguranje
4. Dopunsko osiguranje uz osiguranje života
5. Životna osiguranja iz tačaka 1 i 3 vezana za jedinice investicionih fondova
6. Tontine

[1] Aktuarska matematika je grana primenjene matematike koja se bavi osnovama osiguranja života i imovine, tj. bavi se izračunavanjem tarifa osiguranja lica i imovine, izračunavanjem rezrevi u osiguranju, itd. . Aćimović S. „Ekonomski rečnik“

7. Osiguranje s kapitalizacijom isplate

i podela neživotnih osiguranja na

1. Osiguranje od posledica nezgode
2. Dobrovoljno zdravstveno osiguranje
3. Osiguranje motornih vozila
4. Osiguranje šinskih vozila
5. Osiguranje vazduhoplova
6. Osiguranje plovnih vozila
7. Osiguranje robe u prevozu
8. Osiguranje imovine od požara i drugih opasnosti
9. Ostala osiguranja imovine
10. Osiguranje od odgovornosti zbog upotrebe motornih vozila
11. Osiguranje od odgovornosti zbog upotrebe vazduhoplova
12. Osiguranje od odgovornosti zbog upotrebe plovnih objekata
13. Osiguranje od opšte odgovornosti za štetu
14. Osiguranje kredita
15. Osiguranje jemstva
16. Osiguranje finansijskih gubitaka
17. Osiguranje troškova pravne zaštite
18. Osiguranje pomoći na putovanju

Jedna od specifičnosti životnih osiguranja je to da osiguranik utvrdjuje **osiguranu sumu**, na osnovu koje osiguravajuća kuća odredjuje visinu premije. Tačnije, osigurana suma nema gornje ograničenje kod ovakve vrste osiguranja. Sa druge strane, kod neživotnih osiguranja uvek imamo gornje ograničenje osigurane sume.
Pored ove podele, imamo i podelu prema kriterijumu obaveznosti na

- Dobrovoljna osiguranja
- Obavezna osiguranja

Dobrovoljna osiguranja, kao što i sama reč kaže obuhvata sva osiguranja kod kojih se ugovori sklapaju dobrovoljno uz pristanak klijenta i osiguravajuće kuće (na primer, to je velika većina osiguranja na našem tržištu, dobrovoljno zdravstveno osiguranje, osiguranje motornih vozila, osiguranje života, ...). Dok sa druge strane obavezna osiguranja su sva osiguranja koja su zakonom propisana u cilju zaštite gradjana (kao na primer, zdravstveno osiguranje, penziono osiguranje, osiguranje od odgovornosti zbog upotrebe motornih vozila, ...).

2.2.1 Kasko osiguranje motornih vozila

Pojam kasko osiguranja vezujemo za španski reč kasko koja znači trup broda. Ova reč se u prošlosti upotrebljavala prvo u pomorskom osiguranju, što potvrdjuje razvoj osiguranja najpre u osiguranju brodova. Kasnije se ovaj izraz ustalio u osiguranju transportnih sredstava.
U Republici Srbiji, kasko osiguranje motornih vozila je dobrovoljno osiguranje. Kod naših osiguravajućih kuća osnovno kasko osiguranje obuhvata pokriće većine šteta na motornom vozilu prouzrokovanih za vreme vožnje, u mirovanju ili na parkiralištu. Rizici koji su pokriveni osnovnim kasko osiguranjem su

- saobraćajne nezgode bez obzira na krivicu osiguranika
- pad ili udar nekog predmeta
- požar
- iznenadno hemijsko ili termičko delovanje spolja
- grom
- ekplozija, osim nuklearne eksplozije
- oluja
- grad
- snežna lavina
- pad vazdušne letelice i njenih delova
- manifestacije i demonstracije
- poplave, bujice ili visoke vode
- štete od divljači ili domaćih životinja

Pored osnovnog kasko osiguranja, u našim osiguravajućim kućama se nude dopunska i delimična kasko osiguranja. Dopunska kasko osiguranja se uzimaju uz osnovno kasko osiguranje i obuhvataju osiguranje prtljaga, alata, opreme motornog vozila, itd. . Delimično kasko osiguranje uglavnom podrazumeva osiguranje prednjih, zadnjih i bočnih stakala na motornom vozilu.
Neki od elemenata za utvrdjivanje premije kasko osiguranja vozila kod naših osiguravača su

- vlasnik vozila (fizičko/pravno lice)
- vrsta vozila (putničko/teretno)
- marka vozila
- model i tip vozila
- godina proizvodnje vozila
- snaga motora (u KW)
- zapremina motora

- vrsta goriva ili pogona
- vozilo je novo (da/ne)
- vozilo je kupljeno na lizing/rentu
- broj vrata na vozilu
- datum prve registracije
- mesto osiguranika
- teritorijalno pokriće (RS+Evropa/RS)

Premija kasko osiguranja zavisi i od **učešća** u šteti osiguranika. To je procenat iznosa od moguće štete koju osiguranik premuzima na sebe, pa se u zavisnosti od iznosa procenta, premija umanjuje. Takodje, za „dobre“ vozače (oni koji u prethodnom osiguranom periodu nisu imali naplatu štete od strane osiguravača) tu je premijska olakšica, odnosno **bonus**, dok je za „loše“ vozače (oni koji važe za vozače koji imaju istoriju naplate štete od osiguravača) predvidjeno povećanje premije, odnosno **malus**.
Kasko osiguranje ne važi za one štete koje su nastale usled namere vozača, ukoliko je vozilo bilo transportovano drugim vozilom, usled rata, pobune, nemira, zemljotresa, nuklearnih rizika ili ukoliko je vozač bio pod uticajem alkohola, droga ili drugih psihoaktivnih supstanci. Takodje kasko osiguranje ne važi i ukoliko je vozilo ukradeno od strane člana porodice ili člana domaćinstva ili ukoliko je prilikom udesa za volanom sedela osoba bez vozačke dozvole. Ako se prilikom udesa desi šteta za koju se isplati cela osiguravajuća suma, onda se šteta proglašava totalnom štetom i tada kasko ugovor prestaje da važi.

3 Uopšteni linearni modeli

U velikom broju konstrukcije skoring modela cilj je da se opišu veze izmedju pojava na tržištu na osnovu nekog skupa podataka. To se obično postiže pronalaženjem jednačine koja povezuje veličine koje posmatramo. Neka je Y promenljiva koje želimo da opišemo na osnovu informacija $x_{i1}, x_{i2}, ..., x_{ip} \in \mathbb{R}$, $i = 1, 2, ..., n$ $p \in \mathbb{N}$. Promenljivu Y nazivamo zavisnom promenljivom, dok $x_{i1}, x_{i2}, ..., x_{ip}$ nazivamo nezavisnim promenljivama, prediktorima ili kontrolisanim faktorima.
Problem nalaženja funkcionalne zavisnosti izmedju zavisne i nezavisnih promenljivih se svodi na odabir funkcije koja dobro aproksimira neki skup podataka. U smislu funkcionalne zavisnosti razlikujemo linearne i nelinearne modele. Skoring sistemi kojima ćemo se baviti u ovom radu, baziraju se na linearnim modelima, tačnije uopštenim linearnim modelima, pa ćemo u nastavku detaljnije objasniti ovaj pojam, kao i njegova najvažnija svojstva.
Definicija klasičnog (normalnog) linearnog modela glasi

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$E(Y) = \mu = X\boldsymbol{\beta}\ ,$$

gde je Y vektor realizovanih vrednosti zavisnih promenljivih , X je matrica prediktora , $\boldsymbol{\beta}$ je vektor koeficijenata modela i $\boldsymbol{\varepsilon}$ je vektor reziduala , tačnije

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},\ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}\ .$$

Pretpostavke ovog modela su

(i) Y je merljiva slučajna promenljiva

(ii) $x_{i1}, x_{i2}, ..., x_{ip}$ $i = 1, 2, ..., n$ su neslučajne promenljive sa fiksnim vrednostima i pretpostavlja se da su medjusobno nezavisne

(iii) $E(\varepsilon_i) = 0$, $i = 1, 2, ..., n$

(iv) $Var(\varepsilon_i) = \sigma^2$, $i = 1, 2, ..., n$, tj. pretpostavlja se da su varijanse reziduala konstatnte. Ova osobina se naziva homoskedastičnost.

(v) Reziduali su nekolinearni, tj. $Cov(\varepsilon_i, \varepsilon_j) = 0$, $\forall i, j = 1, 2, ..., n$, $i \neq j$

(vi) Reziduali $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ imaju normalnu raspodelu.

Za razliku od klasičnog linearnog modela, kod uopštenog linearnog modela raspodela zavisne slučajne promenljive ne mora da bude normalna, nego pripada **familiji eksponencijalnih raspodela**, koja je definisana funkcijom gustine na sledeći način:
Za slučajnu promenljivu Y, čija raspodela pripada familiji eksponencijalnih rapsodela, funkcija gustine je data sa

$$\varphi(y) = c(y, \phi) \exp\left[\frac{y\theta - a(\theta)}{\phi}\right].$$

U ovako definisanoj raspodeli a i c su funkcije koje zavise od parametara θ, ϕ, $y \in \mathbb{R}$. Parametar ϕ se zove **disperzioni parametar**, dok se θ zove **kanonički parametar**. Za funkciju $a(\theta)$ pretpostavljamo da je dva puta neprekidno diferencijabilna. Izbor funkcija $a(\theta)$ i $c(y, \phi)$ direktno

odredjuje gustinu $\varphi(y)$, odnosno raspodelu slučajne promenljive Y.

Dakle , uopšteni linearni model se sastoji od

(i) Zavisne slučajne promenljive Y čija raspodela pripada familiji eksponencijalnih raspodela

(ii) Linearnog prediktora, tj.
$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

(iii) Link funkcije $g(\bullet)$ koja povezuje zavisnu promenljivu i linearnog prediktora na sledeći način
$$g(E(Y_i)) = g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \tag{1}$$
gde Y_i predstavlja i-tu komponentu prostog slučajnog uzorka za promenljivu Y. Za link funkciju pretpostavljamo da je monotona i diferencijabilna na svom domenu.

Kod uopštenih linearnih modela često se pojavljuje i pojam **kanoničke link funkcije**. Za kanoničku link funkciju važi $g(\mu_i) = \theta_i$, pa je $\theta_i = X_i\boldsymbol{\beta}$. U ovom radu koristićemo logaritamsku i identičku link funkciju, medjutim u ovakvim modelima se često koriste i stepena, logit, itd. .

Teorema 3.1. *Za svaku slučajnu promenljivu Y, čija raspodela pripada familiji eksponencijalnih raspodela , važi sledeće*

(i) $E(Y) = a'(\theta)$

(ii) $Var(Y) = \phi a''(\theta)$

Dokaz. Za slučajnu promenljivu Y znamo da je njena funkcija gustine data sa
$$\varphi(y) = c(y, \phi) \exp\left[\frac{y\theta - a(\theta)}{\phi}\right] .$$
Prvi i drugi izvod funkcije gustine po parametru θ dati su sa
$$\frac{\partial \varphi}{\partial \theta} = \varphi(y)\left(\frac{y - a'(\theta)}{\phi}\right) \text{ i } \frac{\partial^2 \varphi}{\partial\theta\partial\theta} = \varphi(y)\left(\frac{y - a'(\theta)}{\phi}\right)^2 - \varphi(y)\frac{a''(\theta)}{\phi} .$$
Sada kako je
$$\int \frac{\partial \varphi}{\partial \theta} dy = \frac{\partial}{\partial \theta} \underbrace{\int \varphi dy}_{=1} = 0 \text{ i } \int \frac{\partial^2 \varphi}{\partial\theta\partial\theta} dy = \frac{\partial^2}{\partial\theta\partial\theta} \underbrace{\int \varphi dy}_{=1} = 0 ,$$
imamo
$$0 = \int \varphi(y)\left(\frac{y - a'(\theta)}{\phi}\right) dy = \frac{E(Y) - a'(\theta)}{\phi} \Leftrightarrow E(Y) = a'(\theta)$$
i
$$0 = \int \varphi(y)\left(\frac{y - a'(\theta)}{\phi}\right)^2 dy - \int \varphi(y)\frac{a''(\theta)}{\phi} dy = E((Y - a'(\theta))^2) - \phi a''(\theta) \Leftrightarrow$$
$$Var(Y) = E((Y - E(Y))^2) = E((Y - a'(\theta))^2) = \phi a''(\theta) .$$
□

Primetimo, za disperziju važi još i
$$Var(Y) = \phi a''(\theta) = \phi \frac{\partial a'(\theta)}{\partial \theta} = \phi \frac{\partial \mu}{\partial \theta} = \phi V(\mu) ,$$
što nam govori o vezi očekivane vrednosti i varijanse promenljive. Funkcija $V(\mu)$ naziva se varijansna funkcija.

3.1 Ocena nepoznatih parametara raspodele zavisne promenljive

Nepoznate parametre raspodele slučajne promenljive Y izračunaćemo pomoću metode maksimalne verodostojnosti. Ideja ove metode je da se pomoću realizovanog uzorka odabere ona vrednost nepoznatog parametra koja daje najveću verovatnoću da baš taj realizovani uzorak bude odabran. Pretpostavimo da je naš realizovani uzorak dat sa $(y_1, y_2, ..., y_n)$. Tada za bilo koju raspodelu iz eksponencijalne familije raspodela, funkcija verodostojnosti data je sa

$$L(\phi, \theta; y_1, y_2, ..., y_n) = \varphi(y_1)\varphi(y_2)...\varphi(y_n) = \prod_{i=1}^{n} \varphi(y_i) \ . \tag{2}$$

Dakle, zadatak metode maksimalne verodostojnosti je naći vrednosti ϕ i θ koje maksimiziraju funkciju verodostojnosti, odnosno tražimo

$$\max_{\phi,\theta} \prod_{i=1}^{n} \varphi(y_i) \ .$$

Jedna od glavnih ideja u rešavanju ovog problema je traženje maksimuma logaritmovane funkcije verodostojnosti. Na taj način se umesto proizvoda u jednačini (2) pojavljuje suma, što znatno olakšava dalje nalaženje maksimuma. Iz definicije funkcije gustine znamo da je $\varphi(y_i) \geq 0$ za sve $i = 1, 2, ..., n$, pa zbog monotonosti funkcije ln imamo da funkcije L i $\ln L$ postižu maksimum za iste vrednosti parametara. Dakle, ako logaritmujemo funkciju verodostojnosti datu u (2), dobijamo

$$\ln L(\phi, \theta; y_1, y_2, ..., y_n) = \sum_{i=1}^{n} \ln(\varphi(y_i)) \ .$$

Na osnovu jednačine gustine eksponencijalne familije raspodela, imamo da je

$$\ln L(\phi, \theta; y_1, y_2, ..., y_n) = \sum_{i=1}^{n} \Big\{ \ln c(y_i, \phi) + \frac{y_i\theta - a(\theta)}{\phi} \Big\} = \sum_{i=1}^{n} \Big\{ \ln c(y_i, \phi) \Big\} + \frac{n(\overline{y}\theta - a(\theta))}{\phi},$$

gde je $\overline{y} = \dfrac{1}{n} \sum_{i=1}^{n} y_i$.
Da bismo našli parametre θ i ϕ koji maksimiziraju $\ln L$, dovoljno je da nadjemo njene stacionarne tačke. Ovo sledi iz osobine da je logaritamska funkcija monotono rastuća. Stacionarne tačke dobijamo kada parcijalne izvode funkcije $\ln L$ izjednačimo sa nulom.
Parcijalni izvod funkcije $\ln L$ po parametru θ dat je sa

$$\frac{\partial \ln L}{\partial \theta} = \frac{n(\overline{y} - a'(\theta))}{\phi} = 0,$$

pa se ocenjivač $\hat{\theta}$ dobija rešavanjem diferencijalne jednačine

$$a'(\theta) = \overline{y} \ .$$

Napomena 3.1. *Iz diferencijalne jednačine vidimo da je $E(Y) = a'(\theta) = \overline{y}$, odnosno da je ocena očekivanja slučajne promenljive Y jednaka aritmetičkoj vrednosti realizovanog uzorka.*

Ocena $\hat{\phi}$ se takodje dobija iz parcijalne jednačine kada odgovarajući parcijalni izvod izjednačimo sa nulom, medjutim račun je mnogo komplikovaniji, pa se za ocenjivanje disperzionog parametra obično koristi Pirsonova χ^2 statistika.

3.2 Ocena koeficijenata uopštenog linearnog modela

3.2.1 Metoda maksimalne verodostojnosti

Predjimo sada na ocenjivanje nepoznatih koeficijenata modela (1), odnosno ocenjivanje parametara $\beta_0, \beta_1, ..., \beta_p$. Kako je $g(\mu_i) = g(a'(\theta_i)) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$, vidimo da parametar θ zavisi od koeficijenata modela, pa možemo napisati $\theta_i = \theta_i(\beta_0, \beta_1, ..., \beta_p)$. Koristeći sada pravilo za izvod složene funkcije i ocene parametra θ_i raspodele Y_i, dobijamo parcijalne izvode logaritamske funkcije verodostojnosti po nepoznatim parametrima $\beta_0, \beta_1, ..., \beta_p$

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ln L}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i} \frac{x_{ij}(y_i - \mu_i)}{\phi} = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i} x_{ij}(y_i - \mu_i) = 0, \quad j = 1, 2, ..., p$$

i analogno

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i}(y_i - \mu_i) = 0,$$

gde je $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$.
Parcijalne izvode možemo zapisati u matričnom obliku

$$\frac{1}{\phi} X^T D(\boldsymbol{y} - \boldsymbol{\mu}) = 0 \Leftrightarrow X^T D(\boldsymbol{y} - \boldsymbol{\mu}) = 0, \tag{3}$$

gde su

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

i $D = diag\left(\frac{\partial \theta_1}{\partial \eta_1}, \frac{\partial \theta_2}{\partial \eta_2}, ..., \frac{\partial \theta_n}{\partial \eta_n}\right)$.
Primetimo da je

$$\left(\frac{\partial \theta_i}{\partial \eta_i}\right)^{-1} = \frac{\partial \eta_i}{\partial \theta_i} = \frac{\partial \eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial g(\mu_i)}{\partial \mu_i} \frac{\partial a'(\theta_i)}{\partial \theta_i} = g'(\mu_i) a''(\theta_i),$$

pa je $D = diag\left((g'(\mu_1)a''(\theta_1))^{-1}, (g'(\mu_2)a''(\theta_2))^{-1}, ..., (g'(\mu_n)a''(\theta_n))^{-1}\right)$.
Dalje, kako je $(g'(\mu_i)a''(\theta_i))^{-1} = \left((g'(\mu_i))^2 a''(\theta_i)\right)^{-1} g'(\mu_i)$, matricu D možemo zapisati u obliku $D = QG$, gde su $Q = diag\left(\left((g'(\mu_1))^2 a''(\theta_1)\right)^{-1}, \left((g'(\mu_2))^2 a''(\theta_2)\right)^{-1}, ..., \left((g'(\mu_n))^2 a''(\theta_n)\right)^{-1}\right)$ i $G = diag\left(g'(\mu_1), g'(\mu_2), ..., g'(\mu_n)\right)$.
Tejlorov razvoj funkcije g u tački y_i je dat sa

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i)\ .$$

Ako ovu aproksimaciju prebacimo u matrični oblik, dobijamo

$$G_y \approx G_\mu + G(\boldsymbol{y} - \boldsymbol{\mu}) \Leftrightarrow G(\boldsymbol{y} - \boldsymbol{\mu}) \approx G_y - G_\mu,$$

gde su

$$G_y = \begin{bmatrix} g(y_1) \\ g(y_2) \\ \vdots \\ g(y_n) \end{bmatrix}, \; G_\mu = \begin{bmatrix} g(\mu_1) \\ g(\mu_2) \\ \vdots \\ g(\mu_n) \end{bmatrix} .$$

Sada je

$$X^T D(\boldsymbol{y} - \boldsymbol{\mu}) = X^T Q G(\boldsymbol{y} - \boldsymbol{\mu}) \approx X^T Q(G_y - G_\mu) = X^T Q G_y - X^T Q G_\mu = 0 .$$

Konačno, kako je $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$, za sve $i = 1, 2, ..., n$, imamo da je $G_\mu = X\boldsymbol{\beta}$, pa je ocena koeficijenata metodom maksimalne verodostojnosti data sa

$$\hat{\boldsymbol{\beta}} \approx \left(X^T Q X\right)^{-1} X^T Q G_y .$$

Ocene dobijene metodom maksimalne verodostojnosti su asimptotski nepristrasne i za velike obime uzorka imaju približno normalnu raspodelu, pa je tako

$$\hat{\boldsymbol{\beta}} : \mathcal{N}(\boldsymbol{\beta}, \phi(X^T Q X)^{-1}), \quad n \to \infty .$$

Jednačinu (3) moguće je rešiti i iterativnim postupcima. U ovom radu pokazaćemo primenu iterativnog postupka ponderisanih najmanjih kvadrata.

3.2.2 Iterativni postupak najmanjih kvadrata

Posmatramo jednačinu $X^T Q G(\boldsymbol{y} - \boldsymbol{\mu}) = 0$. Koristeći Tejlorovu aproksimaciju funkcije do prvog reda dobijamo

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(y - \mu_i) .$$

Neka je $z_i = g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$. Primetimo da važi

$$E(z_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

i

$$V(z_i) = \left(g'(\mu_i)\right)^2 V(\mu_i) .$$

Neka je $\hat{\eta}_i^{(0)}$ početna ocena linearnog prediktora η_i i neka je $\hat{\mu}_i^{(0)}$ odgovarajuća vrednost dobijena iz $\hat{\eta}_i^{(0)} = g(\hat{\mu}_i^{(0)})$. Tada je $z_i^{(0)} = \hat{\eta}_i^{(0)} + g'(\mu_i^{(0)})(y_i - \hat{\mu}_i^0)$. Za početnu vrednost $\mu_i^{(0)}$ se obično bira $\mu_i^{(0)} = y_i$. Težinski koeficijent $Q_i^{(0)}$ je definisan sa $Q_i^{(0)} = \dfrac{1}{diag\left[(g'(\mu_i^{(0)}))^2 V(\mu_i^{(0)})\right]}$, gde je sa V definisana varijansna matrica.
Iterativni postupak ponderisanih najmanjih kvadrata je sada dat sa

(1) U svakoj iteraciji računamo vrednost

$$z_i^{l-1} = \hat{\eta}_i^{(l-1)} + g'(\mu_i^{(l-1)})(y_i - \hat{\mu}_i^{(l-1)})$$

i težinske koeficijente

$$Q_i^{(l-1)} = \frac{1}{diag\left[\left((g'(\mu_i^{(l-1)}))^2 V(\mu_i^{(l-1)})\right)\right]} .$$

(2) Formiramo regresiju sa težinama, gde je $\boldsymbol{z}^{(l-1)}$ vektor zavisne promenljive dimenzije $n \times 1$, zatim X je matrica nezavisnih promenljivih dimenzija $(p+1) \times n$, dok matrica Q predstavlja matricu dimenzija $n \times n$ sa težinskim koeficijentima na dijagonali. Koeficijente formirane regresije dobijamo iz sledeće jednakosti $\boldsymbol{b}^{(l)} = \left(X^T Q^{(l-1)} X\right)^{-1} X^T Q^{(l-1)} \boldsymbol{z}^{(l-1)}$.

(3) Korake (1) i (2) ponavljamo sve dok promene u ocenjenim parametrima ne budu dovoljno male.

3.2.3 Pirsonova statistika za ocenu disperzionog parametra

Kao što smo rekli, za ocenu disperzionog parametra ϕ koristimo Pirsonovu χ^2 statistiku. Pirsonova statistika je data sa

$$\chi^2 = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \ .$$

Data statistika ima približno $\chi^2_{n-(p+1)}$ raspodelu, čije je očekivanje $E(\chi^2) = n - (p+1)$. Ocena parametra ϕ je sada

$$\hat{\phi} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \ .$$

3.3 Valdov test za testiranje hipoteza o statističkoj značajnosti koeficijenata u modelu

Kod testiranja hipoteza za pojedinačne ili za grupu koeficijenata uopštenog linearnog modela, hipoteze se pišu u obliku $C\boldsymbol{\beta} = r$, gde se C naziva hipotetička matrica, dok je r vektor datih vrednosti.
Uopšteno, ako se testira hipoteza $H_0(C\boldsymbol{\beta} = r)$ protiv alternativne $H_1(C\boldsymbol{\beta} \neq r)$, onda Valdov[2] test posmatra razliku $C\hat{\boldsymbol{\beta}} - r$. Velike vrednosti te rezlike ukazuju da nulta hipoteza nije tačna.
Ako pretpostavimo da je H_0 tačna, onda je $C\boldsymbol{\beta} = r$, pa kako je za uzorke velikog obima $\hat{\boldsymbol{\beta}} : \mathcal{N}(\boldsymbol{\beta}, \phi(X^T Q X)^{-1})$, imamo da

$$C\hat{\boldsymbol{\beta}} - r : \mathcal{N}(0, \phi C (X^T Q X)^{-1} C^T),$$

pa je Valdova test statisitka data sa

$$(C\hat{\boldsymbol{\beta}} - r)^T \left[\phi C (X^T Q X)^{-1} C^T\right]^{-1} (C\hat{\boldsymbol{\beta}} - r) : \chi^2_q,$$

gde q označava broj stepeni slobode χ^2 raspodele. Broj q je zapravo broj restrikcija na vektoru $\boldsymbol{\beta}$, tj. broj nenula redova matrice C.
U slučaju da testiramo hipotezu $\beta_j = r$, onda je matrica C data kao matrica u koja u j-toj koloni ima sve jedinice, dok su na svim ostalim pozicijama nule. Valdov test za testiranje hipoteze o jednom parametru $H_0(\beta_j = r_j)$ protiv alternativne $H_1(\beta_j \neq r_j)$ dat je statistikom

$$w_j = \frac{(\hat{\beta}_j - r_j)^2}{b_{jj}}, \quad r_j \in \mathbb{R}, \quad j = 0, 1, ..., p$$

koja približno ima χ^2_1 raspodelu. Sa b_{jj} označen je dijagonalni element matrice $B = \phi(X^T Q X)^{-1}$.

Napomena 3.2. *Kada je ϕ nepoznato, zamenjujemo ga ocenjenom vrednošću $\hat{\phi}$.*

Valdov test za uopštene linearne modele je u statističkom softveru „R“ dat funkcijom *wald.test()*.

[2]eng. Wald test - nazvan po Abrahamu Valdu (eng. Abraham Wald) madjarskom matematičaru koji je živeo od 1902.-1950. godine.

3.4 Analiza reziduala

Reziduali kod normalnog linearnog modela su dati sa $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Za njih važe sledeće pretpostavke

- $\varepsilon_i : \mathcal{N}(0, \sigma^2)\ \forall i = 1, 2, ..., n$
- $cov(\varepsilon_i, \varepsilon_j) = 0,\ \forall i \neq j$

Kod uopštenih linearnih modela definicija reziduala je uopštenija kako bi bila primenljiva za sve raspodele iz eksponencijalne familije raspodela.

Pirson reziduali
Pirson reziduali su definisani na sledeći način

$$(r_p)_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

i važi da je

$$\sum_{i=1}^{n} (r_p)_i^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} : \chi\ .$$

Anskombe reziduali
Anskombe reziduale karakteriše funkcija $h(y)$ koja je definisana sa

$$h(y) = \int \frac{dy}{[V(y)]^{1/3}}\ .$$

Za ovu funkciju važi da je $h'(y) = [V(y)]^{-1/3}$. Anskombe reziduali su dati sa

$$(r_A)_i = \frac{h(y_i) - h(\hat{\mu}_i)}{h'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}}\ .$$

Tabela 1: Anskombe reziduali za neke raspodele

Raspodela	$(r_A)_i$
Normalna	$y_i - \mu_i$
Inverzna Gausova	$\frac{(\log(y_i) - \log(\hat{\mu}_i))}{\hat{\mu}_i^{1/2}}$
Poasonova	$\frac{2}{3}(y_i^{2/3}\hat{\mu}_i^{-1/6} - \hat{\mu}_i^{1/2})$
Gama	$3\left(\left(\frac{y_i}{\hat{\mu}_i}\right)^{1/3} - 1\right)$

Za Anskombe reziduale važi da imaju približno standardizovanu normalnu raspodelu. Anskombe reziduali se obično koriste kod uopštenih linearnih modela kada zavisna promenljiva nema normalnu raspodelu.

3.5 Autlajeri i uticajne tačke

Autlajeri predstavljaju opservacije iz uzorka za koje model ne daje dobru aproksimaciju. Drugim rečima to su tačke koje imaju veliki rezidual, pa se često otkrivaju pomoću grafičkog prikaza reziduala modela. Sa druge strane, **uticajne tačke** su opservacije koje imaju veliki uticaj na ocene nepoznatih parametara modela, pa samim tim uticajna tačka menja nagib regresijskog pravca. Autlajer ne mora biti uticajna tačka, a ni uticajna tačka ne mora biti autlajer.
Sve opervacije, za koje se sumnja da su uticajne tačke ili autlajeri, potrebno je ispitati. Analiza tih opervacija se vrši tako što se posmatraju parametri modela sa i bez te sumnjive opservacije. Ako se parametri modela značajno promene, onda možemo smatrati da je opservacija uticajna i ne smemo je zanemariti. Sa druge strane, autlajere koji nisu uticajne tačke, možemo isključiti iz dalje analize. Postoji nekoliko metoda za otkrivanje autlajera i uticajnih tačaka, medju kojima su najpoznatiji Leveridž, Dfbeta i Cook's rastojanje.

Leveridž metod
Matrica H za koju važi

$$\hat{\boldsymbol{\mu}} = H\boldsymbol{\mu}$$

nazivamo „kapa matricom"(eng. Hat matrix). Matrica H je simetrična i idempotentna[3] matrica. Kod uopštenih linearnih modela „kapa matrica" data je sa

$$H = Q^{1/2}X(X^TQX)^{-1}X^TQ^{1/2} .$$

Za i-tu opservaciju iz uzorka, vrednost leveridž predstavlja izvod dijagonalni element matrice H, tj. vrednost $h_{ii} \in [0, 1]$. Drugim rečima, vrednost leveridž predstavlja odstupanje i-te opservacije od preostalih opservacija. Kako je $tr(H) = \sum_{i=1}^{n} h_{ii} = p + 1$, sledi da je prosečna vrednost leveridža za svaku opservaciju jednak $\frac{p+1}{n}$, pa se sve tačke koje imaju vrednost leveridža veću od $\frac{2p+2}{n}$ smatraju potencijalnim autlajerima.

Dfbeta metod
Vrednost dfbeta predstavlja promenu ocenjenog koeficijenta modela kada je i-ta opservacija obrisana. Ako sa $\hat{\beta}_j^{(i)}$ obeležimo ocenu koeficijenata koji je dobijen bez i-te opservacije, onda je vrednost dfbeta data sa

$$D_i = \frac{1}{p}(\hat{\beta}_j - \hat{\beta}_j^{(i)})^T X^T X(\hat{\beta}_j - \hat{\beta}_j^{(i)}) .$$

Sve opservacije za koje je D_i veće od $\frac{2}{\sqrt{n}}$, su potencijalne uticajne tačke i njih je potrebno dodatno ispitati.

Cook's rasojanje
Slično kao i kod dfbeta vrednosti, Cook's rastojanje se koristi kako bi se ispitao uticaj opservacije na parametre modela. Cook's rastojanje je dato statistikom

$$C_i = \frac{(y_i - \hat{\mu}_i)^2}{(p+1)V(\hat{\mu}_i)(1-h_{ii})}\frac{h_{ii}}{1-h_{ii}} .$$

Opservacije i za koje je C_i veće od $\frac{4}{n}$ potrebno je dodatno ispitati.

[3] $H^2 = H$

3.6 Provera reprezentativnosti modela

3.6.1 Testiranje izbora link funkcije

Koriteći Tejlorov razvoj imamo da je

$$g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i) \ .$$

Ako je link funkcija dobra, trebalo bi da važi da je $g(\hat{y}_i) \approx \eta_i = X_i\hat{\boldsymbol{\beta}}$. Grafički, to znači da bi tačke $(g(\hat{\mu}_i) + g'(\hat{(\mu)}_i)(y_i - \hat{\mu}_i), \eta_i)$ trebalo približno da leže na istoj pravoj. Zakrivljenost znači da smo napravili loš izbor link funkcije.

3.6.2 Načini i kriterijumu za izbor modela

Uopšteni linearni modeli imaju veliki broj mogućih izbora modela, što zbog izbora link funkcija, tako i zbog izbora raspodele zavisne promenljive. Izbor modela podrazumeva traženje najreprezentativnijeg modela koji će na najbolji način opisati date podatke. Povećavanjem broja objašnjavajućih promenljivih, povećava se i preciznost modela, tj. smanjuje se razlika $y_i - \hat{y}_i$. Medjutim povećava se i broj nepoznatih parametara modela što dovodi do veće standardne devijacije ocenjenih parametara, tj. smanjujemo njihovu preciznost. Sa druge strane, mali broj prediktora ukazuje na manju standardnu devijaciju ocenjenih parametara, medjutim sa manjim brojem prediktora smanjuje se i preciznost modela. Dakle, potrebno je naći balans izmedju preciznosti modela sa jedne strane i statističke značajnosti ocenjenih koeficijenata, sa druge strane.
Što se tiče načina na koji se biraju modeli, najpoznatije su stepenaste metode. Postoje stepenasta metoda unapred, stepenasta metoda unazad i kombinovana stepenasta metoda. Kod stepenaste metode unapred kreće se od modela bez prediktora , kojem se postupno, po odredjenim kriterijumima dodaje po jedan prediktor u svakom koraku, sve dok se ne dodje do modela kod kojeg, dodavanjem bilo kojeg novog prediktora, smanjujemo reprezentativnost modela. Stepenasta metoda unazad je obrnuta varijanta stepenaste metode unapred. Kod stepenaste metode unazad krećemo od modela u kojem su uključeni svi mogući prediktori iz skupa podataka. Zatim se, po odredjenim kriterijumima uklanja jedan po jedan prediktor sve dok se ne dodje do modela kojem, uklanjanjem bilo kojeg njegovog prediktora, narušavamo njegovu reprezentativnost. Kombinovana stepenasta metoda je kombinacija prethodne dve metode.
Najpoznatiji kriterijumi koji se koriste za uporedjivanje dva modela su

- AIC[4] kriterijum
- BIC[5] kriterijum

AIC kriterijum našeg uopštenog linearnog modela je dat sa

$$AIC = -2\ln L + 2(p+1),$$

dok je BIC dat sa

$$BIC = -2\ln L + (p+1)\ln(n) \ .$$

Primetimo da se vrednosti AIC i BIC kriterijuma smanjuje povećanjem vrednosti $\ln L$, što znači da bolje fitovani modeli imaju manje vrednosti AIC i BIC. To znači, da ako poredimo AIC(BIC)

[4] **A**kaike's **I**nformation **C**riterium
[5] **B**ayesian **I**nformation **C**riterium

vrednosti dva modela, bolji model biti onaj sa manjom vrednosti AIC(BIC) kriterijuma. Sa druge strane vidimo da se kod oba kriterijuma dodaje multiplikovana vrednost $p + 1$, što znači da AIC i BIC kriterijumi „kažnjavaju“ uvodjenje novih prediktora u model.
Većina statističkih softvera imaju ugradjene kriterijume AIC i BIC kod računanja performansi modela.
U statističkom paketu „R“ funkcija *stepAIC()* nalazi najbolji model na osnovu stepenaste metode i AIC kriterijuma.

3.7 Kategorijalne promenljive kod regresionih modela

U prethodnim poglavljima smo videli da na zavisnu promenljivu može uticati faktor preko svojih kategorija (nivoa). Takve nezavisne promenljive se zovu kategorijalne (opisne) promenljive. Unija kategorija neke kategorijalne promenljive čini skup svih mogućih vrednosti te promenljive u okviru posmatrane populacije. Sa druge strane, kategorije su disjunktne, odnosno svaki član populacije pripada tačno jednoj kategoriji posmatrane kategorijalne promenljive. Često se dešava da kategorije nisu numeričkog tipa, na primer, ako posmatramo faktor „pol“, moguće kategorije su „žena“ i „muškarac“. Postavlja se pitanje, kako se u matematičkom modeliranju i analizi interpretiraju ovakve kategorije?
Kategorjalne promenljive su promenljive koje uzimaju jednu od vrednosti iz diskretnog skupa. Diskretni skup je obično oblika $\{0,1\}$, $\{-1,0,1\}$ ili je to skup od nekoliko uzastopnih prirodnih brojeva. Razlikujemo tri vrste kategorijalnih promenljivih, a to su **ordinalne**, **nominalne** i **binarne (dihotomne)**. Kod ordinalnih varijabli kategorije nekog faktora su uredjene, tj. medju kategorijama postoji prirodni poredak, kao na primer kod klase automobila možemo imati 1- niža klasa, 2-srednja klasa i 3-visoka klasa. Sa druge strane, kod nominalnih varijabli nemamo uredjenje, tj. veća (manja) vrednost promenljive ne znači bolju (lošiju) kategoriju. Primer nominalne varijable može biti boja vozila, 1-crna boja, 2-siva boja, 3-bela boja, ... Binarne kategorijalne promenljive imaju samo dve moguće vrednosti (obično $\{0,1\}$) i koriste se kod faktora koji imaju samo dve kategorije. Na primer, ako posmatramo faktor , „da li polisa ima zahtev za odštetu? “, dihotomna promenljiva će uzeti vrednost 1 ako je odgovor „da“, a 0 ako je odgovor „ne“.
Binarne promenljive možemo koristiti i kod kategorijalnih promenljivih koji imaju više od dve kategorije. Na primer, ako posmatramo faktor „osiguranik“ i njegove kategorije „fizičko lice-žena“, „fizičko lice-muškarac“ i „pravno lice“, onda možemo definisati tri binarne promenljive

- $x_1 = \begin{cases} 1 & \text{ako je osiguranik fizičko lice-žena} \\ 0 & \text{inače} \end{cases}$
- $x_2 = \begin{cases} 1 & \text{ako je osiguranik fizičko lice-muškarac} \\ 0 & \text{inače} \end{cases}$
- $x_3 = \begin{cases} 1 & \text{ako je osiguranik pravno lice} \\ 0 & \text{inače} \end{cases}$

3.7.1 Binarne varijable

Rekli smo da binarne (eng. „dummy“) varijable imaju dve moguće vrednosti, 0 ili 1. Njihova interpretacija u regresionim modelima je sledeća:
Pretpostavimo da postoji kategorijalna promenljiva A koja ima p kategorija u nekom regresionom modelu. Tada možemo definisati p binarnih varijabli za tu kategorijalnu promenljivu. Medjutim ako bi sve te binarne varijable uključili u model, onda bi narušili pretpostavku o nezavisnosti objašnjavajućih promenljivih. Tačnije, tada se svaka binarna varijabla može predstaviti preko preostalih $p-1$ binarnih varijabli. Zbog toga za svaku kategorijalnu promenljivu definišemo referentni nivo, odnosno kategoriju sa kojom uporedjujemo sve ostale kategorije u modelu. Binarna varijabla referentnog nivoa ne ulazi u model. Dakle, ako imamo p kategorija neke kategorijalne promenljive, onda za tu kategorijalnu promenljivu definišemo $p-1$ binarnu varijablu koje ulaze u model.

Ispitivanje zavisnosti izmedju binarnih varijabli
Pretpostavimo da želimo da ispitamo nezavisnost dve binarne varijable x_1 i x_2. Tada je tabela kontigencije data sa

Tabela 2: Tabela kontigencije za varijable x_1 i x_2

		x_2		
		0	1	TOTAL
x_1	0	O_{11}	O_{12}	$n_{1\bullet}$
	1	O_{21}	O_{22}	$n_{2\bullet}$
TOTAL		$n_{\bullet 1}$	$n_{\bullet 2}$	n

gde vrednosti O_{ij}, $i, j = 1, 2$ predstavljaju brojeve opservacija iz uzorka za odgovarajuće vrednosti binarnih varijabli.
Zavisnost izmedju dve binarne varijable ispituje se statistikom

$$\phi_{test} = \frac{O_{11}O_{22} - O_{12}O_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}} \in [-1, 1] \ .$$

Sve vrednosti ϕ_{test} bliske 1 (-1) ukazuju na pozitivnu (negativnu) korelaciju izmedju dve varijable x_1 i x_2.

Pogledajmo sledeći primer i praktičnu primenu binarnih varijabli u regresionom modelu.

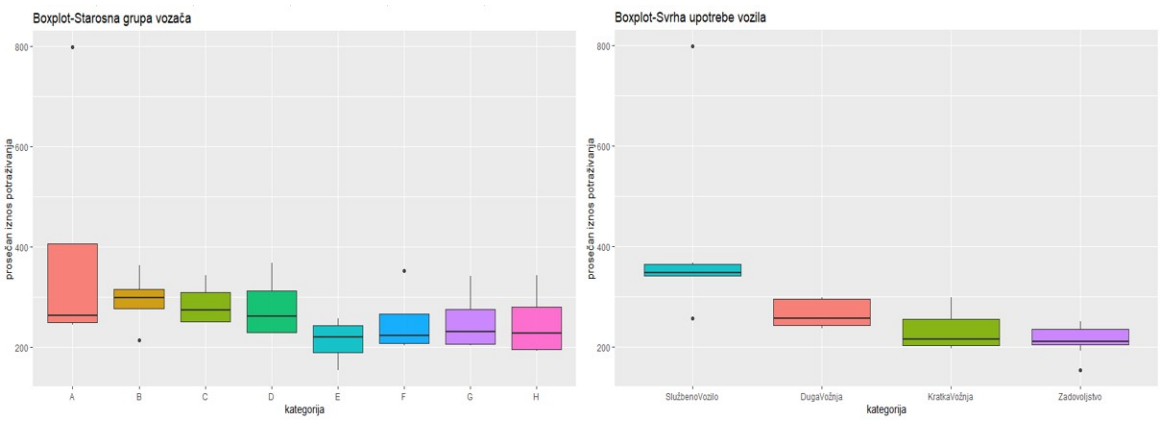
Primer 3.1. *Pomoću gama linearnog modela sa logaritamskom link funkcijom, opisati kako starosna grupa vozača i svrha uoptrebe vozila utiču na prosečan iznos potraživanja. Za izvor podataka koristiti podatke „AutoCollision" koji se nalaze na sajtu* `http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling/BookWebDec2010/data.html`

Neka je prosečan iznos potraživanja dat slučajnom promenljivom Y koja ima gama raspodelu sa očekivanjem μ i neka su date dve kategorijalne promenljive A_1 (starosna grupa vozača) i A_2 (svrha upotrebe vozila). Neka su dalje, sa $A_{11}, ..., A_{18}$ označene kategorije $A, ..., H$ i sa $A_{21}, ..., A_{24}$ označene kategorije službeno vozilo, ..., zadovoljstvo (tabela 3).
Prosečni iznosi potraživanja za svaku kategoriju dati su u vidu box-plot grafika na slici 1.

Tabela 3: Kategorijalne promenljive za primer (3.1)

Kategorijalna promenljiva	**Oznaka**	**Kategorija**	**Oznaka**	**Binarna varijabla**
Starosna grupa vozača	A_1	A	A_{11}	x_{11}
		B	A_{12}	x_{12}
		C	A_{13}	x_{13}
		D	A_{14}	x_{14}
		E	A_{15}	x_{15}
		F	A_{16}	x_{16}
		G	A_{17}	x_{17}
		H	A_{18}	x_{18}
Svrha upotrebe vozila	A_2	**Službeno vozilo**	A_{21}	x_{21}
		Kratka vožnja[6]	A_{22}	x_{22}
		Duga vožnja[7]	A_{23}	x_{23}
		Zadovoljstvo	A_{24}	x_{24}

[6]vožnja do posla koji je udaljen manje od 10 milja
[7]vožnja do posla koji je udaljen više od 10 milja



Slika 1: Box-plot grafici za prosečan iznos potraživanja po kategorijama

Za svaku kategoriju A_{jk}, $j = 1, ..., 8$, $k = 1, ..., 4$, definišemo binarnu varijablu x_{jk} sa

$$x_{jk} = \begin{cases} 1, & \text{ako osiguranik pripada kategoriji } A_{ij} \\ 0, & \text{inače} \end{cases}.$$

Neka su, bez gubljenja opštosti, kategorije A_{11} i A_{21} definisane kao referentni nivoi. Sada definišemo gama model sa

$$g(E(Y)) = g(\mu) = \beta_0 + \beta_1 x_{12} + ... + \beta_7 x_{18} + \beta_8 x_{22} + ... + \beta_{10} x_{24},$$

gde su $\beta_0, \beta_1, ..., \beta_{10} \in \mathbb{R}$ regresioni koeficijenti modela.
Zavisnost izmedju prediktora ispitaćemo statistikom ϕ_{test}. Rezultat je dat sa

$$\phi_{test}(x_{j_1 k_1}, x_{j_2 k_2}) = \begin{cases} 1 & ; j_1 = j_2 \wedge k_1 = k_2 \\ -0,14 & ; j_1 = 1 \wedge j_2 = 1 \wedge k_1 \neq k_2 \\ 0 & ; j_1 = 1 \wedge j_2 = 2 \\ -0,33 & ; j_1 = 2 \wedge j_2 = 2 \wedge k_1 \neq k_2 \end{cases}$$

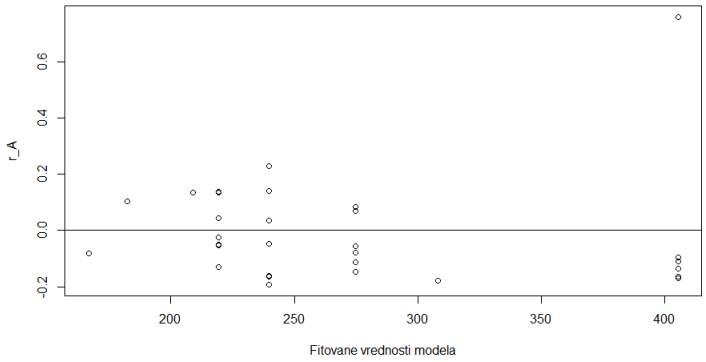
gde su $j_1, j_2 = \{1, 2\}$, $k_1 = \{2, 3, ..., 8\}$ i $k_2 = \{2, 3, 4\}$, pa zaključujemo da nema zavisnosti medju varijablama.
Za ovako definisan gama model, ocene koeficijenata modela i odabir najboljeg modela odredjujemo pomoću statističkog softvera „R“(ugradjenim funkcijama glm() i stepAIC()). Rezultat je model

$$g(\hat{\mu}_i) = 6 - 0,27 x_{i15} - 0,52 x_{i22} - 0,39 x_{i23} - 0,61 x_{i24} \tag{4}$$

sa parametrima datim u tabeli 4. Vidimo da su, primenom stepenaste metode unazad, iz modela izbačene binarne varijable $x_{i12}, x_{i13}, x_{i14}, x_{i16}, x_{i17}$ i x_{i18}. Izbacivanjem ovih varijabli iz modela, odgovarajuće kategorije B, C, D, F, G i H pripojene su referentnom nivou A.
Pogledajmo šta se dešava sa rezidualima ovog modela. U ovom primeru računaćemo Anskombe reziduale i koristićemo Šapiro-Vilkov test za testiranje hipoteze o normalnosti reziduala. Grafički prikaz Anskombe reziduala dat je na slici 2. Šapiro-Vilkova statistika za Anskombe reziduale modela datog jednačinom (4) jednaka je $W_{sw} = 0,76$ čija je p-vrednost jednaka $p_{vred} = 8,818 \cdot 10^{-6}$, pa možemo zaključiti da Anskombe reziduali nemaju normalnu raspodelu.
Zaključak da reziduali nemaju normalnu raspodelu nas dovodi do ispitivanja autlajera i uticajnih

Tabela 4: Ocenjeni koeficijenti gama modela datog jednačinom (4)

Koeficijenti		Ocenjena vrednost	Standardna devijacija	w-vrednost	p-vrednost
β_0	Slobodan član	6,00	0,08	8436,9	$< 0,001$
β_4	E	-0,27	0,12	8,0	$< 0,005$
β_8	Kratka vožnja	-0,52	0,11	33,4	$< 0,001$
β_9	Duga vožnja	-0,39	0,11	18,3	$< 0,001$
β_{10}	Zadovoljstvo	-0,61	0,11	45,7	$< 0,001$
	AIC	351			
	$\hat{\phi}$	0,052			



Slika 2: Anskombe reziduali modela koji je dat jednačinom (4)

tačaka. Na slici 3 prikazane su vrednosti Cook's rastojanja za svaku opservaciju iz uzorka. Sa slike vidimo da je vrednost Cook's rastojanja za četvrtu opservaciju iz uzorka daleko iznad praga koji iznosi $\frac{4}{n} = \frac{4}{32} = 0,125$ (isprekidana crvena linija), pa ćemo tu tačku označiti kao potencijalnu uticajnu tačku i konstruisaćemo model bez te tačke, kako bismo ispitali njen uticaj na model.
Ponovnom konstrukcijom gama modela (na uzorku koji ne sadrži potencijalnu uticajnu tačku) dobijamo model

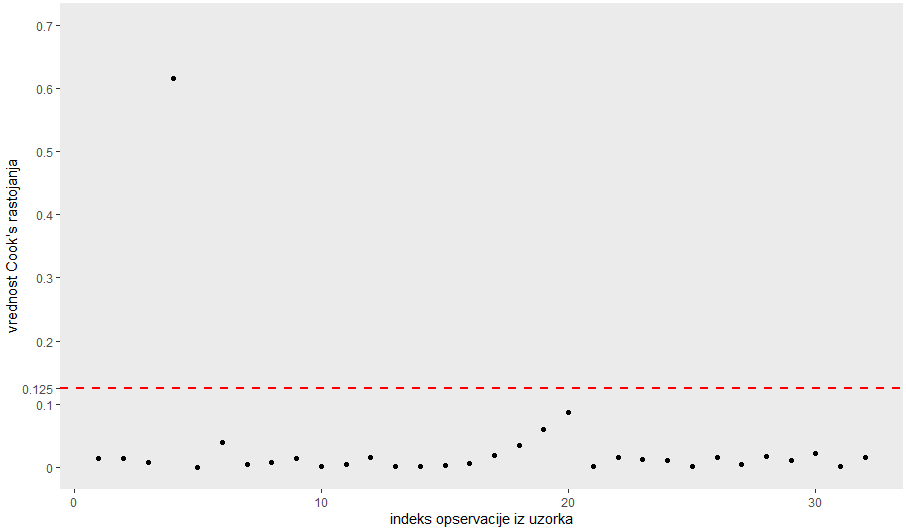
$$g(\hat{\mu}_i) = 5,85 - 0,24x_{i15} - 0,38x_{i22} - 0,24x_{i23} - 0,46x_{i24} \, . \tag{5}$$

sa statistikom datom u tabeli 5. Ako poredimo sa prethodno konstruisanim modelom (tabela 4),

Tabela 5: Ocenjeni koeficijenti gama modela datog jednačinom (5)

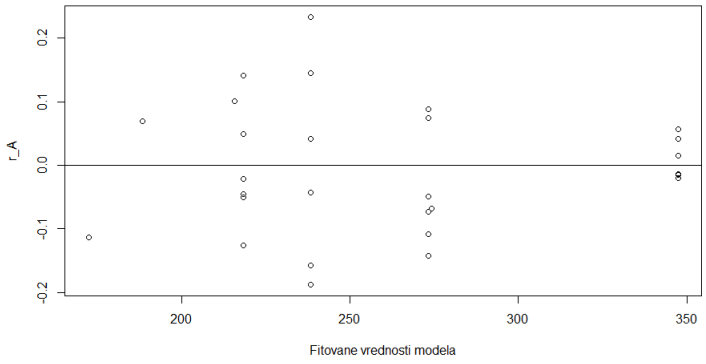
Koeficijenti		Ocenjena vrednost	Standardna devijacija	w-vrednost	p-vrednost
β_0	Slobodan član	5,85	0,04	21863,8	$< 0,001$
β_4	E	-0,24	0,06	18,4	$< 0,001$
β_8	Kratka vožnja	-0,38	0,06	50,2	$< 0,001$
β_9	Duga vožnja	-0,24	0,06	20,5	$< 0,001$
β_{10}	Zadovoljstvo	-0,46	0,06	76,5	$< 0,001$
	AIC	302			
	$\hat{\phi}$	0,013			

uočavamo znatnu promenu u vrednostima ocenjenih koeficijenata modela i njihovim standardnim devijacijama. Standardne devijacije ocenjenih koeficijenata modela koji je konstruisan na uzorku



Slika 3: Vrednosti Cook's rastojanja za svaku opservaciju iz uzorka za model (4)

bez potencijalne uticajne tačke, su manje nego kod modela koji je konstruisan na celom uzorku. To nam govori da su ocenjene vrednosti koeficijenata drugog modela pouzdanije. Takodje, ako posmatramo AIC vrednost, sada ona iznosi 302, što je manje u odnosu na AIC vrednost modela koji je konstruisan na celom uzorku. Sva ova zapažanja nam ukazuju da je četvrta opservacija iz prvobitnog uzorka zapravo uticajna tačka.
Posmatrajmo sada model koji je dat jednačinom (5). Anskombe reziduali ovog modela dati su



Slika 4: Anskombe reziduali modela koji je dat jednačinom (5)

na slici 4. Test statistika Šapiro-Vilk testa normalnosti za Anskombe reziduale sada je jednaka $W_{sw} = 0,98$ čija je p-vrednost jednaka $p_{vred} = 0,79$, pa nemamo razloga da odbacimo hipotezu o normalnoj raspodeli Anskombe raziduala, što smo i želeli da postignemo.
Dakle, zaključujemo da je gama model koji opisuje uticaj starosne grupe vozača i svrhe upotrebe vozila na prosečan iznos potraživanja dat jednačinom (5). Vrednost slobodnog člana u modelu (5,85) predstavlja ocenjenu vrednost modela za sve osiguranike sa službenim vozilima koji pripadaju starosnim kategorijama A, B, C, D, F, G ili H. Njihov ocenjen prosečan iznos potraživanja jednak je $\hat{\mu}_i = e^{5,85} = 347,23$ funti.

4 Skoring modeli kasko osiguranja

Skoring modeli imaju sve veću primenu u industriji osiguranja još od pojave prvih statističkih softvera, koji imaju mogućnost obrade velike količine podataka. Sa porastom broja osiguranika, dolazi i potreba za sistemima kao što su i skoring modeli, koji će pomoći menadžmentu kompanije da donese brzu i efikasnu odluku na tržištu. Još jedna prednost sistema bodovanja u industriji osiguranja je, kao i u drugim oblastima, da i prodavci osiguranja mogu kroz sistem bodovanja razlikovati dobre od loših osiguranika ili na primer, mogu odrediti iznos premije za odredjenog klijenta.

Jedan od čestih primera primene skoring modela u oblasti osiguranja je tarifna analiza. Tarifna analiza predstavlja odredjivanje tarife za odredjenu grupu osiguranika. Cilj tarifne analize je odredjivanje riziko premije[8], kao i utvrdjivanje odnosa rizika izmedju grupa osiguranika. Jedan od načina konstrukcije skoring modela koji se koriste u tarifnoj analizi, je pomoću uopštenih linearnih modela. Tako se, na primer, pomoću uopštenih linearnih modela mogu konstruisati dva skoring modela, jedan koji će definisati skorove za grupe osiguranika prema broju zahteva za odštetu i drugi koji će definisati skorove prema prosečnom iznosu odštete. Na ovaj način, riziko premija za odredjenu grupu osiguranika se dobija kao proizvod skorova u modelu koji opisuje broj zahteva i skorova u modelu koji opisuje prosečan iznos zahteva za odštetu. Primer ovakve tarifne analize dat je u radu [10].
Pored tarifne analize, skoring modeli, u oblasti osiguranja, imaju primenu i u pronalaženju lažnih potraživanja. Lažno potraživanje predstavlja ono potraživanje od strane osiguranika u kojem se, lažnim prikazivanjem činjenica, želi pribaviti protivpravna materijalna korist. Skoring modeli koji se koriste u otkrivanju lažnih potraživanja su veoma slični onima u bankarskom sektoru (kreditni skoring sistemi). Cilj modela je da otkrije ona potraživanja koja imaju veliku verovatnoću da budu lažna. Konstrukcija ovakvih skoring modela bazira se na uopštenim linearnim modelima sa logit link funkcijom. Dobijenim skoring modelom, utvrdjuje se broj bodova za svako novo potraživanje. Ako potraživanje ima broj bodova veći od zadatog praga, onda se smatra da je potraživanje sumnjivo i potrebno ga je detaljno istražiti. Vrednost zadatog praga utvrdjuje menadžment osiguravajućeg društva. Primer konstrukcije ovakve vrste skoring modela dat je u radu [11].

4.1 Skoring model za računanje premije kasko osiguranja

Široku primenu skoring modeli imaju u kasko osiguranju automobila, kako zbog toga što je polisa kasko osiguranja jedna od najprodavanijih polisa osiguravajućeg društva, tako i zbog toga što premija kasko osiguranja tarifnog sistema zavisi od mnogobrojnih faktora. Jedan od najčešćih obrazaca konstrukcije skoring modela premije kasko osiguranja je upotreba uopštenih linearnih modela. U ovom radu predstavićemo tri modela pomoću kojih se može dobiti veoma pouzdan aparat za izračunavanje premije novih polisa kasko osiguranja.
Neka je dat uzorak od $n \in \mathbb{N}$ jednogodišnjih polisa kasko osiguranja, ugovorenih u nekom periodu i neka su date kategorijalne promenljive[9] $A_1, A_2, ..., A_m$, $m \in \mathbb{N}$. Neka je ukupan broj kategorija prve promenljive jednak r_1, druge promenljive r_2, ..., m-te promenljive r_m, gde su $r_1, r_2, ..., r_m \geq 2$. Tabelarni prikaz kategorijalnih promenljivijh i njihovih kategorija dat je u tabeli 6.

Napomena 4.1. *Broj kategorija, dve različite kategorijalne promenljive, ne mora biti jednak.*

Pretpostavimo da za svaku polisu iz uzorka, znamo vrednost premije i vrednosti svake kategorijalne promenljive.

[8] Predstavlja iznos koji osuguravač mora da plati kako bi pokrio potraživanja
[9] Definisane kao u odeljku 3.7

Tabela 6: Prikaz kategorijalnih promenljivih i njihovih kategorija

Kat. promenljive	A_1	A_2	$\cdots$	A_m
Kategorije	A_{11}	A_{12}	$\cdots$	A_{m1}
	A_{21}	A_{22}	$\cdots$	A_{m2}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$A_{1(r_1-1)}$	$A_{2(r_2-1)}$	$\cdots$	$A_{m(r_m-1)}$
	A_{1r_1}	A_{2r_2}	$\cdots$	A_{mr_m}

Za svako $j = 1, 2, ..., m$, $k = 1, 2, ..., r_j$ uvodimo binarnu promenljivu

$$x_{jk} = \begin{cases} 1, & \text{ako polisa pripada kategoriji } A_{jk} \\ 0, & \text{inače} \end{cases} .$$

Ako sa $S_0 \in \mathbb{R}$ označimo vrednost početnog skora, dok sa $S_{jk} \in \mathbb{R}$ označimo vrednost skora kategorije A_{jk}, za sve $j = 1, 2, ..., m$ i $k = 1, 2, ..., r_j$, onda je skoring model definisan jednačinom

$$s = S_0 + \sum_{j=1}^{m} \sum_{k=1}^{r_m} S_{jk} x_{jk}, \tag{6}$$

gde sa s označena ukupna vrednost skora.
Konstrukcija skoring modela bazira se na odgovarajućem uopštenom linearnom modelu. Zavisna promenljiva Y je u ovom slučaju premija kasko osiguranja, dok su prediktori malopre definisane binarne promenljive x_{jk}. Ako su, bez gubljenja opštosti, kategorije $A_{1r_1}, A_{2r_2}, ..., A_{mr_m}$ (kategorije označene crvenom bojom u tabeli 6) označene kao referentni nivoi, onda je uopšteni linearni model dat sa

$$g(E(Y)) = g(\mu) = \eta = \beta_0 + \sum_{j=1}^{m} \sum_{k=1}^{r_j - 1} \beta_{jk} x_{jk}, \tag{7}$$

gde su β_0, $\beta_{jk} \in \mathbb{R}$ nepoznati regresioni koeficijenti modela , dok je sa g označena link funkcija.

Napomena 4.2. *Kao što smo naglasili u poglavlju* (3.7)*, binarne varijable koje odgovaraju referentnim nivoima, ne učestvuju u regresionom modelu, jer bi u suprotnom bila narušena pretpostavka o nezavisnosti izmedju prediktora.*

Zbog jednostavnijeg zapisa u nastavku, definisaćemo funkciju koja će transformisati indekse j i k. Neka je funkcija $\pi : \mathbb{N}^2 \to \mathbb{N}$ definisana na sledeći način

$$\pi : \{j, k\} \mapsto k + \sum_{j_1=1}^{j} (r_{j_1 - 1} - 1), \quad j = 1, 2, ..., m,\ k = 1, 2, ..., r_j - 1,\ r_0 = 1 .$$

Za funkciju π se lako može pokazati da je bijekcija, na skupu $\{\{j, k\}\,;\, j = 1, 2, ..., m\,,\, k = 1, 2, ..., r_j\}$. Sada za $p = \sum_{j=1}^{m} (r_j - 1)$, jednačina (7) je data sa

$$g(E(Y)) = g(\mu) = \eta = \beta_0 + \sum_{l=1}^{p} \beta_l x_l \, . \tag{8}$$

U zavisnosti od raspodele zavisne slučajne promenljive Y razlikujemo lognormalni, gama i normalni skoring model.

4.1.1 Lognormalni skoring model

Kod lognormalnog modela pretpostavljamo da premija Y prati lognormalnu raspodelu sa parametrima μ i $\sigma^2 > 0$. Znamo, ako slučajna promenljiva Y ima lognormalnu raspodelu sa parametrima μ i σ^2, onda slučajna promenljiva $\ln Y$ ima normalnu raspodelu sa parametrima μ i σ^2. Gustina slučajne promenljive $\ln Y$ data je sa

$$\varphi(\ln y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(\ln y - \mu)^2}{2\sigma^2}\right] \ , \sigma^2 > 0 \ .$$

Ako za link funkciju g uzmemo da je identička , onda je lognormalni model, zapravo klasičan normalni regresioni model koji je dat sa

$$g(E(\ln Y)) = g(\mu) = \mu = \eta = \beta_0 + \sum_{l=1}^{p} \beta_l x_l \ . \tag{9}$$

Za ocenu koeficijenata iz jednačine (9) koristićemo metodu maksimalne verodostojnosti. Rekli smo da je ocena koeficijenata uopštenog linearnog modela izračunata metodom maksimalne verodostojnosti data sa

$$\hat{\boldsymbol{\beta}} = \left(X^T Q X\right)^{-1} X^T Q G_y \ .$$

Kako je kod klasičnog normalnog linearnog modela matrica Q zapravo jedinična matrica dimenzije $n \times n$, dobijamo da je

$$\hat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T G_y,$$

gde su

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \ X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \ G_y = \begin{bmatrix} \ln y_1 \\ \ln y_2 \\ \vdots \\ \ln y_n \end{bmatrix} \ .$$

Napomena 4.3. *Sa x_{il}, $i = 1, 2, ..., n$, $l = 1, 2, ..., p$ su označene vrednosti binarnih varijabli x_l za i-tu opservaciju iz uzorka.*

Postupak konstrukcije lognormalnog skoring modela se nastavlja primenom stepenaste metode za odredjivanje regresionog modela koji najbolje opisuje kretanje premije. U svakoj iteraciji stepenaste metode, za ocenu koeficijenata modela, primenjujemo metodu maksimalne verodostojnosti.
Neka je sa

$$\hat{\eta} = \hat{\beta}_0 + \sum_{l=1}^{p} \hat{\beta}_l x_l$$

obeležen regresioni model koji najbolje opisuje kretanje premije. Ako na indeks l primenimo funkciju π^{-1} dobićemo model sa indeksima koji označavaju kategorijalnu promenljivu i kategoriju, tačnije

$$\hat{\eta} = \hat{\beta}_0 + \sum_{j=1}^{m} \sum_{k=1}^{r_j} \hat{\beta}_{jk} x_{jk} \ . \tag{10}$$

Napomena 4.4. *Ako je neka od promenljivih x_l, $l = 1, 2, ..., p$ izbačena posle postupka stepenaste metode, njen odgovarajući koeficijent u jednačini* (10) *imaće vrednost 0.*

Poslednji korak u konstrukciji lognormalnog modela je pretvaranje ocenjenih koeficijenata modela $\hat{\beta}_0, \hat{\beta}_{jk}$, $j = 1, ..., m$, $k = 1, ..., r_j - 1$ u vrednosti $S_0, S_{11}, ..., S_{1r_1}, S_{21}, ...S_{2r_2}, ..., S_{m,r_m}$. Taj proces je opisan sledećim postupkom.

Postupak 1

(1) unutar svakog faktora naći minimalnu vrednost u skupu odgovarajućih negativnih ocenjenih koeficijenata, tj. odrediti vrednosti $\delta_1, \delta_2, ..., \delta_m$ takve da je

$$\delta_j = min\{0, min\{\hat{\beta}_{j1}, ..., \hat{\beta}_{jr_j-1}\}\}$$

za $j = 1, 2, ..., m$

(2) vrednosti skorova S_0 i S_{jk} definišemo na sledeći način

$$S_0 = \hat{\beta}_0 + \sum_{l=1}^{m} \delta_j$$

$$S_{jk} = \hat{\beta}_{jk} + |\delta_j|$$

za sve $j = 1, ..., m$ i $k = 1, ..., r_j$

Konačno, lognormalni skoring model je dat sa

$$s = S_0 + \sum_{j=1}^{m} \sum_{k=1}^{r_j} S_{jk} x_{jk} \ .$$

Ocenjenu vrednost $\hat{\mu}$ dobijamo iz

$$\hat{\mu}_i = \exp[s] \ .$$

Napomena 4.5. *Radi jednostavnosti skoring modela, ponekad se traži od aktuara da se ukupan skor kreće u rasponu od 0 do 100 i da svi skorovi budu celobrojni. Tada u postupku 1 dodajemo još jedan korak, gde ćemo vrednosti dobijene iz koraka (2) podeliti brojem* $\ln b$ *,* $b > 0 \wedge b \neq 1$ *(ovim postupkom prelazimo sa baze e na bazu b) i zatim, količnik zaokružiti na ceo broj.*

4.1.2 Gama skoring model

Konstrukcija gama skoring modela bazirana je na uopštenom linearnom modelu, gde zavisna promenljiva Y prati gama raspodelu, dok je link funkcija logaritamska.

Tvrdjenje 4.1. *Gama raspodela pripada familiji eksponencijalnih raspodela.*

Dokaz. Rekli smo da je eksponencijalna familija raspodela data funkcijom gustine

$$\varphi(y) = c(y,\phi)\exp\left[\frac{y\theta - a(\theta)}{\phi}\right].$$

Za gustinu gama raspodele važi

$$\varphi(y) = \frac{1}{y\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu}\exp\left[-\frac{\nu y}{\mu}\right] = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{(\nu-1)}\exp\left[-\frac{\nu y}{\mu}\right].$$

Logaritmovanjem funkcije gustine dobijamo

$$\begin{aligned} ln\varphi(y) &= \nu\ln\left(\frac{\nu}{\mu}\right) + (\nu-1)\ln y - \ln\Gamma(\nu) - \frac{\nu y}{\mu} \\ &= \frac{\frac{\nu}{\nu}\ln\left(\frac{\nu}{\mu}\right) - \frac{\nu y}{\nu\mu}}{\frac{1}{\nu}} + (\nu-1)\ln y - \ln\Gamma(\nu) \\ &= \frac{\ln\left(\frac{\nu\nu}{\nu\mu}\right) - \frac{\nu y}{\nu\mu}}{\frac{1}{\nu}} + (\nu-1)\ln y - \ln\Gamma(\nu) \\ &= \frac{\ln\left(\frac{\nu}{\nu\mu}\right) - \frac{\nu y}{\nu\mu}}{\frac{1}{\nu}} - \frac{\ln\left(\frac{1}{\nu}\right)}{\frac{1}{\nu}} + \left(\frac{1}{1/\nu} - 1\right)\ln y - \ln\Gamma\left(\frac{1}{1/\nu}\right). \end{aligned}$$

Sada ako uzmemo da je $\theta = -\frac{\nu}{\nu\mu} = -\frac{1}{\mu}$, $\phi = \frac{1}{\nu}$ i $a(\theta) = \ln(-\theta) = \ln\left(\frac{1}{\mu}\right)$ dobijamo

$$\ln\varphi(y) = (-1)\frac{y\theta - a(\theta)}{\phi} - \underbrace{\frac{\ln(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\ln y - \ln\Gamma\left(\frac{1}{\phi}\right)}_{\ln c(y,\phi)},$$

odnosno

$$\varphi(y) = c(y,\phi)\exp\left[\frac{y\theta - a(\theta)}{\phi}\right].$$

Za očekivanje i disperziju dobijamo vrednosti

$$E(Y) = a'(\theta) = \frac{\partial a}{\partial\theta} = -\frac{1}{\theta} = \mu$$

$$Var(Y) = \phi a''(\theta) = \frac{1}{\nu}\frac{\partial^2 a}{\partial\theta\partial\theta} = \frac{1}{\nu\theta^2} = \frac{1}{\nu}\mu^2,$$

pa sledi da gama raspodela pripada familiji eksponencijalnih raspodela. □

Gama uopšteni linearni model je dat sa

(1) Zavisna promenljiva Y ima gama raspodelu sa gustinom

$$\varphi(y) = \frac{1}{y\Gamma(\nu)} \left(\frac{\nu y}{\mu}\right)^{\nu} \exp\left[-\frac{\nu y}{\mu}\right]$$

(2) Link funkcija je data sa

$$g(E(Y)) = g(\mu) = ln\mu = \eta,$$

gde je η data jednačinom (8)

Za ovaj uopšteni linearni model, ocena koeficijenata modela dobijena metodom maksimalne verodostojnosti, približno je jednaka

$$\hat{\boldsymbol{\beta}} \approx \left(X^T Q X\right)^{-1} X^T Q G_y \, .$$

Kako je

$$((g'(\mu))^2 a''(\theta))^{-1} = \frac{1}{\mu \ln b}\left(-\mu^2\right) = -\frac{\ln b}{\mu},\ i = 1, 2, ..., n$$

dobijamo da je matrica Q dimenzije $n \times n$ data sa $Q = diag\left(-\frac{\ln b}{\mu}, -\frac{\ln b}{\mu}, ..., -\frac{\ln b}{\mu}\right)$.
Sada, analognim metodama i postupcima kao i kod lognormalnog modela dolazimo do gama skoring modela koji je dat sa

$$s = S_0 + \sum_{j=1}^{m} \sum_{k=1}^{r_j} S_{jk} x_{jk},$$

pa je ocenjena očekivana premija gama skoring modela data sa

$$\hat{\mu} = b^s \, .$$

4.1.3 Normalni skoring model

Konstrukcija normalnog skoring modela bazirana je na modelu gde zavisna promenljiva Y, ima normalnu raspodelu, dok je link funkcija logaritamska.

Tvrdjenje 4.2. *Normalna raspodela pripada familiji eksponencijalnih raspodela.*

Dokaz. Pokazaćemo da je funkcija gustine normalne raspodele data u obliku

$$\varphi(y) = c(y, \phi) \exp\left[\frac{y\theta - a(\theta)}{\phi}\right]$$

i da su $E(Y) = a'(\theta)$ i $Var(Y) = a''(\theta)$ za sve $i = 1, 2, ..., n$.
Logaritmovanjem gustine normalne raspodele dobijamo

$$\begin{aligned}\ln(\varphi(y)) &= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2} = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y^2 - 2y\mu + \mu^2)}{2\sigma^2} \\ &= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{y2}{2\sigma^2} - \frac{\mu^2 - 2y\mu}{2\sigma^2} = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} \, .\end{aligned}$$

Sada ako uzmemo da je $\theta = \mu$, $a(\theta) = \frac{1}{2}\theta^2$ i $\phi = \sigma^2$, dobijamo

$$\ln(\varphi(y)) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{\sigma^2} + \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} = \underbrace{-\frac{1}{2}\ln(2\pi\phi) - \frac{y^2}{2\phi}}_{\ln c(y,\phi)} + \frac{y\theta - a(\theta)}{\phi},$$

pa je

$$\ln(\varphi(y)) = c(y,\phi)\exp\left[\frac{y\theta - a(\theta)}{\phi}\right] .$$

Dalje, za normalnu raspodelu važi

$$E(Y) = \mu = \theta = \frac{\partial a}{\partial \theta} = a'(\theta)$$

$$Var(Y) = \sigma^2 = \phi = \phi \cdot 1 = \phi\frac{\partial^2 a}{\partial\theta\partial\theta} = \phi a''(\theta) ,$$

iz čega sledi da normalna raspodela pripada familiji eksponencijalnih raspodela. □

Sada definišemo normalni model sa

(1) Zavisna promenljiva Y ima normalnu raspodelu sa gustinom

$$\varphi(y) = \frac{1}{\sqrt{a\pi\sigma^2}}\exp\left[-\frac{(y-\mu)}{2\sigma^2}\right]$$

(2) Link funkcija je data jednačinom

$$g(E(Y)) = g(\mu) = \log_b \mu = \eta,$$

gde je η data jednačinom (8)

Dakle, normalni model je u stvari dat kao uopšteni linearni model sa zavisnom promenljivom koja ima normalnu raspodelu i logaritamskom link funkcijom. U poglavlju (3.1) rekli smo da je ocena parametara modela metodom maskimalne verodostojnosti data sa

$$\hat{\boldsymbol{\beta}} \approx \left(XQX^T\right)^{-1} XQG_y . \tag{11}$$

Kako je sada

$$((g'(\mu))^2 a''(\theta))^{-1} = \left(\frac{1}{\mu\ln b}\right)^{-1} = \mu\ln b \quad , i = 1, 2, ..., n$$

imamo da je matrica Q dimenzije $n \times n$ data sa $Q = diag(\mu_1 \ln b, \mu_2 \ln b, ..., \mu_n \ln b)$.

Sada, primenjujemo postupke analogno kao kod lognormalnog modela, pa dobijamo normalni skoring model koji je dat sa

$$s = S_0 + \sum_{j=1}^{m}\sum_{k=1}^{r_j} S_{jk}x_{jk} .$$

Ocenjena očekivana premija i-te polise normalnog skoring modela je data sa

$$\hat{\mu} = b^s .$$

5 Primena skoring modela za odredjivanje premije kasko osiguranja

U prethodnom poglavlju videli smo konstrukciju tri skoring modela koji su se bazirali na pretpostavci da premija Y prati lognormalnu, gama, odnosno normalnu raspodelu. U cilju praktične primene ova tri modela na pravim podacima, predstavićemo studiju [12] Norisure Ismail[10] i Abdula Aziza Džemejna[11], proferose univerziteta u Kebangasanu (Malezija). Pored poredjenja modela, ova studija će nam pokazati i koliko ocenjene premije odstupaju od stvarne vrednosti i na koji način se modeli mogu popraviti tako da ta odstupanja budu manja.

Pre nego što damo informacije o podacima koji su korišćeni u studiji, definisaćemo pojam klase osiguranika. Neka su date kategorijalne promenljive i njihove kategorije kao u tabeli 6. Pojam klase će označavati skup osiguranika koji pripada kategorijama $\{A_{1\bullet}, A_{2\bullet}, ..., A_{m\bullet}\}$, gde je sa $A_{j\bullet}$ predstavljena tačno jedna kategorija j-te kategorijalne promenljive, za $j = 1, 2, ..., m$. Broj klasa definisanih na ovaj način jednak je $r_1 \cdot r_2 \cdot ... \cdot r_m$.

Za potrebe ove studije korišćeni su podaci polisa kasko osiguranja jedne malezijske osiguravajuće kuće koje su grupisane u klasama prema kategorijalnim promenljivama i njihovim kategorijama kojima se želi opisati premija. Kategorijalne promenljive i njihove kategorije su date u tabeli 5. Iz tabele vidimo da je broj faktora $m = 5$, dok je ukupan broj kategorija jednak 16, tačnije broj kategorija prvog faktora jednak $r_1 = 2$, drugog $r_2 = 2$, trećeg $r_3 = 3$, četvrtog $r_4 = 4$ i petog $r_5 = 5$, pa je broj klasa jednak $n = 2 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 240$. Za svaku klasu osiguranika data je vrednost prosečne godišnje premije - μ_i (vrednosti u valuti Malezijski Ringit-RM) i vrednost izloženosti - e_i. Vrednost izloženosti jedne klase data je u vidu ukupnog broja godina trajanja svih polisa u okviru te klase.

Tabela 7: Kategorijalne promenljive i njihove kategorije uključene u konstrukciju modela studije [12]

Kategorijalna promenljiva	**Kategorija**	**Skor**	**Binarna varijabla**
Obim pokrića polise	Standardno	S_{11}	x_{11}
	Sveobuhvatno	S_{12}	x_{12}
Marka vozila	Strana marka	S_{21}	x_{21}
	Lokalna marka	S_{22}	x_{22}
Pol osiguranika	Pravno lice	S_{31}	x_{31}
	Fizičko lice-žena	S_{32}	x_{32}
	Fizičko lice-muškarac	S_{33}	x_{33}
Starost vozila	2-3	S_{41}	x_{41}
	4-5	S_{42}	x_{42}
	6+	S_{43}	x_{43}
	0-1	S_{44}	x_{44}
Regija	Severna Malezija	S_{51}	x_{51}
	Istočna Malezija	S_{52}	x_{52}
	Južna Malezija	S_{53}	x_{53}
	Ostrvo Borneo	S_{54}	x_{54}
	Centralna Malezija	S_{55}	x_{55}

Primetimo da u ovoj studiji ne postoji informacija o premiji za svaku polisu osiguranja, nego se koristi informacija o prosečnoj godišnjoj premiji za svaku klasu osiguranika. Na taj način, obim

[10] Norisura Ismail (eng. Noriszura Ismail) je profesorka na Univerzitetu Kebangsaan u Maleziji koja se izmedju ostalog bavi aktuarstvom, modelima rizika i primenjenom statistikom.

[11] Abdul Aziz Džemejn (eng. Abdul Aziz Jemain) je profesor statistike na Univerzitetu Kebangsaan u Maleziji.

uzorka ne predstavlja broj polisa, nego broj klasa osiguranika, što je u ovom slučaju 240. Vrednost izloženosti će imati ulogu težinskog faktora u regresiji, ali više o tome u nastavku.

Skoring model u ovoj studiji je dat sa

$$s = S_0 + S_{11}x_{11} + S_{12}x_{12} + ... + S_{54}x_{54} + S_{55}x_{55},$$

gde s_i predstavlja ocenjen ukupan skor osiguranika i-te klase, S_0 predstavlja parametar početnog skora, zatim $S_{11}, S_{12}, ..., S_{45}, S_{55}$ predstavljaju parametre skorova kategorija i konačno sa x_{ijk} su označene vrednosti binarne promenljive x_{jk} za i-tu klasu.
Kategorije „Sveobuhvatno“,„Lokalna marka“, „Muškarac“, „0-1“ i „Centralna Malezija“ (označene crvenom bojom u tabeli (5)) su definisani kao referentni nivoi.

5.1 Primer lognormalnog skoring modela

Neka slučajna promenljiva Y, koja predstavlja premiju, ima lognormalnu raspodelu sa očekivanjem μ i disperzijom $\sigma^2 e^{-1}$, gde je sa e obeležen težinski koeficijent (koeficijent izloženosti). Tada, za slučajnu promenljivu $\log_b Y$ ($b > 0$, $b \neq 1$ je baza logaritma), važi $\log_b Y : \mathcal{N}(\mu, \sigma^2 e^{-1})$. Kako bi se krajnja vrednost skora kretala u rasponu od 0 do 100, za logaritamsku osnovu b uzeta je vrednost $1,1$.

Tabela 8: Ocenjeni koeficijenti lognormalnog modela

Koeficijenti		Ocenjena vrednost	Standardna devijacija	p-vrednost
β_0	Slobodan član	78.81	0.26	0.00
β_{11}	Stadnardna polisa	-14.52	0.43	0.00
β_{21}	Strana marka	4.23	0.26	0.00
β_{31}	Pravno lice	-9.25	0.53	0.00
β_{32}	Fizičko lice-žena	-4.30	0.28	0.00
β_{42}	4-5 godina	-1.17	0.33	0.02
β_{43}	6+godina	-1.56	0.30	0.01
β_{51}	Severna Malezija	0.84	0.29	0.04
β_{54}	Ostrvo Borneo	-4.48	0.45	0.00

Lognormalni model dat je sa

$$g(E(\log_b Y)) = g(\mu) = \mu = \eta = \beta_0 + \sum_{j=1}^{5} \sum_{k=1}^{r_j - 1} \beta_{jk} x_{jk} \ .$$

Primenom metode maskimalne verodostojnosti za ocenu koeficijenata modela i primenom ANOVA[12] metode za odabir najboljeg modela, dobijen je model

$$\hat{\eta} = 78.81 - 14.52x_{11} + 4.23x_{21} - 9.25x_{31} - 4.30x_{32} - 1.17x_{42} - 1.56x_{43} + 0.84x_{51} - 4.18x_{54}, \quad (12)$$

čiji je koeficijent determinacije jednak sa $R^2 = 89,3\%$.
Karakteristike ocenjenog lognormalnog modela date su u tabeli 8. Iz tabele vidimo da su svi koeficijenti statistički značajni na nivou značajnosti od 95% ($p < 0.05$). Takodje, primetimo da binarne

[12] Analiza varijanse

promenljive x_{i41}, x_{i52} i x_{i53} nisu ušle u najbolji lognormalni model. To znači da su njihove kategorije pripojene odgovarajućim referentnim nivoima. Konkretno, kategorija „2-3“ je pripojena refentnom nivou „0-1“ i kategorije „Istočna Malezija“ i „Južna Malezija“ su pripojene referentnom nivou „Centralna Malezija“.
Interpretacija slobodnog člana iz tabele 8 je ta da ona označava skor onih osiguranika koji ne pripadaju ni jednoj od kategorija iz najboljeg modela, tj. kategorija prve kolone tabele 8. Tačnije, β_0 je skor osiguranika koji pripadaju samo kategorijama koji su označeni kao referentni nivoi ili kategorijama koje su pripojene referentnim nivoima.
Finalni korak u konsktrukciji lognormalnog skoring modela je primena postupka 1, s tim što je postupku dodat i treći korak, koji podrazumeva zaokruživanje skorova na cele brojeve. U tabeli 9 date su konačne vrednosti skorova kategorija lognormalnog skoring modela.
Dakle, naš konačni lognormalni skoring model je dat sa

Tabela 9: Ocenjene, modifikovane i konačne vrednosti skorova po kategorijama

Kategorije	**Koeficijenti**	**Modifikovana vrednost skora**	**Konačna vrednost skora**
Slobodan član (minimalni skor)	78.81	49.30	49
Obim pokrića police			
Standardno	-14.52	0.00	0
Sveobuhvatno	0.00	14.52	15
Marka vozila			
Strana marka	4.23	4.23	4
Lokalna marka	0.00	0.00	0
Pol osiguranika			
Pravno lice	-9.25	0.00	0
Fizičko lice-žena	-4.30	4.95	5
Fizičko lice-muškarac	0.00	9.25	9
Starost vozila			
2-3	0.00	1.56	2
4-5	-1.17	0.39	0
6+	-1.56	0.00	0
0-1	0.00	1.56	2
Regija			
Severna Malezija	0.84	5.02	5
Istočna Malezija	0.00	4.18	4
Južna Malezija	0.00	4.18	4
Ostrvo Borneo	-4.18	0.00	0
Centralna Malezija	0.00	4.18	4

$$s = 49 + 0 \cdot x_{11} + 15x_{12} + 4x_{21} + 0 \cdot x_{22} + 0 \cdot x_{31} + 5x_{32} + 9x_{33} + 2x_{41} + 0 \cdot x_{42} + 0 \cdot x_{43} + 2x_{44} + 5x_{51} + 4x_{52} + 4x_{53} + 0 \cdot x_{54} + 4x_{55} \, .$$

Primetimo početni skor S_0 predstavlja minimalni skor, odnosno minimalnu premiju kasko osiguranja. Minimalni skor 49, odnosno premiju kasko osiguranja od $1,1^{49} \approx 107$ RM će plaćati svi osiguranici koji imaju sve sledeće osobine

- kupili su standardnu polisu osiguranja
- osigurali su vozilo lokalne marke

- kupuju polisu u svojstvu pravnog lica
- osigurano vozilo je starije od 4 godine
- stanovnik je malezijskog dela ostrva Borneo

Sa druge strane, maksimalni skor iznosi 84, što je u novcu oko 3 000 RM.
Posmatrajmo tri osiguranika koji pripadaju istim kategorijama u okviru svih kategorijalnih promenljivih, osim promenljive „Pol osiguranika“. Neka je prvi osiguranik kupio polisu kao pravno lice, drugi kao fizičko lice-žena, a treći kao fizičko lice-muškarac i neka su im izračunati skorovi, redom, s_1, s_2, s_3. Tada važi sledeće

$$y_3 = 1,1^{s_3} = 1,1^{(s_1+9)} = 1.1^{s_1} \cdot 1.1^9 \approx 2.36 \cdot y_1$$

$$y_3 = 1.1^{s_3} = 1.1^{(s_2+5)} = 1.1^{s_2} \cdot 1.1^5 \approx 1.46 \cdot y_2,$$

gde su y_1, y_2 i y_3 odgovarajuće premije skorova, redom, s_1, s_2 i s_3.
Prednost skoring modela u analizi podataka je ta što se na jednostavan način mogu interpretirati odnosi kategorija unutar faktora. Tako na primer, vidimo da kategorija fizička lica-muškarci plaćaju oko 2.36 puta veću premiju od pravnih lica, odnosno oko 1.46 puta veću premiju od kategorije fizičkih lica-žene. Na isti način dolazimo i do podatka da su premije sveobuhvatne polise kasko osiguranja oko 4.18 puta veće od premija standardnih polisa kasko osiguranja.

Korekcija lognormalnog skoring modela.

Za validaciju dobijenog skoring modela, posmatrana je ukupna vrednost stvarne i fitovane premije. Vrednosti su date u tabeli 10. Iz tabele vidimo da je ukupna vrednost dobijene fitovane premije manja za 560 380 RM, pa je neophodna korekcija modela. Autori predlažu da se fitovana premija pomnoži korektivnim faktorom, koji bi izjednačio ukupne vrednosti fitovane i stvarne vrednosti. Kako je ukupan premijski racio (predstavlja količnik ukupne vrednosti fitovane premije i ukupne vrednosti stvarne premije, videti tabelu 10) jednak $0,998$, za korektivni faktor se uzima broj $\dfrac{1}{0,998} \approx 1,002$.

Tabela 10: Ukupna vrednost, razlika i racio prihoda od fitovane i stvarne premije

		Vrednost
Ukupna vrednost izloženosti	$\sum_{i=1}^{240} e_i$	170.749
Ukupna vrednost fitovanih premija	$\sum_{i=1}^{240} e_i\hat{y}_i$	RM 275.269.816
Ukupna vrednost stvarnih premija	$\sum_{i=1}^{240} e_i y_i$	RM 275.830.196
Razlika ukupne vrednosti stvarne i fitovane premije	$\sum_{i=1}^{240} e_i(\hat{y}_i - y_i)$	-RM 560.380
Ukupan premijski racio	$\dfrac{\sum_{i=1}^{240} e_i\hat{y}_i}{\sum_{i=1}^{240} e_i y_i}$	0,998

5.2 Primer gama skoring modela

Pretpostavimo da Y ima gama raspodelu sa očekivanjem μ i disperzijom $e^{-1}\sigma^2\mu^2$. Gama model je dat sa

$$g(E(Y)) = g(\mu) = \ln \mu = \eta = \beta_0 + \sum_{j=1}^{5} \sum_{k=1}^{r_j - 1} \beta_{jk} x_{jk} \ .$$

Ocene parametara gama modela dobijene su iterativnom postupkom ponderisanih najmanjih kvadrata. Dobijeni gama model je dat sa

$$\hat{\eta} = 78.89 - 13.71x_{11} + 4.19x_{21} - 8.55x_{21} - 4.25x_{32} - 1.17x_{42} \ . - 1.73x_{43} + 0.81x_{51} - 4.21x_{54} \ .$$

U tabeli 11 date su statistike ocenjenih vrednosti koeficijenata najboljeg gama modela. Primetimo

Tabela 11: Ocenjeni koegicijenti gama modela

Koeficijenti		Ocenjena vrednost	Standardna devijacija	p-vrednost
β_0	Slobodan član	78.89	0.02	0.00
β_{11}	Stadnardna polisa	-13.71	0.03	0.00
β_{21}	Strana marka	4.19	0.02	0.00
β_{31}	Pravno lice	-8.55	0.04	0.00
β_{32}	Fizičko lice-žena	-4.25	0.02	0.00
β_{42}	4-5 godina	-1.17	0.02	0.00
β_{43}	6+godina	-1.73	0.02	0.00
β_{51}	Severna Malezija	0.81	0.02	0.00
β_{54}	Ostrvo Borneo	-4.21	0.03	0.00

da su potpuno iste kategorije uključene u gama model, kao i kod lognormalnog modela. Isto kao i kod lognormalnog modela i u ovom slučaju su kategorije „2-3“, „Istočna Malezija“ i „Južna Malezija“ pripojene svojim referentnim nivoima.
Svi koeficijenti iz tabele 11 su statistički značajni, medjutim primetimo da je standardna devijacija mnogo manja nego kod lognormalnog modela. Iz toga možemo zaključiti da je gama model pouzdaniji od lognormalnog modela.
Preostaje nam primena proširenog postupka 1, kao i kod lognormalnog modela. U tabeli 12 prikazane su konačne vrednosti gama skoring modela.
Dakle, konačni gama skoring model je dat sa

$$s = 49 + 0 \cdot x_{11} + 14x_{12} + 4x_{21} + 0 \cdot x_{22} + 0 \cdot x_{31} + 4x_{32} + 9x_{33} + 2x_{41} + 1 \cdot x_{42} + 0 \cdot x_{43} + 2x_{44} + 5x_{51} + 4x_{52} + 4x_{53} + 0 \cdot x_{54} + 4x_{55} \ .$$

Najmanji skor je kao i kod lognormalnog modela 49, dok je maksimalna moguća vrednost gama skoring modela jednaka 83, što znači da se premija kreće u rasponu od 107 RM do 2 726 RM.

Tabela 12: Ocenjene, modifikovane i konačne vrednosti gama skoring modela po kategorijama

Kategorije	**Koeficijenti**	**Modifikovana vrednost skora**	**Konačna vrednost skora**
Slobodan član (minimalni skor)	78.89	50.69	49
Obim pokrića polise			
Standardno	-13.71	0.00	0
Sveobuhvatno	0.00	13.71	14
Marka vozila			
Strana marka	4.19	4.19	4
Lokalna marka	0.00	0.00	0
Pol osiguranika			
Pravno lice	-8.55	0.00	0
Fizičko lice-žena	-4.25	4.30	4
Fizičko lice-muškarac	0.00	8.55	9
Starost vozila			
2-3	0.00	1.73	2
4-5	-1.17	0.56	1
6+	-1.73	0.00	0
0-1	0.00	1.73	2
Regija			
Severna Malezija	0.81	5.02	5
Istočna Malezija	0.00	4.21	4
Južna Malezija	0.00	4.21	4
Ostrvo Borneo	-4.21	0.00	0
Centralna Malezija	0.00	4.21	4

5.3 Primer normalnog skoring modela

Na kraju ako pretpostavimo da Y prati normalnu raspodelu sa očekivanjem μ i disperzijom $e^{-1}\sigma^2$, onda je normalni model dat sa

$$g(E(Y)) = g(\mu) = \ln \mu = \eta = \beta_0 + \sum_{j=1}^{5} \sum_{k=1}^{r_j - 1} \beta_{jk} x_{jk} \, .$$

Tabela 13: Ocenjeni koeficijenti normalnog modela

Koeficijenti		Ocenjena vrednost	Standardna devijacija	p-vrednost
β_0	Slobodan član	79.02	0.01	0.00
β_{11}	Stadnardna polisa	-12.79	0.05	0.00
β_{21}	Strana marka	4.02	0.01	0.00
β_{31}	Pravno lice	-7.40	0.03	0.00
β_{32}	Fizičko lice-žena	-4.03	0.01	0.00
β_{42}	4-5 godina	-1.17	0.01	0.00
β_{43}	6+godina	-2.10	0.01	0.00
β_{51}	Severna Malezija	0.49	0.01	0.00
β_{54}	Ostrvo Borneo	-4.01	0.03	0.00

Primenom iterativne metode ponderisanih najmanjih kvadrata dobijene su ocene normalnog modela koji je dat sa

$$\hat{\eta} = 79.02 - 12.79x_{11} + 4.02x_{21} - 7.40x_{21} - 4.03x_{32} - 1.17x_{42} - 2.10x_{43} + 0.49x_{51} - 4.01x_{54} \, .$$

Statistički parametri koeficijenata normalnog modela dati su u tabeli 13. Vidimo da su svi koeficijenti statistički značajni, dok su standardne devijacije koeficijenata mnogo manje od lognormalnog modela, pa zaključujemo da je i normalni model pouzdaniji od lognormalnog modela.
Primetimo da i u normalnom modelu učestvuju iste one kategorije kao i u prethodna dva modela. Na taj način smo ponovo kategorije „2-3“, „Istočna Malezija“ i „Južna Malezija“ pridružili njihovim odgovarajućim referentnim nivoima.
Preostaje nam još da prikažemo rezultate posle primene postupka 1 sa zaokruživanjem na cele brojeve. U tabeli 14 prikazane su konačne vrednosti skorova svih kategorija normalnog skoring modela. Najmanja moguća premija primenom normalnog skoring modela je 53, dok je maksimalna moguća premija 84, odnosno, premija se kreće u rasponu od 156 RM do 3000 RM.

Tabela 14: Ocenjene, modifikovane i konačne vrednosti skorova kategorija normalnog skoring modela

Kategorije	**Koeficijenti**	**Modifikovana vrednost skora**	**Konačna vrednost skora**
Slobodan član (minimalni skor)	79.02	52.72	53
Obim pokrića police			
Standardno	-12.79	0.00	0
Sveobuhvatno	0.00	12.79	13
Marka vozila			
Strana marka	4.02	4.02	4
Lokalna marka	0.00	0.00	0
Pol osiguranika			
Pravno lice	-7.40	0.00	0
Fizičko lice-žena	-4.03	3.37	3
Fizičko lice-muškarac	0.00	7.40	7
Starost vozila			
2-3	0.00	2.10	2
4-5	-1.17	0.93	1
6+	-2.10	0.00	0
0-1	0.00	2.10	2
Regija			
Severna Malezija	0.49	4.50	5
Istočna Malezija	0.00	4.01	4
Južna Malezija	0.00	4.01	4
Ostrvo Borneo	-4.01	0.00	0
Centralna Malezija	0.00	4.01	4

6 Istraživanje mogućnosti primene skoring modela u Republici Srbiji

Način odredjivanja cene kasko osiguranja u osiguravajućim društvima u Srbiji je u domenu poslovne tajne. Medjutim, iz raznih kalkulatora kasko osiguranja, koji su dostupni na sajtovima domaćih osiguravajućih kuća vidimo da premija zavisi od nekolicine faktora. Neki od tih faktora su: vrednost vozila, starost vozila, marka vozila, rizici koje premija pokriva, teritorijalno pokriće, itd. Intuitivno, na osnovu ovih činjenica, skoring modeli mogu naći primenu u odredjivanju premije kasko osiguranja.
Mogućnost primene skoring modela u Srbiji ispitaćemo istraživanjem na podacima jedne domaće osiguravajuće kuće. Cilj istraživanja je konstrukcija skoring modela na dobijenim podacima i provera valjanosti modela, odnosno potrebno je izvesti zaključak o tome da li se skoring modeli mogu primeniti u osiguravajućim društvima koje posluju u Republici Srbiji.
Podaci sadrže informacije 107 799 jednogodišnjih polisa kasko osiguranja putničkih vozila koje su zaključene u periodu od 2015. do 2019. godine. Ove podatke ćemo podeliti u dve grupe

I podaci za testiranje - nasumično odabranih 5 390 (5%) polisa iz početnog skupa podataka

II podaci za kontrukciju - preostalih 102 409 polisa

6.1 Istraživačka analiza podataka

Svaka od 107 799 polisa osiguranja sadrži informacije o zaključenoj godišnjoj premiji (zavisna promenljiva), kao i informacije o 9 kategorijalnih varijabli (nezavisne promenljive) pomoću kojih želimo oceniti premiju. Nezavisne promenljive sadržane u podacima su

(i) **Obim polise** - predstavlja informaciju o raznim doplatama osiguranika (npr. malus, doplata za neiskusne vozače, itd.).Moguće vrednosti su d1 i d2.

(ii) **Snaga motora** - predstavlja informaciju o snazi motora osiguranog vozila. Varijabla ima 4 moguće vrednosti sn1, sn2, sn3 i sn4.

(iii) **Vlasnik polise** - predstavlja informaciju o osiguraniku polise. Moguće vrednosti su Z - osiguranik je fizičko lice - žena, M - osiguranik je fizičko lice - muškarac i PL - osiguranik je pravno lice.

(iv) **Starost vozila** - predstavlja informaciju o starosti vozila u trenutku zaključivanja polise. Moguće vrednosti su st1 koja predstavlja sva vozila starosti do 4 godine, st2 koja predstavlja sva vozila starosti od 5 do 8 godina i st3 su vozila starosti 9 godina i starija.

(v) **Region** - predstavlja informaciju o regionu u kojem je ugovorena polisa. Moguće vrednosti su BG - Beograd i OST - ostatak Srbije.

(vi) **Osigurana suma vozila** - predstavlja iznos osiguranog rizika. Najčešće je to cena osiguranog vozila, medjutim u zavisnosti od paketa kupljenog kasko osiguranja, ovaj iznos može biti i uvećan za osiguranje dodatnih rizika i umanjen ako se osiguranik odluči za osiguranje samo odredjenih rizika (na primer, dodatne opreme automobila). Varijabla ima 12 mogućih vrednosti A, B, C, ..., K, L.

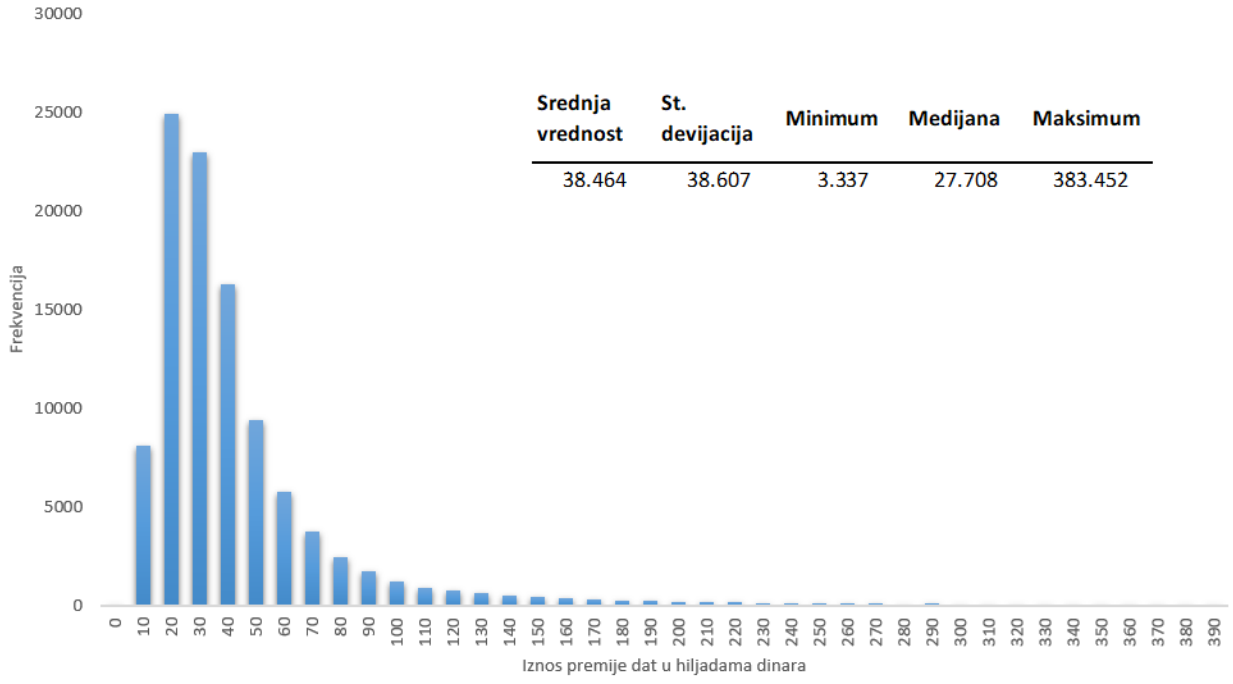
(vii) **Zapremina motora** - predstavlja informaciju o zapremini motora osiguranog vozila. Varijabla ima 4 moguće vrednosti z1, z2, z3 i z4.

(viii) **Bonus** - daje nam informaciju o tome za koliko je osnovna premija smanjena. Moguće vrednosti su b1, b2 i b3.

(ix) **Učešće** - daje nam odgovor na pitanje da li je ugovoreno učešće osiguranika u eventualnoj šteti. Moguće vrednosti su u1, u2 i u3.

6.2 Deskriptivna statistika podataka

U nastavku posmatramo podskup podataka koji koristimo za konstrukciju modela. Na slici 5 prikazane su frekvencije zaključene premije. Grafik frekvencije iznosa premije nam pokazuje da je najveći broj polisa (oko 25 000) ima iznos premije izmedju 10 000 i 20 000 dinara. Takodje, čak 85% polisa ima premiju manju od 60 000 dinara. Ova činjenica kao i zapažanje da je prosečna vrednost premije veća od medijane za skoro 11 000 dinara , nam ukazuje da grafik frekvencije iznosa premije obrazuje pozitivno asimetričnu raspodelu. Zbog toga su raspodele poput normalne, lognormalne i gama raspodele (kao što je navedeno i u studiji [12]), sa odgovarajućim parametrima, idealne za modeliranje premije.

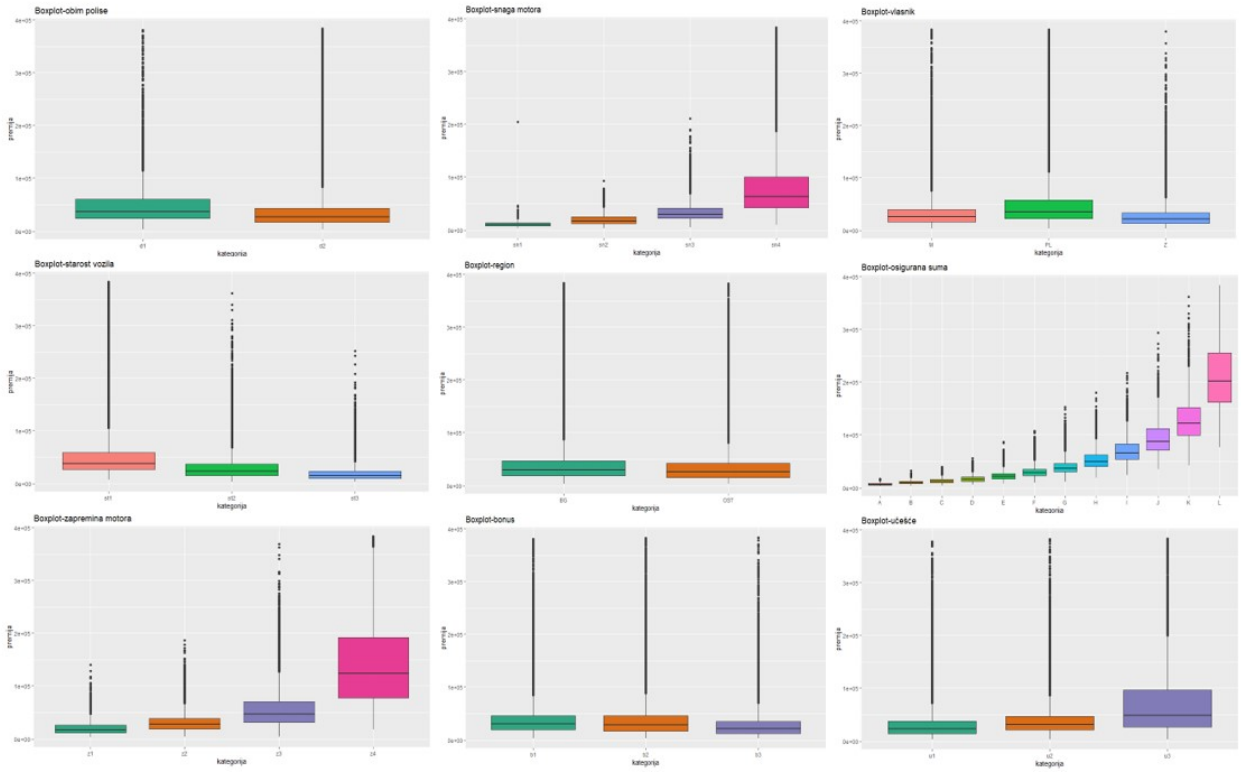


Slika 5: Grafički prikaz frekvencije zavisne promenljive

Što se tiče nezavisnih varijabli, njihova deskriptivna statistika u odnosu na premiju data je u tabeli 15. Skup nezavisnih varijabli čini 9 kategorijalnih promenljivih sa svojim kategorijama koje su definisane u drugoj i trećoj koloni tabele 15. Kategorijalna promenljiva osigurana suma ima najveći broj kategorija (12), dok su kategorijalne promenljive obim polise i region one sa najmanjim brojem kategorija (2).

Tabela 15: Deskriptivna statistika kategorija nezavisnih varijabli

KAT. PROMENLJIVA	KRITERIJUM	KATEGORIJA	BROJ POLISA	SR. VREDNOST	ST. DEVIJACIJA	MINIMUM	MEDIJANA	MAKSIMUM
Obim polise	sa doplatom	d1	4.024	54.508	54.253	3.755	37.351	380.935
	bez doplate	d2	98.385	37.808	37.685	3.337	27.374	383.452
Snaga motora(kW)	[0, 49]	sn1	7.846	11.534	5.675	3.337	10.457	204.682
	(49, 69]	sn2	27.567	19.444	9.664	3.363	17.355	92.940
	(69, 104]	sn3	44.898	33.464	16.485	5.052	30.116	211.437
	(104, +∞)	sn4	22.098	81.913	59.835	10.089	63.206	383.452
Vlasnik polise	žena	Z	20.547	27.303	24.831	3.409	21.086	379.430
	muškarac	M	45.215	33.772	31.018	3.363	26.057	383.450
	pravno lice	PL	36.647	50.511	48.892	3.337	34.800	383.452
Starost vozila (br. godina)	[0, 4]	st1	46.481	53.348	47.415	6.871	37.927	383.452
	[4, 8]	st2	32.758	30.845	25.984	3.543	23.315	361.284
	(8, +∞)	st3	23.170	19.378	15.253	3.337	15.098	251.901
Region	Beograd	BG	49.806	40.898	40.405	3.409	29.682	383.452
	ostatak Srbije	OST	52.603	36.160	36.676	3.337	25.721	382.330
Osigurana suma (RSD)	[0, 130.000)	A	1.087	6.985	2.632	3.337	6.259	16.639
	[130.000, 230.000)	B	5.545	10.454	3.908	3.543	9.766	32.466
	[230.000, 360.000)	C	10.901	13.177	4.983	4.129	12.126	39.460
	[360.000, 490.000)	D	9.845	17.108	6.160	5.645	15.894	56.018
	[490.000, 760.000)	E	16.663	22.864	8.058	7.483	21.466	86.964
	[760.000, 1.110.000)	F	18.774	30.429	10.338	9.335	28.517	108.110
	[1.110.000, 1.600.000)	G	17.949	39.939	13.731	11.963	37.154	152.755
	[1.600.000, 2.030.000)	H	8.045	53.226	18.280	18.693	49.452	180.234
	[2.030.000, 2.790.000)	I	6.616	70.961	24.107	24.607	65.765	217.392
	[2.790.000, 3.440.000)	J	2.449	95.013	33.311	35.333	87.815	293.577
	[3.440.000, 5.020.000)	K	2.471	130.499	43.383	42.416	122.627	361.284
	[5.020.000, +∞)	L	2.064	212.928	66.115	77.064	201.818	383.452
Zapremina motora (ccm)	[0, 1399]	z1	37.373	19.981	11.289	3.337	17.290	139.949
	(1399, 1899]	z2	34.778	31.012	16.435	3.977	27.790	186.585
	(1899, 2899]	z3	25.681	56.844	36.944	4.516	46.921	368.861
	(2899, +∞)	z4	4.577	142.887	81.177	18.243	124.114	383.452
Bonus	do 40% osnovne premije	b1	29.338	40.702	37.996	3.618	30.322	380.750
	od 40% do 60% osnovne premije	b2	55.128	39.088	38.436	3.354	28.457	382.299
	veće od 60% osnovne premije	b3	17.943	32.890	39.590	3.337	21.508	383.452
Učešće	bez učešća	u1	57.315	31.768	30.879	3.337	23.579	378.487
	učešće od 5%	u2	35.966	40.047	33.021	3.623	31.216	382.330
	učešće od 10% i veće	u3	9.128	74.276	69.754	3.767	48.551	383.452



Slika 6: Grafički prikaz premije po kategorijama nezavisnih promenljivih

Frekvencija iznosa premije po kategorijama prikazana je preko box-plot grafika na slici 6.
U tabeli 16 prikazane su sve kategorijalne varijable i kategorije sa njihovim matematičkim oznakama, kao i oznake odgovarajućih binarnih varijabli. Za svaku kategoriju A_{jk}, x_{jk} je binarna varijabla koja ima vrednosti 0 ili 1, u zavisnosti od toga da li polisa zadovoljava odgovarajući kriterijum posmatrane kategorije. Za svaku kategorijalnu promenljivu definisaćemo referentni nivo. Bez gubljenja opštosti, neka je za svaku kategorijalnu promenljivu za referentni nivo odabrana ona kategorija sa najvećim brojem polisa (četvrta kolona u tabeli 15). Dakle, za referentne nivoe biramo kategorije d2, sn3, M, st1, OST, F, z1, b2 i u1.

Tabela 16: Kategorijalne promenljive i kategorije koje učestvuju u kontrukciji skoring modela

Kat. promenljiva (oznaka)	**Broj kategorija - r_j**	**Kategorija**	**Oznaka kategorije**	**Skor**	**Bin. varijabla**
obim polise(A_1)	2	d1	A_{11}	S_{11}	x_{11}
		d2	A_{12}	S_{12}	x_{12}
snaga motora(A_2)	4	sn1	A_{21}	S_{21}	x_{21}
		sn2	A_{22}	S_{22}	x_{22}
		sn3	A_{23}	S_{23}	x_{23}
		sn4	A_{24}	S_{24}	x_{24}
vlasnik polise(A_3)	3	Z	A_{31}	S_{31}	x_{31}
		M	A_{32}	S_{32}	x_{32}
		PL	A_{33}	S_{33}	x_{33}
starost vozila(A_4)	3	st1	A_{41}	S_{41}	x_{41}
		st2	A_{42}	S_{42}	x_{42}
		st3	A_{43}	S_{43}	x_{43}
region(A_5)	2	BG	A_{51}	S_{51}	x_{51}
		OST	A_{52}	S_{52}	x_{52}
osigurana suma(A_6)	12	A	A_{61}	S_{61}	x_{61}
		B	A_{62}	S_{62}	x_{62}
		C	A_{63}	S_{63}	x_{63}
		D	A_{64}	S_{64}	x_{64}
		E	A_{65}	S_{65}	x_{65}
		F	A_{66}	S_{66}	x_{66}
		G	A_{67}	S_{67}	x_{67}
		H	A_{68}	S_{68}	x_{68}
		I	A_{69}	S_{69}	x_{69}
		J	A_{610}	S_{610}	x_{610}
		K	A_{611}	S_{611}	x_{611}
		L	A_{612}	S_{612}	x_{612}
zapremina motora(A_7)	4	z1	A_{71}	S_{71}	x_{71}
		z2	A_{72}	S_{72}	x_{72}
		z3	A_{73}	S_{73}	x_{73}
		z4	A_{74}	S_{74}	x_{74}
bonus(A_8)	3	b1	A_{81}	S_{81}	x_{81}
		b2	A_{82}	S_{82}	x_{82}
		b3	A_{83}	S_{83}	x_{83}
učešće(A_9)	3	u1	A_{91}	S_{91}	x_{92}
		u2	A_{92}	S_{92}	x_{92}
		u3	A_{93}	S_{93}	x_{93}

Ispitivanje zavisnosti izmedju binarnih varijabli

Pre definisanja modela, proverićemo korelaciju izmedju objašnjavajućih varijabli. Zavisnost varijabli proveravamo korišćenjem tabela kontigencije i računanjem vrednosti ϕ_{test} (videti odeljak 3.7). Rezultati su dati na slici 7. Iz korelacione matrice vidimo da je najveća korelacija (pozitivna ili negativna) ostvarena izmedju kategorija sn2 i z3 ($\phi_{test} = 0,58$), medjutim sve vrednosti ϕ_{test} su oko nule, pa zaključujemo da ne postoji značajnija korelacija medju nezavisnim promenljivama.

	x_{i11}	x_{i21}	x_{i22}	x_{i24}	x_{i31}	x_{i33}	x_{i41}	x_{i43}	x_{i51}	x_{i61}	x_{i62}	x_{i63}	x_{i64}	x_{i65}	x_{i67}	x_{i68}	x_{i69}	$x_{i6\,10}$	$x_{i6\,11}$	$x_{i6\,12}$	x_{i72}	x_{i73}	x_{i74}	x_{i81}	x_{i83}	x_{i92}	x_{i93}
x_{i11}	1,00	-0,01	0,03	0,00	-0,06	0,15	-0,01	-0,04	0,06	-0,01	-0,01	-0,01	0,00	0,03	-0,02	-0,01	0,00	-0,01	0,01	0,03	0,00	-0,01	0,02	0,09	-0,01	0,03	0,06
x_{i21}	-0,01	1,00	-0,17	-0,15	0,10	-0,07	0,06	0,18	-0,07	0,22	0,27	0,26	0,06	-0,06	-0,13	-0,08	-0,08	-0,05	-0,05	-0,04	-0,17	-0,17	-0,06	0,03	-0,01	-0,11	-0,06
x_{i22}	0,03	-0,17	1,00	-0,32	0,13	-0,06	0,02	0,10	-0,02	0,02	0,13	0,18	0,15	0,10	-0,11	-0,15	-0,16	-0,09	-0,10	-0,09	-0,08	-0,33	-0,13	0,07	-0,02	-0,08	-0,12
x_{i24}	0,00	-0,15	-0,32	1,00	-0,14	0,14	0,01	-0,10	0,01	-0,05	-0,13	-0,17	-0,14	-0,13	0,01	0,10	0,24	0,23	0,28	0,27	-0,32	0,58	0,41	-0,07	0,03	0,06	0,24
x_{i31}	-0,06	0,10	0,13	-0,14	1,00	-0,37	0,03	0,07	0,00	0,03	0,08	0,09	0,05	0,01	-0,04	-0,05	-0,06	-0,04	-0,05	-0,05	-0,07	-0,13	-0,08	0,10	-0,03	0,00	-0,07
x_{i33}	0,15	-0,07	-0,06	0,14	-0,37	1,00	-0,09	-0,17	0,06	-0,04	-0,10	-0,11	-0,08	-0,04	0,05	0,06	0,08	0,07	0,09	0,09	0,03	0,08	0,10	0,05	0,05	-0,08	0,04
x_{i41}	-0,01	0,06	0,02	0,01	0,03	-0,09	1,00	-0,37	-0,03	-0,07	-0,14	0,01	0,15	0,23	-0,08	-0,10	-0,10	-0,06	-0,08	-0,09	-0,04	0,03	0,01	0,09	-0,10	-0,02	0,00
x_{i43}	-0,04	0,18	0,10	-0,10	0,07	-0,17	-0,37	1,00	-0,08	0,19	0,40	0,35	0,15	-0,01	-0,22	-0,15	-0,13	-0,08	-0,08	0,08	-0,04	-0,02	-0,02	-0,05	0,15	-0,09	-0,05
x_{i51}	0,06	-0,07	-0,02	0,01	0,00	0,06	-0,03	-0,08	1,00	-0,03	-0,05	-0,05	-0,02	0,00	0,04	0,02	0,01	0,00	0,00	0,01	0,02	-0,02	0,00	0,06	-0,05	0,06	-0,02
x_{i61}	-0,01	0,22	0,02	-0,05	0,03	-0,04	-0,07	0,19	-0,03	1,00	-0,02	-0,04	-0,03	-0,05	-0,05	-0,03	-0,03	-0,02	-0,02	-0,01	-0,05	-0,06	-0,02	-0,01	0,03	-0,05	-0,03
x_{i62}	-0,01	0,27	0,13	-0,13	0,08	-0,10	-0,14	0,40	-0,05	-0,02	1,00	-0,08	-0,08	-0,11	-0,11	-0,07	-0,06	-0,04	-0,04	-0,03	-0,05	-0,12	-0,05	-0,01	0,07	-0,08	-0,05
x_{i63}	-0,01	0,26	0,18	-0,17	0,09	-0,11	0,01	0,35	-0,05	-0,04	-0,08	1,00	-0,11	-0,15	-0,16	-0,10	-0,09	-0,05	-0,05	-0,05	-0,06	-0,14	-0,07	0,02	0,03	-0,01	-0,06
x_{i64}	0,00	0,06	0,15	-0,14	0,05	-0,08	0,15	0,15	-0,02	-0,03	-0,08	-0,11	1,00	-0,14	-0,15	-0,10	-0,09	-0,05	-0,05	-0,05	0,01	-0,10	-0,07	0,03	0,00	-0,07	-0,05
x_{i65}	0,03	-0,06	0,10	-0,13	0,01	-0,04	0,23	-0,01	0,00	-0,05	-0,11	-0,15	-0,14	1,00	-0,20	-0,13	-0,12	-0,07	-0,07	-0,06	0,05	-0,06	-0,09	0,06	-0,03	-0,01	-0,06
x_{i67}	-0,02	-0,13	-0,11	0,01	-0,04	0,05	-0,08	-0,22	0,04	-0,05	-0,11	-0,16	-0,15	-0,20	1,00	-0,13	-0,12	-0,07	-0,07	-0,07	0,05	0,05	-0,04	-0,02	-0,02	0,07	0,00
x_{i68}	-0,01	-0,08	-0,15	0,10	-0,05	0,06	-0,10	-0,15	0,02	-0,03	-0,07	-0,10	-0,10	-0,13	-0,13	1,00	-0,08	-0,05	-0,05	-0,04	0,04	0,10	0,01	-0,04	-0,01	0,06	0,03
x_{i169}	0,00	-0,08	-0,16	0,24	-0,06	0,08	-0,10	-0,13	0,01	-0,03	-0,06	-0,09	-0,09	-0,12	-0,12	-0,08	1,00	-0,04	-0,04	-0,04	-0,01	0,16	0,07	-0,05	0,00	0,06	0,06
$x_{i6\,10}$	-0,01	-0,05	-0,09	0,23	-0,04	0,07	-0,06	-0,08	0,00	-0,02	-0,04	-0,05	-0,05	-0,07	-0,07	-0,05	-0,04	1,00	-0,02	-0,02	-0,06	0,14	0,10	-0,03	0,00	0,03	0,06
$x_{i6\,11}$	0,01	-0,05	-0,10	0,28	-0,05	0,09	-0,08	-0,08	0,00	-0,02	-0,04	-0,05	-0,05	-0,07	-0,07	-0,05	-0,04	-0,02	1,00	-0,02	-0,10	0,14	0,20	-0,04	0,01	0,01	0,13
$x_{i6\,12}$	0,03	-0,04	-0,09	0,27	-0,05	0,09	-0,09	0,08	0,01	-0,01	-0,03	-0,05	-0,05	-0,06	-0,07	-0,04	-0,04	-0,02	-0,02	1,00	-0,10	0,02	0,45	-0,04	0,02	-0,02	0,22
x_{i72}	0,00	-0,17	-0,08	-0,32	-0,07	0,03	-0,04	-0,04	0,02	-0,05	-0,05	-0,06	0,01	0,05	0,05	0,04	-0,01	-0,06	-0,10	-0,10	1,00	-0,41	-0,16	0,00	-0,01	0,03	-0,08
x_{i73}	-0,01	-0,17	-0,33	0,58	-0,13	0,08	0,03	-0,02	-0,02	-0,06	-0,12	-0,14	-0,10	-0,06	0,05	0,10	0,16	0,14	0,14	0,02	-0,41	1,00	-0,13	-0,06	0,03	0,08	0,13
x_{i74}	0,02	-0,06	-0,13	0,41	-0,08	0,10	0,01	-0,02	0,00	-0,02	-0,05	-0,07	-0,07	-0,09	-0,04	0,01	0,07	0,10	0,20	0,45	-0,16	-0,13	1,00	-0,04	0,03	-0,03	0,25
x_{i81}	0,09	0,03	0,07	-0,07	0,10	0,05	0,09	-0,05	0,06	-0,01	-0,01	0,02	0,03	0,06	-0,02	-0,04	-0,05	-0,03	-0,04	-0,04	0,00	-0,06	-0,04	1,00	-0,29	0,14	0,03
x_{i83}	-0,01	-0,01	-0,02	0,03	-0,03	0,05	-0,10	0,15	-0,05	0,03	0,07	0,03	0,00	-0,03	-0,02	-0,01	0,00	0,00	0,01	0,02	-0,01	0,03	0,03	-0,29	1,00	-0,16	-0,03
x_{i92}	0,03	-0,11	-0,08	0,06	0,00	-0,08	-0,02	-0,09	0,06	-0,05	-0,08	-0,01	-0,07	-0,01	0,07	0,06	0,06	0,03	0,01	-0,02	0,03	0,08	-0,03	0,14	-0,16	1,00	-0,23
x_{i93}	0,06	-0,06	-0,12	0,24	-0,07	0,04	0,00	-0,05	-0,02	-0,03	-0,05	-0,06	-0,05	-0,06	0,00	0,03	0,06	0,06	0,13	0,22	-0,08	0,13	0,25	0,03	-0,03	-0,23	1,00

Slika 7: Zavisnost binarnih varijabli prikazana kroz vrednost ϕ_{test}

6.3 Skoring model 1 - konstrukcija skoring modela pomoću klasa osiguranika

Slično kao u studiji [12] kontruisaćemo skoring model u kojem se koriste klase osiguranika. Za konstrukciju ovog modela koristićemo kategorijalne varijable snaga motora, starost vozila, obim polise, region i vlasnik polise. Detaljan prikaz ovih varijabli i njihovih kategorija dat je u tabeli 16. Kako ove kategorijalne promenljive imaju, redom, 2, 4, 3, 3 i 2 kategorije, imamo da ove kategorijalne promenljive obrazuju ukupno $2 \cdot 4 \cdot 3 \cdot 3 \cdot 2 = 144$ klasa osiguranika. U tabeli 24, koja se nalazi u plilogu ovog rada, prikazana je srednja vrednost premije i vrednost izloženosti za svaku klasu osiguranika. Kako su u podacima sve polise jednogodišnje, izloženost klase je prikazana kroz broj polisa. Primetimo, broj polisa klase $K7$ je jednak 0, pa ćemo tu klasu izuzeti iz dalje konstrukcije modela.
Skoring model koji želimo dobiti je dat sa

$$s = S_0 + \sum_{j=1}^{5} \sum_{k=1}^{r_j} S_{jk} x_{jk}, \tag{13}$$

gde

- s predstavlja ukupnu vrednost skora
- S_0 označava vrednost početnog skora (minimalan skor)
- S_{jk} predstavlja vrednost skora kategorije A_{jk}, $j = 1, 2, ..., 5$, $k = 1, ..., r_j$
- x_{jk} predstavlja vrednost binarne promenljive A_{jk}-te kategorije , tačnije

  $$x_{jk} = \begin{cases} 1, & \text{osiguranik ima osobinu } A_{jk} \\ 0, & \text{inače} \end{cases}$$

Cilj nam je, dakle, da odredimo vrednosti početnog skora S_0 i skorove kategorija $S_{11}, S_{12}, ..., S_{52}$. Skorove ćemo odrediti pomoću odgovarajućeg gama linearnog modela sa logaritamskom link funkcijom.
Kategorije sn3, st2, d2, OST i PL su označene kao referentni nivoi (u tabeli 16 označene crvenom bojom), pa ako sa Y označimo slučajnu promenljivu koja predstavlja premiju osiguranja, onda gama model ima oblik

$$g(E(Y)) = g(\mu) = \ln \mu = \eta = X\boldsymbol{\beta}, \tag{14}$$

gde

(i) Y ima gama raspodelu sa očekivanjem μ i disperzijom $e^{-1}\sigma^2\mu^2$

(ii) vektor $\boldsymbol{\beta}$ je vektor koeficijenata modela, dok je X vektor nezavisnih binarnih varijabli čiji je prvi element jednak 1, tj. $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_9 \end{bmatrix}$ i $X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{51} \end{bmatrix}$

(iii) g je logaritamska funkcija, tačnije $g(\mu) = \ln \mu$

Za potrebe računanja koeficijenata modela (14), koristimo statistički softver „R“ i njegovu ugradjenu funkciju glm(), koja koeficijente uopštenih linearnih modela, na osnovu uzorka, računa pomoću iterativne metode ponderisanih najmanjih kvadrata.
Primenom stepenaste metode unazad, gde kao kriterijume koristimo vrednosti AIC i $adj.R^2$, tražimo model koji najbolje ocenjuje srednju vrednost premija klasa osiguranika. Nivo značajnosti će u ovom radu biti 95%. Počevši od modela koji sadrži sve binarne promenljive, u svakoj iteraciji stepenaste metode, koeficijente modela računamo pomoću softvera.
Posle primene stepenaste metode, dobijamo sledeći model

$$\hat{\eta} = 10,58 - 0,28x_{11} - 0,8x_{21} - 0,47x_{22} + 0,84x_{24} + 0,12x_{33} - 0,4x_{42} - 0,68x_{43} \ .$$

Dakle, primenom stepenaste metode unazad, iz modela su odstranjene binarne varijable x_{51} i x_{31}, što znači da je kategorijalna promenljiva region odstranjena iz dalje konstrukcije skoring modela, dok je kategorija Z pripojena odgovarajućem referentnom nivou, tj. kategoriji M.
Za ovaj model je $adj.R^2 = 98,62\%$, što je veoma visoka vrednost i govori nam da je model dobro konstruisan.
U tabeli 17 prikazani su ocenjeni koeficijenti, njihove standardne greške, kao i vrednosti Valdovog testa za testiranje hipoteze o jednom parametru. Standardne greške ocena koeficijenata su male, što nam govori o pouzdanosti ocena. Kako su p-vrednosti Valdovog testa manje od 0,05, zaključujemo da su sve ocene koeficijenata statistički značajne na nivou značajnosti od 95%.

Tabela 17: Ocenjeni koeficijenti gama modela, njihove standardne greške i odgovarajuće vrednosti skorova kategorija

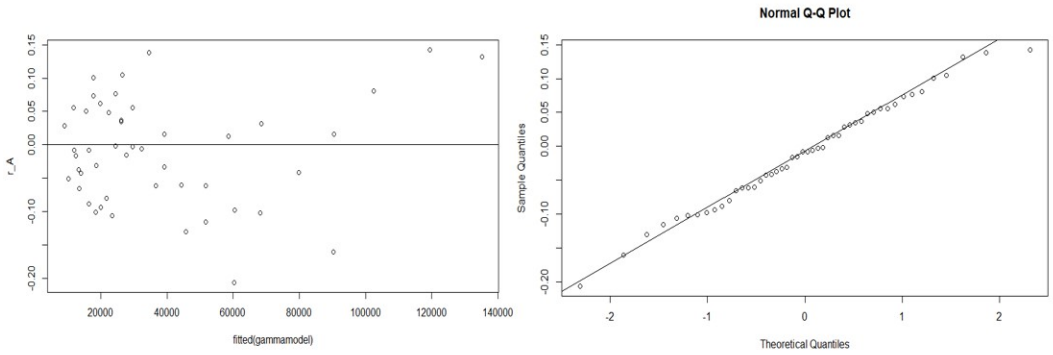
Kategorija	**Koeficijent**	**Ocena**	**St. Greška**	**w-vrednost**	**p vrednost**	**Modifikovana ocena**	**Skor**
Odsečak	β_0	10,58	0,01	437400	<0,001	9,10	65
d1	β_1	0,28	0,03	46	<0,001	0,28	2
d2						0,00	0
sn1	β_2	-0,80	0,02	627	<0,001	0,00	0
sn2	β_3	-0,47	0,01	579	<0,001	0,33	2
sn3						0,80	6
sn4	β_4	0,84	0,02	1611	<0,001	1,63	12
M, Z						0,00	0
PL	β_6	0,12	0,01	52	<0,001	0,12	1
st1						0,68	5
st2	β_7	-0,40	0,01	467	<0,001	0,28	2
st3	β_8	-0,68	0,02	1014	<0,001	0,00	0

Analiza reziduala gama modela

U analizi reziduala gama modela, koristićemo Anskombe reziduale. Za dobro definisane i ocenjene modele smo rekli da Anskombe reziduali imaju približno normalnu raspodelu. Anskombe reziduali gama modela dati su sa

$$(r_A)_i = 3\left(\left(\frac{y_i}{\hat{\mu}_i}\right)^{1/3} - 1\right) \ .$$

Grafički prikaz Anskombe reziduala dat je na slici 8. Na levom grafiku vidimo da vrednosti rezi-



Slika 8: Grafički prikaz anskombe reziduala i njihov Q-Q grafik

duala ne odstupaju previše od nule, dok na desnom grafiku (Q-Q grafik) vidimo da reziduali imaju približno normalnu raspodelu. Medjutim, kako je naš uzorak malog obima, normalnost reziduala ćemo ispitati i Šapiro-Vilk testom. Test statistiku Šapiro-Vilk testa za reziduale gama modela računamo u softveru „R“ i dobijamo $W = 0,98943$, kao i odgovarajuću p-vrednost koja iznosi $0,94$. Kako je p-vrednost veća od zadate $0,05$ zaključujemo da se raspodela reziduala statistički značajno ne razlikuje od normalne raspodele.

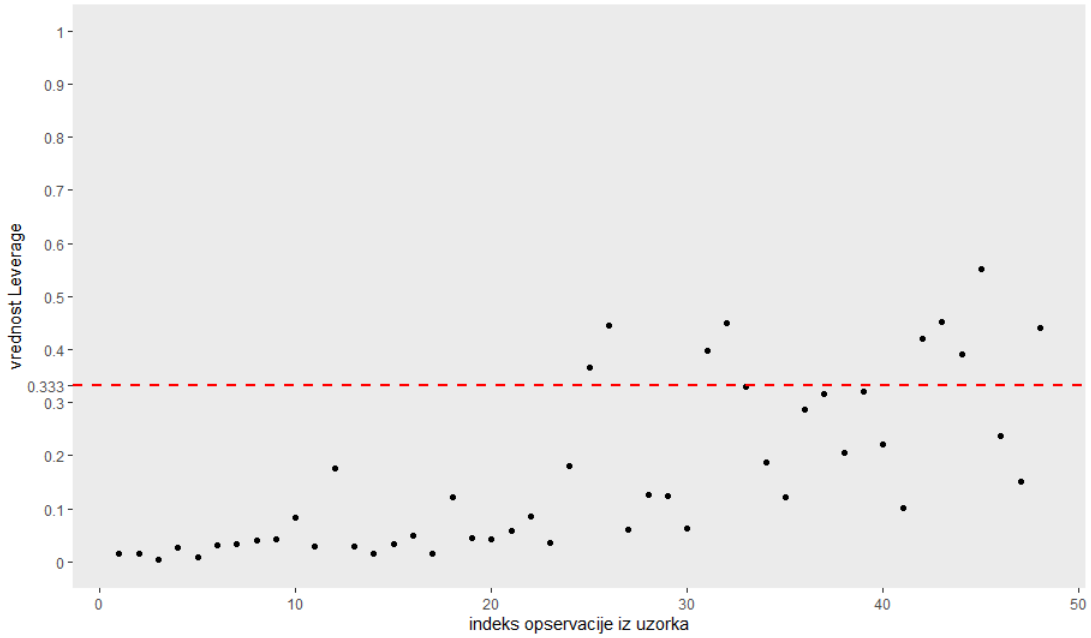
Autlajeri i uticajne tačke

Detektovanje uticajnih tačaka sprovodimo Leveridž metodom. Na slici 9 prikazane su vrednosti Leveridž analize (y-osa) uticajnih tačaka za svaku opservaciju (x-osa) iz uzorka. Isprekidana linija označava granicu $\frac{2p+2}{n} = 0,333$. Sve opservacije čija je vrednost Leveridža veća od granice $0,333$, tj. to su tačke na grefiku koje su iznad isprekidane linije, predstavljaju potencijalne uticajne tačke. Dodatnom analizom modela, koja je podrazumevala konstrukciju modela sa i bez ovih tačaka, utvrdjeno je da potencijalne uticajne tačke ne utiču na vrednosti koeficijenata i parametara modela.

Sada prelazimo na pretvaranje ocena gama modela u skorove modela datog jednačinom 13. Kolona modifikovana ocena u tabeli 17 predstavlja vrednosti skorova kategorija koji su dobijeni iz ocenjenih vrednosti odgovarajućih koeficijenata gama modela, kada se na njih primeni postupak 1, koji je opisan u teorijskom delu. Kolona skor se dobija kada se prethodna kolona podeli brojem $\ln 1,15$ i zaokruži na ceo broj. Na ovaj način smo dobili pregledniji skoring model, jer se sada skorovi kreću u rasponu od 65 do 85. Deljenjem sa $\ln 1,15$ mi ustvari menjamo bazu logaritamske funkcije, sa exp na $1,15$, koja je prirodnija za naše podatke.

Korekcija skoring modela

Pogledajmo sada u kakvom su odnosu ocenjena i stvarna premija. U tabeli 18 vidimo da je ukupna fitovana premija manja od stvarne premije za nešto više od 122 miliona dinara. Medjutim, kako je ukupna stvarna premija skoro 4 milijarde dinara, možemo konstatovati da model dobro fituje premiju kasko osiguranja. U prilog tome je i premijski racio, tj. odnos fitovane i stvarne premije, koji je približno 1,031. U cilju poboljšanja modela fitovanu premiju je potrebno pomnožiti sa koeficijentom korekcije koji je jednak recipročnoj vrednosti premijskog racija. Na taj način, će ukupan premijski racio biti 1, dok će razlika ukupne fitovane vrednosti i ukupne stvarne vrednosti premije



Slika 9: Vrednosti Leveridž analize za svaku opservaciju uzorka

biti 0, što je od velike važnosti za menadžment osiguravajućeg društva, jer se korišćenjem modela ne gubi na prihodu.

Tabela 18: Ukupna vrednost, razlika i racio prihoda od fitovane i stvarne premije

		Vrednost
Ukupna vrednost izloženosti	$\sum_{i=1}^{144} e_i$	102 409
Ukupna vrednost fitovanih premija	$\sum_{i=1}^{144} e_i \hat{y}_i$	4 061 106 521
Ukupna vrednost stvarnih premija	$\sum_{i=1}^{144} e_i y_i$	3 939 090 348
Razlika ukupne vrednosti stvarne i fitovane premije	$\sum_{i=1}^{144} e_i(\hat{y}_i - y_i)$	122 016 173
Ukupan premijski racio	$\frac{\sum_{i=1}^{144} e_i \hat{y}_i}{\sum_{i=1}^{144} e_i y_i}$	1,031

6.4 Lognormalni i normalni skoring model

Lognormalni i normalni skoring model podrazumeva pretpostavku da slučajna promenljiva Y ima lognormalnu i normalnu raspodelu, redom. Etape konstrukcije lognormalnog i normalnog skoring modela na podacima za konstrukciju je potpuno analogan prethodnoj kontrukciji, pa ćemo navesti samo glavne rezultate i zaključke ovih skoring modela.
U tabeli 19 prikazani su dobijeni rezultati lognormalnog skoring modela. Ocene koeficijenata lognormalnog modela su gotovo identične onima koje smo dobili u gama modelu, s tim što je standardna greška ocena lognormalnog modela nešto veća. Minimalna razlika u ocenama modela implicira da gama i lognormalni skoring model imaju potpuno iste skorove. U tabeli 20 date su vrednosti normal-

Tabela 19: Ocenjeni koeficijenti lognormalnog modela, njihove standardne greške i odgovarajuće vrednosti skorova kategorija

Kategorija	**Koeficijent**	**Ocena**	**St. Greška**	**w-vrednost**	**p vrednost**	**p vrednost**	**Skor**
Odsečak	β_0	10,57	0,02	432238	<2e-16	9,10	65
d1	β_1	0,27	0,04	44	<2e-11	0,27	2
d2						0,00	0
sn1	β_2	-0,79	0,03	619	<6e-16	0,00	0
sn2	β_3	-0,47	0,02	571	<2e-16	0,33	2
sn3						0,79	6
sn4	β_4	0,83	0,02	1584	<2e-16	1,63	12
M, Z						0,00	0
PL	β_6	0,12	0,02	51	<2e-16	0,12	1
st1						0,68	5
st2	β_7	-0,40	0,02	461	<2e-16	0,28	2
st3	β_8	-0,68	0,02	1004	<2e-16	0,00	0
	$adj.R^2$	99, 29%					

nog skoring modela. Ocene parametara normalnog modela se nešto više razlikuju od prethodna dva konstruisana modela. Medjutim, te razlike nisu dovele do većih odstupanja u vrednostima skorova kategorija, ako posmatramo sva tri modela. Jedina razika u odnosu na gama i lognormalni skoring model je u vrednosti skora kategorije sn3, tačnije kod normalnog skoring modela ta vrednost je niža za 1 i sada iznosi 5.
Ova tri konstruisana skoring modela su nam pokazala da pretpostavka o različitim raspodelama premije ne daje značajnu razliku u krajnjim skoring modelima, kao što je to bio slučaj u studiji [12].

Tabela 20: Ocenjeni koeficijenti normalnog modela, njihove standardne greške i odgovarajuće vrednosti skorova kategorija

Kategorija	**Koeficijent**	**Ocena**	**St. Greška**	**w-vrednost**	**p vrednost**	**Modifikovana ocena**	**Skor**
Odsečak	β_0	10,54	0,04	376441	<2e-16	9,04	65
d1	β_1	0,30	0,03	131	<2e-16	0,30	2
d2						0,00	0
sn1	β_2	-0,74	0,11	52	<6e-13	0,00	0
sn2	β_3	-0,43	0,04	153	<2e-16	0,31	2
sn3						0,74	5
sn4	β_4	0,89	0,02	2669	<2e-16	1,63	12
M, Z						0,00	0
PL	β_6	0,17	0,02	125	<2e-16	0,17	1
st1						0,75	5
st2	β_7	-0,45	0,02	577	<2e-16	0,30	2
st3	β_8	-0,75	0,04	507	<2e-16	0,00	0
	$adj.R^2$	99, 3%					

6.5 Skoring model 2 - konstrukcija skoring modela pomoću svake pojedinačne polise osiguranja

Za razliku od prethodnog skoring modela koji smo konstruisali na osnovu formiranih klasa osiguranika, ovaj skoring model ćemo kontruisati na osnovu svake pojedinačne polise iz podataka za konstrukciju modela. To znači da je uzorak, pomoću kojeg ćemo kontruisati model, obima 102 409. Razlika u odnosu na prethodno kontruisani skoring model je i broj kategorijalnih varijabli koje ćemo koristiti u konstrikciji modela. Naime, u ovom modelu ćemo iskoristiti sve kategorijalne varijable koje su dostupne u podacima, tj. iskoristićemo sve kategorijalne varijable date u tabeli 16. Skoring model je dat sa

$$s = S_0 + \sum_{j=1}^{9} \sum_{k=1}^{r_j} S_{jk} x_{jk}, \tag{15}$$

gde

- s predstavlja vrednost skora za i-tu polisu
- S_0 predstavlja vrednost početnog skora
- S_{jk} je vrednost skora kategorije A_{jk}, $j = 1, 2, ..., 9$, $k = 1, ..., r_j$
- x_{jk} predstavlja vrednost binarne promenljive A_{jk}-te kategorije, tačnije

  $$x_{jk} = \begin{cases} 1, & \text{osiguranik ima osobinu } A_{jk} \\ 0, & \text{inače} \end{cases}$$

Kao i u kontrukciji prethodnog modela i u ovom modelu ćemo vrednosti skorova dobiti uz pomoć gama linearnog modela sa logaritamskom link funkcijom.
Ako sa Y označimo slučajnu promenljivu koja predstavlja premiju osiguranja, onda naš model ima oblik

$$g(E(Y)) = g(\mu) = \ln \mu = \eta = X\boldsymbol{\beta}, \tag{16}$$

gde

(i) Y ima gama raspodelu sa očekivanjem μ i disperzijom $\sigma^2\mu^2$

(ii) vektor $\boldsymbol{\beta}$ je vektor koeficijenata modela, dok je X vektor vrednosti binarnih varijabli, tačnije

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{27} \end{bmatrix} \text{ i } X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{93} \end{bmatrix} .$$

Važno je napomenuti da vektor X sadrži samo binarne varijable za kategorije koje nisu referentni nivoi.

(iii) g je logaritamska link gunkcija, tj. $g(\mu) = \ln \mu$

Pomoću datog uzorka i softvera „R“, računamo koeficijente modela. Najbolji model, koji je dobijen stepenastom metodom unazad, dat je sa

$$\begin{aligned}\hat{\eta} = 10,11 &+ 0,26x_{11} - 0,25x_{21} - 0,12x_{22} + 0,08x_{24} + 0,04x_{33} + 0,04x_{41} + 0,11x_{43} + 0,04x_{51}\\ &- 1,21x_{61} - 0,89x_{62} - 0,7x_{63} - 0,49x_{64} - 0,26x_{65} + 0,25x_{67} + 0,47x_{68} + 0,69x_{69}\\ &+ 0,91x_{610} + 1,16x_{611} + 1,54x_{612} + 0,11x_{72} + 0,31x_{73} + 0,59x_{74} + 0,25x_{81}\\ &- 0,2x_{83} - 0,09x_{92} - 0,09x_{93}\end{aligned}$$

sa vrednostima $adj.R^2 = 87,5\%$ i AIC =2 097 460.
Dakle, stepenastom metodom unazad iz modela je uklonjena samo varijabla x_{31}, tj. varijabla koja odgovara kategoriji Z. Primetimo da varijable x_{92} i x_{93} (koje pripadaju istoj kategorijalnoj promenljivoj) imaju gotovo isti uticaj na zavisnu promenljivu. Zbog toga, uvešćemo u model novu binarnu varijablu definisanu sa

$$x_{i94} = \begin{cases} 1, & \text{ako je } x_{i92} = 1 \text{ ili } x_{i93} = 1 \\ 0, & \text{inače} \end{cases}.$$

Sa druge strane, primetimo da su koeficijenti koji stoje uz promenljive x_{33}, x_{41} i x_{51} zanemarljivi,

Tabela 21: Ocenjeni koeficijenti gama modela, njihove standardne greške i dobijene vrednosti skorova po kategorijama

Kategorija	**Koeficijent**	**Ocena**	**St. Greška**	**w-vrednost**	**p vrednost**	**Modifikovana ocena**	**Skor**
Odsečak	β_0	10,16	0,003	15.256.403	<0,001	8,37	60
d1	β_1	0,29	0,004	5.487	<0,001	0,29	2
d2						0,00	0
sn1	β_2	-0,25	0,004	4.242	<0,001	0,00	0
sn2	β_3	-0,12	0,002	2.640	<0,001	0,13	1
sn3						0,25	2
sn4	β_4	0,08	0,003	663	<0,001	0,32	2
st1,st2						0,00	0
st3	β_8	0,11	0,003	1.741	<0,001	0,11	1
A	β_{10}	-1,24	0,009	22.067	<0,001	0,00	0
B	β_{11}	-0,92	0,005	36.201	<0,001	0,32	2
C	β_{12}	-0,73	0,004	39.326	<0,001	0,52	4
D	β_{13}	-0,52	0,003	24.911	<0,001	0,73	5
E	β_{14}	-0,27	0,003	10.961	<0,001	0,97	7
F						1,24	9
G	β_{15}	0,26	0,003	10.540	<0,001	1,50	11
H	β_{16}	0,49	0,003	22.727	<0,001	1,74	12
I	β_{17}	0,72	0,004	39.526	<0,001	1,96	14
J	β_{18}	0,94	0,006	30.126	<0,001	2,18	16
K	β_{19}	1,20	0,006	46.586	<0,001	2,44	17
L	β_{20}	1,59	0,007	60.476	<0,001	2,84	20
z1						0,00	0
z2	β_{21}	0,11	0,002	2.922	<0,001	0,11	1
z3	β_{22}	0,30	0,003	10.087	<0,001	0,30	2
z4	β_{23}	0,57	0,005	11.698	<0,001	0,57	4
b1	β_{24}	0,26	0,002	21.560	<0,001	0,46	3
b2						0,20	1
b3	β_{25}	-0,20	0,002	9.292	<0,001	0,00	0
u1						0,10	1
u2,u3	β_{28}	-0,10	0,002	3.682	<0,001	0,00	0

tj. nemaju značajan uticaj na krajnju vrednost $\hat{\eta}$, a samim tim i ocenjene premije. Zbog toga, ove promenljive ćemo izuzeti iz kontrukcije gama modela. Tačnije ovim postupkom izbacivanja vari-

jabli, mi se oslobadjamo „nepotrebnih“ varijabli u modelu, tj. varijabli čiji koeficijenti neće imati konkretan doprinos u skoring modelu, jer će odgovarajući skorovi ovih koeficijenata biti jednaki 0. Sada ćemo još jednom oceniti koeficijente gama modela, ali ovog puta bez varijabli x_{92}, x_{93}, x_{33}, x_{41} i x_{51} i sa uključenom novom varijablom x_{94}. Jednačina ovog gama modela data je sa

$$\begin{aligned}\hat{\eta} = {} & 10,16 + 0,29x_{11} - 0,25x_{21} - 0,12x_{22} + 0,08x_{24} + 0,11x_{43} - 1,24x_{61} - 0,92x_{62} \\ & - 0,73x_{63} - 0,52x_{64} - 0,27x_{65} + 0,26x_{67} + 0,49x_{68} + 0,72x_{69} + 0,94x_{610} \\ & + 1,2x_{611} + 1,59x_{612} + 0,11x_{72} + 0,3x_{73} + 0,57x_{74} + 0,26x_{81} - 0,2x_{83} - 0,1x_{94}\end{aligned} \tag{17}$$

sa vrednostima $adj.R^2 = 87,25\%$ i $AIC = 2\,099\,050$. Postupkom izbacivanja suvišnih promenljvih i uvodjenjem nove promenljive, uočavamo minimalnu promenu u merama $adj.R^2$ i AIC . Tačnije $adj.R^2$ se smanjio za $0,25\%$, dok se AIC povećao za 1 590, pa ovaj postupak možemo smatrati opravdanim. Drugim rečima, ovim postupkom smo uspeli značajno da pojednostavimo model, zarad minimalnog smanjenja kvaliteta modela.
Parametri modela (17) su dati u tabeli 21. Ako posmatramo ocenjene vrednosti koeficijenata nezavisnih promenljivih, vidimo da osigurana suma vozila ima najveći uticaj na krajnju vrednost modela. Razlika krajnje vrednosti modela vozila iz L i A kategorije je približno 2,83 (uz fiksirane ostale varijable). Sa druge strane, vidimo da kategorijalne promenljive starost vozila i učešće imaju najmanji uticaj na krajnji rezultat modela.
Standardne greške svih ocenjenih koeficijenata modela su manje od 0,01, pa zaključujemo da su ocenjene vrednosti koeficijenata stabilne.
Dobijene vrednosti Valdovog testa i odgovarajućih p-vrednosti (četvrta i peta kolona iz tabele 21) nam pokazuju da su svi koeficijenti statistički značajni za model (na nivou značajnosti od 95%).
Poslednje dve kolone tabele 21 predstavljaju modifikovane vrednosti ocenjenih koeficijenata modela i konačne vrednosti skorova kategorija. U koloni mod. ocena date su vrednosti dobijene iz postupka 1, dok su vrednosti iz kolone skor dobijaju kada kolonu mod. ocena podelimo sa $\ln 1,15$. Poslednja kolona zapravo predstavlja vrednosti skorova za skoring model (15).

Analiza reziduala modela

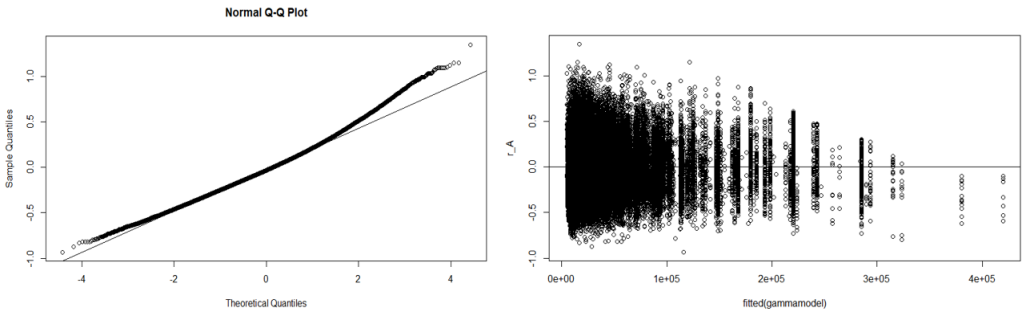
Pogledajmo šta se dešava sa rezidualima našeg modela. Za potrebe računanja reziduala gama skoring modela koristićemo Anskombe reziduale. Zbog rada sa uzorkom koji ima veliki obim, normalnost Anskombe reziduala testiraćemo grafičkim metodom. Za gama uopštene linearne modele, Anskombe reziduali su dati sa

$$(r_A)_i = 3\left(\left(\frac{y_i}{\hat{\mu}_i}\right)^{1/3} - 1\right), \quad i = 1, 2, ..., 102409\ .$$

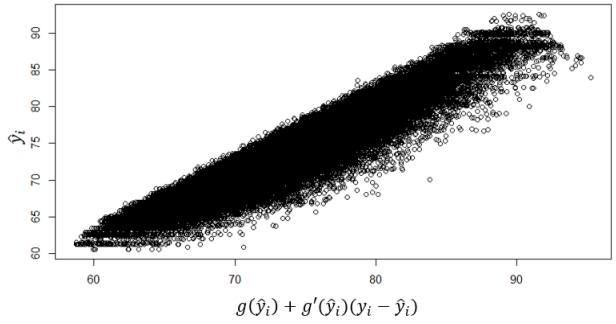
Na slici 10 vidimo da reziduali imaju približno normalnu raspodelu sa očekivanjem 0. Q-Q grafik nam pokazuje da grafik normalne raspodele blago iskrivljen udesno.

Test izbora link funkcije

Na početku konstrukcije skoring modela, pretpostavili smo da premija ima gama raspodelu, dok je za link funkciju uzeta funkcija ln. Opravdanost ovakvog izbora link funkcije leži u testiranju link funkcije, gde tačke $(g(\hat{y}_i) + g'(\hat{y}_i)(y_i - \hat{y}_i), \hat{y}_i)$, $i = 1, 2, ..., 102409$, približno leže na istoj pravoj. Na slici 11 vidimo da je opravdan izbor logaritamske link funkcije.



Slika 10: Grefički prikaz anskombe reziduala, njihova gustina i Q-Q grafik normalne raspodele



Slika 11: Grafički prikaz testa link funkcije

Korekcija skoring modela

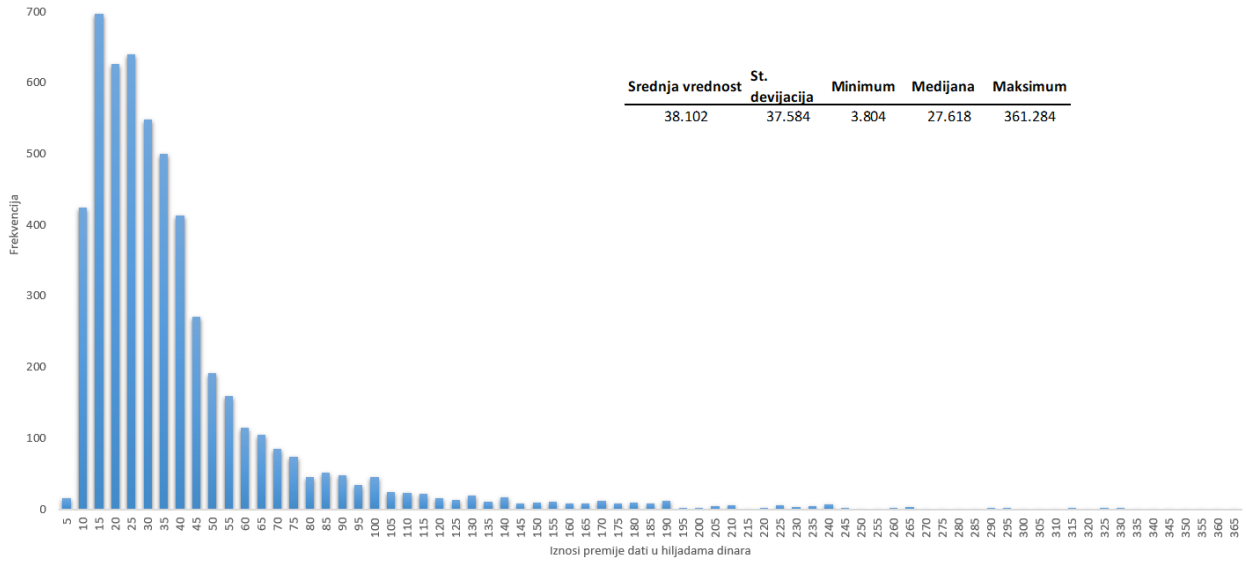
Odnos stvarne i fitovane premije dat je u tabeli 22. Iz tabele takodje vidimo da je ukupna premija manja od ukupne vrednosti fitovane premije. Iz tog razloga korigovaćemo model tako što ćemo fitovanu premiju pomožiti recipročnom vrednošću ukupnog premijskog racija, tj. brojem $\frac{1}{0,981}$. Na ovaj način ćemo izjedačiti ukupnu vrednost premije i ukupnu vrednost fitovane premije.

Tabela 22: Ukupna vrednost, razlika i racio prihoda od fitovane i stvarne premije

		Vrednost
Ukupna vrednost izloženosti	$\sum_{i=1}^{102409} e_i$	102 409
Ukupna vrednost fitovanih premija	$\sum_{i=1}^{102409} e_i \hat{y}_i$	3 864 980 772
Ukupna vrednost stvarnih premija	$\sum_{i=1}^{102409} e_i y_i$	3 939 090 348
Razlika ukupne vrednosti stvarne i fitovane premije	$\sum_{i=1}^{102409} e_i(\hat{y}_i - y_i)$	−74 109 575
Ukupan premijski racio	$\frac{\sum_{i=1}^{102409} e_i \hat{y}_i}{\sum_{i=1}^{102409} e_i y_i}$	0,981

6.6 Testiranje modela

Posmatrajmo skup podataka za testiranje modela. Skup podataka za testiranje modela sadrži 5 390 jednogodišnjih polisa kasko osiguranja. Frekvencija i deskriptivna statistika premije data je na slici 12.



Slika 12: Frekvencija i deskriptivna statistika premije iz skupa podataka za testiranje modela

Na ovom skupu podataka ćemo testirati dva konstruisana gama skoring modela[13]. U tabeli 23 dato je poredjenje ukupnih vrednosti premija kroz razliku stvarne i fitovane premije i kroz premijski racio, tj. količnik fitovane i stvarne premije. Dobijene vrednosti nam pokazuju da skoring model 2 mnogo bolje fituje ukupnu vrednost premije.

Tabela 23: Stvarne i fitovane vrednosti premije podataka za testiranje modela

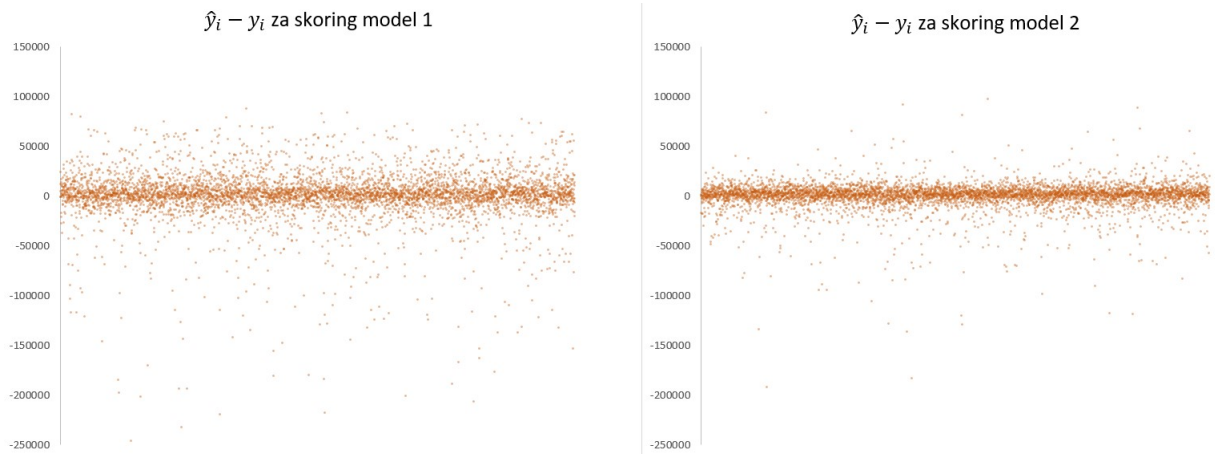
	Stvarna premija	Skoring model 1	Skoring model 2
Ukupna vrednost premije	205.369.048	206.591.487	205.147.075
Razlika ukupne vrednosti fitovane i stvarne premije		1.222.439	-221.973
Ukupan premijski racio		1,006	0,999

Pogledajmo sada kako skoring modeli fituju premiju svake polise ponaosob. Na slici 13 vidimo da su vrednosti $\hat{y}_i - y_i$ za skoring model 2 znatno bliže skoncentrisane oko 0, nego iste vrednosti za skoring model 1. Ovu tvrdnju dokazuje i činjenica da je zbir $\sum_{i=1}^{5930} |\hat{y}_i - y_i|$ kod skoring modela 1 jednak 76 587 617, dok je kod skoring modela 2 niži i jednak 40 877 660. To znači da skoring model 2 bolje fituje podatke i u slučaju ako posmatramo svaku pojedinačnu polisu.

Testiranje dva konstuisana skoring modela nam je pokazalo da oba modela veoma dobro fituju ukupnu vrednost premije. Primetimo da kao i pri konstrukciji modela, poredeći ih sa ukupnom premijom, skoring model 1 daje nešto veće vrednosti ukupne premije, dok je ukupna vrednost premije skoring modela 2 nešto niža. Sa druge strane, ako gledamo svaku pojedinačnu polisu, tu je moć

[13] koristimo korigovane skoring modele

skoring modela 2 mnogo bolja, što se i moglo očekivati, uzevši u obzir to da u skoring modelu 2 učestvuje veći broj kategorijalnih promenljivih.



Slika 13: Razlika fitovane i stvarne vrednosti premije za svaku polisu

7 Zaključak

Matematički modeli imaju veliku primenu u poslovanju osiguravajućih kuća, pre svega zbog ogromne količine podataka koje one poseduju. Razvojem brzih statističkih softvera koji postaju sve pristupačniji, modeli imaju veliki značaj u donošenju brzih i efektivnih odluka menadžmenta osiguravajućeg društva. Skoring modeli su jedna vrsta matematičkih modela koji imaju široku primenu u različitim oblastima. Najpoznatiji su kreditni skoring modeli, koji, izmedju ostalog, imaju ulogu u definisanju „loših“ klijenata, tačnije klijenata za koje se smatra da imaju malu verovatnoću da neće imati kašnjenja u otplati kredita banci. Sa druge strane, u industriji osiguranja skoring modeli imaju široku primenu kako u procenama potraživanja osiguranika, tako i u detekciji sumnjivih potraživanja.

Ovaj rad se bavi primenom skoring modela u oblasti računanja i analize premije kasko osiguranja. Glavna primena skoring modela je odredjivanje profitabilnosti odredjenih kategorija osiguranika, medjutim skoring modeli se mogu primeniti i kao jednostavni kalkulatori, gde osiguranici mogu proceniti svoje individualne rizike. Kako se premija kasko osiguranja može modelirati lognormalnom, gama ili na primer, normalnom raspodelom, najpogodniji skoring modeli za modeliranje premije su oni koji su bazirani na uopštenim linearnim modelima. Prvi deo rada posvećen je teorijskoj osnovi uopštenih linearnih modela. Tu je definisana eksponencijalna familija raspodela, zatim varijansna funkcija, kao i statistike za procenu adekvatnosti i dijagnostike modela.

Kroz primere modela, koji su dati u studiji [12], zaključujemo da su skoring modeli, koji su bazirani na uopštenim linearnim modelima, pogodni za modeliranje premije kasko osiguranja, ako se odabere pogodna raspodela. U predstavljenim modelima za raspodelu zavisne promenljive, tj. premije kasko osiguranja uzete su lognormalna, gama i normalna raspodela. Na osnovu dobijenih skorova u modelima, može se proceniti rizičnost odredjene grupe osiguranika, a time i njihova profitabilnost. Šesti odeljak ovog rada posvećen je istraživanju mogućnosti primene skoring modela u Republici Srbiji. Primena skoring modela u našoj zemlji je još uvek u teorijskoj fazi, gde se osiguravajuće kuće retko odlučuju na primenu ovakvih modela. Zato je istraživanje podrazumevalo konstrukciju gama skoring modela na podacima jedna domaće osiguravajuće kuće. Izvodjenje zaključaka u konstrukciji modela zasnovano je na dobijenoj statističkoj analizi koristeći program „R“. Prvi gama skoring model, koji je konstruisan pomoću definisanih klasa osiguranika, konstruisan je na osnovu prethodne detaljne analize teorijskog dela i studije [12]. Cilj konstrukcije ovog modela bio je da se napravi paralela u odnosu na skoring modele date u studiji. Primećeno je da najveći uticaj na krajnju vrednost skora imaju snaga motora i starost vozila , dok pol osiguranika i doplate nemaju značajan uticaj. Sa druge strane, region, čiji su skorovi u malezijskim podacima pokazali da stanovnici ostrva Borneo plaćaju znatno manju premiju, kod nas nema značajnijih razlika medju osiguranicima iz Beograda i onima iz ostatka Srbije.

Zatim se prešlo na konstrukcjiu drugog gama skoring modela koji je sadržao sve dostupne kategorijalne promenljive iz podataka za konstrukciju i koji je konstruisan na osnovu svake polise ponaosob. Cilj ove konstrukcije je bio da se model uporedi prethodno konstruisanim gama skoring modelom i da se uvide odnosi medju kategorijama svake kategorijalne promenljive. Na osnovu dobijenih skorova kategorija, zaključak je da osigurana suma vozila ima najveći uticaj na premiju kasko osiguranja, dok region ima najmanji uticaj na konačnu premiju kasko osiguranja. Zapravo, možemo zaključiti da vrednost osigurane sume najdirektnije odredjuje vrednost premije kasko osiguranja u Srbiji, dok ostale kategorijalne promenljive imaju samo korektivnu ulogu.

Validacija i testiranje dva konstruisana gama skoring modela podrazumevala je testiranje modela na skupu podataka za testiranje modela. Testiranje je pokazalo da drugi gama skoring model bolje fituje kako ukupnu vrednost premije, tako i svaku polisu zasebno. Medjutim, prednost modela konstruisanog na osnovu klasa osiguranika je, pored manjeg broja prediktora i ta da za ovaj model nije

potrebno znati kolika je zaključena premija za svaku polisu, vec je dovoljno znati srednju vrednost premije za svaku klasu osiguranika, što često može znatno da uprosti proces modeliranja.
Konačno, zaključujemo da se skoring modeli mogu primeniti u našoj zemlji.

8 Prilog

8.1 Osobine ocena napoznatih parametara

Neka je $(Y_1, Y_2, ..., Y_n)$ prost slučajni uzorak obima n, $n \in \mathbb{N}$ za obeležje Y i neka je $\theta \in \Omega \subset \mathbb{R}^{\geq}$, $m \in \mathbb{N}$ nepoznati parametar u raspodeli tog obeležja.

Definicija 8.1. *Statistika $\hat{\theta} \in \mathbb{R}^{\geq}$ je nepristrasna ocena za parametar θ ako je $E(\hat{\theta}) = \theta$, a asimptotski nepristrasna ocena ako važi $E(\hat{\theta}) \to \theta$, kada $n \to \infty$.*

Definicija 8.2. *Statistika $\hat{\theta}$ je konzistentna ocena parametra θ ako $\hat{\theta}$ konvergira u verovatnoći ka θ, odnosno ako za svako $\varepsilon > 0$ važi*

$$P\{|\hat{\theta} - \theta| \geq \varepsilon\} \to 0, \text{ kada } n \to \infty \ .$$

Konzustentnost možemo ispitati preko srednje kvadratne greške u oznaci MSE[14]

$$\lim_{n\to\infty} MSE(\hat{\theta}) = \lim_{n\to\infty} \left(Var(\hat{\theta}) + (\theta - E(\hat{\theta}))^2 \right) \ .$$

Definicija 8.3. *Statistika $\hat{\theta} \in \mathbb{R}^{\geq}$ je stabilna ocena nepoznatog parametra θ ako je ona nepristrasna i konzistentna.*

Definicija 8.4. *Ocena $\hat{\theta}_1$ je efikasnija od ocene $\hat{\theta}_2$ nepoznatog parametra θ ako je $D(\hat{\theta}_1) < D(\hat{\theta}_2)$.*

Napomena 8.1. *Donju granicu disperzije nepristrasnih ocena daje* ***Rao-Kramerova nejednakost***

$$D(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial \log \varphi(y;\theta)}{\partial \theta}\right)^2\right]},$$

gde je sa φ označena gustina slučajne promenljive Y. Ocena čija je disperzija jednaka toj donjoj granici naziva se ***efikasna ocena****.*

8.2 Normalna raspodela

Neka je data normalna slučajna promenljiva Y sa očekivanjem μ ($\mu \in \mathbb{R}$) i disperzijom σ^2 ($\sigma^2 > 0$) ili matematički $Y : \mathcal{N}(\mu, \sigma^2)$. Tada je gustina ovakve slučajne promenljive data sa

$$\varphi_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \ .$$

Parametar μ (parametar lokacije) predstavlja sredinu slučajne promenljive Y, dok σ^2 (parametar skaliranja) predstavlja njenu varijansu.

Teorema 8.1. *Neka slučajne primenljive Y_i, $i = 1, 2, ..., n$ imaju normlane $\mathcal{N}(\mu_i, \sigma_i^2)$ raspodele i neka su $\alpha_i \in \mathbb{R}$. Tada slučajna promenljiva $Y = \sum_{i=1}^n \alpha_i X_i$ ima normalnu $\mathcal{N}(\mu, \sigma^2)$ raspodelu, gde je*

$$\mu = \sum_{i=1}^n \alpha_i \mu_i \quad i \quad \sigma^2 = \sum_{i=1}^n (\alpha_i \sigma_i)^2 \ .$$

[14] eng. **M**ean **S**tandard **E**rror

Veze sa drugim raspodelama

Neka su $Y_1, Y_2, ..., Y_n : \mathcal{N}(\mu, \sigma^2)$ nezavisne slučajne promenljive i neka je $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Tada važi

(1) $\frac{\sqrt{n}(\overline{Y}-\mu)}{\sigma} : \mathcal{N}(0,1)$

(2) $\frac{\sqrt{n}(\overline{Y}-\mu)}{\widetilde{\sigma}} : t_{n-1}$, gde smo sa t_{n-1} obeležili Studentovu raspodelu sa $n-1$ stepeni slobode

Za nezavisne slučajne promenljive $Z_1, Z_2, ..., Z_n : \mathcal{N}(0,1)$ važi

$$Z_1^2 + Z_2^2 + ... + Z_n^2 : \chi_n^2,$$

gde smo sa χ_n^2 obeležili hi-kvadratnu raspodelu sa n stepeni slobode.
Za nezavisne slučajne promenljive $W_1 : \chi_n^2$ i $W_2 : \chi_m^2$ važi

$$\frac{W_1/n}{W_2/m} : F_{n,m},$$

gde je sa $F_{n,m}$ data Fišerova rapsodela sa n i m stepeni slobode.

Šapiro-Vilkov test normalnosti uzorka

Šapiro-Vilkov[15] test normalnosti uzorka koristi se za uzorke obima od 12 do 5000. Neka je dat uzorak $y_1, ..., y_n \in \mathbb{R}$, pri čemu je $y_1 \leq y_2 \leq ... \leq y_n$. Šapiro-Vilkova statistika data je sa

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

gde je $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ i gde su a_i konstante koje su odredjene formulom

$$(a_1, a_2, ..., a_n) = \frac{\boldsymbol{m}^T V^{-1}}{\left(\boldsymbol{m}^T V^{-1} V^{-1} \boldsymbol{m}\right)^{1/2}},$$

pri čemu je $\boldsymbol{m} = \begin{bmatrix} m_1 & m_2 & \cdots m_n \end{bmatrix}^T$, gde m_i predstavlja očekivanu vrednost i-te uredjene statistike iz uzorka koji potiče iz standardne normalne raspodele, dok sa druge strane, V predstavlja kovarijacionu matricu uredjenih statistika.
Ako je nulta hipoteza tačna, tj. uzorak ima normalnu raspodelu, onda statistika $\ln(1-W)$ ima približno normalnu raspodelu.
Ovaj test je u softveru „R“ dat komandom shapiro.test().

Q-Q grafik

Ako želimo da ispitamo normalnost uzorka velikog obima, poželjno je koristiti grafičku metodu. Jedan od grafičkih prikaza koji nam približno može odgovoriti na pitanje da li posmatrani uzorak ima željenu raspodelu (u ovom slučaju normalnu), jeste tzv., Q-Q grafik.
Q-Q grafik je grafik koji poredi sortiran uzorak sa teorijskim uzorkom, u ovom slučaju, normalne raspodele. Ukoliko te raspodele odgovaraju jedna drugoj, tačke na grafiku će praktično biti na jednoj pravoj. U slučaju malog odstupanja, intuitivno se može smatrati da posmatrani uzorak ima normalnu raspodelu. Primeri Q-Q grafika dati su na slikama 8 i 10.

[15] eng. Shapiro-Wilk

8.3 Lognormalna raspodela

Neka je data slučajna promenljiva Y za koju važi $Y : \mathcal{N}(\mu, \sigma^2)$, $\sigma^2 > 0$. Tada slučajna promenljiva $\widetilde{Y} = \exp[Y]$ ima lognormalnu raspodelu sa parametrima μ i σ^2 i matematički se zapisuje $\widetilde{Y} : \mathcal{LN}(\mu, \sigma^2)$
Gustina slučajne promenljive $\widetilde{Y}$ data je sa

$$\varphi_{\widetilde{Y}}(\exp[y]) = \frac{1}{\exp[y]\,\sigma\sqrt{2\pi}} - \exp\left[\frac{(y-\mu)^2}{2\sigma^2}\right],\ y \in \mathbb{R}\,.$$

Očekivanje slučajne promenljive $\widetilde{Y}$ dato je sa

$$E(\widetilde{Y}) = \exp\left[(\mu + \frac{\sigma^2}{2})\right],$$

dok je njena disperzija data sa

$$D(\widetilde{Y}) = \exp\left[(2\mu + \sigma^2)\right](\exp\left[\sigma^2\right] - 1)\,.$$

8.4 Gama raspodela

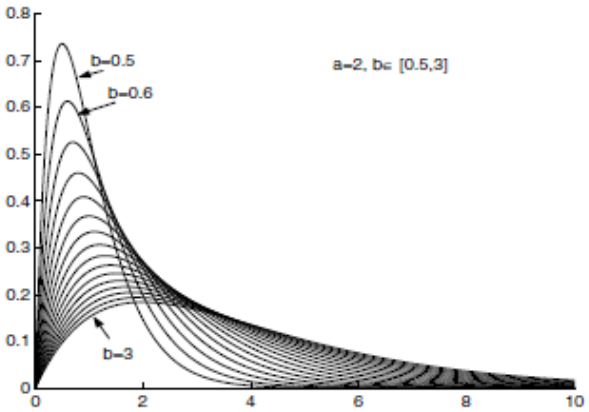
Osnovna dvoparametarska Gama raspodela data je gustinom

$$\varphi(y) = \frac{1}{y\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu} \exp\left[-\frac{\nu y}{\mu}\right], y \geq 0, \nu > 0, \frac{\mu}{\nu} > 0,$$

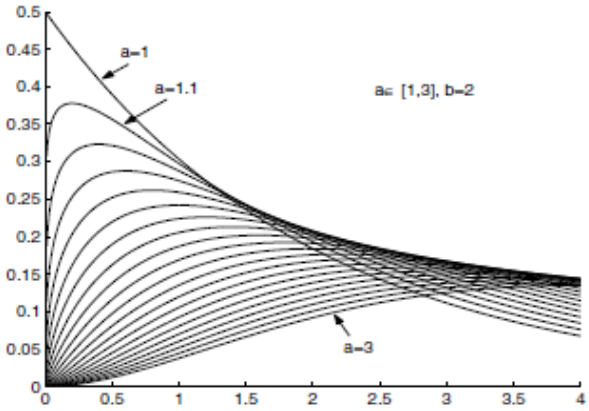
gde je $\Gamma(a)$ gama funkcija defeinisana sa $\Gamma(a) = \int_0^\infty \exp[-t]\, t^{a-1} dt$.
Ako slučajna promenljiva Y ima gama raspodelu, onda je

$$E(Y) = \mu, \quad Var(Y) = \frac{\mu^2}{\nu}\,.$$

Na slikama 14 i 15 prikazano je ponašanje gustine Gama raspodele u zavisnosti od promene parametara $a = \nu$ i $b = \frac{\mu}{\nu}$.
Funkcija dvoparametarske Gama raspodele data je sa



Slika 14: Prikaz ponašanja gustine Gama raspodele za fiksiran parametar b



Slika 15: prikaz ponašanja gustine Gama raspodele za fiksiran parametar a

$$F(y) = \frac{1}{\Gamma(a)}\gamma(a, \frac{y}{b}), y \geq 0,$$

gde je $\gamma(a, y) = \int_0^y t^{a-1}e^{-t}dt$ nepotpuna gama funkcija.

Tabela 24: Podaci za konstrukciju gama skoring modela po klasama osiguranika

klasa	**obim polise**	**snaga motora**	**vlasnik polise**	**starost vozila**	**region**	**sr. vred. premije** (μ_i)	**izloženost** (e_i)
K1	d1	sn1	Z	st1	BG	23.653	1
K2					OST	22.993	5
K3				st2	BG	15.180	5
K4					OST	17.902	18
K5				st3	BG	11.464	10
K6					OST	10.275	5
K7			M	st1	BG	0	0
K8					OST	18.349	5
K9				st2	BG	13.801	2
K10					OST	15.063	12
K11				st3	BG	12.531	19
K12					OST	15.111	7
K13			PL	st1	BG	29.119	73
K14					OST	30.620	8
K15				st2	BG	19.565	42
K16					OST	19.333	27
K17				st3	BG	14.454	7
K18					OST	12.089	15
K19		sn2	Z	st1	BG	31.417	23
K20					OST	30.044	37
K21				st2	BG	18.061	27
K22					OST	19.224	16
K23				st3	BG	15.572	28
K24					OST	12.341	20
K25			M	st1	BG	37.517	31
K26					OST	30.453	45
K27				st2	BG	21.763	36
K28					OST	19.740	29
K29				st3	BG	14.524	49
K30					OST	16.551	31
K31			PL	st1	BG	34.584	496
K32					OST	33.558	121
K33				st2	BG	25.546	180
K34					OST	22.313	97
K35				st3	BG	14.594	46
K36					OST	18.630	48
K37		sn3	Z	st1	BG	51.409	25
K38					OST	39.020	33
K39				st2	BG	32.605	15
K40					OST	36.360	13
K41				st3	BG	31.714	11
K42					OST	28.016	20
K43			M	st1	BG	51.909	69
K44					OST	41.462	71
K45				st2	BG	43.455	76
K46					OST	35.733	40
K47				st3	BG	27.267	58
K48					OST	24.948	40
K49			PL	st1	BG	61.417	548

K50					OST	49.021	115
K51				st2	BG	39.601	186
K52					OST	39.938	101
K53				st3	BG	36.628	29
K54					OST	26.890	82
K55		sn4	Z	st1	BG	173.180	11
K56					OST	123.468	11
K57				st2	BG	68.892	7
K58					OST	93.272	7
K59				st3	BG	29.661	5
K60					OST	65.311	2
K61			M	st1	BG	128.949	30
K62					OST	137.420	53
K63				st2	BG	74.812	34
K64					OST	76.324	41
K65				st3	BG	44.658	26
K66					OST	62.349	12
K67			PL	st1	BG	149.171	256
K68					OST	161.096	167
K69				st2	BG	68.261	107
K70					OST	93.116	54
K71				st3	BG	67.416	20
K72					OST	57.245	28
K73	d2	sn1	Z	st1	BG	21.458	64
K74					OST	19.003	99
K75				st2	BG	12.381	345
K76					OST	11.755	694
K77				st3	BG	9.311	481
K78					OST	9.184	939
K79			M	st1	BG	18.677	91
K80					OST	17.862	105
K81				st2	BG	12.299	520
K82					OST	11.055	846
K83				st3	BG	9.689	532
K84					OST	8.878	1085
K85			PL	st1	BG	17.470	156
K86					OST	18.801	192
K87				st2	BG	12.114	277
K88					OST	12.768	453
K89				st3	BG	10.195	266
K90					OST	9.245	440
K91		sn2	Z	st1	BG	28.006	1193
K92					OST	25.638	1207
K93				st2	BG	16.902	1324
K94					OST	15.590	1423
K95				st3	BG	12.091	1184
K96					OST	12.204	1412
K97			M	st1	BG	27.272	1293
K98					OST	25.156	1567
K99				st2	BG	16.793	1775
K100					OST	15.797	2104
K101				st3	BG	12.205	1734
K102					OST	12.146	2348
K103			PL	st1	BG	27.982	2245

K104					OST	26.536	1879
K105				st2	BG	18.379	1015
K106					OST	17.612	1248
K107				st3	BG	14.441	384
K108					OST	12.962	872
K109		sn3	Z	st1	BG	38.926	2138
K110					OST	36.391	1899
K111				st2	BG	26.499	1324
K112					OST	25.363	1113
K113				st3	BG	19.399	689
K114					OST	20.396	700
K115			M	st1	BG	38.732	4788
K116					OST	37.249	4880
K117				st2	BG	28.017	3248
K118					OST	27.015	3631
K119				st3	BG	21.966	1889
K120					OST	20.947	2843
K121			PL	st1	BG	40.980	5743
K122					OST	42.932	3552
K123				st2	BG	31.031	1465
K124					OST	31.569	1848
K125				st3	BG	24.984	598
K126					OST	22.606	1018
K127		sn4	Z	st1	BG	107.573	463
K128					OST	78.192	417
K129				st2	BG	55.129	369
K130					OST	51.153	356
K131				st3	BG	41.857	180
K132					OST	38.051	179
K133			M	st1	BG	96.326	1742
K134					OST	87.466	1964
K135				st2	BG	56.197	1766
K136					OST	53.961	1798
K137				st3	BG	40.236	896
K138					OST	39.742	954
K139			PL	st1	BG	111.269	3498
K140					OST	110.708	3072
K141				st2	BG	74.670	1124
K142					OST	67.644	1520
K143				st3	BG	53.159	419
K144					OST	44.826	510

Literatura

[1] Avdalović, V. (2007) Osiguranje, *Beogradska Bankarska Akademija*, Beograd

[2] Marović, B., Avdalović, V. (2003) Osiguranje i upravljanje rizikom, *Biografika*, Subotica

[3] Zakon o osiguranju, *Službeni glasnik RS*, br. 139/2014 i 44/2021

[4] Goldburd, M., Khare, A., Tevet, D., Guller, D. (2019) Generalized Linear Models For Insurance Rating, *Casualty Acturial Society*, Arlington, Virginia, Second Edition

[5] Dobson, A., J. (2002) An Introduction to Generalized Linear Models, *Chapman & Hall/CRC*, Second Edition

[6] Frees W. E. (2010) Regression Modelling with Acturial and Financial Application, *Cambridge University Press*

[7] Lužanin, Z. (2018) Beleške sa predavanja iz ekonometrije *Prirodno-matematički fakultet u Novom Sadu*, Novi Sad

[8] Rajter-Ćirić, D. (2009) Verovatnoća *Prirodno-matematički fakultet u Novom Sadu*, Novi Sad, Drugo dopunjeno izdanje

[9] de Jong, P., Heller, G.Z. (2008) Generalized Linear Models for Insurance Data. *Cambridge University Press*, New York

[10] Pešta, M., Petrová, B., Procházka, J., Smolárová, T., Zimmermann, P., (2016) EXERCISES FOR NON-LIFE INSURANCE. *University of Economics, Charles University*, Prague, 1-27

[11] Moon, H., Pu, Y., Ceglia, C., (2019) A Predictive Modeling for Detecting Fraudlent Automobile Insurance Claims. *Theoretical Economics Letters*, **9**, 1886-1900

[12] Ismail, N., Jemain A. A. (2008) Construction of Insurance Scoring System using Regression Models. *Journal of Modern Applied Statistical Methods*, Vol. 7 : Iss. 2, Article 25

[13] Vučenović, T. (2016) Uopšteni linearni modeli sa primenama u aktuarstvu. *Prirodno-matematički fakultet u Novom Sadu*, Novi Sad

[14] Olhsson, E., Johansson, B. (2010). Non-Life Insurance Pricing with Generalized Linear Models, *Springer*

[15] https://online.stat.psu.edu/stat501/lesson/8 (12.04.2022)

Kratka biografija autora

Albert Koložvari je rodjen 10. septembra 1993. godine u Zrenjaninu. Osnovnu školu „Miloš Crnjanski¨ u Srpskom Itebeju je završio 2008. godine. Iste godine upisao je prirodno-matematički Zrenjaninske Gimnazije. Nakon odbranjenog maturskog rada i završene srednje škole 2012. godine, upisao je Prirodno-matematički fakultet u Novom Sadu, smer diplomirani profesor matematike. Osnovne studije završio je 13. jula 2017. godine. U julu 2013. godine, na istom fakultetu, položio je prijemni ispit i upisao se na master akademske studije, smer primenjena matematika, modul matematika finansija. Položio je sve ispite na master studijama koji su predvidjeni planom i programom zaključno sa septembrom 2020. godine i time stekao uslov za odbranu master rada.

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: Albert Koložvari
AU
Mentor: dr Zorana Lužanin
MN
Naslov rada: Primena skoring modela u izračunavanju premije kasko osiguranja
NR
Jezik publikacije: srpski(latinica)
JP
Jezik izvoda: srpski/engleski
JI
Zemlja publikovanja: Srbija
ZP
Uže geografsko područje: Vojvodina
UGP
Godina: 2022
GO
Izdavač: Autorski reprint
IZ
Mesto i adresa: Novi Sad, Departman za matematiku i informatiku, Prirodno-matematički fakultet, Trg Dositeja Obradovića 4
MA
Fizički opis rada: 7/70/0/24/15/0/1
(broj poglavlja/strana/lit. citata/tabela/slika/grafika/priloga)
FO
Naučna oblast: Matematika
NO
Naučna disciplina: Aktuarska matematika
ND
Predmetna odrednica/Ključne reči: Osiguranje, eksponencijalna familija raspodela, uopšteni linearni modeli, skoring modeli za računanje premije kasko osiguranja
PO
UDK:

Čuva se: Biblioteka Departmana za matematiku i informatiku Prirodno-matematičkog fakulteta u Novom Sadu
ČU
Važna napomena:
VN
Izvod: Cilj rada je analiza mogućnosti upotrebe skoring modela u Republici Srbiji. Na početku rada, predstavljeni su osnovni pojmovi osiguranja. Zatim, prikazana je teorijska osnova uopštenog linearnig modeliranja. Kroz primer, prikazana je način implementacije kategorijalnih promenljivih u regresionim modelima. Potom je predstavljena teorijska osnova tri skoring modela koja će se koristiti u radu. Primere primene tih modela dati su kroz publikaciju na podacima jedne Malezijske osiguravajuće kuće. Na kraju rada, testirana je mogućnost primene skoring modela na podacima jedne domaće osiguravajuće kuće, kroz konstukciju nekoliko skoring modela. Svi rezultati, dobijeni su uz pomoć statističkog softvera „R“ i programa „Excel“.
IZ
Datum prihvatanja teme od strane NN veća: 12.1.2022.
DP
Datum odbrane:
DO
Članovi komisije:

Predsednik: dr Andreja Tepavčević, redovni profesor
Mentor: dr Zorana Lužanin, redovni profesor
Član: dr Sanja Rapajić, redovni profesor

KO

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE AND MATHEMATICS
KEY WORD DOCUMENTATION

Accesion number:
ANO
Identification number:
INO
Document type: Monograph type
DT
Type of record: Printed text
TR Contents code: Master's thesis
CC
Author: Albert Kološvari
AU
Mentor: Zorana Lužanin, PhD
MN
Title: Aplication of scoring system in calculating casco insurance premium
TI
Language of text: Serbian(latin)
LT
Language of abstract: Serbian/English
LA
Country of publication: Serbia
CP
Locality of publication: Vojvodina
LP
Publication year: 2022
PY
Publisher: Author's reprint
PU
Publication place: Novi Sad, Department of Mathematics and Informatics, Faculty of Science and Mathematics, University of Novi Sad, Trg Dositeja Obradovića 4
PP
Physical description: 7/70/0/24/15/0/1
(chapters/pages/literature/tables/pictures/graphics/appendices)
PD
Scientific field: Mathematics
SF
Scientific discipline: Actirial science
SD
Subject/Key words: Insurance, exponential distribution family, general linear models, scoring models for calculating casco insurance premiums
SKW
UC:
Holding data: The Library of the Department of Mathematics and Informatics, Faculty of Science

and Mathematics, University of Novi Sad
HD
Note:
N
Abstract: The purpose of this thesis is to look at the possibility of using scoring models in Serbia. At the beginning of the paper, the basic concepts of insurance are presented. Then, the theoretical basis of general linear modeling is presented. The use of categorical variables in regression models is demonstrated using an example. Then, the theoretical basis of three scoring models that will be used in the paper is presented. Examples of the application of these models are given through a publication on the data of a Malaysian insurance company. The ability to apply scoring models to the data of one domestic insurance business was examined at the end of the article through the building of several scoring models. All results were obtained with the help of statistical software "R"and the program "Excel".
AB
Accepted by Scientific Board on: 12.1.2022.
ASB
Defended:
DE
Thesis defend board:

President: Andreja Tepavčević, PhD, full professor, Faculty of Science and Mathematics, University of Novi Sad
Mentor: Zorana Lužanin, PhD, full professor, Faculty of Science and Mathematics, University of Novi Sad
Member: Sanja Rapajić, PhD, full professor, Faculty of Science and Mathematics, University of Novi Sad

DB