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

Tatjana Vučenović

Uopšteni linearni modeli sa primenama u aktuarstvu

- Master rad-

Mentor:
dr Dora Seleši

Novi Sad, 2016.

Sadržaj

1 Uvod

Tema master rada su *uopšteni linearni modeli* i njihove primene u polju *aktuarstva*. Aktuarska matematika je grana primenjene matematike koja izučava i obrađuje matematičke osnove osiguranja. Osiguranje označava sigurnost i njegova svrha je da obezbedi zaštitu od rizika. Stručnjaci koji se bave izračunavanjem premija u osiguranju nazivaju se aktuarima. Termin ,,aktuar" se koristi za opisivanje lica koje prikuplja i analizira statističke podatke sa ciljem ocenjivanja rizika i izračunavanja premije osiguranja.
U današnje vreme, aktuar je stručnjak koji poseduje znanja iz opšte matematike, finansijske matematike, matematičke statistike, osiguranja, stohastičke analize, finansija i koji koristi svoja znanja za kvantitativnu analizu u osiguranju.
Drugim rečima, aktuarska matematika je matematika osiguranja.
Uopšteni linearni model je kao što i sam naziv kaže, uopštenje linearnog modela. To je jedna velika klasa statističkih modela za povezivanje zavisne promenljive sa linearnom kombinacijom nezavisnih promenljivih. Da bi se razumela struktura ulm[1] od koristi je osvrnuti se i razmotriti klasičan linearan model.
Svrha uopštenog linearnog modela je da se izrazi veza između posmatrane zavisne promenljive Y i nezavisnih promenljivih X koje zovemo prediktori. Zavisna promenljiva Y pripada takozvanoj eksponencijalnoj familiji raspodela, koja je jedna šira klasa raspodela u koju spadaju normalna, gama, Poasonova i mnoge druge poznate raspodele. Varijansa promenljive Y ne mora da bude konstantna kao kod klasičnog linearnog modela, već je dopustivo da varira sa promenama srednje vrednosti .

U drugom odeljku ovog rada opisuju se osnovni pojmovi uopštenog linearnog modela i formuliše se matematički model. Definiše se eksponencijalna familija raspodela, metode za ocene nepoznatih parametara modela, zatim statistike za procenu adekvatnosti modela tj. fitovanja i još neke dijagnostike modela.

Treći odeljak rada se bazira na modeliranju neprekidnih promenljivih koje se javljaju u primeni tj. aktuarstvu. Često su to iznosi zahteva za odštetu na polisi osiguranja. Varijable koje su pogodne za modeliranje takvih podataka su gama raspodela i inverzna Gausova. Primeri su obrađeni u statističkom paketu SPSS, analizirani su rezultati i izvedeni zaključci.

Četvrti odeljak je modeliranje kategorijalnih podataka pomoću uopštenog linearnog modeliranja. Ukratko su obrađene nominalna i ordinalna regresija, kao i logistička regresija kao specijalan slučaj kada ishodna promenljiva prima samo dve vrednosti. Takođe su dati i primeri.

Peti odeljak govori o Poasonovoj regresiji koja je najpogodnija raspodela za modeliranje podataka koji su celi nenegativni brojevi tj. prirodni brojevi. Kako su kod ove raspodele matematičko očekivanje i disperzija teorijski jednake, često se u praksi dešava da to nije slučaj već da je varijansa uzorka mnogo veća od očekivanja (overdispersion).

[1] uopšteni linearni model

U tom slučaju Poasonova raspodela nije prikladna za modeliranje, a problem se prevazilazi koristeći negativnu binomnu raspodelu.

Poslednji deo rada je kratak teorijski osvrt na naprednije modele koji predstavljaju proširenja uopštenog linearnog modela.
Nadalje, oznaka za transponovanje biće $'$ a za izvod $\cdot$ kao npr. $\dot{a},\ \ddot{a},\ \dot{g}$.

2 Uopšteno linearno modeliranje

2.1 Osnovni pojmovi i postavka modela

Uopšteno linearno modeliranje podrazumeva uspostavljanje veze između dve ili više promenljivih. Promenljiva koju treba objasniti naziva se zavisna ili ishodna promenljiva, a promenljive koje je objašnjavaju zovemo nezavisne promenljive.[2]
Linearna vrsta veze u kome zavisna promenljiva ima normalnu raspodelu i važi

$$E(Y_i) = \mu_i = x_i'\beta \text{ i } Y_i \sim N(\mu_i, \sigma^2)$$

naziva se klasičan ili normalan linearan model. Pritom su Y_i, $i = 1, 2, \ldots, n$ međusobno nezavisne slučajne promenljive, β je vektor regresionih koeficijenata tj.
$\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ i $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip-1})'$, a $x_i'\beta$ je skalarni proizvod tih vektora. Ovaj model je bazičan za analizu podataka koji imaju neprekidnu raspodelu.
Ulm[3] se razlikuje od klasičnog linearnog modela po tome što

- Raspodela zavisne promenljive ne mora da bude normalna (Gausova), i pripada eksponencijalnoj familiji raspodela
- transformacija srednje vrednosti zavisne promenljive linearno je povezana sa nezavisnim promenljivama.

Model je definisan preko sledeće dve jednačine

$$f(y) = c(y, \phi)exp\left\{\frac{y\theta - a(\theta)}{\phi}\right\} \tag{1}$$

$$g(\mu) = x'\beta \tag{2}$$

Jednačina (1) govori nam da je raspodela zavisne iz eksponencijalne familije[4], a jednačina (2) ukazuje na linearnu vezu između transformacije srednje vrednosti i nezavisnih promenljivih.
Parametar ϕ se naziva disperzioni parametar a parametar θ kanonički.
Funkcija g se naziva **link** funkcija i treba da bude monotona i diferencijabilna.
U sledećoj tabeli su prikazane link funkcije koje se najčešće koriste u modeliranju.

[2]U literaturi se za nezavisne promenljive koriste još nazivi: prediktori, faktori rizika, kovarijable
[3]Uopšten linearan model
[4]eksponencijalna familija raspodela biće objašnjena u ovom odeljku

Link funkcija	$g(\mu)$
identička	μ
logaritam	$ln(\mu)$
stepena	μ^2
kvadratni koren	$\sqrt{\mu}$
logit	$ln\frac{\mu}{1-\mu}$

Ako važi da je $g(\mu) = \theta$ tj $g = (\dot{a})^{-1}$, tada je link g kanonički. Onda je

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

Izbor funkcija $a(\theta)$ i $c(y, \phi)$ određuje koja konkretna raspodela je u pitanju. Dakle, uopšteno linearno modeliranje karakterišu sledeće komponente :

- Zavisne varijable $Y_1, Y_2, \ldots, Y_n$ imaju istu raspodelu verovatnoća koja pripada eksponencijalnoj familiji.
- x je vektor vrednosti nezavisnih promenljivih, odgovarajuće dimenzije a β je vektor nepoznatih parametara modela .
- g je monotona link funkcija tako da je

  $$g(\mu_i) = x_i'\beta$$

  gde je $\mu_i = E(Y_i)$.

2.2 Eksponencijalna familija raspodela i varijansna funkcija

Za svaku raspodelu verovatnoća slučajne promenljive koja se može napisati u obliku

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

kažemo da pripada eksponencijalnoj familiji. U okviru te familije važe sledeće osobine

$$E(Y) = \dot{a}(\theta) \tag{3}$$

$$Var(Y) = \phi\ \ddot{a}(\theta) \tag{4}$$

Sada ćemo postupno pokazati kako se dobijaju jednačine (3) i (4).

Dokaz: Neka $\dot{f}(y)$ i $\ddot{f}(y)$ predstavljaju prvi i drugi izvod funkcije f po parametru θ . Tada dobijamo

$$\dot{f}(y) = f(y)\left\{\frac{y - \dot{a}(\theta)}{\phi}\right\} \text{ i } \ddot{f}(y) = f(y)\left\{\frac{y - \dot{a}(\theta)}{\phi}\right\}^2 - f(y)\frac{\ddot{a}(\theta)}{\phi}$$

Sada prethodne dve jednačine integralimo po y i dobijamo

$$0 = \frac{E(Y) - \dot{a}(\theta)}{\phi} \text{ i } 0 = \frac{E[(y - \dot{a}(\theta))^2]}{\phi^2} - \frac{\ddot{a}(\theta)}{\phi}$$

Leve strane jednačina su 0 jer je

$$\int \dot{f}(y)dy = \frac{\partial}{\partial\theta}\int f(y)dy \text{ i } \int \ddot{f}(y)dy = \frac{\partial^2}{\partial\theta^2}\int f(y)dy$$

uzimajući činjenicu da je $\int f(y)dy = 1$, i da redosled integracije i diferencijacije može biti zamenjen. Dakle, imamo da je $E(Y) - \dot{a}(\theta) = 0$ i $E[(y - \dot{a}(\theta))^2] - \phi\ddot{a}(\theta) = 0$ tj.

$$E(Y) = \dot{a}(\theta) \text{ i } Var(Y) = \phi\ddot{a}(\theta)$$

što je i trebalo pokazati. U slučaju da je zavisna promenljiva diskretnog tipa integraciju zamenjuje sumiranje.
Mnoge ,,poznate" raspodele su članovi te familije: binomna, Poasonova, normalna, gama, inverzna Gausova itd. Logaritmovanjem jednačine (1) dobijamo sledeći izraz

$$ln\{f(y)\} = ln\{c(y,\phi)\} + \frac{y\theta - a(\theta)}{\phi} \tag{5}$$

Sada ćemo pokazati da Poasonova i normalna raspodela pripada pomenutoj familiji raspodela, a za ostale navedene raspodele pokazuje se slično.

Poasonova raspodela: Neka promenljiva Y ima Poasonovu raspodelu u oznaci $Y \sim P(\mu)$. Njena raspodela verovatnoće je $f(y) = \frac{e^{-\mu}\mu^y}{y!}$; $\mu > 0$ i $y = 0, 1, 2, 3, 4 \ldots$
Primenom jednačine (5) dobijamo

$$ln\{f(y)\} = -\mu + yln(\mu) - ln(y!)$$

$$\Rightarrow \theta = ln(\mu) \rightsquigarrow \mu = e^\theta$$
$$a(\theta) = \mu \rightsquigarrow a(\theta) = e^\theta$$

Parametar $\phi = 1, E(Y) = \dot{a}(\theta) = e^\theta = \mu$, $Var(Y) = \ddot{a}(\theta) = e^\theta = \mu$
Dakle, dobili smo $E(Y) = Var(Y) = \mu$ što je i trebalo pokazati.

Poasonova raspodela spada u diskretne raspodele i ima važnu ulogu u modeliranju celobrojnih vrednosti tj. podataka čije su vrednosti celi brojevi. Modelira se broj događaja od interesa u određenom vremenskom periodu. Na primer, broj dece u porodici, broj tropskih ciklona za vreme trajanja njihove sezone, broj neispravnih komponenti u računaru, broj zahteva za odštetu među polisama osiguranja itd.

Normalna raspodela[5] $N(\mu, \sigma^2)$

Ova raspodela ima najveći značaj među raspodelama verovatnoća slučajnih promenljivih koje su apsolutno neprekidne. Koristi se da se modeliraju podaci koji imaju simetričnu raspodelu. Široko je u upotrebi iz više razloga. Mnoge pojave u prirodi se modeliraju ovom raspodelom, kao što su visina ljudi, krvni pritisak itd. U slučaju da podaci koje analiziramo nemaju normalnu raspodelu prosečna vrednost ili suma vrednosti slučajnog uzorka će imati približno normalnu raspodelu jer to obezbeđuje centralna granična teorema. Mnoge statističke teorije su razvijene za normalnu raspodelu, pa ako podaci nemaju normalnu raspodelu onda se primeni jedna vrsta transformacije kao što je logaritamska i korena tako da dobijeni podaci imaju približno normalnu distribuciju.

Njena gustina raspodele je

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad y \in \mathbb{R}.$$

Primenimo logaritamsku transformaciju i dobijamo

$$ln\{f(y)\} = -ln(\sigma) - \frac{(y-\mu)^2}{2\sigma^2}$$

Ovde smo zanemarili konstantu $-ln(\sqrt{2\pi})$ i sređivanjem datog izraza formiramo oblik jednačine (5).

Sada je $ln\{f(y)\} = -ln(\sigma) - \frac{y^2/2}{\sigma^2} + \frac{y\mu - \mu^2/2}{\sigma^2}$. Prva dva člana u datom izrazu predstavljaju $ln\{c(y,\phi)\}$ a iz trećeg očigledno sledi da je parametar $\theta = \mu$, $a(\theta) = \frac{\theta^2}{2}$ i parametar $\phi = \sigma^2$. Sada je $E(Y) = \theta = \mu$ i $Var(Y) = \sigma^2 \cdot 1 = \sigma^2$.
Time smo pokazali da normalna raspodela pripada eksponencijalnoj familiji raspodela.

Varijansna funkcija

$$\ddot{a}(\theta) = \frac{\partial \dot{a}(\theta)}{\partial \theta} = \frac{\partial \mu}{\partial \theta} \equiv V(\mu) \tag{6}$$

Varijansna funkcija data jednačinom (6) ukazuje na vezu između srednje vrednosti i varijanse promenljive. Kod uopštenog linearnog modeliranja očekivana vrednost slučajne promenljive (μ) je povezana sa nezavisnim promenljivima, pa kako se one menjaju direktno se menja i vrednost μ. Kako μ varira tako varira i varijansa kroz $V(\mu)$. Dakle, možemo zaključiti da je model koji na neki način povezuje očekivanu vrednost μ sa objašnjavajućim promenljivima istovremeno i model koji povezuje varijansu slučajne promenljive sa objašnjavajućim promenljivima.

[5]Normalnu raspodelu je uveo nemački matematičar Gaus (1777–1855) pa se još naziva i Gausova raspodela.

Pokazali smo da je kod normalne raspodele

$a(\theta) = \dfrac{\theta^2}{2} \quad \Rightarrow \ddot{a}(\theta) = 1$ pa je $V(\mu) = 1$, konstantna varijansna funkcija znači da varijansa slučajne promenljive ne varira sa očekivanom vrednošću. Takva slučajna promenljiva se naziva **homoskedastična**. Slučaj kada uslov homoskedastičnosti nije ispunjen naziva se **heteroskedastičnost**.

Kod Poasonove slučajne promenljive

$a(\theta) = e^{\theta} \Rightarrow \ddot{a}(\theta) = e^{\theta} = \mu$ pa je $V(\mu) = \mu$, stoga promene srednje vrednosti direktno imaju uticaj na varijansu.

2.3 Različite metode za ocene nepoznatih parametara modela

Metoda maksimalne verodostojnosti

Osnovna ideja ove metode je da se pomoću izabranog uzorka $(x_1, x_2, \ldots, x_n)$ odabere ona vrednost nepoznatog parametra θ koja daje najveću verovatnoću da baš taj uzorak bude odabran. Sada ćemo dati definiciju funkcije verodostojnosti da bismo mogli formulisati metodu maksimalne verodostojnosti.

Definicija 2.3.1. Funkcija verodostojnosti za prost slučajan uzorak $(X_1, X_2, \ldots, X_n)$ na osnovu realizovanog uzorka $(x_1, x_2, \ldots, x_n)$ obima n je

$$L(\theta; x_1, x_2, \ldots, x_n) = \begin{cases} p(x_1, \theta) \cdot p(x_2, \theta) \cdots p(x_n, \theta) & \text{za diskretnu raspodelu} \\ \varphi(x_1, \theta) \cdot \varphi(x_2, \theta) \cdots \varphi(x_n, \theta) & \text{za neprekidnu raspodelu} \end{cases}$$

Kod diskretne raspodele $p(x_i, \theta) = P_\theta\{X_i = x_i\}$. Oznaka P_θ znači da se verovatnoće izračunavaju tako što se izabere ona raspodela iz dopustivog skupa raspodela koja odgovara baš parametru θ.
Familija raspodela $\{F(x, \theta), x \in \mathbb{R}, \theta \in \Theta\}$, gde je Θ skup parametara koji može biti jednodimenzionalan ili višedimenzionalan kada je nepoznato nekoliko parametara se naziva **dopustiva familija raspodela** za obeležje X, a skup Θ se naziva **dopustiv skup parametara**.

Definicija 2.3.2. Neka je $L(\theta)$ funkcija verodostojnosti za prost slučajni uzorak $(X_1, X_2, \ldots, X_n)$ na osnovu realizovanog uzorka $(x_1, x_2, \ldots, x_n)$[6].
Ako je $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ u skupu dopustivih parametara Θ, tada je statistika $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$ ocena parametra θ dobijena metodom maksimalne verodostojnosti i ta ocena maksimizira funkciju $L(\theta)$. Realizovana vrednost $\hat{\theta} = h(x_1, x_2, \ldots, x_n)$ je ocena maksimalne verodostojnosti za θ na osnovu uzorka $(x_1, x_2, \ldots, x_n)$.

[6] $L(\theta) = L(\theta; x_1, x_2, \ldots, x_n)$

Maksimum funkcije verodostojnosti se može odrediti rešavanjem sledeće jednačine

$$\frac{dL(\theta)}{d\theta} = 0 \tag{7}$$

Međutim , u mnogim situacijama je lakše naći maksimum prirodnog logaritma funkcije verodostojnosti , a kako funkcije $L(\theta)$ i $lnL(\theta)$ postižu maksimum za istu vrednost parametra θ rešavamo sledeću jednačinu

$$\frac{d\ ln[L(\theta)]}{d\theta} = 0 \tag{8}$$

Logaritam funkcije verodostojnosti je pogodno koristiti iz više razloga. Neki od njih su :

- Logaritam je monotona transformacija pa funkcije L i $\ln(L)$ postižu maksimum za istu vrednost odgovarajućeg parametra. Traženje maksimuma funkcije $l = ln[L(\theta)]$ ekvivalentno je traženju minimuma funkcije $-l$.
- Ponašanje funkcije L može biti takvo da je numerički veoma teško naći maksimum. Stoga ćemo za rešavanje jednačine (8) koristiti Njutn-Rapson metod.
- Koristeći log transformaciju znak za proizvod zamenjuje se sumom što u određenoj meri olakšava traženje izvoda.

Vratimo se sada eksponencijalnoj familiji raspodela. Pretpostavićemo da su observacije y_i nezavisne i u daljem tekstu neka je $L(\theta, \phi; y_1, y_2, \ldots, y_n) = f(\theta, \phi, y_1) \cdot f(\theta, \phi, y_2) \cdots f(\theta, \phi, y_n)$ tj.

$$L(\theta, \phi) = \prod_{i=1}^{n} f(\theta, \phi, y_i)$$

Kada primenimo logaritamsku transformaciju na funkciju verodostojnosti dobijamo

$$l(\theta, \phi) = \sum_{i=1}^{n} ln\{f(\theta, \phi, y_i)\}$$

Na osnovu jednačine (5) sada je $l(\theta, \phi) = \sum_{i=1}^{n} \left\{ \ln c(\phi, y_i) + \frac{y_i\theta - a(\theta)}{\phi} \right\}$

$$l(\theta, \phi) = \sum_{i=1}^{n} ln\{c(y_i, \phi)\} + \frac{n\{\overline{y}\theta - a(\theta)\}}{\phi}$$

Diferenciranjem funkcije l po θ i izjednačavanjem dobijenog izraza sa 0 sledi

$$\frac{n(\overline{y} - \dot{a}(\theta))}{\phi} = 0 \quad \Rightarrow \dot{a}(\theta) = \overline{y}$$

Iz dobijenog izraza dobijamo ocenjivač za parametar θ u oznaci $\hat{\theta}$. Pomoću ove metode možemo dobiti i ocenjivač za disperzioni parametar ϕ, ali je u praksi to teže pa se koristi metoda momenata.

Cilj nam je da ocenimo nepoznate parametre β_j koji su povezani sa θ, tako da je $\theta_i = \theta_i(\beta)$ pa tražimo parcijalne izvode funkcije $l(\theta, \phi)$ koristeći pravilo za izvod složene funkcije.

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j}$$

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - \dot{a}(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi} \qquad \text{i} \qquad \frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \; ; \qquad \text{gde je } \eta_i = x_i' \beta \; .$$

Dakle, imamo $\eta_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{i,p-1}\beta_{p-1}$ koji se naziva **linearni prediktor**. Odavde vidimo da je $\dfrac{\partial \eta_i}{\partial \beta_j} = x_{ij}$.

Sada je $\dfrac{\partial l}{\partial \beta_j} = \dfrac{1}{\phi} \displaystyle\sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i}(y_i - \mu_i) x_{ij}$; $\dfrac{\partial l}{\partial \beta_j} = 0 \Leftrightarrow \displaystyle\sum_{i=1}^{n} \frac{\partial \theta_i}{\partial \eta_i} x_{ij}(y_i - \mu_i) = 0$. Dobijenu jednačinu zapišemo u matričnom zapisu

$$X' D (y - \mu) = 0 \tag{9}$$

Preciznije

$D = diag(\partial\theta_1/\partial\eta_1, \partial\theta_2/\partial\eta_2, \ldots, \partial\theta_n/\partial\eta_n)$; $\quad y = (y_1, y_2, \ldots, y_n)'$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p-1} \end{bmatrix}$$

$$\left(\frac{\partial \theta_i}{\partial \eta_i} \right)^{-1} = \frac{\partial \eta_i}{\partial \theta_i} = \frac{\partial \eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial g(\mu_i)}{\partial \mu_i} \frac{\partial \dot{a}(\theta_i)}{\partial \theta_i} = \dot{g}(\mu_i) \ddot{a}(\theta_i).$$

Matrica D na dijagonali ima elemente $\{\dot{g}(\mu_i)\ddot{a}(\theta_i)\}^{-1}$. U jednačini (9) β je implicitno dato. Matricu D možemo napisati kao proizvod dijagonalnih matrica W i G gde je W dijagonalna sa elementima $[\{\dot{g}(\mu_i)\}^2 V(\mu_i)]^{-1}$, a G sa $\dot{g}(\mu_i)$. Novi oblik jednačine (9) je

$$X' W G (y - \mu) = 0 \tag{10}$$

Koristeći Tejlorovu aproksimaciju funkcije g, možemo dobiti aproksimaciju parametra β. Dalje je

$g(y_i) \approx g(\mu_i) + \dot{g}(\mu_i)(y_i - \mu_i) \quad \Rightarrow g(y) \approx g(\mu) + G(y - \mu) \Rightarrow G(y - \mu) \approx g(y) - g(\mu)$

Ovaj rezultat uvrstimo u jednačinu (10) pa dobijemo $X'Wg(y) - X'Wg(\mu) \approx 0$

$g(\mu) = X\beta \;\Rightarrow X'WX\beta \approx X'Wg(y)$ pa je

$$\hat{\beta} \approx (X'WX)^{-1} X'Wg(y) \tag{11}$$

Osobine ocenjivača maksimalne verodostojnosti

Pretpostavimo da smo ocenili nepoznati parametar θ u oznaci $\hat{\theta}$. Sada se možemo pitati kakve osobine ima. Ocenjivač $\hat{\theta}$ zavisi od vrednosti izabranog uzorka $(y_1, y_2, \ldots, y_n)$ što znači da je on **slučajna** promenljiva. Dve važne osobine ocenjivača su njegova **pristrasnost** i **varijansa**. Ocenjivač $\hat{\theta}$ je nepristrasan ako je $E(\hat{\theta}) = \theta$.
Varijansa ocenjivača ukazuje na njegovu preciznost. Nepristrasan ocenjivač koji ima malu varijansu ukazuje na visoku preciznost tj. dobijena ocena je veoma blizu stvarnoj vrednosti. Ako je $\hat{\theta}$ ocenjivač dobijen metodom maksimalne verodostojnosti, onda su poželjne osobine ovog ocenjivača sledeće :

- **asimptotska nepristrasnost** $\lim_{n\to\infty} E(\hat{\theta}) = \theta$

- **konzistentnost** Kako raste obim uzorka n raspodela $\hat{\theta}$ teži ka raspodeli za θ. $\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \epsilon) = 1$. Konzistentnost možemo ispitivati i preko srednje kvadratne greške u oznaci mse[7]

$$\lim_{n\to\infty} mse(\hat{\theta}) = \lim_{n\to\infty} [Var(\hat{\theta}) + (\theta - E(\hat{\theta}))^2] = 0$$

- **minimalna varijansa** U klasi svih nepristrasnih ocenjivača, $\hat{\theta}$ ima najmanju varijansu. Tada kažemo da je on **najefikasniji** ocenjivač.

$$Var(\hat{\theta}) \leq Var(\tilde{\theta})$$

- **invarijantnost** Ako je h neka monotona funkcija tada je $h(\hat{\theta})$ ocenjivač maksimalne verodostojnosti za $h(\theta)$.

Važno je napomenuti da kada se radi sa malim obimom uzorka n ocenjivači dobijeni ovom metodom su pristrasni. Ocenjivači maksimalne verodostojnosti su **asimptotski nepristrasni** i za veliki obim uzorka n imaju približno **normalnu** raspodelu.

$$\hat{\beta} \sim N(\beta, \phi(X'WX)^{-1})$$

Ako je parametar ϕ nepoznat takođe ga je potrebno oceniti. To je moguće uraditi metodom maksimalne verodostojnosti ili metodom momenata.

[7]mse od eng.-mean standard error, u prevodu znači srednja kvadratna greška.

Njutn-Rapson metod i Fišer skoring metod

Jednačina (10) nema eksplicitno rešenje, pa je potrebno koristiti se numeričkim metodama. Formira se niz iteracija koji u velikom broju slučajeva konvergira ka rešenju. Sada ćemo ukratko opisati **Njutn-Rapson** metod koji je široko upotrebljiv i tzv. **Fišer** skoring metod.
Njutn-Rapson metod je iterativni metod za rešavanje jednačine oblika
$f(x) = 0$. Startna početna vrednost je $x^{[0]}$, a sledeće iteracije su date sa jednačinom

$$x^{[n+1]} = x^{[n]} - f'(x^{[n]})^{-1} f(x^{[n]}) \tag{12}$$

gornji indeks [n] označava koja je iteracija u pitanju. U našem slučaju f biće zamenjeno sa $\dot{l}$ gde je $\dot{l}(\beta)$ vektor parcijalnih izvoda $\partial l/\partial\beta_j$ koji se naziva **skor** vektor.
Dakle, na osnovu jednačina (8) i (10) stacionarne tačke koje su ujedno i tačke minimuma za funkciju $-l$ tj. tačke maksimuma za funkciju l mogu se **približno** odrediti Njutn-Rapsonovim postupkom koji je dat sa

$$\beta^{[n+1]} = \beta^{[n]} - [\ddot{l}(\beta^{[n]})]^{-1} \dot{l}(\beta^{[n]}) \tag{13}$$

$\ddot{l}(\beta)$ je matrica sa elementom $\partial^2 l/(\partial\beta_j \partial\beta_k)$ na mestu (j,k), koja se naziva **Hesijan** matrica, a $\dot{l}(\beta^{[n]})$ označava gradijent funkcije. Uslov za maksimum je da je Hesijan negativno definitna matrica u tački (vektoru) tj. proizvod $z'\ddot{l}(\beta)z < 0$ za svaki nenula vektor z. Procedura ponavljanja izračunavanja skor vektora i Hesijan matrice da bi se ažurirala jednačina (13) se naziva **Njutn-Rapson** iteracija.

Fišer skoring metod dobijamo tako što se matrica $\ddot{l}(\beta)$ u jednačini (13) zameni sa $E[\ddot{l}(\beta)]$. Pritom se matrica $-E[\ddot{l}(\beta)]$ naziva **Fišer**-ova matrica informacija. Može se pokazati da važi sledeća jednakost

$$E[\ddot{l}(\beta)] = -E[\dot{l}(\beta)\dot{l}(\beta)']$$

Da bismo pokazali da je Fišerova matrica informacija jednaka sa $E[\dot{l}(\beta)\dot{l}(\beta)']$ prvo treba pokazati da je $E[\dot{l}(\beta)] = 0$.

Dokaz 1: $\dot{l}(\beta) = \dfrac{\partial \ln f(y)}{\partial\beta} = \dfrac{1}{f(y)} \dfrac{\partial f(y)}{\partial\beta}$. Sada je

$$E[\dot{l}(\beta)] = \int \dot{l}(\beta) f(y) dy = \int \frac{1}{f(y)} \frac{\partial f(y)}{\partial\beta} f(y) dy$$

$$= \int \frac{\partial f(y)}{\partial\beta} dy = \frac{\partial}{\partial\beta} \int f(y) dy = 0$$

Dokaz 2: Prvo pretpostavimo da je β skalar . Na osnovu dokaza 1 važi da je $\frac{\partial}{\partial\beta}E[\dot{l}(\beta)] = 0$. Sada je

$$0 = \frac{\partial}{\partial\beta}\int \dot{l}(\beta)f(y)dy = \int \frac{\partial}{\partial\beta}[\dot{l}(\beta)f(y)]dy = \int [\ddot{l}(\beta)f(y) + \dot{l}(\beta)\frac{\partial f(y)}{\partial\beta}]dy = \int \ddot{l}(\beta)f(y)dy + \int [\dot{l}(\beta)]^2 f(y)dy$$

Dakle

$$E[\ddot{l}(\beta)] + E[(\dot{l}(\beta))^2] = 0$$

Ako je β vektor onda je

$$-E[\ddot{l}(\beta)] = E[\dot{l}(\beta)\dot{l}(\beta)']$$

što je i trebalo pokazati.

Preciznije

$$E\left(\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) = -E\left[\left(\frac{\partial l}{\partial\beta_j}\right)\cdot\left(\frac{\partial l}{\partial\beta_k}\right)\right]$$

U matričnom zapisu je $\dot{l}(\beta) = \frac{1}{\phi}X'D(y-\mu)$, pa je

$$E[\dot{l}(\beta)\dot{l}(\beta)'] = \phi^{-2}X'\underbrace{DE[(y-\mu)(y-\mu)']D}_{W}X = \phi^{-1}X'WX \tag{14}$$

Sada ako umesto $\ddot{l}(\beta)$ i $\dot{l}(\beta)$ u jednačinu (13) ubacimo $E[\ddot{l}(\beta)]$ i $\frac{1}{\phi}X'WG(y-\mu)$ dobijamo

$$\beta^{[n+1]} = \beta^{[n]} + (X'WX)^{-1}X'WG(y-\mu)$$

Sređivanjem izraza dalje je

$$\beta^{[n+1]} = (X'WX)^{-1}X'W[X\beta^{[n]} + G(y-\mu)]$$

Inverz Fišerove matrice informacija je za veliko n približno kovarijansna matrica za $\hat{\beta}$. Prednost korišćenja očekivane vrednosti matrice Hesijan je to što je ona uvek pozitivno definitna pa nema problema sa konvergencijom niza iteracija ka rešenju jednačine. Sa druge strane, Fišer skoring metod sporije konvergira nego Njutn-Rapson metod.

Iterativna metoda najmanjih kvadrata

Ova numerička metoda predstavlja uobičajenu proceduru za rešavanje jednačine

$$X'WG(y-\mu)=0$$

gde je potrebno numeričkim postupkom aproksimirati vektor regresionih koeficijenata β koji je implicitno dat u prethodnoj jednačini. Podsetimo se da je

$$\beta=\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_{p-1}\end{bmatrix}$$

Ukratko, ovaj algoritam se zasniva na sledećim koracima :

1. Potrebno je linearizovati link funkciju g koristeći Tejlorovu aproksimaciju funkcije do prvog reda. Tako dobijamo

$$g(y)\approx g(\mu)+\dot{g}(\mu)(y-\mu)$$

Nadalje, neka je $g(\mu)+\dot{g}(\mu)(y-\mu)=z$.

2. Neka je $\hat{\eta}_0$ trenutna ocena linearnog prediktora η, i neka je $\hat{\mu}_0$ odgovarajuća fitovana vrednost dobijena iz odnosa $\hat{\eta}_0=g(\hat{\mu}_0)$. Formiramo sledeću zavisnu varijablu koju nazivamo ,,prilagođena"

$$z_0=\hat{\eta}_0+(y-\hat{\mu}_0)(\frac{\partial\eta}{\partial\mu})_0$$

gde je izvod od link funkcije izračunat u $\hat{\mu}_0$.

3. Definišemo težinsku matricu W tako da je $W_0^{-1}=(\frac{\partial\eta}{\partial\mu})^2V_0$ gde je V varijansna funkcija tako da je $V_0=V(\hat{\mu}_0)$. Prisetimo se da smo matricu W i ranije definisali da bismo dobili oblik jednačine (10).

4. Formiramo regresiju sa težinama gde je zavisna promenljiva z a prediktori su $x_1,x_2,\ldots,x_{p-1}$, koristeći težine iz matrice W_0. Kao rezultat dobijamo novu ocenu vektora β u oznaci $\hat{\beta}_1$ a samim tim dobijamo i ocenu za linearni prediktor u oznaci $\hat{\eta}_1$.

5. Koraci $1-4$ se ponavljaju sve dok promene u ocenjenim parametrima ne budu dovoljno male.

Za modele kod kojih je link kanonički ova metoda je zapravo Njutn-Rapson metod .

Metoda momenata

Neka je X slučajna promenljiva koja predstavlja neko obeležje. Njena familija raspodela je $\{F_X(x, \theta_1, \theta_2, \ldots, \theta_k), x \in \mathbb{R}, (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta)\}$.[8] Dakle, vidimo da raspodela obeležja X zavisi od k parametara θ. Neka je $m_r = E(X^r)$ r-ti momenat slučajne promenljive X. On zavisi od nepoznatih parametara, pa pišemo $m_r = f_r(\theta_1, \theta_2, \ldots, \theta_k)$. Sada, neka je $(X_1, X_2, \ldots, X_n)$ prost slučajan uzorak za obeležje X i M_r r-ti uzorački momenat.

$$M_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r \ .$$

$$M_j = f_j(\theta_1, \theta_2, \ldots, \theta_k), \quad j = 1, 2, \ldots, k.$$

Prethodni sistem jednačina predstavlja izjednačavanje uzoračkog j-tog momenta sa teorijskim. Važno je napomenuti da imamo onoliko jednačina koliko je nepoznatih parametara. Ako ovaj sistem ima jedinstveno rešenje dobijamo ocene $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$. Kažemo da su te ocene dobijene **metodom momenata**.

Ocene su dobijene tako što su momenti populacije zamenjeni sa uzoračkim momentima. U slučaju eksponencijalne familije raspodela funkcija raspodele verovatnoća $f(y)$ zavisi od jednog ili dva parametra koji su u opštem slučaju nepoznati. Ti nepoznati parametri, u ovom slučaju θ i ϕ se ocenjuju na osnovu uzorka obima n , i pri tome se pretpostavlja da svako zapažanje y_i dolazi iz iste raspodele. Ova metoda ,,pronalazi" θ i ϕ tako da je populaciona srednja vrednost i varijansa jednaka uzoračkoj.

Tačnije

$$\dot{a}(\theta) = \overline{y} \quad i \quad \phi\ddot{a}(\theta) = \hat{\sigma^2}$$

Pirson χ^2 statistika za ocenu parametra ϕ

U praksi je ϕ obično nepoznato, pa ga je potrebno oceniti. Jedan od načina jeste pomoću **Pirson**-ove χ^2 statistike. Kod uopštenog linearnog modeliranja statistika je definisana na sledeći način

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{Var(Y_i)} = \frac{1}{\phi} \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Data statistika je približno χ^2_{n-r} raspodeljena pa je $E(\chi^2) \approx n - r$, gde je r broj ocenjenih parametara β. Sada je ocena za parametar ϕ

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

[8] Θ je dopustiv skup parametara

2.4 Intervali poverenja i predikcije

Predikcije se bave sa problemima kao što su verovatne vrednosti nekih posmatranih pojava, ili određenih parametara modela. Uz predviđenu vrednost nekog parametra mora da ide i mera preciznosti. Ocenjivači **maksimalne verodostojnosti** su asimptotski nepristrasni i za veliki obim uzorka n imaju približno normalnu raspodelu. Matematički to zapisujemo

$$\hat{\beta} \approx N(\beta, \phi(X'WX)^{-1})$$

Ovde je $cov(\hat{\beta}) = \phi(X'WX)^{-1}$ kovarijansna matrica ocenjivača $\hat{\beta}$. Označimo sa F Fišerovu matricu informacija. Važi da je $F = [cov(\hat{\beta})]^{-1}$ tj. $F = \phi^{-1}(X'WX)$. Za konstrukciju intervala poverenja potrebna nam je Fišerova matrica informacija.
Defininisana je kao $F = -E[I]$ gde je I matrica sa elementima $\partial^2 l/\partial\beta_j\partial\beta_k$.
Da bi odredili matricu I diferenciramo $\partial l/\partial\beta_j$ po β_k .

Već smo pokazali da je $\dfrac{\partial\theta_i}{\partial\eta_i} = [\ddot{a}(\theta_i)\dot{g}(\mu_i)]^{-1}$, pa je $\dfrac{\partial l}{\partial\beta_j} = \displaystyle\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi\ddot{a}(\theta_i)\dot{g}(\mu_i)}$
Koristeći činjenicu da je $V(\mu_i) = \ddot{a}(\theta_i)$ dobijamo

$$\frac{\partial l}{\partial\beta_j} = \sum_{i=1}^{n} \frac{1}{\phi}\frac{(y_i - \mu_i)x_{ij}}{V(\mu_i)\dot{g}(\mu_i)} \tag{15}$$

Sada je

$$\frac{\partial}{\partial\beta_k}\left(\frac{\partial l}{\partial\beta_j}\right) = \sum_{i=1}^{n} \frac{1}{\phi}\frac{\partial}{\partial\mu_i}\left[\frac{y_i - \mu_i}{V(\mu_i)\dot{g}(\mu_i)}\right] x_{ij}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k}$$

Ranije smo pokazali da je $\dfrac{\partial\eta_i}{\partial\beta_k} = x_{ik}$ i $\dfrac{\partial\mu_i}{\partial\eta_i} = \dfrac{1}{\dot{g}(\mu_i)}$. Daljim sređivanjem ovog izraza na kraju dobijamo

$$I = -\sum_{i=1}^{n} x_{ij}c_i x_{ik}$$

gde je $c_i = \phi^{-1}\dfrac{1}{V(\mu_i)\dot{g}(\mu_i)^2}\left(1 + (y_i - \mu_i)\dfrac{[V(\mu_i)\ddot{g}(\mu_i) + \dot{V}(\mu_i)\dot{g}(\mu_i)]}{V(\mu_i)\dot{g}(\mu_i)}\right)$.
Neka je C dijagonalna sa elementima c_i tj. $C = diag(c_i)$. Sada je $I = -X'CX$.
$F = -E(I) = X'E(C)X$. Kako je $E(Y_i) = \mu_i$ očekivana vrednost matrice C biće matrica koja na dijagonali ima elemente $\phi^{-1}/V(\mu_i)[\dot{g}(\mu_i)]^2$ a to je matrica $\phi^{-1}W$ koju smo već definisali. Otuda je

$$F = \phi^{-1}X'WX$$

Raspodelu za $\hat{\beta}$ možemo zapisati

$$\hat{\beta} \approx N(\beta; F^{-1})$$

Neka je $B = F^{-1}$ i neka je $b_{j,k}$ element matrice B. 95% interval poverenja za β_j je $\hat{\beta}_j \pm 1.96\sqrt{b_{jj}}$, gde je 1.96 kvantil reda 0.975 N raspodele a b_{jj} dijagonalni element matrice B. Interval poverenja za ocenjeni parametar se koristi da ukaže na preciznost ocene. Kako je link funkcija $g(\mu) = x'\beta$ i ako smo ocenili parametre $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1})$ metodom maksimalne verodostojnosti sada znamo i ocenu za link funkciju g $\hat{g}(\mu) = x'\hat{\beta}$, a kako je g monotona funkcija na osnovu osobine invarijantnosti sledi da je $\hat{g}(\mu)$ ocenjivač maksimalne verodostojnosti za $g(\hat{\mu})$. Sada je

$$g(\hat{\mu}) = x'\hat{\beta}$$

Izračunavanje intervala poverenja za μ zahteva da znamo raspodelu ocenjivača $\hat{\mu}$. $Var(x'\hat{\beta}) = x'Var(\hat{\beta})x = \phi x'(X'WX)^{-1}x$. Aproksimativno, interval poverenja za μ je (μ_l, μ_u) gde je

$$g(\mu_l) = g(\hat{\mu}) - z\sqrt{Var(g(\hat{\mu}))}$$

Tj. $g(\mu_l) = x'\hat{\beta} - z\sqrt{Var(x'\hat{\beta})} \rightsquigarrow\ g(\mu_l) = x'\hat{\beta} - z\sqrt{\phi x'(X'WX)^{-1}x}$

$g(\mu_u) = x'\hat{\beta} + z\sqrt{\phi x'(X'WX)^{-1}x}$, gde je z kvantil normalne raspodele.

Ako je na primer g ln funkcija tada je $g(\hat{\mu}) = \ln\hat{\mu} = x'\hat{\beta} \Rightarrow \quad \hat{\mu} = e^{x'\hat{\beta}}$.

Napomenimo da su x objašnjavajuće promenljive i da su nam njihove vrednosti poznate. Interval poverenja je tačan u slučaju kada je link identička funkcija i raspodela zavisne promenljive je normalna , u svim ostalim slučajevima je aproksimacija. Tačnost te aproksimacije raste sa povećanjem obima uzorka.
Dakle, (μ_l, μ_u) je interval poverenja za očekivanu vrednost slučajne promenljive Y ako su date vrednosti x -sa.

2.5 Devijansa uopštenog linearnog modela

Prilagođenost uopštenog linearnog modela podacima se može proceniti na osnovu tzv. devijanse. Neka je $\hat{l}$ vrednost funkcije l u tački maksimuma tj.

$$\hat{l} = \sum_{i=1}^{n} \left\{ lnc(y_i, \phi) + \frac{y_i\hat{\theta}_i - a(\hat{\theta}_i)}{\phi} \right\}$$

i neka je $\tilde{l}$ vrednost **zasićenog** modela. Zasićen model ima onoliko nepoznatih parametara koliko imamo i observacija tj. n. U takvom modelu ocenjivač maksimalne verodostojnosti za θ je takav da je $\dot{a}(\theta) = y$. Neka je $\tilde{\theta}$ ocenjivač maksimalne verodostojnosti za θ kod zasićenog modela. Sada imamo da je $\dot{a}(\tilde{\theta}_i) = y_i$ a funkcija $\tilde{l}$ je

$$\tilde{l} = \sum_{i=1}^{n} \left\{ ln(y_i, \phi) + \frac{y_i\tilde{\theta}_i - a(\tilde{\theta}_i)}{\phi} \right\}$$

Dakle, kod zasićenog modela svaka observacija y_i fituje model savršeno tj. $y_i = \hat{y}_i$.

Devijansa je definisana na sledeći način

$$D \equiv 2(\tilde{l} - \hat{l})$$

Dakle, devijansa predstavlja meru odstupanja fitovanog modela od zasićenog. Velika vrednost devijanse ukazuje na loše fitovanje. Kada je model ,,dobar" vrednost za $\hat{l}$ treba da bude blizu vrednosti za $\tilde{l}$. Uvrštavanjem izraza za $\tilde{l}$ i $\hat{l}$ u formulu za devijansu dobijamo

$$D = 2\sum_{i=1}^{n}\left\{\frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - a(\tilde{\theta}_i) + a(\hat{\theta}_i)}{\phi}\right\}$$

Kako normalna raspodela pripada eksponencijalnoj familiji raspodela, za nju možemo pokazati da devijansa predstavlja rezidualnu sumu kvadrata.

normalna raspodela : Već smo pokazali da je u ovom slučaju $a(\theta) = \dfrac{\theta^2}{2}$ pa je $\dot{a}(\theta) = \theta$, $\tilde{\theta}_i = y_i$ i $\hat{\theta}_i = \hat{\mu}_i$ i $\phi = \sigma^2$. Izraz pod sumom u formuli za devijansu je[9] $y_i(y_i - \hat{\mu}_i) - \dfrac{y_i^2 - \hat{\mu_i^2}}{2} = \dfrac{(y_i - \hat{\mu}_i)^2}{2}$. Sada je

$$D = \frac{2}{\sigma^2}\sum_{i=1}^{n}\left\{\frac{(y_i - \hat{\mu}_i)^2}{2}\right\} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

Na isti način se može pokazati kako izgleda izraz za devijansu kod drugih članova eksponencijalne familije raspodela. Izraz **,,scaled deviance"** se odnosi na skaliranu devijansu a definisana je kao $D^* = \dfrac{1}{\phi}D$. Za binomnu i Poasonovu raspodelu, devijansa i skalirana devijansa su jednake jer je $\phi = 1$. Raspodela za devijansu D je χ^2 pod pretpostavkom da je fitovan model korektan i da je n veliko. $D \approx \chi^2_{n-p}$ gde je n obim uzorka a p broj ocenjenih parametara modela.

Kod statistika koje imaju χ^2 raspodelu njihova očekivana vrednost je broj stepeni slobode. Stoga je očekivana vrednost devijanse $n - p$. Prilikom analize u kojoj se ispituje koliko je fitovan model ,,dobar" koristi se i statistika $D/(n-p)$; ukoliko se dobije vrednost koja je mnogo veća od 1 to ukazuje na loše fitovan model. Kada je disperzioni parametar ϕ nepoznat i ocenjen, ne možemo tvrditi da je devijansa χ^2 raspodeljena. Kod Poasonove raspodele parametar ϕ je poznat tj. $\phi = 1$. Dakle, nije ga potrebno oceniti pa je χ^2 raspodela korisna. Međutim, ako razmatramo normalnu raspodelu njen disperzioni parametar je σ^2. U nekim slučajevima taj parametar je poznat i tada devijansa ima χ^2 raspodelu. Ali ako je potrebno oceniti σ^2 onda se ne možemo pouzdati u χ^2 raspodelu devijanse. Zbog te činjenice se možemo zapitati da li je devijansa pogodna statistika za proveru adekvatnosti modela.

[9] u izrazu smo izostavili ϕ

Druga važna upotreba devijanse služi za poređenje dva ,,konkurentska" modela .
Pretpostavimo da analiziramo dva modela, od kojih je jedan jednostavniji od drugog. Neka je devijansa složenijeg modela označena sa D_1 sa n_1 stepeni slobode a devijansa jednostavnijeg modela je D_2 sa n_2 stepeni slobode. ,,Jednostavniji" model ima manje nepoznatih parametara od složenijeg modela, pa će zato imati veći broj stepeni slobode tj. $n_2 > n_1$, i imaće veću devijansu od složenog modela. Za poređenje ova dva modela koristimo razliku devijansi $D_2 - D_1$ koja ima $\chi^2_{n_2-n_1}$ raspodelu. Ova test statistika je korisna za testiranje značajnosti parametara koji su uključeni u jedan model ali ne i u drugi. Napomenimo da su parametri jednog modela podskup parametara drugog (složenijeg) modela tj. modeli su ,,ugnježdeni".

2.6 Testiranje hipoteza

Kod uopštenog linearnog modeliranja testiranje hipoteza za pojedinačne parametre ili za grupu parametara može se sprovesti na više načina. Sada ćemo navesti neke od njih koje se najčešće upotrebljavaju. Hipoteze se pišu u formi $C\beta = r$ gde je C tzv. hipotetička matrica a r je vektor datih vrednosti. Postoje tri glavna pristupa za testiranje hipoteze $C\beta = r$. Neka je $\hat{\beta}$ ocenjivač maksimalne verodostojnosti za β kod modela bez restrikcija i neka je $\tilde{\beta}$ ocenjivač maksimalne verodostojnosti za β kada se funkcija l maksimizira prema restrikciji $C\beta = r$. Nadalje, neka je $\hat{l}$ vrednost funkcije l u tački $\hat{\beta}$ i $\tilde{l}$ vrednost funkcije l u tački $\tilde{\beta}$. Podsetimo, funkcija l prestavlja logaritam funkcije verodostojnosti.

Wald test

Metoda maksimalne verodostojnosti za ocenu parametara nekog modela nam pruža ocenjene parametre i ocene za standardne greške tih ocena. Ako je $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ nepoznati parametar modela onda je ocena $\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_{p-1}})$.
Može se pokazati da kako raste obim uzorka n varijansa ocenjivača teži ka $\phi(X'WX)^{-1}$.

$$\lim_{n\to\infty} Var(\hat{\beta}) = \phi(X'WX)^{-1}.$$

Standardna greška ocenjivača u oznaci $\sigma_{\hat{\beta}}$ predstavlja kvadratni koren iz varijanse ocenjivača tačnije

$$\sigma^2_{\hat{\beta}} = diag[cov(\hat{\beta})]$$

U praksi je vrlo često potrebno oceniti parametar ϕ pa dobijamo ocenjenu standardnu grešku ocenjivača $\hat{\beta}$ u oznaci $\hat{\sigma}_{\hat{\beta}}$. Ako pretpostavimo da je obim uzorka veliki možemo testirati hipotezu o jednom parametru tj. $\beta_j = r$, tada je matrica C vektor vrsta koji na poziciji j ima jedinicu a sve ostale vrednosti su nule. Wald statistika je

$$z = \frac{(\hat{\beta}_j - r)^2}{\phi\psi_j}$$

gde je z je slučajna promenljiva koja ima χ^2_1 raspodelu. Ta statistika se naziva **Wald** test. ψ_j je dijagonalni element matrice $(X'WX)^{-1}$.

Kada je ϕ nepoznato zamenjuje se sa ocenom $\hat{\phi}$.
Uopšte, ako se testira hipoteza $C\beta = r$, meri se razdaljina $C\hat{\beta}$ od r. Ako je razlika $C\hat{\beta} - r$ velika to ukazuje da hipoteza nije tačna. Kako je $\hat{\beta} \sim N(\beta, \phi(X'WX)^{-1})$ sledi da je

$$C\hat{\beta} - r \sim N(0, \phi C(X'WX)^{-1}C')$$

Wald statistika za testiranje hipoteze $C\beta = r$ je

$$(C\hat{\beta} - r)'\{\phi C(X'WX)^{-1}C'\}^{-1}(C\hat{\beta} - r) \sim \chi^2_q$$

q je broj stepeni slobode χ^2 raspodele. Ako je p_1 broj nepoznatih parametara u modelu bez restrikcija, a p_2 broj nepoznatih parametara u modelu sa restrikcijama tada je $q = p_1 - p_2$. U praksi je matrica W zamenjena sa ocenom pa je zato raspodela za Wald statistiku približno χ^2.

Test odnosa verodostojnosti

Neka je L_1 funkcija verodostojnosti koju maksimiziramo pod pretpostavkom da model nema restrikcija, tj. model bez postavljanja hipoteza i sa L_0 označimo funkciju verodostojnosti koju maksimiziramo tako da neki parametri zadovoljavaju vrednosti koje su date nultom hipotezom koja se testira. Odnos verodostojnosti je statistika

$$-2\log(L_0/L_1) = -2[\log(L_0) - \log(L_1)] = -2(l_0 - l_1) \tag{16}$$

Vrednost ove statistike je uvek nenegativna i može se pokazati da ova statistika data jednačinom (16) ima približno χ^2 raspodelu kako obim uzorka raste. Označimo sa p_1 broj nepoznatih parametara modela 1 i p_2 broj nepoznatih parametara u modelu 2. Onda statistika odnosa verodostojnosti ima $\chi^2_{p_1-p_2}$ raspodelu. Izuzetak je slučaj kada imamo linearna ograničenja za parametre. Ako je hipoteza $C\beta = r$ tačna onda statistika ima χ^2_q raspodelu gde je $q = p_1 - p_2$ tj. q je broj restrikcija u vektoru β.
Ako je vrednost statistike $-2\log(L_0/L_1)$ mala tj. blizu nule onda je model sa restrikcijama jednako dobar kao i model bez restrikcija. To ukazuje da ne treba da odbacujemo nultu hipotezu $H_0 : C\beta = r$. Oblast odbacivanja za test je gornji rep χ^2_q raspodele.
χ^2_q raspodela za pomenutu statistiku uključuje disperzioni parametar ϕ koji je često nepoznat. U tom slučaju ga je potrebno oceniti. χ^2 raspodela je i dalje pogodna ukoliko je ocenjivač za ϕ konzistentan. Napomenimo da su ocenjivači maksimalne verodostojnosti konzistentni. Statistika odnosa verodostojnosti može da se izrazi kao razlika devijansi modela bez restrikcija i modela sa restrikcijama ako se koristi ista ocena za ϕ u obe funkcije logaritma verodostojnosti.

Skor test

Neka je $\tilde{\beta}$ ocenjivač maksimalne verodostojnosti za parametar β u modelu sa restrikcijama tj. kada se funkcija l maksimizira prema ograničenju da je $C\beta = r$. Hoćemo da testiramo hipotezu $H_0 : C\beta = r$. Neka su L_1 i L_0 funkcije verodostojnosti pod hipotezama H_1 i H_0. H_1 je tzv. alternativna hipoteza tj. $H_1 : C\beta \neq r$.
Skor test se bazira na izvodu funkcije l u tački $\tilde{\beta}$. Taj izvod u oznaci $\dot{l}(\tilde{\beta})$ se naziva **skor**. Ako je skor tj. nagib veliki to ukazuje da treba da odbacimo hipotezu H_0. Ako je izvod funkcije blizu 0, to znači da smo blizu maksimuma. Kod uopštenog linearnog modeliranja skor je vektor $\dot{l}(\beta) = \phi^{-1}X'WG(y - \mu)$ i važi da je

$$E[\dot{l}(\beta)] = 0 \;\; Var[\dot{l}(\beta)] = E[\dot{l}(\beta)\dot{l}(\beta)'] = \phi^{-1}X'WX$$

Ako je $\dot{l}(\tilde{\beta})$ blizu 0 nemamo razloga da odbacimo hipotezu H_0 .

Skor test statistika je

$$\dot{l}(\tilde{\beta})'[Var\{\dot{l}(\beta)\}]^{-1}\dot{l}(\tilde{\beta})$$

gde je $\tilde{\beta}$ ocenjivač maksimalne verodostojnosti funkcije L_0.
$\dot{l}(\tilde{\beta}) = \phi^{-1}X'WG(y - \tilde{\mu})$, gde je $\tilde{\mu}$ $E(y)$ izračunato u $\tilde{\beta}$ i $Var[\dot{l}(\beta)] = \phi^{-1}(X'WX)$.
Skor test statistika ima približno χ^2_q raspodelu gde je q kao i ranije broj stepeni slobode. Raspodela je manje tačna ukoliko se ϕ i W zamene sa ocenama. Treba i napomenuti da je matrica X punog ranga tj. da ona sadrži i vrednosti onih objašnjavajućih promenljivih za čije se koeficijente pretpostavlja da su nule u hipotezi $C\beta = r$. Oblast odbacivanja za test je ,,gornji rep" χ^2_q raspodele tj. suviše velike vrednosti **skor** statistike.

2.7 Analiza reziduala

Analiza reziduala se koristi da se ispita adekvatnost fitovanja modela u odnosu na izbor link funkcije, varijansne funkcije i članova u linearnom prediktoru. Koriste se da se proveri podesnost izabrane raspodele i za neke nepredviđene vrednosti koje zahtevaju dalje istraživanje. Podsetimo se da su reziduali kod normalnog linearnog modela $\hat{\epsilon}_i = y_i - \hat{y}_i$ i treba da zadovoljavaju sledeće pretpostavke :

- normalnost – znači da greše ϵ_i imaju normalnu raspodelu
- homoskedastičnost – $Var(\epsilon_i) = const.$ za $\forall i = 1, 2, \ldots, n.$
- odsustvo autokorelacije – $cov(\epsilon_i, \epsilon_j) = 0$ za $\forall i \neq j$

Kod uopštenog linearnog modeliranja reziduali nisu normalno distribuirani, niti imaju konstantnu varijansu osim u slučaju kada je u pitanju normalna raspodela. U ovom slučaju reziduali nisu samo razlika između opaženih i fitovanih vrednosti već je njihova definicija uopštenija. Dakle, kod ulm[10] zahtevamo širu definiciju reziduala koja će biti primenljiva za sve raspodele koje mogu da zamene normalnu.

[10]uopšteno linearno modeliranje

Sada ćemo definisati različite forme uopštenih reziduala koje se koriste, i matricu Hat koja će koristiti u definisanju nekih reziduala. Tipovi reziduala o kojima će biti reči zavise od tipa odgovarajućeg testa za testiranje hipoteza. Pirson reziduali su povezani sa Pirson χ^2 i sa Wald testom. Reziduali devijanse su povezani sa devijansom kao merom fitovanja modela i sa testom odnosa verodostojnosti, skor reziduali su povezani sa skor testom itd.

Hat matrica u oznaci H

Ocenjena očekivana vrednost zavisne promenljive kod normalnog linearnog modela je

$$\widehat{E(y_i)} = \hat{\mu}_i = \hat{y}_i$$

Fitovane vrednosti su linearne funkcije od opaženih vrednosti tj.

$$\hat{y} = Hy$$

a matrica H se zove ,,kapa matrica" ili možemo reći Hat. H je simetrična i idempotentna matrica.
Kod proste linearne regresije je $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$ odakle je $H = X(X'X)^{-1}X'$.
Kod uopštenog linearnog modeliranja H matrica je nešto složenijeg oblika

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

Pirson reziduali

Pirson reziduali su definisani na sledeći način

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

Važi da je

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

Za Poasonovu raspodelu suma kvadrata reziduala je zapravo Pirsonova χ^2 statistika pa otuda naziv Pirson reziduali.

Anskombe rezidual

Raspodela Pirson reziduala r_p kod promenljivih koje nisu normalno raspodeljene je često primetno iskošena, što utiče na to da ti reziduali neće imati osobine slične kao reziduali normalnog linearnog modela. Zato definišemo funkciju h koja će imati približno normalnu raspodelu.

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

Anskombe reziduali se baziraju na razlici $h(y_i) - h(\hat{\mu}_i)$.

Za funkciju h važi da je

$$\dot{h}(y) = [V(y)]^{-1/3}$$

Pritom je $V(\mu)$ varijansna funkcija za raspodelu zavisne promenljive Y.

Funkcija h je transformacija koja normalizuje funkciju verovatnoće zavisne promenljive Y ali ne stabilizuje varijansu pa je izraz $h(y) - h(\hat{\mu})$ potrebno skalirati tj. podeliti sa kvadratnim korenom varijanse $h(y)$ tj. sa standardnom devijacijom od $h(y)$.
Aproksimacija prvog reda za standardnu devijaciju od $h(y)$ je $\dot{h}(\hat{y})\sqrt{V(\hat{y})}$. Sada dobijamo

$$\frac{h(y_i) - h(\hat{\mu_i})}{\dot{h}(\hat{\mu_i})\sqrt{V(\hat{\mu_i})}}$$

Ovi reziduali se zovu **Anskombe** reziduali, standardizovani su i imaju približno normalnu raspodelu. Sada ćemo pokazati kako ovi reziduali izgledaju kada se u obzir uzmu konkretne raspodele.

Primer: Posmatrajmo Poasonovu raspodelu tj. $Y \sim P(\mu)$. Ranije smo pokazali da je $V(\mu) = \mu$ pa je $\dot{h}(y) = y^{-1/3}$ i $h(y) = \int y^{-1/3}dy = \frac{3}{2}y^{2/3}$. Anskombe reziduali u oznaci r_A su

$$(r_A)_i = \frac{3}{2} \cdot \frac{y_i^{2/3} - \hat{\mu_i}^{2/3}}{\hat{\mu_i}^{1/6}}$$

Primer: Neka sada Y ima gama raspodelu, $Y \sim G(\mu, \nu)$. Može se pokazati da je kod ove raspodele varijansna funkcija $V(\mu) = \mu^2$, pa je otuda $\dot{h}(y) = y^{-2/3}$ i $h(y) = \int y^{-2/3}dy = 3y^{1/3}$. Anskombe reziduali su

$$(r_A)_i = \frac{3y_i^{1/3} - 3\hat{\mu_i}^{1/3}}{\hat{\mu_i}^{-2/3} \cdot \hat{\mu_i}} = 3 \cdot \frac{y_i^{1/3} - \hat{\mu_i}^{1/3}}{\hat{\mu_i}^{1/3}}$$

Primer: inverzna Gausova raspodela, $Y \sim IG(\mu, \sigma^2)$. $V(\mu) = \mu^3$, $V(y) = y^3$ sledi da je $\dot{h}(y) = \frac{1}{y}$, $h(y) = \int \frac{dy}{y} = \ln(y)$. Anskombe reziduali su

$$(r_A)_i = \frac{\ln(y_i) - ln(\hat{\mu_i})}{\hat{\mu_i}^{-1} \cdot \hat{\mu_i}^{3/2}} = \frac{\ln(y_i) - \ln(\hat{\mu_i})}{\sqrt{\hat{\mu_i}}}$$

Primer: normalna raspodela , $Y \sim N(\mu, \sigma^2)$. Varijansna funkcija je konstantna, $V(\mu) = 1$ sledi $\dot{h}(y) = 1$ i $h(y) = \int 1 \cdot dy = y$. Reziduali su

$$r_i = y_i - \hat{\mu_i}$$

Reziduali devijanse

Pokazali smo da je izraz za devijansu kod članova eksponencijalne familije raspodela sledeći

$$D = \sum_{i=1}^{n} 2 \cdot \left\{ \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - a(\tilde{\theta}_i) + a(\hat{\theta}_i)}{\phi} \right\}$$

Definišemo reziduale devijanse u oznaci $(r_d)_i$ za $i = 1, 2, \ldots, n$ tako da je ispunjeno

$$(r_d)_i^2 \equiv \frac{2[y_i(\tilde{\theta}_i - \hat{\theta}_i) - a(\tilde{\theta}_i) + a(\hat{\theta}_i)]}{\phi} \quad \text{i} \quad \sum_{i=1}^{n} (r_d)_i^2 = D$$

Znak od $(r_d)_i$ zavisi od znaka $y_i - \hat{\mu}_i$ tj. $(r_d)_i = sign(y_i - \hat{\mu}_i)\sqrt{(r_d)_i^2}$.

Ako je model ,,dobar" i ako je n dovoljno veliko devijansa ima približno χ^2_{n-p} raspodelu. Stoga je očekivana vrednost devijanse $n - p$ pa svaki slučaj i treba da doprinosi devijansi približno $(n - p)/n \approx 1$. Ako je $|(r_d)_i|$ mnogo veće od 1 onda to znači da slučaj i pogoršava fitovanje. Model ima neke nedostatke ili pogrešno unete vrednosti.

Reziduali devijanse i anskombe reziduali su numerički veoma slični iako se u matematičkom zapisu prilično razlikuju. Kako anskombe reziduali imaju približno normalnu raspodelu, značilo bi da su i reziduali devijanse približno normalno raspodeljeni. Ovo je korisna činjenica ako softver koji se upotrebljava za analizu ne produkuje Anskombe reziduale. Reziduali devijanse mogu da se napišu u standardizovanom obliku na sledeći način

$$(r_d)_{i_s} = \frac{(r_d)_i}{\sqrt{1 - h_{ii}}} = \frac{sign(y_i - \hat{\mu}_i)\sqrt{(r_d)_i^2}}{\sqrt{1 - h_{ii}}}$$

Gde je h_{ii} dijagonalni element Hat matrice. Umesto $(r_d)_i^2$ možemo pisati d_i da bi izbegli komplikovan tehnički zapis. Sada je

$$(r_d)_{i_s} = \frac{sign(y_i - \hat{\mu}_i)\sqrt{d_i}}{\sqrt{1 - h_{ii}}}$$

Skor reziduali

Skor reziduali su povezani sa skor testom za testiranje hipoteze o parametrima modela. Metodom maksimalne verodostojnosti ocene parametara se dobijaju rešavanjem jednačina oblika

$$U_i = \frac{\partial l}{\partial \theta_i} = 0$$

koje se nazivaju **skor** jednačine, gde je θ nepoznati parametar. Standardizovani skor reziduali su

$$(r_i)_s = \frac{U_i}{\sqrt{(1 - h_{ii})}}$$

gde su h_{ii} dijagonalni elementi matrice H .

2.8 Neke dijagnostike modela

Autlajeri i uticajne tačke

Autlajeri su observacije tj. opažanja za koje model ne daje dobru aproksimaciju. Drugim rečima to su pojedinačne observacije koje na neki način odstupaju iz skupa podataka tj. opažanja. Njihova opažena vrednost je veoma različita od vrednosti koja je predviđena regresijskim modelom. To su tačke koje imaju veliki rezidual. Oni mogu biti posledica ekstremne vrednosti zavisne promenljive ili jedne ili više nezavisnih promenljivih. Veoma često se otkrivaju koristeći grafički prikaz tzv. plot. Observacije koje imaju veliki uticaj na ocene nepoznatih parametara modela nazivaju se **uticajne** observacije. Opažanje je uticajno ako menja nagib regresijskog pravca. Takve observacije ne moraju da budu autlajeri, u tom smislu da ta zapažena vrednost ne odstupa od našeg skupa podataka. Takođe ni autlajeri ne moraju da budu uticajne informacije. Ako se uoči observacija koja puno odstupa od fitovanog modela potrebno je utvrditi da li je ona uticajna. Analiza se vrši tako što se posmatra model sa i bez tog podatka, pa ako se vrednosti ocenjenih parametara modela puno promene možemo zaključiti da je ta observacija značajna i ne možemo je zanemariti. Podaci koji su autlajeri a nisu uticajne tačke mogu se isključiti iz analize. Sada ćemo definisati neke pojmove koji će nam koristiti za otkrivanje autlajera i uticajnih tačaka.

Leveridž : Za observaciju i leveridž predstavlja izvod fitovane vrednosti $\hat{\mu}_i$ po y_i tj. $\dfrac{\partial \hat{\mu}_i}{\partial y_i}$. Leveridž je jedna od indikacija koliko je posmatrana observacija uticajna. Taj izvod je zapravo dijagonalni element matrice H tj. h_{ii}, gde su $h_{ii} \in [0, 1]$. Predstavlja meru odstupanja posmatrane observacije od preostalih zapažanja u prostoru nezavisnih promenjivih. Tačka sa visokim leveridžom je zapažanje sa ekstremnom vrednosti nezavisne promenljive. Trag matrice H u oznaci $tr(H)$ jeste suma njenih dijagonalnih elemenata tj. $tr(H) = \sum_{i=1}^{n} h_{ii}$. Važi da je $tr(H) = p$ gde je p broj nepoznatih parametara modela. Iz ovoga sledi da je prosečan leveridž za svaku observaciju p/n. Observacije koje imaju veći leveridž od p/n, na primer $2p/n$, treba dodatno ispitati.

Cook's rastojanje i Dfbeta – Dfbeta predstavlja promenu ocenjenog parametra kada je observacija i obrisana tj. kada se ne uzima u obzir pri analizi. Definišemo je kao

$$D_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_i)' X' X (\hat{\beta} - \hat{\beta}_i)$$

Cook's rastojanje se koristi da se ispita u kojoj meri svaka observacija utiče na kompletan skup ocenjenih parametara modela. Ocene sa i bez svake observacije mogu da se porede koristeći sledeću statistiku

$$C_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_{(i)})' X' W X (\hat{\beta} - \hat{\beta}_{(i)})$$

$\hat{\beta}_{(i)}$ predstavlja ocenu parametra β bez i-te observacije. Data statistika meri kvadratno rastojanje između $\hat{\beta}$ i $\hat{\beta}_{(i)}$.

Kako izračunavanje $\hat{\beta} - \hat{\beta}_{(i)}$ zahteva ponovo fitovanje modela bez $i-$ te observacije da bi se ovo izbeglo koristi se sledeća aproksimacija koja predstavlja kombinaciju leveridža i reziduala

$$C_i \approx \frac{h_{ii}}{p(1-h_{ii})} \cdot (r_p)_i^2$$

gde je p broj parametara modela a h_{ii} dijagonalni elementi matrice H. Podsetimo, matrica H kod uopštenog linearnog modeliranja je

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$$

Još jedan od načina da se ispita da li je observacija uticajna jeste da se izračuna devijansa modela bez te observacije. Ako je promena u devijansi velika to ukazuje da je ta observacija od značaja.

Rezidualni grafik

Najčešće se koriste sledeći tipovi rezidualnih grafika :

1. Crtaju se reziduali naspram fitovanih vrednosti $\hat{\eta}$. Ovaj grafik treba da prikaže uzorak gde reziduali imaju konstantu srednju vrednost 0 i konstantnu varijansu. Odstupanje od ovih navedenih pravila ukazuje na neke nedostatke kao što su : pogrešno izabrana link funkcija, pogrešan izbor nezavisnih tj. objašnjavajućih promenljivih, izostavljanje ne-linearnih članova kod linearnog prediktora itd.

2. Crtaju se reziduali naspram nezavisnih promenljivih. Treba da se dobije model po kome reziduali imaju osobine navedene u 1. Odstupanje od ovog modela može da ukazuje na pogrešan izbor link funkcije, nekorektan izbor ili propust nekih nelinearnih članova linernog prediktora.

3. Crtaju se reziduali po redosledu zabeleženih observacija iz skupa podataka. Na ovaj način se može detektovati moguća zavisnost između observacija.

4. Grafik normalne raspodele reziduala crta vrednosti reziduala naspram njihovih očekivanih vrednosti. Dati su sa

$$\phi^{-1}[(i-3/8)/(n+1/4)]$$

gde je ϕ^{-1} inverz funkcije standardne normalne raspodele, n je obim uzorka a i je redosled observacije tj. opažanja. Grafik treba da prikazuje **pravu** liniju sve dok pretpostavljamo da reziduali imaju približno normalnu raspodelu.

5. Takođe uz pomoć reziduala može se detektovati izostavljena kovarijabla x. Ovo se radi na sledeći način. Fituje se model sa x kao zavisnom promenljivom i koristi se isti model kao za y. Zatim, za oba modela treba sačuvati nestandardizovane reziduale i iste prikazati na grafiku , jedne naspram drugih. Ako je moguće uspostaviti neku vrstu veze između njih, to ukazuje da x treba da bude uključena u modelu.

Provera link funkcije

Da bismo proverili link funkciju potrebna nam je tzv. ,,prilagođena" zavisna varijabla koju obeležavamo sa z_i.

$$z_i = g(\hat{\mu}_i)$$

Koristeći Tejlorov razvoj prvog reda za $g(y_i)$ dobijamo

$$g(y_i) \approx g(\mu_i) + \dot{g}(\mu_i)(y_i - \mu_i) \approx x_i'\beta$$

Grafički , crtaju se $z_i + \dot{z}_i(y_i - \hat{\mu}_i)$ naspram $\hat{\eta}_i = x_i'\hat{\beta}$. Grafik treba da prikazuje tačke koje ,,leže" na približno **pravoj** liniji, a stroga zakrivljenost govori o tome da smo pogrešno odabrali link funkciju.

Transformacija promenljivih

Parcijalni rezidualni grafik može da se koristi da se otkrije da li je potrebno transformisati neku od objašnjavajućih promenljivih. Kod uopštenog linearnog modeliranja parcijalni reziduali su definisani kao

$$\dot{g}(\hat{\mu}_i)(y_i - \hat{\mu}_i) + x_{ij}\hat{\beta}_j$$

Drugačije, možemo ih zapisati kao

$$p_r = z - \hat{\eta} + \hat{\gamma}x$$

gde je z prilagođena zavisna promenljiva, $\hat{\eta}$ je fitovani linearni prediktor , x je nezavisna promenljiva a γ je ocena parametra za odgovarajuću nezavisnu promenljivu. Na grafičkom prikazu crtaju se parcijalni reziduali naspram promenljive x. Ako se dobije približno linearna veza između ovih podataka transformacija nezavisne promenljive nije potrebna. Pojava krive linije na grafiku znači da promenljiva x treba da bude transformisana.

Izbor modela

U eksponencijalnoj familiji raspodela širok je skup mogućih dostupnih modela, tako da je veoma važno izabrati najbolji model. Izbor modela podrazumeva traženje najjednostavnijeg ,,razumnog" modela koji će na adekvatan način da opiše registrovane podatke. Svaka objašnjavajuća promenljiva koja je dodata u model poboljšava fitovanje, a samim tim se povećava i broj nepoznatih parametara modela koje je potrebno oceniti i opada greška $y_i - \hat{y}_i$ fitovanja. Veliki broj parametara u modelu ukazuje da je fitovanje blizu registrovanih vrednosti podataka ali u ovom slučaju ocene parametara $\hat{\beta}_j$ za $j = 0, 1, 2, \ldots, p-1$ imaju nisku preciznost tj. visoku standardnu devijaciju. Tako, model sa malim broj parametara β dovodi do manje dobrog fitovanja modela. U ovom slučaju ocene parametara imaju visoku preciznost tj. malu standardnu devijaciju. Broj stepeni slobode tj. broj ocenjenih parametara obezbeđuje odgovarajuću meru složenosti za skup modela. Cilj je naći model koji će da balansira odnos između broja parametara i ,,dobrote" fitovanja.

Mogu se razlikovati dva različita tipa izbora modela :

1. Može da se razmatra složeniji model koji sadrži puno parametara. Iz takvog modela možemo ,,izvući" submodel eliminacijom nekih parametara. Takođe moguće je i poći od jednostavnijeg modela pa mu dodati neke parametre.

2. Mogu se razmatrati nekoliko različitih funkcija modela obično sa različitim skupom parametara.

Najpoznati kriterijumi za izbor modela su sledeći

- AIC (Akaike's Information criterium)
- BIC (Bayesian Information criterium)

Oni su definisani na sledeći način

$$AIC \equiv -2l + 2p \qquad BIC \equiv -2l + p\ln(n)$$

gde je l logaritam funkcije verodostojnosti a p je broj nepoznatih parametara u modelu tj. broj parametara koje je potrebno oceniti. Dobro fitovanje znači velika vrednost za funkciju verodostojnosti pa stoga malu vrednost za $-2l$. Sada ćemo zbog jednostavnosti pokazati kako izgleda AIC i BIC kriterijum kod normalnog linearnog modela koji je specijalan slučaj uopštenog linearnog modela. Isti princip se primenjuje u slučaju da je u pitanje uopšteno linearno modeliranje. U ovom slučaju zavisna promenljiva $Y \sim N(\mu, \sigma^2)$. Funkcija gustine za datu raspodelu je

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Kod normalnog linearnog modela je $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1}$ pa je $\mu_i = \mu_i(\beta)$ i $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_{p-1} x_{ip-1}$. Zatim je

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$$

Funkcija verodostojnosti

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(y_i; \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}} \tag{17}$$

Logaritmovanjem jednačine (17) dobijamo logaritam verodostojnosti l

$$l(\beta, \sigma^2) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{2\sigma^2} \tag{18}$$

Kada smo ocenili parametre beta sada je

$$l(\hat{\beta}, \sigma^2) = -n \ln(\sqrt{2\pi}) - n \ln(\sigma) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$

Izraz $-2l$ nam je potreban da bismo izračunali vrednost AIC i BIC kriterijuma pa je kod ovog modela

$$-2l = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

gde su prva dva člana u izrazu za $l(\hat{\beta}, \sigma^2)$ obično izostavljena. Označimo sumu u prethodnoj jednačini sa S pa dobijamo sledeći zapis

$$-2l = \frac{S}{\sigma^2}$$

Sada je $AIC = \frac{S}{\sigma^2} + 2p$ i $BIC = \frac{S}{\sigma^2} + p \ln(n)$

Dobro fitovanje podrazumeva malu vrednost za $S = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$.

Član $\frac{S}{\sigma^2}$ predstavlja pristrasnost modela a drugi član je član za varijansu kod oba kriterijuma.

Model sa velikim brojem parametara u modelu i sa dobrim fitovanjem ima malu vrednost za $\frac{S}{\sigma^2}$ ali veću vrednost za varijansni deo. Obrnuto, model sa malim brojem parametara ima veću vrednost za $\frac{S}{\sigma^2}$ ali mali varijansni deo. Bira se model koji ima najnižu vrednost AIC ili BIC kriterijuma. Da bi se mogle porediti vrednosti AIC i BIC kriterijuma, analize moraju biti bazirane na istom skupu observacija.

Standardni **statistički softveri** isključuju iz analize bilo koji slučaj kojem nedostaje vrednost neke nezavisne promenljive. Dakle, treba izvršiti odabir modela na skupu slučajeva sa svim vrednostima objašnjavajućih promenljivih. U većini softvera dostupni su takozvani ,,stepwise" algoritmi tj. stepwise regresija (Postepena regresija) koja služi za odabir modela.

3 Modeliranje neprekidnih varijabli pomoću ulm

3.1 Neprekidne varijable u aktuarstvu

Neprekidne varijable koje su od interesa osiguravajućim kompanijama su **iznos** zahteva za odštetu i **vreme** koje protekne između podnošenja zahteva za odštetu i njegove isplate.
Iznos zahteva za odštetu (**Claim amount**) je primer neprekidne varijable u aktuarstvu. Neprekidne promenljive se još nazivaju i ,,interval" promenljive da se ukaže na to da mogu da uzimaju bilo koju vrednost iz intervala na realnoj osi.
Vreme koje protekne od podnošenja zahteva pa do njegove isplate u literaturi je poznato kao **,,settlement delay"** . To je takođe neprekidna varijabla i obično se izražava kao ceo broj dana ili meseci. Ako imamo skup polisa sa zahtevima za odštetu onda se procenat zahteva koji su isplaćeni brže nego posmatrani konkretan zahtev u analizi podataka osiguranja naziva **operativno vreme**.
Na primer, ako je operativno vreme nekog zahteva za odštetu 20% to onda znači da je taj zahtev plaćen u prvih 20% zajedno sa grupom njemu sličnih. Zahtev koji je poslednji plaćen ima operativno vreme 100%.
Neprekidne varijable u aktuarstvu su obično nenegativne i iskošene udesno. Takve varijable se mogu modelirati na sledeća dva načina :

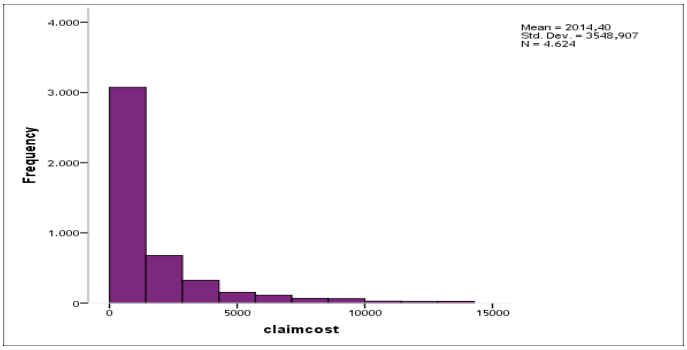
- Transformisati zavisnu varijablu Y tako da transformacija ima normalnu raspodelu i dalje nastaviti rad sa normalnim modelom

  $$g(Y) \sim N(\mu, \sigma^2)$$

  gde je $\mu = x'\beta$.

- Pomoću uopštenog linearnog modeliranja, koristeći raspodele kao što su gama i inverzna Gausova.

Na sledećoj slici je prikazan histogram jedne neprekidne promenljive tj. iznos zahteva za odštetu i date frekvencije. Od ukupno 67856 polisa osiguranja njih 4624 ima zahtev za odštetu. U pitanju su polise osiguranja vozila.



Slika 3.1.1 Frekvencija iznosa zahteva (claimcost) za odštetu

Sa slike se vidi da je neprekidna varijabla, iznos zahteva za odštetu, nenegativna i iskošena udesno.

3.2 Gama regresija

Gama slučajna promenljiva je najčešće najpogodnija za fitovanje promenljivih kao sto je iznos zahteva za odštetu (eng.claim cost, claim amount) ili pak godišnji prihod osiguranika (eng.annual income).

Neka je $Y \sim G(\mu, \nu)$. Njena funkcija verovatnoće je

$$f(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{y\nu}{\mu}\right)^{\nu} e^{-y\nu/\mu} , \quad y > 0$$

i važi da je $E(Y) = \mu$, $Var(Y) = \dfrac{\mu^2}{\nu}$.

Male vrednosti parametra ν utiču na to da je raspodela dosta iskošena udesno tj. da ima dugačak ,,rep" udesno.

Uopšten linearni model sa gama raspodelom zavisne promenljive je dat sa

$$Y \sim G(\mu, \nu), \quad g(\mu) = x'\beta$$

Najčešće se za link funkciju bira log.

Primer: Osiguranje vozila

Analizirani skup podataka obuhvata 67856 polisa osiguranja vozila od kojih 4624 ima zahtev za odštetu. Raspodela iznosa zahteva za odštetu je predstavljena na slici 3.1.1. Cilj statističke analize je da se utvrdi da li je gama pogodna raspodela za modeliranje iznosa zahteva za odštetu.

Zavisna promenljiva je iznos zahteva (claim cost) .

Nezavisne promenljive su : Starost vozača (agecat) je kategorijalna promenljiva koja ima šest kategorija a to su : 1 (najmlađi), 2, 3, 4, 5, 6 (najstariji) .

Pol (gender), koja je binarna promenljiva (male/female) .

Tip vozila koja je isto kategorijalna sa 13 kategorija (bus, minibus, sedan . . .).

U analizu je uključeno $n = 4624$ polise osiguranja što čini 6,8% od ukupnog broja polisa. Na osnovu statističke analize dobijeni su sledeći rezultati

statistika	vrednost	br. stepeni slobode df	vrednost/df
devijansa	7243	4605	1.57
skalirana devijansa	5523.7	4605	1.199
skalirana Pirson χ^2	10275.6	4605	-
logaritam verodostojnosti	-39611.16	-	- -
AIC	79262.313	-	-
BIC	79391.09	-	-

Tabela 3.2.1 Procena adekvatnosti modela

Iz tabele se vidi da vrednost devijanse (skalirana) iznosi 1.199 tj. približno 1.2 . To govori o tome da gama raspodela nije najprikladnija raspodela za modeliranje iznosa zahteva za odštetu jer je vrednost devijanse malo veća od jedinice, ali je prihvatljiva za modeliranje. Idealno bi bilo da je vrednost 1 tj. da jako malo odstupa od jedinice.

Fitovani model[11] je

$$\hat{\mu} = e^{7.683 + 0.313x_1 + \cdots - 0.07x_5 - 0.166x_6 - 0.412x_7 + \cdots + 0.146x_{18}}$$

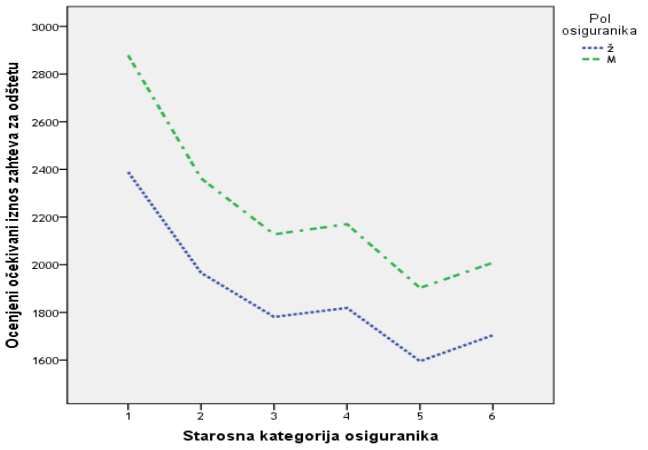
gde su x_1 do x_5 indikator varijable za promenljivu agecat tj. starosna kategorija vozača, x_6 je indikator varijabla za ženski pol, x_7 do x_{18} su indikator varijable za tip vozila tj. veh-body. Ocenjeni očekivani iznos zahteva za odštetu ako je u pitanju bazna kategorija (agecat=6, gender=M, vehbody=UTE) je $e^{7.683} \approx 2171.12$. Ako osiguranik (vozač) pripada nekoj drugoj grupi kao na primer agecat=1, gender=F, veh-body=bus onda se vrednost 2171.12 množi sa $e^{0.313} \cdot e^{-0.166} \cdot e^{-0.412}$ što očitavamo iz tabele dobijene statističkom analizom u programu SPSS[12].
Sada je ocenjeni očekivani iznos zahteva
2171.12 × 1.367 × 0.847 × 0.662=1664.15 . Slično, za polisu osiguranja gde je starosna kategorija 2 tj. agecat=2, pol vozača je muški i vozilo je kamion (truck) ocenjeni očekivani iznos zahteva za odštetu je

$$\hat{\mu} = e^{7.683+0.125+0+0.146}$$

Tačnije

$$\hat{\mu} = 2171.12 \times 1.133 \times 1 \times 1.157 \approx 2846.08$$

Po ovom modelu, na osnovu analiziranog uzorka možemo zaključiti da su žene koje pripadaju najmlađoj starosnoj kategoriji i čiji je tip vozila autobus manje rizične od muškaraca starosne kategorije 2 čiji je tip vozila kamion jer je njihova očekivana vrednost zahteva za odštetu manja tj. 1664.15 <2846.08 .



Slika 3.2.1 Predviđene vrednosti za iznos zahteva za odštetu u zavisnosti od starosti i pola osiguranika

[11]Baza podataka za ovaj primer je dostupna na sajtu http://www.acst.mq.edu.au/GLMsforInsuranceData/
[12]Tabela je data u prilogu na kraju rada

Sa slike se vidi da je za žene (plava linija) manji očekivani iznos zahteva za odštetu nego za muškarce (zelena linija) bez obzira na starosnu kategoriju .

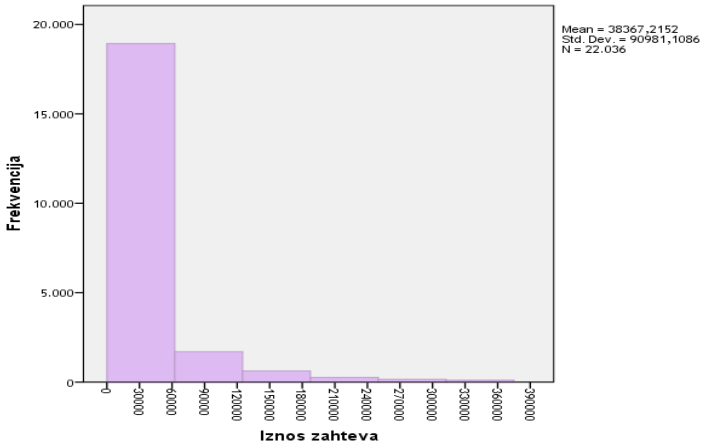
	Wald χ^2	df	p vrednost
odsečak	6608.13	1	< 0.0001
starosna kategorija	36.424	5	< 0.0001
pol	21.993	1	< 0.0001
tip vozila	33.505	12	0.001

Tabela 3.2.2 Testiranje koeficijenata modela

Sve promenljive u modelu su statistički značajne jer je p vrednost manja od 0.05 . Test odnosa verodostojnosti poredi fitovani model sa modelom koji sadrži samo odsečak, tj. on testira da li su svi β koeficijenti osim odsečka jednaki nuli. U ovom primeru vrednost te statistike iznosi 103.53 a p vrednost je manja od 0.0001 što znači da odbacujemo hipotezu da su svi koeficijenti u postavljenom modelu jednaki nuli.

Primer: Osiguranje u slučaju telesnih povreda

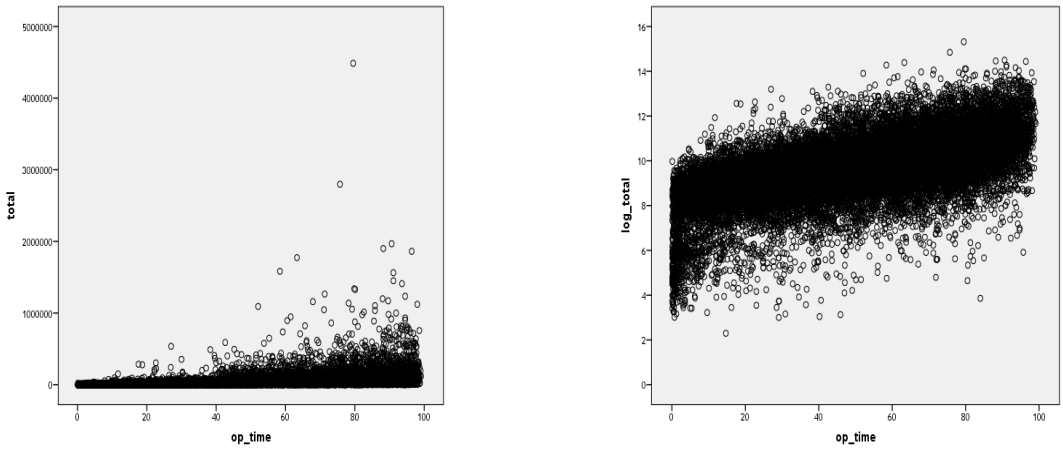
Skup podataka sadrži informacije o 22036 isplaćenih zahteva za odštetu. Ti zahtevi za odštetu su nastali usled nesreća koje su se dogodile u periodu jul 1989–jun 1999. godine.



Slika 3.2.2 Histogram iznosa zahteva za odštetu

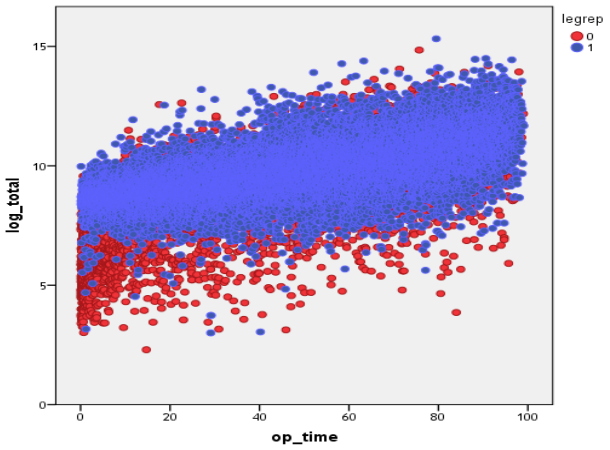
Od ukupno 22036 polisa osiguranja samo 1825 polisa ima zahtev za odštetu preko 100 000$. Promenljiva koja predstavlja plaćeni iznos zahteva (zavisna) se kreće u intervalu od 10-4490000. Od nezavisnih promenljivih razmatramo **stepen telesne povrede**, koja je kategorijalna promenljiva i uzima vrednosti od 1 (nema povrede), 2, 3, 4, 5, 6 (fatalna) i 9 što znači da nije zabeležen podatak.
Legalna reprezentacija tj. zastupanje koja je binarna varijabla (ne/da) .
Operativno vreme koja je neprekidna promenljiva.

Primer: Hoćemo da ispitamo vezu između iznosa zahteva i operativnog vremena. Kako su obe promenljive neprekidne možemo iskoristiti scatterplot da ih prikažemo.



Slika 3.2.3 Operativno vreme naspram iznosa zahteva za odštetu

Prva slika prikazuje operativno vreme naspram iznosa zahteva za odštetu (total). Sa slike se vidi da veza nije linearna. Međutim, ako se napravi logaritamska transformacija promenljive iznosa zahteva za odštetu (log_total) i grafički predstavi naspram operativnog vremena dobija se veza koja sada liči na linearnu. Cilj logaritamske transformacije je stabilizacija varijanse. Tj. na prvoj slici varijansa raste sa operativnim vremenom i srednjom vrednošću dok je na drugoj slici približno konstantna.



Slika 3.2.4 Operativno vreme naspram logaritma iznosa zahteva (log_total)

Na ovoj slici bojom su predstavljeni zahtevi za odštetu kod kojih osiguranik ima pravno zastupanje (plava boja) i crvena kod kojih nema legalne reprezentacije tj. pravnog zastupanja. Zahtevi za odštetu kod kojih osiguranik nema advokata tj. pravnog zastupnika imaju manje operativno vreme od ostalih što znači da su isplaćeni brže u odnosu na ostale.

Pošto je iznos zahteva za odštetu (total) neprekidna promenljiva iskošena udesno za modeliranje je pogodno uzeti gama raspodelu, sa link funkcijom log, a nezavisne promenljive su operativno vreme koja je neprekidna i legalna reprezentacija koja je kategorijalna tj. binarna promenljiva.

Uopšten linearan model je

$$Y \sim G(\mu, \nu)\ ; \quad \ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

x_1 je promenljiva koja predstavlja operativno vreme, x_2 je legalna reprezentacija tj. pravno zastupanje (uzima vrednost 1 ako ima pravnog zastupanja) a $x_1 x_2$ interakcija između te dve promenljive.
Uz pomoć softvera SPSS dobijamo sledeće rezultate

	odsečak	operativno vreme	pravno zastupanje NE/DA	op.vreme × pr.zast(da)
$\hat{\beta}$	8.212	0.038	0/0.467	-0.005
$se(\hat{\beta})$	0.021	0.0004	-/0.027	0.0005
$e^{\hat{\beta}}$	3684.25	1.039	1/1.595	0.995
χ^2	151573.44	8971.87	-/294.224	94.99
p vrednost	<0.0001	<0.0001	-/< 0.0001	< 0.0001

Tabela 3.2.3 Ocene parametara modela

Sada je očekivana vrednost iznosa zahteva za odštetu na osnovu uzorka

$$\hat{\mu} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2}$$

Za polise osiguranja kod kojih je prisutno pravno zastupanje za 59.5% je veći očekivani iznos zahteva za odštetu, što vidimo iz tabele rezultata.
Ako je x_2=0 tada je

$$\hat{\mu} = e^{8.212 + 0.038 x_1}$$

a u slučaju da je x_2=1 tada je

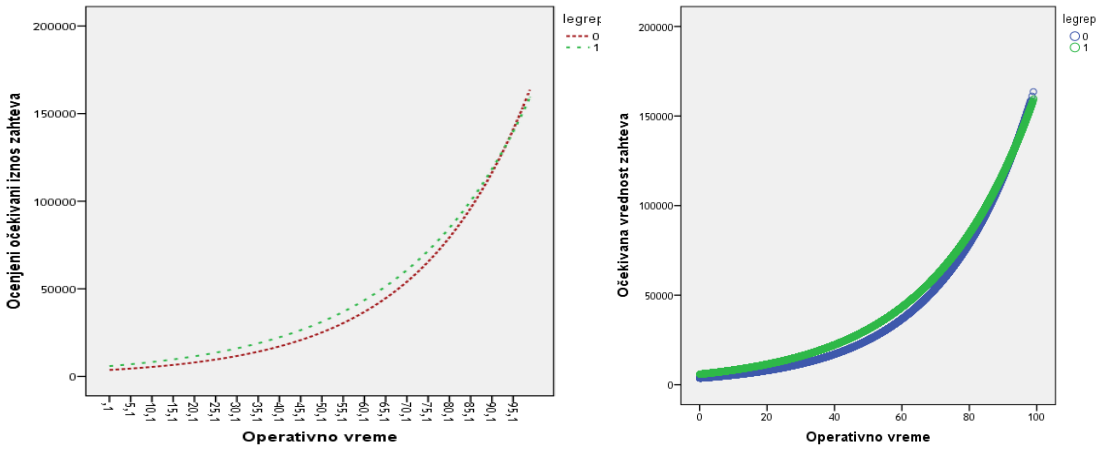
$$\hat{\mu} = e^{8.212 + 0.038 x_1 + 0.467 - 0.005 x_1}$$

Ocenjena vrednost za parametar ν je $\hat{\nu}$=0.999 .

	Wald χ^2	df	p vrednost
odsečak	385196.985	1	<0.0001
legrep	294.224	1	<0.0001
op.vreme	19509.369	1	<0.0001
legrep∗ op.vreme	94.989	1	<0.0001

Tabela 3.2.4 Testiranje značajnosti koeficijenata modela

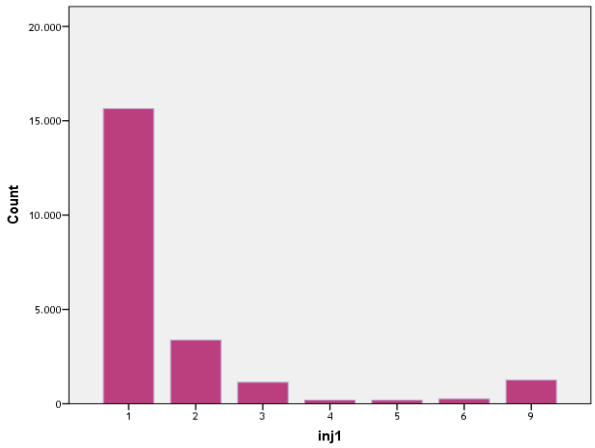
Rezultati analize pokazuju da su sve nezavisne promenljive koje su unete u model statistički značajne jer je p vrednost manja od 0.05. Vrednost devijanse iznosi 25411.69 a broj stepeni slobode te statistike je 22032 pa je D/df=1.15 . Tabela je data u prilogu rada.



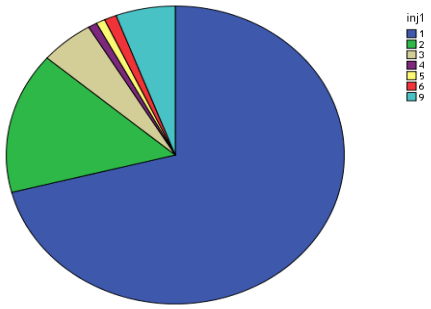
Slika 3.2.5 Očekivani iznos zahteva za odštetu naspram operativnog vremena

Sa slike možemo zaključiti da sa porastom operativnog vremena eksponencijalno raste i očekivani iznos zahteva za odštetu. Tj. ,,teži" zahtevi za odštetu su kasnije isplaćeni.

Na sledećoj slici je prikazana kategorijalna promenljiva stepen telesne povrede (inj1) i njihove frekvencije (count).



Slika 3.2.6 Stepen telesne povrede i njihove frekvencije



Slika 3.2.7 Kružni dijagram za stepen telesne povrede

Najveći procenat (71%) obuhvataju telesne povrede kodirane sa 1 tj. najblaži tip ozlede, dok veoma ozbiljne povrede kategorije 5 i fatalne povrede obuhvataju samo 0,9% i 1,2% slučajeva u posmatranom uzorku.

3.3 Inverzna Gausova regresija

Inverzna Gausova raspodela je neprekidna raspodela data sa funkcijom gustine

$$f(y) = \frac{1}{\sqrt{2\pi y^3}\sigma} exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}, \qquad y > 0$$

U oznaci $Y \sim IG(\mu, \sigma^2)$ i sa osobinama da je

$$E(Y) = \mu \quad Var(Y) = \sigma^2\mu^3$$

IG raspodela je više iskošena nego gama raspodela i ima oštriji ,,rep" udesno. Zato se koristi u situacijama ekstremne kosine.
Za modeliranje neprekidnih varijabli u aktuarstvu kao što je iznos zahteva za odštetu, veličina zahteva koju osiguravajuća kompanija treba da plati tj. gubitak uglavnom se koriste te neprekidne varijable poput IG raspodele, gama, weibull, Pareto itd.

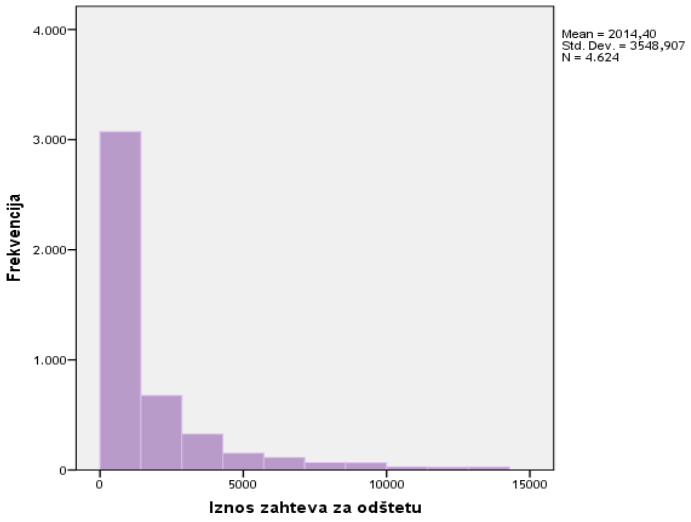
Uopšteno linearno modeliranje sa IG raspodelom je

$$Y \sim IG(\mu, \sigma^2) \quad g(\mu) = x'\beta$$

gde se za funkciju g načešće bira log.

Primer: Osiguranje vozila

Skup podataka obuhvata 67856 polisa osiguranja vozila od kojih 4624 ima zahtev za odštetu. Iznosi zahteva za odštetu kreću se između 0-57000 (0 ako polisa nema zahtev za odštetu).
Na sledećoj slici je prikazan histogram iznosa zahteva za odštetu.



Slika 3.3.1 Histogram iznosa zahteva za odštetu

Zbog jasnijeg pregleda histograma maksimalna vrednost na horizontalnoj osi iznosi 15000$. Sa prikaza je izostavljeno 65 zahteva čija se vrednost kreće između 15000-57000.
U ovom primeru za modeliranje iznosa zahteva za odštetu (claimcost) koristićemo IG raspodelu sa link funkcijom ln. Nezavisne promenljive u modelu biće :

- Starost osiguranika (agecat) -kategorijalna promenljiva sa vrednostima : 1 (najmlađi) , 2, 3, 4, 5, 6 (najstariji).
- Pol osiguranika (gender) koja je takođe kategorijalna promenljiva sa vrednostima F (ženski) i M (muški).
- Oblast stanovanja osiguranika (area) - kategorijalna promenljiva koja uzima vrednosti A, B, C, D, E, F.

Kako je $g(\mu) = \ln(\mu)$ imamo

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{10} x_{10} + \beta_{11} x_{11}$$

gde su x_1, x_2, x_3, x_4, x_5 indikator varijable za starosnu grupu osiguranika, x_6 indikator varijabla koja ukazuje na ženske osobe, i $x_7, x_8, x_9, x_{10}, x_{11}$ indikator promenljive za oblast stanovanja osiguranika.
Treba napomenuti da se kod kategorijalnih promenljivih bira jedna vrednost (nivo) koji se naziva bazni nivo na osnovu koga se upoređuju vrednosti ostalih i izvode zaključci analize. U ovom primeru za starosnu grupu vozača to je vrednost 6, za pol je M, i za oblast stanovanje F.
Pomoću softvera SPSS dobijeni su rezultati analize prikazani na sledećoj strani

	Wald χ^2	df	p vrednost
odsečak	53988.56	1	<0.0001
starost osiguranika	16.013	5	0.007
pol	9.465	1	0.002
oblast stanovanja	12.558	5	0.028

Tabela 3.3.1 Testiranje koeficijenata modela

Sve nezavisne promenljive su statistički značajne jer je p vrednost manja od 0.05.

statistika	vrednost	st.slobode-df	vrednost/df
skalirana devijansa	4624	4612	1.003
skalirana Pirson χ^2	4897.19	4612	1.06
log.verodostojnosti	-38568.16		
AIC	77162.32		
BIC	77246.03		

Tabela 3.3.2 Statistike za adekvatnost fitovanja

Vrednost devijanse tj. statistike za procenu adekvatnosti fitovanja modela je 1.003 što ukazuje na to da je model dobar i da je inverzna Gausova raspodela bolja od gama raspodele za modeliranje iznosa zahteva za odštetu.
Napomena : U analizu je uključeno n=4624 polise osiguranja vozila jer one imaju **pozitivan** zahtev za odštetu. Preostale 63232 polise su isključene jer one nemaju zahtev za odštetu tj. $y = 0$. Broj nepoznatih parametara u ovom primeru je p =12 $(\beta_0, \beta_1, \ldots, \beta_{11})$ pa je zato broj stepeni slobode za statistike, $df = n - p$ i iznosi 4612.

parametar	$\hat{\beta}$	$se(\hat{\beta})$	$e^{\hat{\beta}}$	Wald χ^2	p vrednost
odsečak	7.898	0.15	2691.7	2892.7	<0.0001
agecat=1	0.32	0.12	1.37	7.39	0.007
agecat=2	0.16	0.1	1.17	2.57	0.11
agecat=3	0.07	0.096	1.07	0.49	0.48
agecat=4	0.06	0.096	1.06	0.42	0.52
agecat=5	-0.054	0.103	0.95	0.27	0.6
agecat=6	0		1		
gender=F	-0.153	0.05	0.86	9.46	0.002
gender=M	0		1		
area=A	-0.35	0.13	0.7	7.87	0.005
area=B	-0.38	0.13	0.68	9.13	0.003
area=C	-0.28	0.125	0.75	5.1	0.024
area=D	-0.38	0.14	0.68	7.74	0.005
area=E	-0.21	0.15	0.81	2.12	0.14
area=F	0		1		

Tabela 3.3.3 Ocene parametara modela

U poslednjoj tabeli agecat je starosna kategorija osiguranika, gender je pol a area je oblast stanovanja osiguranika. Sada je ocenjen model

$$\hat{\mu} = e^{7.898 + 0.32x_1 + 0.16x_2 + \cdots - 0.054x_5 - 0.153x_6 - 0.35x_7 + \cdots - 0.21x_{11}}$$

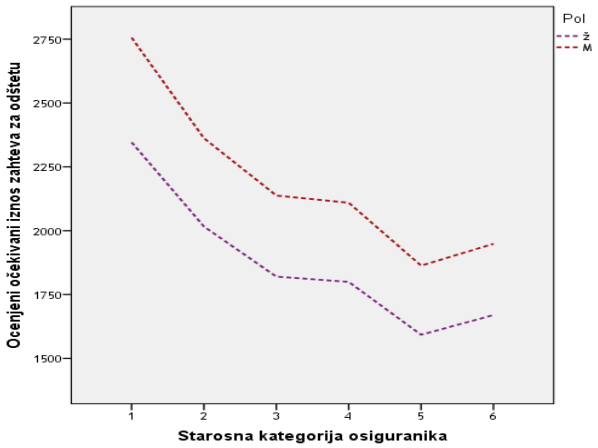
Za osiguranike koji su u starosnoj kategoriji 6 (agecat=6), muškarci, i koji stanuju u oblasti F očekivani iznos zahteva za odštetu je $\hat{\mu} = e^{7.898} = 2691.7$
Na primer, ako je osiguranik žena koja pripada starosnoj kategoriji 1 (agecat=1) i stanuje u oblasti D (area=D) tada je očekivani iznos zahteva za odštetu

$$\hat{\mu} = e^{7.898} \cdot e^{0.32} \cdot e^{-0.153} \cdot e^{-0.38}$$

Dobije se da je $\hat{\mu} = 2156.52$.
Ako posmatramo grupu muških osiguranika (vozača) koji su u starosnoj grupi 1 (agecat=1), najmlađi, i stanuju u oblasti F tada je očekivani iznos zahteva za odštetu

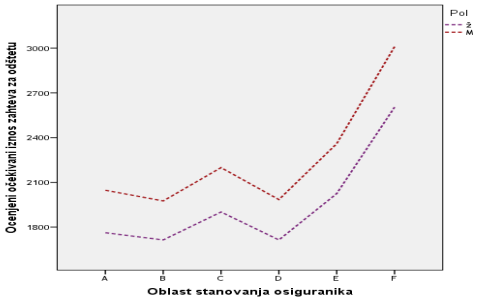
$$\hat{\mu} = e^{7.898} \cdot e^{0.32} = 3687.63$$

Za 37% je veći očekivani iznos zahteva nego za vozače starosne kategorije 6.



Sika 3.3.2 Očekivani iznos zahteva za odštetu prema starosnoj kategoriji i polu osiguranika

Možemo zaključiti da je za žene (ljubičasta linija) manji očekivani iznos zahteva za odštetu nego za muškarce (crvena linija) bez obzira na starosnu kategoriju.



Slika 3.3.3 Očekivani iznos zahteva za odštetu prema oblasti stanovanja i polu osiguranika

Bez obzira na pol, osiguranici koji stanuju u oblasti A, B ili D imaju manji očekivani iznos zahteva u poređenju sa ostalim oblastima.

Napomena : inverzan Gausov model kao i gama model za iznos zahteva se koristi pod uslovom da polisa ima zahtev za odštetu koji je veći od nule. Na primer, ocenjeni očekivani iznos zahteva 2156.52 za polise sa agecat=1, gender=F i area=D primenjuje se na polise koje ,,naprave" pozitivan zahtev za odštetu. Ako je iznos zahteva 0 za polise kod kojih nema zahteva za odštetu onda je ukupan očekivani iznos zahteva za sve polise iz navedene grupe bitno niži nego 2156.52. Tj. ako se u analizu uzmu i polise čija je vrednost zahteva 0 onda će ukupna očekivana vrednost zahteva biti manja.

4 Modeliranje kategorijalnih varijabli pomoću ulm

Kategorijalne promenljive uzimaju jednu vrednost iz diskretnog skupa vrednosti. Na primer, ako posmatramo tip vozila skup mogućih vrednosti je : automobil, autobus, kamion, voz, bicikl itd. Osoba može biti muško ili žensko, zatim status zaposlenja osobe itd.
Najjednostavniji primer kategorijalne promenljive je binarna promenljiva što znači da može da uzima samo dve vrednosti koje su obično kodirane sa 0 i 1. Sa 1 se obeležava realizacija događaja od interesa kao npr. ako polisa osiguranja sadrži zahtev za odštetu ubeležićemo vrednost 1 a u suprotnom vrednost 0. Poznati su još termini ,,uspeh" i ,,neuspeh".
Razlikuju se dve klase kategorijalnih promenljivih a to su

- Nominalne
- Ordinalne

Kategorije ovih promenljivih se obično obeležavaju numerički tj. brojevima.
Ordinalne promenljive su one kod kojih su kategorije uređene kao na primer kod stepena obrazovanja možemo imati : 1–završena osnovna škola, 2–završena srednja škola, 3–završena visoka škola, 4–završen fakultet.
Nominalne su one kod kojih nemamo takvo uređenje i vrednost 2 ne znači bolje od 1. Na primer, boja automobila pa imamo vrednosti 1–plava, 2–crna, 3–crvena itd. Primer nominalne promenljive je tip zahteva za odštetu kod osiguranja kuće (požar, krađa, šteta od oluje). Primer ordinalne promenljive u osiguranju u slučaju saobraćajne nesreće je stepen povrede (nema povrede, povreda, fatalna povreda).

Posmatrajmo kategorijalnu promenljivu sa k kategorija koje su označene sa $1, 2, \ldots, k$. Zavisnu promenljivu Y definišemo kao skup od $k-1$ indikator promenljive Y_j gde $j = 1, 2, \ldots, k-1$. Y_j prima vrednost 1 ako promenljiva Y pripada kategoriji j, u suprotnom uzima vrednost 0. Kažemo da je promenljiva $Y = (Y_1, Y_2, \ldots, Y_{k-1})'$ multivarijativna.

Neka je dato n nezavisnih observacija zavisne Y i neka je n_j broj koji pokazuje koliko puta se realizovala kategorija j tj. $n_j = \sum y_j$.

Zajednička raspodela za $n_1, n_2, \ldots, n_k$ je multinominalna

$$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \cdots \cdot n_k!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_k^{n_k}$$

gde je π_j verovatnoća da se realizuje kategorija j i pritom važi

$$\sum_{j=1}^{k} \pi_j = 1, \qquad n = \sum_{j=1}^{k} n_j$$

Kod ove raspodele važe sledeće osobine :

1. $E(N_j) = n\pi_j$
2. $Var(N_j) = n\pi_j(1 - \pi_j)$
3. $cov(N_j, N_k) = -n\pi_j\pi_k$

Binarnu slučajnu promenljivu definišemo kao

$$Z : \begin{pmatrix} 0 & 1 \\ 1-\pi & \pi \end{pmatrix}$$

Moguća su samo dva ishoda slučajne promenljive, koji se obično kodiraju sa 0 i 1. Ako se događaj od interesa realizovao slučajna promenliva Z će primiti vrednost 1, a u suprotnom uzeće vrednost 0. π je verovatnoća da se događaj realizuje tj.

$$P(Z = 1) = \pi \qquad \text{i} \qquad P(Z = 0) = 1 - \pi$$

Binarne slučajne promenljive su najjednostavniji slučaj kategorijalnih promenljivih. Primer binarne promenljive u aktuarstvu je da li polisa osiguranja sadrži zahtev za odštetu ili ne.
Cilj će biti objasniti kategorijalnu (binarnu) slučajnu promenljivu u terminima nezavisnih promenljivih x.
Dakle

$$Z \sim B(1, \pi)$$

Pa je $E(Z) = \pi$ i $Var(Z) = \pi(1 - \pi)$. Ako je raspodela zavisne promenljive binarna onda je **uopšteni linearni model** dat sa

$$Z \sim B(1, \pi) \qquad \text{i} \qquad g(\pi) = x'\beta$$

Količnik $\dfrac{\pi}{1-\pi}$ se naziva **„odnos šansi"** (eng. odds ratio) i predstavlja odnos dve verovatnoće.
Binarna raspodela je specijalan slučaj binomne tj. $B(n, \pi)$ kada je $n = 1$.
Ako je $Y \sim B(n, \pi)$ onda slučajna promenljiva $\dfrac{Y}{n}$ predstavlja proporciju tj. procenat uspešno realizovanih događaja.

Njena numerička obeležja su

$$E\left(\frac{Y}{n}\right) = \pi \quad Var\left(\frac{Y}{n}\right) = \frac{\pi(1-\pi)}{n}$$

Kada se radi sa binarnim podacima važno je razlikovati grupisane i negrupisane podatke. Pretpostavimo da posmatramo N različitih podgrupa jedne grupe, i da su Y_i slučajne promenljive tj. Y_i broji koliko puta se realizovao događaj od interesa u podgrupi i.

$$Y_i \sim B(n_i, \pi_i)$$

Ukupno opažanja u analizi je $n_{total} = n_1 + n_2 + \cdots + n_N$.
Observacije u okviru jedne grupe (klase) su $x_{i1}, x_{i2}, \ldots, x_{ip}$. Tabelarno to izgleda

Podgrupa				
	1	2	...	N
Uspešno	Y_1	Y_2	...	Y_N
Neuspešno	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Ukupno	n_1	n_2	...	n_N

Kada su binarni podaci ovako grupisani u klase onda je zavisna promenljiva oblika Y_i/n_i , za $Y_i \in [0, n_i]$. Slučajna promenljiva $\frac{Y_i}{n_i}$ predstavlja proporciju uspešno realizovanih događaja.

4.1 Logistički model

Neka je raspodela ishodne promenljive binarna. Model je dat sa

$$Y \sim B(1, \pi) \qquad \text{i} \qquad \ln\left(\frac{\pi}{1-\pi}\right) = x'\beta$$

Dakle, modelira se verovatnoća realizacije događaja od interesa π u funkciji nezavisnih promenljivih x-sa. Link funkcija $g(\pi) = ln\left(\frac{\pi}{1-\pi}\right)$ se naziva **logit** i obezbeđuje da vrednost verovatnoće π bude u intervalu $(0, 1)$ jer je

$$\frac{\pi}{1-\pi} = e^{x'\beta} \rightsquigarrow \pi = \frac{e^{x'\beta}}{1+e^{x'\beta}}$$

Kako je $0 < e^{x'\beta} < 1 + e^{x'\beta}$, očigledno je $0 < \pi < 1$.

Možemo reći da Bernulijeva raspodela i logit link definišu **logističku** regresiju.
Ovaj model je široko upotrebljiv za analizu podataka kod kojih je zavisna promenjiva binarna ili binomna.
Pretpostavimo sada da u modelu imamo samo jednu nezavisnu promenljivu x i hoćemo da interpretiramo kakav uticaj ima parametar β na odnos šansi.

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x \rightsquigarrow \frac{\pi}{1-\pi} = e^{\alpha+\beta x}$$

Ako se promenljiva x poveća za jednu jedinicu, onda je odnos šansi

$$\frac{\pi}{1-\pi} = e^{\alpha+\beta(x+1)} = e^{\beta} \cdot e^{\alpha+\beta x}$$

Tj. povećanje x-sa za jednu jedinicu dovodi do promene odnosa ,,šansi" za e^{β}.
Pritom ako je

- $\beta > 0$ onda se odnos ,,šansi" poveća za e^{β}
- $\beta < 0$ onda se odnos ,,šansi" smanji za e^{β}
- $\beta = 0$ onda se odnos ,,šansi" ne menja.

Pored logit linka koji je ,,najpopularniji" kada se radi sa Bernulijevim varijablama koriste se još i sledeći :

Probit link $g(\pi) = \phi^{-1}(\pi)$ gde je ϕ funkcija standardne normalne raspodele tj.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy$$

Tako je $\phi^{-1}(\pi) = x'\beta$ pa je predviđena vrednost $\pi = \phi(x'\beta)$. Ocenjena vrednost za π je $\hat{\pi} = \phi(x'\hat{\beta})$.

Komplementarni log-log link $g(\pi) = ln\{-ln(1-\pi)\}$. Sada je

$$ln\{-ln(1-\pi)\} = x'\beta \rightsquigarrow ln(1-\pi) = -e^{x'\beta} \rightsquigarrow 1-\pi = exp(-e^{x'\beta})$$

Odavde vidimo da je predviđena vrednost za π , $\hat{\pi} = 1 - exp(-e^{x'\hat{\beta}})$

Dakle

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

Ovaj model je široko upotrebljiv kod podataka čija je zavisna promenljiva binarna ili binomna. Ocene parametara β dobijamo metodom maksimalne verodostojnosti, a mere za procenu slaganja modela sa podacima su **devijansa** i **Pirson** χ^2 statistika.

Kako je devijansa kod ovog modela

$$D = -2\sum_{i=1}^{n} \{\hat{\pi}_i \ln \frac{\hat{\pi}_i}{1-\hat{\pi}_i} + \ln(1-\hat{\pi}_i)\}$$

Vidimo da ova statistika zavisi od y_i samo preko $\hat{\pi}_i$ pa zato ne daje informacije o tome koliko dobro predviđene vrednosti $\hat{\pi}_i$ fituju y_i. Stoga devijansa nije dobra mera za procenu adekvatnosti fitovanja kod modela sa binarnim podacima.
S druge strane, **Pirson** χ^2 statistika definisana kao

$$\sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

Ova statistika ima približno χ^2_{n-p} raspodelu i asimptotski je jednaka devijansi. Za razliku od devijanse ova statistika zavisi od stvarnih vrednosti y_i, ali ipak se smatra nepouzdanom merom fitovanja.

Jedan od načina da se ispitaju performanse modela za binarne podatke je preko tabele klasifikacije. Izračunavaju se fitovane verovatnoće $\hat{\pi}_i$ i svaki slučaj i (npr. svaka polisa i) se klasifikuje kao ,,ima zahtev za odštetu" ili ,,nema zahtev za odštetu", a u zavisnosti od toga da li je vrednost $\hat{\pi}_i$ veća ili manja od unapred zadate vredosti koja se naziva **,,prag"** u oznaci c. Važiće da je $\hat{y}_i = 1$ ako je $\hat{\pi}_i > c$ i $\hat{y}_i = 0$ ako je $\hat{\pi}_i \leq c$. Najčešće se za c uzima vrednost 0.5. Tabela klasifikacije je dimenzija 2×2 i poredi stvarne vrednosti sa predviđenim. Za tabelu klasifikacije se vezuju dva bitna pojma, a to su

- **Senzitivnost**
- **Specifičnost**

Senzitivnost predstavlja relativnu frekvenciju tj. odnos predviđenog broja realizovanih događaja i broja događaja koji su se realizovali u stvarnosti. To je verovatnoća da je predviđena vrednost zavisne promenljive 1 ukoliko je zaista zavisna promenljiva primila vrednost 1.

$$P\{\hat{y} = 1|y = 1\}$$

Specifičnost predstavlja odnos neuspešno realizovanih događaja na osnovu modela i stvarnog broja nerealizovanih događaja. To je verovatnoća da je predviđena vrednost zavisne promenljive 0 , ukoliko je stvarna vrednost zavisne promenljive 0 tj.

$$P\{\hat{y} = 0|y = 0\}$$

Idealna situacija je kada su senzitivnost i specifičnost blizu 1. Kako vrednost za tzv. prag raste tako je manje događaja od interesa predviđeno modelom pa senzitivnost opada a specifičnost raste. Ako je zadata vrednost za prag 0 onda je jasno da senzitivnost ima vrednost 1, a specifičnost 0.
Senzitivnost i specifičnost su obično dati u procentima.
Ne mogu se upoređivati modeli na bazi mera izvedenih iz tabele klasifikacije, jer ove mere ne možemo posmatrati nezavisno od raspodela verovatnoća u uzorcima na kojima su bazirani.

ROC (Receiver Operating Characteristic) je kriva koja grafički prikazuje senzitivnost naspram specifičnosti za svaku zadatu vrednost ,,praga". Obično se na x osi beleže vrednosti za 1 – specifičnost, a na y osi vrednosti za senzitivnost. Sve ROC krive startuju u tački (0,0) i završavaju u tački (1,1), jer ako je senzitivnost 0 onda je specifičnost 1 i obratno. ,,Dobra" **ROC** kriva je monotono rastuća i brzo raste ka vrednosti 1. Površina ispod ROC krive (AUC) je mera koja predstavlja prediktivnu sposobnost modela. To je mera sposobnosti modela u razdvajanju subjekata koji su iskusili događaj koji se posmatra u odnosu na one koji nisu. AUC je prihvaćena izvedena mera za ROC krivu.

Kao opšte pravilo koristimo da je :

- $AUC = 0.5$ nema razdvajanja
- $AUC \in [0.5, 0.7)$ loše razdvajanje
- $AUC \in [0.7, 0.8)$ prihvatljivo razdvajanje
- $AUC \in [0.8, 0.9)$ odlično razdvajanje
- $AUC \geq 0.9$ savršeno razdvajanje

Maksimalna vrednost za AUC je 1.
Ova kriva je zapravo ,,razumno" sredstvo za poređenje performansi različitih modela.

Kada se procenjuje adekvatnost modela za binarne i binomne podatke treba uzeti u obzir i pojam "**overdispersion**" tj. prevelika disperzija, koja je prilično uobičajena u praksi. Pojam se odnosi na situaciju kada opažene vrednosti Y imaju mnogo veću disperziju od očekivane $n\pi(1-\pi)$ podrazumevajući binomni model. Ta pojava se javlja usled neadekvatne specifikacije modela tj. pogrešan izbor link funkcije, usled izostavljanja nekih relevantnih objašnjavajućih promenljivih u modelu ili su podaci Y_i korelisani. Jedan od načina da se to prevaziđe je da se u model uključi disperzioni parametar ϕ tako da je $Var(Y) = \phi n\pi(1-\pi)$. U slučaju korelacije pristupa se modeliranju koje je dizajnirano za korelisane podatke. Takođe, da bi se utvrdila adekvatnost modela crtaju se reziduali (pirson i devijansni) naspram nezavisnih promenljivih.

4.2 Primer logističkog modela u osiguranju vozila

Podaci koji će se analizirati su realni i obuhvataju 67856 polisa osiguranja vozila od kojih 4624 ima zahtev za odštetu.[13] Zavisna promenljiva je pojava zahteva za odštetu. Pretpostavimo sada najjednostavniji slučaj tj. da verovatnoću zahteva za odštetu modeliramo pomoću samo jedne nezavisne promenljive x, koja je neprekidna i neka je x vrednost vozila koja se nalazi između \$0-\$350000. Nezavisnu promenljivu grupisaćemo u kategorije (grupe) u zavisnosti od vrednosti vozila.

Grupa	**Vrednost vozila** $(\cdot 10^3)$\$	**Broj polisa**
1	≤ 25	54971 (81.01%)
2	25-50	11439 (16.86%)
3	50-75	1265 (1.86%)
4	75-100	104 (0.15%)
5	100-125	44 (0.06%)
6	> 125	33 (0.05%)

Tabela 4.2.1 Vrednost vozila kao kategorijalna promenljiva sa ukupnim brojem polisa u svakoj grupi

[13]Baza podataka za ovaj primer je dostupna na sajtu http://www.acst.mq.edu.au/GLMsforInsuranceData/.

Sada možemo da postavimo model

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

x_j je indikator varijabla za grupu j. Izostavljena kategorija je grupa 1 i ona predstavlja baznu tj. referentnu grupu (kategoriju) što znači da će se ,,meriti" razlika između grupe 1 i svake druge grupe.
Uz pomoć programskog paketa SPSS dobijamo rezultate koji su prikazani u sledećoj tabeli.

	odsečak (konstanta)	kategorija I	kategorija II	kategorija III	kategorija IV	kategorija V	kategorija VI
$\hat{\beta}$	-2.647	0	0.174	0.102	-0.571	-0.397	-0.818
$e^{\hat{\beta}}$	0.07	1	1.19	1.11	0.56	0.67	0.44
$se(\hat{\beta})$	0.017	0	0.039	0.11	0.510	0.724	1.016
χ^2	23799.7	-	19.93	0.87	1.25	0.30	0.65
p	<0.0001	-	<0.0001	0.352	0.2627	0.5834	0.4205

Tabela 4.2.2 Ocene parametara modela

Interpretacija modela i rezultata : Zavisna promenljiva u ovom modelu je pojava zahteva za odštetu, koja ima Bernulijevu raspodelu. Verovatnoća da polisa osiguranja ima zahtev za odštetu se modelira u funkciji nezavisne promenljive koju smo podelili u kategorije. Vrednosti $\hat{\beta}$ su ocenjeni parametri β logističkog regresionog modela koji su neophodni da bi se dobile predviđene vrednosti $\hat{\pi}$. $se(\hat{\beta})$ predstavlja standardnu grešku ocene $\hat{\beta}$. Vrsta $e^{\hat{\beta}}$ predstavlja procenu ,,**mogućnosti**" zahteva za odštetu za polisu iz određene kategorije a u poređenju sa referentnom kategorijom.

Na primer, ,,mogućnost" zahteva za odštetu za polise iz kategorije II je za 19% veća nego za polise iz bazne kategorije I. Takođe za polise iz kategorije V ,,mogućnost" zahteva za odštetu je za 33% manja nego za polise iz kategorije I.
Ovo znači da su za vozila čija je cena u opsegu \$ 25000-50000\$ za 19% veće ,,mogućnosti" pojave zahteva za odštetu nego za vozila čija je cena manja od \$25000\$. Analogno se interpretiraju i ostali slučajevi.

4.3 Nominalna regresija

Nominalni regresioni model je dat sa

$$logit(\pi_j) = ln\frac{\pi_j}{\pi_k} = \theta_j + x'\beta_j \qquad j = 1, 2, \ldots, k-1$$

gde je $\pi_j = P(Y = j)$ a Y je nominalna zavisna promenljiva sa k kategorija, gde je bazni nivo kategorija k. Mora da bude ispunjeno da je

$$\hat{\pi_1} + \hat{\pi_2} + \cdots + \hat{\pi_k} = 1$$

Verovatnoće π_j su povezane sa nezavisnim promenljivima i modelira se odnos dve verovatnoće π_j i π_k. Sada je

$$\frac{\pi_j}{\pi_k} = e^{\theta_j + x'\beta_j}$$

$\Rightarrow \pi_j = \pi_k e^{\theta_j + x'\beta_j} \qquad j = 1, 2, \ldots, k-1.$
$\sum_{i=1}^{k} \pi_i = 1 \Rightarrow \sum_{i=1}^{k-1} \pi_i + \pi_k = 1 \Rightarrow \sum_{i=1}^{k-1} \pi_k e^{\theta_i + x'\beta_i} + \pi_k = 1$. Daljim sređivanjem izraza se dobija

$$\pi_k \cdot \left(1 + \sum_{i=1}^{k-1} e^{\theta_i + x'\beta_i}\right) = 1$$

$$\pi_k = \frac{1}{1 + \sum_{i=1}^{k-1} e^{\theta_i + x'\beta_i}}$$

Statistike za procenu adekvatnosti fitovanja su iste kao kod logističkog modela tj. može se koristiti devijansa, test odnosa verodostojnosti ili Pirsonova statistika. Sve imaju približno χ^2_{n-p} raspodelu pod pretpostavkom da model fituje dobro. Mogu se koristiti i reziduali da se proceni adekvatnost modela.

Često nije lako direktno interpretirati parametre β_j. Zato je pogodno koristiti odnos ,,šansi" jer nam to omogućava lakšu interpretaciju parametara.
Posmatrajmo model u kome zavisna promenljiva ima k kategorija i u kome imamo samo jednu nezavisnu promenljivu x koja je binarna i prima vrednost $x = 1$ (,,uspeh") ili $x = 0$ (,,neuspeh").
Odnos ,,šansi" za kategoriju j u odnosu na referentnu kategoriju k definisan je kao

$$OR_j = \frac{\pi_{ju}}{\pi_{jn}} \Big/ \frac{\pi_{ku}}{\pi_{kn}}$$

gde je π_{ju} verovatnoća da zavisna promenljiva primi vredost j ako je $x = 1$ i π_{jn} je verovatnoća da zavisna promenliva primi vrednost j ako je $x = 0$.

Model je

$$ln\left(\frac{\pi_j}{\pi_k}\right) = \beta_{0j} + \beta_{1j}x$$

za $j = 1, 2, \ldots, k-1$
Sada je

$$ln(OR_j) = ln\left(\frac{\pi_{ju}}{\pi_{ku}}\right) - ln\left(\frac{\pi_{jn}}{\pi_{kn}}\right) = \beta_{1j}$$

Iz ove jednačine sledi da je $\widehat{OR_j} = e^{\hat{\beta_{1j}}}$.
Ako je $OR_j = 3$ onda to znači da je za kategoriju j tri puta veća ,,šansa" da je $x = 1$ nego za referentnu kategoriju k. To ne znači da je verovatnoća tri puta veća. Ako je $OR_j \in (0, 1)$ onda je za kategoriju j manja „ šansa za ,,uspeh" $(x = 1)$ nego za kategoriju k.

4.4 Ordinalna regresija

Neka je zavisna promenljiva Y ordinalna sa k kategorija. Definišemo je na sledeći način

$$Y = j \qquad \text{ako je} \qquad \theta_{j-1} \leq Y^* \leq \theta_j, \quad j = 1, 2, 3, \ldots, k$$

Ovde je Y^* neprekidna promenljiva a $\theta_0, \theta_1, \theta_2, \ldots, \theta_k$ takozvane ,,prag" vrednosti.
Model za ordinalnu promenljivu Y se formuliše pomoću kumulativnih verovatnoća τ_j.

$$\tau_j = P(Y \leq j) = P(Y^* < \theta_j) \quad j = 1, 2, \ldots, k.$$

Cilj je povezati kumulativne verovatnoće τ_j sa nezavisnim promenljivima modela.
Ako zapišemo neprekidnu promenljivu Y^* u obliku $Y^* = -x'\beta + \epsilon$ tako da je $E(\epsilon) = 0$ onda je

$$\tau_j = P(Y^* < \theta_j) = P(-x'\beta + \epsilon < \theta_j) \Rightarrow P(\epsilon < \theta_j + x'\beta)$$

Dakle, da bismo izračunali kumulativne verovatnoće τ_j potrebno je da znamo raspodelu za ϵ.

Kumulativni logistički model

Pretpostavimo da slučajna promenljiva ϵ ima standardnu logističku raspodelu tj.

$$P(\epsilon \leq x) = \frac{1}{1 + e^{-x}}$$

Sada je $\tau_j = P(\epsilon \leq \theta_j + x'\beta) = \dfrac{1}{1 + e^{-(\theta_j + x'\beta)}}$, a $1 - \tau_j = \dfrac{e^{-(\theta_j + x'\beta)}}{1 + e^{-(\theta_j + x'\beta)}}$.
Sledi da je

$$\frac{\tau_j}{1 - \tau_j} = e^{\theta_j + x'\beta} \dashrightarrow \ln\left(\frac{\tau_j}{1 - \tau_j}\right) = \theta_j + x'\beta \quad j = 1, 2, \ldots, k-1.$$

Poslednja jednačina predstavlja kumulativni logistički model. U njoj odsečak θ_j zavisi od j dok β ne zavisi od j tj. isti je kod svih $k-1$ jednačina. Vektor x ne sadrži jedinicu koja odgovara odsečku.
Kumulativni logit model može se zapisati i na sledeći način

$$\ln\left(\frac{\pi_1+\pi_2+\cdots+\pi_j}{\pi_{j+1}+\pi_{j+2}+\cdots+\pi_k}\right)=\beta_{0j}+\beta_{1j}x_1+\cdots+\beta_{p-1,j}x_{p-1}$$

gde je k broj kategorija zavisne promenljive. Ako pretpostavimo da parametri β ne zavise od j onda dobijamo takozvani model proporcionalnih ,,šansi" u kome je efekat nezavisnih promenljivih isti za svaku kategoriju j na logaritamskoj skali. Samo će parametar koji predstavlja odsečak da zavisi od j. Sada model izgleda

$$\ln\left(\frac{\pi_1+\pi_2+\cdots+\pi_j}{\pi_{j+1}+\pi_{j+2}+\cdots+\pi_k}\right)=\beta_{0j}+\beta_1x_1+\beta_2x_2+\cdots+\beta_{p-1}x_{p-1}$$

Ovaj model je uobičajen za modeliranje kod ordinalne regresije i ugrađen je u statističke pakete.

Kumulativni probit model

Neka slučajna promenljiva ϵ ima normalnu raspodelu. Probit model je definisan preko sledeće jednačine

$$\phi^{-1}(\tau_j)=\theta_j+x'\beta \qquad j=1,2,\ldots,k-1$$

Kumulativne verovatnoće su $\tau_j=\phi(\theta_j+x'\beta)$, gde je ϕ funkcija standardne normalne raspodele. Dakle, verovatnoća da promenljiva Y bude u grupi j (primi vrednost j) ili manje od j je vrednost funkcije ϕ u tački $\theta_j+x'\beta$.

Primer : Baza podataka za ovaj primer sadrži informacije o sudaru iz 2004. godine u Novom Južnom Velsu, Australija. U skupu podataka je ukupno 82659 vozača a od toga je u analizu uključeno 76341($\approx 92.3\%$). Izostavljeni su vozači koji imaju manje od 17 godina, čija je starost nepoznata i oni čiji tip vozila nije među navedenim kategorijama. Promenljive su sledeće :

1. Starosna kategorija vozača
   - 1=17-20 god.
   - 2=21-25 god.
   - 3=26-29 god.
   - 10=30-39 god(4 je kodirano kao 10 i predstavlja baznu kategoriju)
   - 5=40-49 god.
   - 6=50-59 god.
   - 7=60+

2. Pol vozača
    - Ženski (F)
    - Muški (M)
3. Tip putničkog vozila
    - 10= automobil (bazna kategorija)
    - 2= laki kamion
    - 4= autobus/teški kamion
    - 6= motocikli
4. Stepen povrede
    - 1= bez povrede
    - 2= povreda
    - 3= smrtni ishod
5. Broj -frekvencija sudara

Za ordinalnu regresiju najpoznatiji model i najčešće korišćen je kumulativni logistički model koji je korišćen za analizu u ovom primeru.
Zavisna promenljiva Y je stepen povrede, koja je ordinalna sa 3 nivoa.

j	Stepen povrede
1	bez povrede
2	povreda
3	smrtni ishod

Tabela 4.4.1 Zavisna promenljiva i njene kategorije

Nezavisne promenljive u modelu su : starosna kategorija vozača, pol vozača i tip putničkog vozila. Kako su sve nezavisne promenljive kategorijalne treba izabrati baznu kategoriju za svaku od njih. U ovom primeru to će biti starosna kategorija 30-39 god. kodirana kao 10, muški pol i tip vozila automobil.

Kumulativni logistički model je

$$\ln\left(\frac{\tau_j}{1-\tau_j}\right) = \theta_j + x'\beta \qquad j = 1, 2$$

Pomoću SPSS dobijaju se rezultati prikazani u sledećoj tabeli

Parametar	$\hat{\beta}$	$se(\hat{\beta})$	$e^{\hat{\beta}}$	Wald χ^2	p vrednost
odsečak (j=1)	0.47	0.021	1.6	494.09	< 0.0001
odsečak (j=2)	5.049	0.045	155.89	12518.28	< 0.0001
Tip putničkog vozila					
tip vozila 2	0.151	0.027	1.163	31.23	< 0.0001
tip vozila 4	0.297	0.037	1.345	66.1	< 0.0001
tip vozila 6	2.449	0.057	11.577	1858.57	< 0.0001
tip vozila 10	0	0	1		
Starosna kategorija vozača					
star.kategorija 1	-0.179	0.032	0.836	30.73	<0.0001
star.kategorija 2	-0.112	0.032	0.894	12.23	<0.0001
star.kategorija 3	-0.058	0.036	0.944	2.54	0.11
star.kategorija 5	0.055	0.03	1.057	3.37	0.066
star.kategorija 6	0.07	0.033	1.072	4.47	0.035
star.kategorija 7	0.15	0.034	1.162	19.09	< 0.0001
star.kategorija 10	0	0	1		
Pol vozača					
Žensko (F)	0.172	0.033	1.188	26.45	<0.0001
Muško (M)	0	0	1		
Interakcije					
[star.kategorija 1]*F	0.129	0.053	1.138	5.90	0.015
[star.kategorija 2]*F	0.118	0.052	1.125	5.11	0.024
[star.kategorija 3]*F	0.042	0.06	1.043	0.49	0.485
[star.kategorija 5]*F	0.028	0.49	1.029	0.34	0.56
[star.kategorija 6]*F	0.018	0.056	1.019	0.11	0.741
[star.kategorija 7]*F	-0.148	0.06	0.862	6.22	0.013
[star.kategorija 10]*F	0	0	1		

Tabela 4.4.2 Ocene parametara modela

Fitovane jednačine za $j = 1$

$$\ln\left(\frac{\hat{\tau_1}}{1-\hat{\tau_1}}\right) = 0.47 + 0.151x_1 + 0.297x_2 + \cdots + 0.018x_{15} - 0.148x_{16}$$

i za $j = 2$

$$\ln\left(\frac{\hat{\tau_2}}{1-\hat{\tau_2}}\right) = 5.049 + 0.151x_1 + 0.297x_2 + \cdots + 0.018x_{15} - 0.148x_{16}$$

Gde su x_1, x_2, x_3 indikator promenljive za tip putničkog vozila, $x_4, x_5, \ldots, x_9$ su indikator promenljive za starosnu kategoriju vozača, x_{10} indiktor promenljiva za ženski pol, $x_{11}, x_{12}, \ldots, x_{16}$ indikator promenljive za interakciju starosne kategorije sa polom.

Stepen povrede manji od j za $j = 1$ znači da nije bilo povrede, a za $j = 2$ može biti da je prošlo sa povredom ili bez povrede.

Preciznije

$$\tau_1 = P(Y \leq 1) = P(Y = 1)$$

$$\tau_2 = P(Y \leq 2) = P(Y = 1) + P(Y = 2)$$

Interpretacija parametara

Ako posmatramo vozače starosne kategorije 1 (17-20 godina) iz tabele možemo videti da je $e^{\hat{\beta}} = 0.836$. To znači da je ,,šansa" da oni imaju stepen povrede j ili manji za 16.4% manja za vozače ove kategorije nego za vozače automobila koji predstavljaju baznu kategoriju. Tj. ,,šansa" da imaju sudar manjeg stepena povrede je manja nego za vozače automobila. Ovo zapravo znači da je za vozače starosne kategorije 1 veća verovatnoća da izazovu teži oblik sudara.

Ako posmatramo ženske osobe kao vozače kod njih je $e^{\hat{\beta}} = 1.188$. To znači da je ,,šansa" da imaju sudar stepena j ili manji za 18.8% veća nego za muške vozače. To znači da je veća verovatnoća da će žene vozači imati blaži oblik sudara. Možemo reći da su manje rizične nego muškarci.

Ako posmatramo ženske osobe starosne kategorije 3 (26-29 god.) ,,šansa" da njihov stepen povrede u sudaru bude j ili manji je $1.188 \times 0.944 \times 1.043 \approx 1.17$, a u poređenju sa baznim nivom. Tj. za 17% je veća u odnosu na baznu kategoriju koju predstavljaju muškarci starosti od 30-39 godina. Stoga možemo pretpostaviti da će one imati blaži oblik sudara nego pomenuti muški vozači.

5 Poasonova regresija

5.1 Poasonov model

Poasonova raspodela je uobičajena raspodela za modeliranje celobrojnih (counts) podataka. Neka je Y slučajna promenljiva koja broji koliko puta se realizovao događaj od interesa, npr. broj zahteva za odštetu u okviru jedne grupe polisa osiguranja sa trajanjem ω_i i neka je μ_i očekivana vrednost ako je $\omega_i = 1$. Njena raspodela verovatnoća je data sa

$$f(y;\mu) = \frac{(\omega\mu)^y e^{-\omega\mu}}{y!} \quad y = 0, 1, 2, \ldots$$

gde je μ prosečan broj realizovanih događaja. Pritom je $E(Y) = Var(Y) = \omega\mu$.

Parametar μ se obično predstavlja kao stopa. Na primer, kod sudara motornih vozila parametar μ može predstavljati prosečan broj sudara na 1000 stanovnika, ili na 1000 motornih vozila ili prosečan broj sudara na 1000 km pređenog puta. Takođe tu treba uključiti i vremensku komponentu tj. vremenski period za koji se ta stopa izračunava. Stopa sudara kod motornih vozila se obično računa na nivou jedne godine. Na primer, broj sudara na nivou jedne godine ako je pređeni put 1000km itd.

Neka je $N(t)$ broj zahteva za odštetu za individualnu polisu osiguranja, dok traje vremenski period [0,t] tako da je $N(0) = 0$. Stohastički proces $\{N(t); t \geq 0\}$ se naziva proces zahteva za odštetu. Ako su ispunjene određene pretpostavke, od kojih jedna zahteva da zahtevi nisu grupisani onda je taj proces Poasonov. Na osnovu toga možemo pretpostaviti da broj zahteva u okviru individualne polise dok traje neki vremenski period ima Poasonovu raspodelu.

Promenljiva od interesa u osiguranju je frekvencija tj. učestalost zahteva za odštetu $Z_i = \dfrac{Y_i}{\omega_i}$. Kaže se da prati relativnu Poasonovu raspodelu. Neka su Z_1 i Z_2 frekvencije zahteva za dve različite grupe polisa koje imaju trajanje ω_1 i ω_2 i obe promenljive prate relativnu Poasonovu raspodelu sa očekivanjem μ. Ako spojimo te dve grupe polisa dobićemo novu promenljivu za frekvenciju zahteva, Z koja predstavlja težinsku aritmetičku sredinu

$$Z = \frac{\omega_1 Z_1 + \omega_2 Z_2}{\omega_1 + \omega_2}$$

Kako je $\omega_1 Z_1 + \omega_2 Z_2$ suma dve nezavisne promenljive koje imaju Poasonovu raspodelu onda i sama ima istu raspodelu, a Z prati relativnu Poasonovu raspodelu sa ekspozicijom $\omega_1 + \omega_2$. Koristeći elementarna pravila za očekivanje lako se pokaže da je $E(Z) = \mu$.
Efekat nezavisnih promenljivih na zavisnu promenljivu Y se modelira posredstvom parametra μ.
Neka su $Y_1, Y_2, \ldots Y_n$ nezavisne slučajne promenljive tako da je

$$E(Y_i) = \mu_i = n_i \theta_i$$

Na primer, u osiguranju Y_i može biti broj zahteva za odštetu za određenu marku i model automobila. Taj broj će zavisiti od broja osiguranih automobila tog tipa (n_i) i nekih drugih faktora koji utiču na θ_i kao što je starost automobila ili lokacija na kojoj su korišćeni. Može se modelirati i broj saobraćajnih nezgoda sa takozvanom ekspozicijom koja u ovom slučaju može da bude broj registrovanih motornih vozila. Ako je ekspozicija konstanta onda ona nije toliko relevantna za model.
Podaci se mogu predstaviti u obliku tabele klasifikacije u slučaju kada su objašnjavajuće promenljive kategorijalne i nema ih puno. Onda će zavisna promenljiva biti frekvencija ili broj u svakoj ćeliji tabele. Sve promenljive koje definišu tabelu su nezavisne promenljive. Za ovako prestavljene podatke pogodno je koristiti log-linearne modele.
Zavisnost θ_i od nezavisnih promenljivih x_i se modelira na sledeći način

$$\theta_i = e^{x_i' \beta}$$

Uopšteni linearni model je dat sa sledeće dve jednačine

$$E(Y_i) = \mu_i = n_i e^{x_i' \beta}; \quad Y_i \sim P(\mu_i)$$

$$\ln(\mu_i) = \ln(n_i) + x_i' \beta$$

Član $\ln(n_i)$ se naziva **,,offset"** i to je poznata konstanta koja je ukjučena u postupak ocenjivanja parametara modela.
Porast nezavisne promenljive x_k za jednu jedinicu imaće **multiplikativni** efekat na vrednost μ u iznosu od e^{β_k}.
Za testiranje hipoteza o parametrima β_j mogu se koristiti Wald test, skor test ili test odnosa verodostojnosti koji su detaljno objašnjeni u odeljku 2.6
Za interval poverenja koristi se sledeća statistika

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim N(0,1)$$

Fitovane vrednosti su $\hat{Y}_i = \hat{\mu}_i = n_i e^{x_i'\hat{\beta}}$, $i = 1, 2, \ldots, n$. $e_i = \widehat{E(Y_i)} = \hat{Y}_i$ pa dobijamo da je $e_i = n_i e^{x_i'\hat{\beta}}$. Pošto je $Var(Y) = E(Y)$ onda je ocena standardne greške za Y_i data sa $\sqrt{e_i}$. Neka su o_i opažene vrednosti promenljive Y_i.
Pirson reziduali su dati sa

$$(r_p)_i = \frac{o_i - e_i}{\sqrt{e_i}}$$

Suma kvadrata ovih reziduala daje Pirson χ^2 statistiku pomoću koje se procenjuje koliko je dobro fitovanje. Dakle

$$\chi^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

Reziduali mogu biti transformisani u sledeći oblik

$$(r_p)_i = \frac{o_i - e_i}{\sqrt{e_i(1 - h_i)}}$$

gde je h_i i-ti dijagonalni element ,,Hat" matrice koji se naziva leveridž.
Devijansa je detaljno objašnjena u odeljku 2.5 pa se može pokazati da je za Poasonov model data sa

$$D = 2\sum_{i=1}^{n}[o_i \log(o_i/e_i) - (o_i - e_i)]$$

Reziduali devijanse su dati sa $d_i = sign(o_i - e_i)\sqrt{2[o_i \log(o_i/e_i) - (o_i - e_i)]}$
$i = 1, 2, \ldots, n$, tako da važi $D = \sum_{i=1}^{n} d_i^2$.
Statistike D i χ^2 koriste se kao mere za adekvatnost fitovanog modela. Mogu da se porede sa χ^2_{n-p} raspodelom koja ima $n - p$ stepeni slobode, gde je p broj ocenjenih parametara. Mogu se koristiti i statistika odnos verodostojnosti i pseudo R^2. Statistika

$$C = 2[l(b) - l(b_{min})]$$

Bazirana je na poređenju između maksimalne vrednosti funkcije l za minimalan model tj. model bez nezavisnih varijabli koji je dat sa $\ln(\mu_i) = \ln(n_i) + \beta_0$ i maksimalne vrednosti funkcije l za model sa p parametara koji je dat sa $\ln(\mu_i) = \ln(n_i) + x_i'\beta$. Statistika C predstavlja test za testiranje hipoteze da je $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, i poredi se sa χ^2_{p-1} raspodelom.
Pseudo R^2 statistika je

$$R^2 = \frac{[l(b_{min}) - l(b)]}{l(b_{min})}$$

5.2 Primer - osiguranje automobila

U ovom primeru svrha analize je da se razume uticaj nekih karakteristika samih vozača i tipova njihovih vozila na broj saobraćajnih nesreća. Ta veza između promenljivih obezbeđuje osnovu da se odredi cena za pokriće u osiguranju.
Podaci su iz opšteg osiguravajućeg drustva u Singapuru.[14]
Za analizu su korišćene promenljive kao što su starost i tip automobila, kao i promenljive koje karakterišu vozača tj. osiguranika poput pola, starosti i ranijeg iskustva u vezi sa saobraćajnim nezgodama.

Promenljiva	Opis promenljive
tip vozila	tip vozila koji je osiguran, auto(A) ili neko drugo vozilo (O)
starost vozila	godine starosti vozila, grupisane u 5 kategorija
pol osiguranika	muško (M) ili žensko (F)
starost	godine starosti osiguranika grupisane u 6 kategorija
NCD (no claims discount)	veća vrednost ove promenljive bolji raniji podaci o nesrećama osiguranika

Tabela 5.2.1 Opisi promenljivih

Broj zahteva	Frekvencija	Procenat	Kumulativni procenat
0	6996	93.5%	93.5%
1	455	6.1%	99.6%
2	28	0.4%	99.9%
3	4	0.1%	100%

Tabela 5.2.2 Frekvencije broja zahteva za odštetu

[14]Više o ovoj organizaciji pogledati na www.gia.org.sg, a baza podataka za ovaj primer je dostupna na www.instruction.bus.wisc.edu

Broj zahteva za odštetu --→	0	1	2	3	Ukupno
tip vozila A	3555 92.5%	271 7.1%	15 0.4%	1 0.0%	3842 (51.3%)
tip vozila O	3441 94.5%	184 5.1%	13 0.4%	3 0.1%	3641(48.7%)

Tabela 5.2.3 Frekvencije zahteva za odštetu prema tipu vozila

Kategorija automobila ima manje ,,iskustvo" u vezi sa zahtevima za odštetu nego kategorija O.
Vozila koja su označena sa O uglavnom obuhvataju vozila za zaposlene, teretna vozila, komercijalna vozila.

Starosna kategorija vozila	**Broj zahteva**				
	0	1	2	3	Ukupno
2 (0-2 god.)	4069 92.4%	313 7.1%	20 0.4%	4 0.1%	4406
3 (3-5 god.)	708 91.8%	59 7.7%	4 0.5%	0	771
4 (6-10 god.)	872 94.4%	49 5.3%	3 0.3%	0	924
5 (11-15 god.)	1133 97.3%	30 2.6%	1 0.1%	0	1164
6 (16 god. i više)	214 98.2%	4 1.8%	0	0	218

Tabela 5.2.4 Broj zahteva za odštetu prema starosti vozila

Sledeća tabela prikazuje polise koje nemaju zahtev za odštetu prema starosnoj grupi osiguranika ako je tip vozila automobil.

	broj zahteva =0	Ukupan broj polisa
22-25 god.	131 (92.9%)	141
26-35 god.	1354 (91.7%)	1476
36-45 god.	1413(93.2%)	1516
46-55 god.	503(93.8%)	536
56-65 god.	140(89.2%)	157
66 i stariji	15(88.2%)	17

Tabela 5.2.5 Osiguranici sa brojem zahteva 0

Neka je za ovaj primer zavisna promenljiva broj zahteva za odštetu i može da uzima vrednosti 0, 1, 2, 3. Za modeliranje ćemo koristiti Poasonovu raspodelu koja je prikladna za ovakav tip podataka, sa link funkcijom ln.
Zavisna promenljiva : Broj zahteva za odštetu

Nezavisne promenljive su :

- Pol osiguranika
- Starost vozila
- NCD

Od ukupno 7483 osiguranika, 700 su žene što čini 9.4%, a njih 6783 (90.6%) su muškarci.
Koristeći program SPSS dobijaju se sledeći rezultati

Parametar	$\hat{\beta}$	$se(\hat{\beta})$	$e^{\hat{\beta}}$	Wald χ^2	p vrednost
odsečak	-2.118	0.0794	0.12	711.782	<0.0001
Pol osiguranika					
žene	-0.16	0.1523	0.852	1.104	0.293
muškarci	0	0	1		
Starosna kategorija vozila					
starosna kategorija 2	0	0	1		
starosna kategorija 3	0.001	0.1426	1.001	0.00	0.995
starosna kategorija 4	-0.388	0.1524	0.679	6.479	0.011
starosna kategorija 5	-1.202	0.1899	0.301	40.067	<0.0001
starosna kategorija 6	-1.578	0.504	0.206	9.781	0.002
Vrednost za NCD					
NCD=0	0	0	1		
NCD=10	-0.311	0.1248	0.732	6.23	0.013
NCD=20	-0.477	0.1303	0.62	13.428	<0.0001
NCD=30	-0.385	0.1918	0.68	4.027	0.045
NCD=40	-0.733	0.2412	0.48	9.25	0.002
NCD=50	-0.677	0.1336	0.51	25.642	<0.0001

Tabela 5.2.6 Ocene parametara modela

Po ovom modelu, očekivani broj zahteva za odštetu je dat jednačinom

$$\hat{\mu} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta_{10}} x_{10}}$$

gde je x_1 indikator promenljiva za ženski pol, x_2, x_3, x_4, x_5 su indikator promenljive za starosnu kategoriju vozila i $x_6, x_7, x_8, x_9, x_{10}$ su indikator promenljive za vrednosti NCD.

Očekivani broj zahteva za odštetu za muškarce čije je vozilo staro 0-2 god. i koji imaju NCD=0 je

$$\hat{\mu} = e^{-2.118} = 0.12$$

a za žene čije je vozilo iste starosti i koje takođe imaju vrednost NCD=0

očekivani broj zahteva je

$$\hat{\mu} = e^{-2.118} \cdot e^{-0.16} = 0.12 \cdot 0.852 = 0.102$$

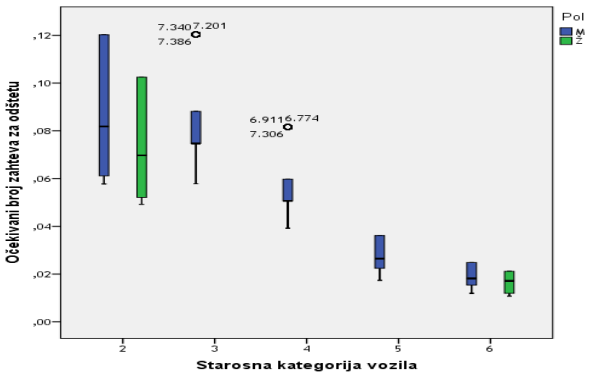
Na osnovu modela procenjuje se da vozila starosne kategorije 4 imaju za 32.1% manji očekivani broj zahteva za odštetu nego vozila starosne kategorije 2. Isto tako, za žene očekivani broj zahteva za odštetu manji je za 14.8% nego za muškarce, u datom uzorku.
Za žene čije je vozilo starosne kategorije 4 i koje imaju vrednost NCD=40 očekivani broj zahteva je

$$\hat{\mu} = e^{-2.118} \cdot e^{-0.16} \cdot e^{-0.388} \cdot e^{-0.733} = 0.12 \cdot 0.852 \cdot 0.679 \cdot 0.48 = 0.033$$

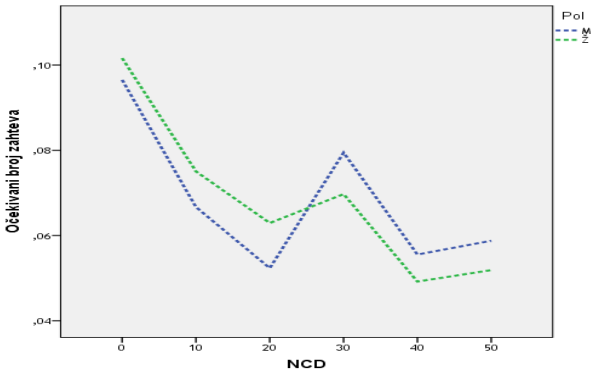
Za muškarce čije je vozilo starosne kategorije 5 i koji imaju vrednost NCD=20 očekivani broj zahteva za odštetu a na osnovu analiziranog uzorka obima 7483 je

$$\hat{\mu} = e^{-2.118} \cdot e^{0} \cdot e^{-1.202} \cdot e^{-0.477} = 0.12 \cdot 1 \cdot 0.301 \cdot 0.62 = 0.022$$

Na isti način se interpretiraju ostali slučajevi.



Slika 5.2.1 Očekivani broj zahteva prema starosnoj kategoriji vozila i polu osiguranika



Slika 5.2.2 Očekivani broj zahteva prema vrednosti NCD i polu osiguranika

Rezultati su dobijeni na osnovu uzorka obima 7483. Zabeleženi realni podaci o broju zahteva za odštetu su prikazani u sledećoj tabeli.

Broj zahteva	**Frekvencija**	**Procenat**
0	6996	93.5%
1	455	6.1%
2	28	0.4%
3	4	0.1%
Ukupno: 7483 polise		

Tabela 5.2.7 Frekvencija broja zahteva za odštetu

Na osnovu statističke analize dobijeno je da je vrednost devijanse D=2784.91 a broj stepeni slobode te statistike je 7472. Otuda je D/df=0.373 što je dobra vrednost za devijansu.
Test odnosa verodostojnosti poredi fitovani model sa modelom koji sadrži samo odsečak. Vrednost ove statistike je 102.294 a p vrednost je manja od 0.0001 što znači da odbacujemo hipotezu da su svi koeficijenti u fitovanom modelu jednaki nuli.

	Wald χ^2	df	p vrednost
odsečak	527.636	1	<0.0001
pol	1.104	1	0.293
starosna kategorija vozila	52.751	4	<0.0001
NCD	35.922	5	<0.0001

Tabela 5.2.8 Testiranje značajnosti pojedinačnih koeficijenata

Iz tabele vidimo da promenljiva pol nije statistički značajna jer je p=0.293 .

5.3 Pojam overdispersion

Slučajne varijacije između kupaca polisa osiguranja i osiguranih objekata i efekat nezavisnih promenljivih koje nisu uključene u model dovodi do pojma ,,overdispersion" tj. prevelika disperzija. Varijansa observacija u okviru grupe polisa je veća nego varijansa za pretpostavljenu Poasonovu raspodelu.
Uobičajeni način da se modelira ova pojava je da se srednja vrednost kod Poasonove raspodele posmatra kao slučajna promenljiva.
Posmatrajmo niz nezavisnih slučajnih promenljivih $\Lambda_1, \Lambda_2, \ldots$ koje su distribuirane na intervalu $(0, \infty)$, i neka su $X_1, X_2, \ldots$ nezavisne slučajne promenljive takve da je za dato $\Lambda_i = \lambda_i$, X_i ima Poasonovu raspodelu sa parametrom λ_i.
Onda važi da je $E(X_i|\Lambda_i) = \Lambda_i$ i $Var(X_i|\Lambda_i) = \Lambda_i$. Takođe za promenljive X_i je $E(X_i) = E[E(X_i|\Lambda_i)] = \lambda_i$ i varijansa

$$Var[X_i] = E[Var(X_i|\Lambda_i)] + Var[E(X_i|\Lambda_i)] = E[\Lambda_i] + Var[\Lambda_i]$$

Ako se uspostavi veza između očekivanja i disperzije za Λ_i

$$Var[\Lambda_i] = \nu E[\Lambda_i] \quad \nu > 0$$

Ako je sada X_i broj zahteva za odštetu tako da je $E(X_i) = \omega_i \mu_i$, na osnovu prethodnog izraza za varijansu dobijamo

$$Var[X_i] = (1+\nu)E[\Lambda_i] \text{ pa je}$$
$$Var[X_i] = (1+\nu)\omega_i \mu_i$$

Posmatrajući sada frekvenciju zahteva $Y_i = \dfrac{X_i}{\omega_i}$ imamo $Var[Y_i] = Var[X_i]/\omega_i^2$

$$Var[Y_i] = \frac{(1+\nu)\omega_i \mu_i}{\omega_i^2} = \frac{(1+\nu)\mu_i}{\omega_i}$$

za $\phi = 1 + \nu$, $Var[Y_i] = \dfrac{\phi \mu_i}{\omega_i}, \quad \phi > 0$
Ako pretpostavimo da Λ ima gama raspodelu i zadovoljava da je $Var[\Lambda] = \nu E[\Lambda]$ onda se može pokazati da X ima negativnu binomnu raspodelu tako da je

$$P(X_i = x_i) = \frac{\Gamma(\omega_i \mu_i/\nu + x_i)}{\Gamma(\omega_i \mu_i/\nu) x_i!} \left(\frac{1}{1+\nu}\right)^{\omega_i \mu_i/\nu} \left(\frac{\nu}{1+\nu}\right)^{x_i}$$

5.4 Negativna binomna regresija

Ovaj model regresije se koristi za modeliranje broja zahteva za odštetu kada Poasonov model nije odgovarajući. To se dešava kada je varijansa opaženih podataka dosta veća nego očekivana vrednost. Ta pojava se naziva ekstra Poasonova varijacija ili ,,overdispersion".
Ta pojava se modelira koristeći složenu Poasonovu raspodelu. U tom slučaju slučajna promenljiva od interesa ima Poason raspodelu tj. $Y \sim P(\Lambda)$ gde je Λ slučajna promenljiva što uzrokuje da je varijansa veća od očekivane kada je parametar λ fiksiran.

Pretpostavimo sada da je Λ pozitivna i neprekidna sa funkcijom verovatnoće $g(\Lambda)$, tako da je $g(\Lambda) = 0$ za $\lambda < 0$. Funkcija verovatnoće za promenljivu Y je data sa

$$f(y) = \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!} g(\lambda) d\lambda$$

gde za $\Lambda = \lambda$, $Y \sim P(\lambda)$. Pogodan izbor za $g(\Lambda)$ je gama raspodela $G(\mu, \nu)$. Znajući da je

$$Y|(\Lambda = \lambda) \sim P(\lambda) \Rightarrow f(Y|\Lambda = \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Bezuslovna raspodela za Y je

$$f(y) = \int_0^\infty f(y|\lambda) g(\lambda) d\lambda$$

Neka $\Lambda \sim G(\mu, \nu)$

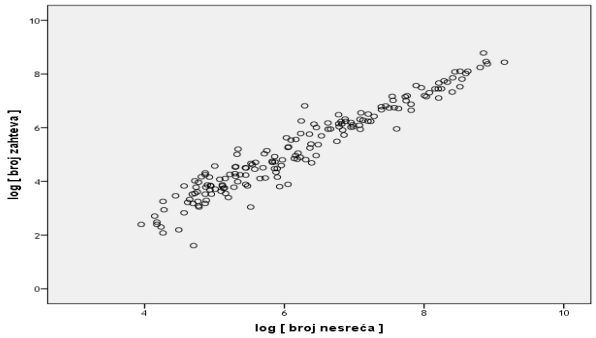
$$f(y) = \int_0^\infty \frac{e^{-\lambda}\lambda^y}{y!}\frac{\lambda^{-1}}{\Gamma(\nu)}\left(\frac{\lambda\nu}{\mu}\right)^\nu e^{-\lambda\nu/\mu}d\lambda = \frac{1}{y!\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^\nu \int_0^\infty \lambda^{y+\nu-1}e^{-\lambda(1+\nu/\mu)}d\lambda$$

$$= \frac{\Gamma(\nu+y)}{y!\Gamma(\nu)}\left(\frac{\nu}{\nu+\mu}\right)^\nu \left(\frac{\mu}{\mu+\nu}\right)^y$$

$y = 0, 1, 2, \ldots$
Stavljajući da je $\kappa = 1/\nu$ ovo je negativna binomna raspodela $NB(\mu, \kappa)$.
Osim gama raspodele za promenljivu Λ u literaturi za aktuarstvo javljaju se još inverzna Gausova raspodela i uopštena inverzna Gausova raspodela.
Uopštena inverzna Gausova raspodela je pogodna za modeliranje ali je mana to što je njeno izračunavanje složenije. Zato je inverzna Gausova raspodela nešto jedostavnija za izračunavanje. Tako se dobija složena raspodela Poason-inverzna Gausova koja je više iskošena nego negativna binomna i samim tim pogodnija za modeliranje raspodele frekvencije zahteva za odštetu. Međutim, veoma često je ipak negativna binomna raspodela izbor za modeliranje velike disperzije prebrojivih podataka (overdispersion) zbog njene dostupnosti u standardnim softverima i jednostavnijeg izračunavanja.
Primer : Jedna vrsta osiguranja nosi naziv osiguranje ,,treće" strane (eng.third party insurance). To je osiguranje koje je obavezno za vlasnike vozila u Australiji. Ono osigurava vlasnike od povreda koje su izazvane drugim vozačima, putnicima ili pešacima u saobraćajnim nezgodama. Posmatrani skup podataka beleži broj zahteva za odštetu ,,treće" strane u periodu od 12 meseci između 1984. i 1986. godine. Obuhvaćene su 176 geografske regije u Novom Južnom Velsu Australija.
Promenljive od interesa jesu broj zahteva za odšetu (claims), broj nesreća (accidents), ki-broj ubijenih ili povređenih u neseći, veličina populacije (population size).
Analizom promenljive claims tj. broj zahteva dobijamo sledeće vrednosti

N	176
Srednja vrednost	≈ 587
Varijansa	$\approx 1.03 \cdot 10^6$

Tabela pokazuje da je varijansa opaženih vrednosti mnogo veća od srednje vrednosti, pa zato Poason model nije pogodan za modeliranje broja zahteva za odštetu.



Slika 5.4.1 Veza između promenljivih log[broj nesreća] i log[broj zahteva]

Sa prethodne slike uočava se linearna veza između dve transformisane promenljive.
Negativni binomni regresioni model sa link funkcijom log je

$$Y \sim NB(\mu, \kappa), \quad \ln(\mu) = \ln(n) + \beta_1 + \beta_2 \ln(z)$$

y je zavisna promenljiva koja predstavlja broj zahteva za odštetu ,,treće" strane u jednoj oblasti, z je nezavisna promenljiva broj saobraćajnih nezgoda u jednoj oblasti i n je veličina populacije jedne oblasti. $\ln(n)$ predstavlja ,,offset" termin. Pomoću programa SPSS dobijaju se rezultati analize

	$\hat{\beta}$	$se(\hat{\beta})$	$e^{\hat{\beta}}$	Wald χ^2	p vrednost
odsečak	-6.954	0.17	0.001	1733.65	<0.0001
log_acc	0.254	0.026	1.289	94.87	<0.0001
κ	0.17	0.02			

Tabela 5.4.1 Ocene nepoznatih parametara modela

Sada je očekivani broj zahteva za odštetu, u određenoj geografskoj regiji

$$\hat{\mu} = ne^{-6.954 + 0.254ln(z)}$$

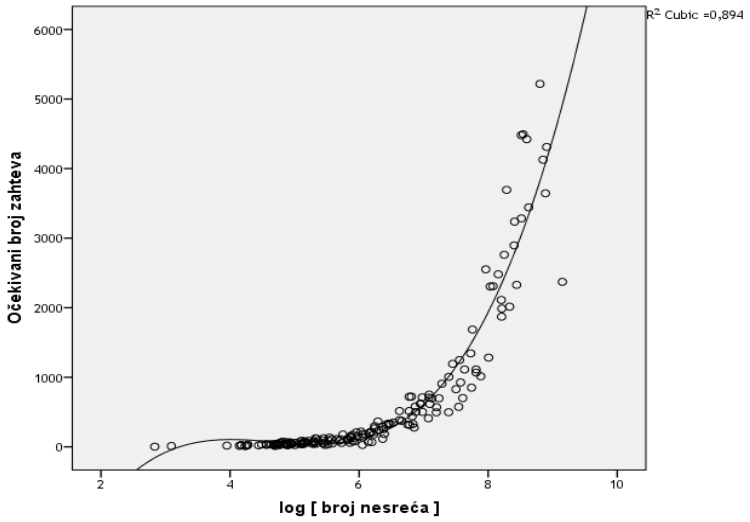
$$\frac{\hat{\mu}}{n} = e^{-6.954} \cdot e^{0.254ln(z)}$$

Ako se broj nesreća z poveća za faktor b u bz onda se očekivani broj zahteva u jednoj oblasti $\hat{\mu}$ poveća za faktor $e^{0.254ln(b)} = b^{0.254}$. Na primer, ako broj nesreća poraste za 20% onda je $b = 1.2$ pa se ocenjena očekivana stopa zahteva $\frac{\hat{\mu}}{n}$ poveća za $1.2^{0.254}$, što približno iznosi 1,05 tj. povećanje je 5%.
Ako se broj nesreća poveća za 25% onda je faktor $b = 1.25$ a vrednost $\frac{\hat{\mu}}{n}$ se poveća za $1.25^{0.254} \approx 1.06$ tj. povećanje je za oko 6% . Devijansa modela iznosi 192.3 a broj stepeni slobode te statistike je 173. Vrednost D/df=1.11 što pokazuje da je model dobar.
Parametar κ ukazuje na oblik raspodele. Ako $\kappa \mapsto 0$ onda $NB(\mu, \kappa)$ aproksimira $P(\mu)$, a kada je $\kappa = 0$ onda je $E(Y) = Var(Y)$ zbog osobine da je kod nb[15] raspodele

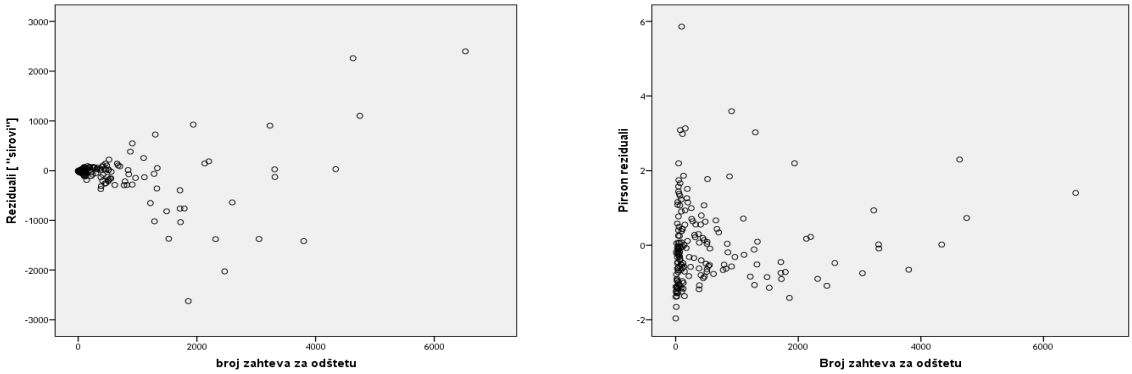
$$E(Y) = \mu \quad Var(Y) = \mu(1 + \kappa\mu)$$

Kada je κ veliko onda negativna binomna raspodela ima dugačak ,,rep" udesno.

[15]negativna binomna



Slika 5.4.2 Veza između logaritma broja nesreća i očekivanog broja zahteva za odštetu



Slika 5.4.3 Izgled "sirovih" i Pirson reziduala

Napomena : Kada su prebrojivi podaci (eng.counts) prevedeni u frekvencije onda se takvi podaci ne mogu modelirati sa Poasonovom i negativno binomnom raspodelom. Bolje je da se modeliraju ,,sirovi" podaci sa članom offset koji zapravo predstavlja korekciju za broj izloženih riziku. Na primer, u zdravstvenom osiguranju od interesa su modeli za frekvenciju bolesti i tu je važan broj osoba koji su izloženi riziku od oboljevanja, tj. taj broj predstavlja offset.

6 Proširenja uopštenog linearnog modela

Neki od modela koji predstavljaju proširenje ulm[16] su

- dupli uopšteni linearni modeli
- uopšteni aditivni modeli
- modeli za srednju vrednost i disperziju

Dupli uopšteni linearni model (d.ulm) je najjednostavniji model za modeliranje srednje vrednosti i disperzionog parametra. Kod uopštenog linearnog modeliranja parametar ϕ je konstanta, i model za srednju vrednost $g(\mu) = x'\beta$ je implicitno i model za varijansu jer je $Var(Y) = \phi V(\mu)$.

Dupli uopšteni linearni model je dat jednačinama

$$g(\mu) = x'\beta \qquad \text{i} \qquad h(\phi) = z'\gamma$$

z je vektor nezavisnih promenljivih koji može da sadrži iste promenljive kao vektor x, a γ je vektor parametara, ϕ je disperzioni parametar.
Ovaj model je uveden od strane Smyth-a 1989.godine, a koristio mu je u osiguranju za zahteve za odštetu.

6.1 Uopšteni aditivni model

Klasičan linearan model se zasniva na regresionoj funkciji oblika

$$\mu = E(y|x_1, x_2, \ldots, x_n) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

Kod uopštenog linearnog modeliranja imamo malo drugačiju formu regresije

$$g(\mu) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

Ovaj model je linearan po koeficijentima β_j a ne neophodno po nezavisnim promenljivima x_j. Umesto promenljive x_j u regresiji se može javiti x_j^2, x_j^3, $\sin(x_j)$, ali i dalje se radi o linearnom modeliranju.
Uopšteni aditivni model (uam) je proširenje uopštenog linearnog modela (ulm).
Model je definisan na sledeći način

$$g(\mu) = \beta_0 + \sum_{j=1}^{n} \beta_j m_j(x_j)$$

[16] ulm-uopšteni linearni model

Funkcije m_j su glatke funkcije i zavise od x_j i ne moraju da budu iste za različite nezavisne promenjive tj. $m_i(x_i) \neq m_k(x_k)$.
Na primer, pretpostavimo da osiguravajuća kompanija ima veliku bazu podataka i želi da predvidi verovatnoću zahteva za odštetu π. Onda može koristiti model

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + m(z)$$

Ovaj model se naziva semi-parametarski model jer sistematska komponenta sadrži parametarski $\beta_0 + \sum_{j=1}^{n} \beta_j x_j$ deo i neparametarski deo $m(z)$.
Leva strana jednačine predstavlja logit link funkciju, a desna strana se sastoji od linearne komponente $\beta_0 + \sum_{j=1}^{n} \beta_j x_j$ gde su x_j nezavisne promenljive kao što su pol, tip vozila ili kuće a sve u zavisnosti od toga kakva je polisa osiguranja i o kakvom zahtevu za odštetu se radi. Dodatna promenljiva z je neka neprekidna promenljiva kao npr. starost osiguranika i ona je stavljena u model zbog mogućnosti nelinearnih efekata. Za funkciju $m(z)$ jedan od pogodnih izbora je polinomni oblik.

Uopšteni aditivni modeli su korisni za istraživanje podataka i u početnim fazama izgradnje modela kada funkcionalan oblik modela još nije očigledan. Izlazni rezultat kod ovog modela ne sadrži eksplicitnu jednačinu za fitovanu vrednost zavisne promenljive. Analitičari koriste ove rezultate za predlaganje transformacija i za dobijanje fitovanih vrednosti i predikcija.
Uopšteno aditivno fitovanje dakle obezbeđuje informacije za otkrivanje veza između promenljivih iako te veze obično nisu izražene preko funkcionalnog oblika.

Uopšteni aditivni modeli za parametre mere (varijansa) i oblik raspodele

Kod ovih modela raspodela zavisne promenljive ne mora da bude u eksponencijalnoj familiji, već to može biti bilo koja raspodela za koju se može izračunati prvi i drugi izvod. Pomoću ovog modela mogu se pored srednje vrednosti i disperzije zavisne promenljive, modelirati i asimetrija raspodele (skewness) i oštrina krive raspodele (kurtosis). Model je oblika

$$g(\rho) = x_1'\beta + s(x_2) + x_3'\gamma$$

gde je ρ parametar koji se modelira, x_1, x_2, x_3 su vektori nezavisnih promenljivih, β je fiksirano a γ je slučajan efekat, a s je vektor glatkih funkcija za svaku komponentu vektora x_2.

Ovaj model je veoma uopšten i obuhvata sve prethodne modele kao što su normalan linearan, uopšten linearni, duplo uopšten linearni, uopšteni aditivni.

6.2 Nula prilagođen inverzan Gausov model

U odeljku 3. objašnjeno je modeliranje neprekidnih promenljivih kao što je iznos zahteva za odštetu, i pokazano je da su adekvatne raspodele za modeliranje gama i inverzna Gausova. Međutim ovi modeli su bili pogodni samo za **pozitivne** zahteve za odštetu tj. za one koji su veći od nule.
Do sada, za verovatnoću zahteva za odštetu i iznos zahteva za odštetu pravljeni su odvojeni modeli čiji rezultati bi se potom kombinovali u cilju predikcije očekivanog iznosa zahteva.

Nula prilagođen inverzan Gausov model ,,stapa" ta dva modela u jedan.

Pretpostavimo da promenjiva Y predstavlja iznos zahteva za odštetu. Raspodela ove promenjive biće mešovita diskretna-neprekidna.

Nula prilagođena inverzna Gausova raspodela je

$$f(y) = \begin{cases} 1 - t\pi, & y = 0 \\ t\pi h(y), & y > 0 \end{cases}$$

t predstavlja meru izloženosti riziku, π je verovatnoća pozitivnog zahteva za odštetu a $1 - \pi$ verovatnoća da je $y = 0$. $h(y)$ je gustina raspodele za pozitivan zahtev.

Diskretan deo ovog modela objašnjava da li postoji ili ne zahtev za odštetu i ima Bernulijevu raspodelu $B(1, \pi)$. Raspodela $h(y)$ je $IG(\mu, \sigma^2)$ i važi da je

$$E(Y) = t\pi\mu \qquad Var(Y) = t\pi\mu^2(1 - t\pi + \mu\sigma^2)$$

Prednost ovog modela je da se pomoću njega modelira iznos zahteva za odštetu bezuslovno o zahtevu.

7 Zaključak

Cilj ovog master rada je objasniti šta je to uopšten linearan model i kako se može primeniti u oblasti aktuarstva. Prvo je definisana i detaljno objašnjena teorijska osnova na kojoj se temelji model počevši od eksponencijalne familije raspodela, varijansne funkcije, statistika za procenu adekvatnosti fitovanja modela kao i druge dijagnostike modela poput reziduala i autlajera. Pokazano je da je normalan (klasičan) linearan model koji se često koristi u praksi, samo specijalan slučaj uopštenog linearnog modela.

Uopšteno linearno modeliranje omogućava da se modeliraju podaci koji nemaju normalnu raspodelu već je njihova raspodela binomna, gama, Poasonova itd. Takvo modeliranje ima veliki značaj kako u aktuarstvu tako i u mnogim drugim istraživanjima (medicinskim, biološkim, ekonomskim). U aktuarstvu pomaže analitičarima pri donošenju važnih odluka.

Kroz primere u ovom radu, možemo zaključiti da su uopšteni linearni modeli ako se izabere pogodna raspodela, adekvatni za modeliranje podataka kao što su iznosi zahteva za odštetu koji su od velikog interesa osiguravajućim kućama i kompanijama. Na osnovu procenjenog očekivanog iznosa zahteva za odštetu kompanija može da odredi visinu premije za svoje osiguranike. Tj. procenjuje rizičnost pojedinca ili određene grupe.

Takođe pomoću logističkog regresionog modela može se ocenjivati verovatnoća da polisa osiguranja ima pozitivan zahtev za odštetu.

Za primere je korišćena realna baza podataka koja je dostupna na sajtu
http://www.businessandeconomics.mq.edu.au/ i na http://instruction.bus.wisc.edu/
Izvođenje zaključaka kod primera zasnovano je na dobijenoj statističkoj analizi koristeći program SPSS. Pored SPSS korisno je poznavati program SAS i R za modeliranje.

U Prilogu rada izloženi su samo neki izlazni rezultati iz programa SPSS jer su rezultati za svaki primer prikazani u obliku tablica u samom radu i interpretirani su.

Prilog

Tabela koja prikazuje rezultat statističke analize za primer dat na strani 29. odeljak 3.2 gama regresija- osiguranje vozila.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	7,683	,0932	7,500	7,866	6791,794	1	,000	2171,571	1808,912	2606,937
[agecat=1]	,313	,0796	,157	,469	15,432	1	,000	1,367	1,170	1,598
[agecat=2]	,125	,0715	-,015	,265	3,063	1	,080	1,133	,985	1,304
[agecat=3]	,029	,0698	-,108	,165	,168	1	,682	1,029	,897	1,180
[agecat=4]	,050	,0694	-,086	,186	,526	1	,468	1,052	,918	1,205
[agecat=5]	-,072	,0759	-,221	,076	,911	1	,340	,930	,802	1,079
[agecat=6]	0[a]	.	.	.	.	.	.	1	.	.
[gender=F]	-,166	,0355	-,236	-,097	21,993	1	,000	,847	,790	,908
[gender=M]	0[a]	.	.	.	.	.	.	1	.	.
[veh_body=BUS]	-,412	,3883	-1,173	,349	1,125	1	,289	,662	,309	1,418
[veh_body=CONVT]	,166	,6653	-1,138	1,470	,062	1	,803	1,181	,321	4,350
[veh_body=COUPE]	,234	,1563	-,072	,541	2,250	1	,134	1,264	,931	1,717
[veh_body=HBACK]	-,023	,0795	-,179	,132	,087	1	,769	,977	,836	1,142
[veh_body=HDTOP]	,002	,1232	-,240	,243	,000	1	,989	1,002	,787	1,275
[veh_body=MCARA]	-1,095	,3143	-1,711	-,479	12,142	1	,000	,334	,181	,619
[veh_body=MIBUS]	,260	,1892	-,111	,631	1,885	1	,170	1,297	,895	1,879
[veh_body=PANVN]	,002	,1619	-,316	,319	,000	1	,991	1,002	,729	1,376
[veh_body=RDSTR]	-1,108	,8132	-2,702	,486	1,858	1	,173	,330	,067	1,625
[veh_body=SEDAN]	-,127	,0783	-,280	,027	2,609	1	,106	,881	,756	1,027
[veh_body=STNWG]	-,083	,0790	-,238	,071	1,117	1	,291	,920	,788	1,074
[veh_body=TRUCK]	,146	,1264	-,102	,394	1,332	1	,248	1,157	,903	1,483
[veh_body=UTE]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	1,311[b]	,0234	1,266	1,358						

Dependent Variable: claimcost
Model: (Intercept), agecat, gender, veh_body

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Rezultat statističke analize za primer- telesne povrede, strana 33. odeljak 3.2 gama regresija.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	8,212	,0211	8,170	8,253	151573,441	1	,000	3684,250	3535,046	3839,751
op_time	,038	,0004	,038	,039	8971,877	1	,000	1,039	1,038	1,040
[legrep=1]	,467	,0272	,413	,520	294,224	1	,000	1,595	1,512	1,682
[legrep=0]	0[a]	.	.	.	.	.	.	1	.	.
[legrep=1] * op_time	-,005	,0005	-,006	-,004	94,989	1	,000	,995	,994	,996
[legrep=0] * op_time	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	,999[b]	,0084	,983	1,016						

Dependent Variable: total
Model: (Intercept), op_time, legrep, legrep * op_time

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	385196,985	1	,000
op_time	19509,369	1	,000
legrep	294,224	1	,000
legrep * op_time	94,989	1	,000

Dependent Variable: total
Model: (Intercept), op_time, legrep, legrep * op_time

Goodness of Fit[a]

	Value	df	Value/df
Deviance	25411,692	22032	1,153
Scaled Deviance	25436,176	22032	
Pearson Chi-Square	53582,436	22032	2,432
Scaled Pearson Chi-Square	53634,062	22032	
Log Likelihood[b]	-245326,074		
Akaike's Information Criterion (AIC)	490662,147		
Finite Sample Corrected AIC (AICC)	490662,150		
Bayesian Information Criterion (BIC)	490702,149		
Consistent AIC (CAIC)	490707,149		

Dependent Variable: total
Model: (Intercept), op_time, legrep, legrep * op_time

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing

Rezultat statistične analize za primer osiguranje vozila, dat na strani 36. odeljak 3.3 inverzna Gausova regresija.

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	53988,564	1	,000
agecat	16,013	5	,007
gender	9,465	1	,002
area	12,558	5	,028

Dependent Variable: claimcost
Model: (Intercept), agecat, gender, area

Goodness of Fit[a]

	Value	df	Value/df
Deviance	6,377	4612	,001
Scaled Deviance	4624,000	4612	
Pearson Chi-Square	6,753	4612	,001
Scaled Pearson Chi-Square	4897,194	4612	
Log Likelihood[b]	-38568,160		
Akaike's Information Criterion (AIC)	77162,319		
Finite Sample Corrected AIC (AICC)	77162,398		
Bayesian Information Criterion (BIC)	77246,026		
Consistent AIC (CAIC)	77259,026		

Dependent Variable: claimcost
Model: (Intercept), agecat, gender, area

a. Information criteria are in small-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	7,898	,1468	7,610	8,186	2892,716	1	,000	2691,752	2018,551	3589,471
[agecat=1]	,319	,1172	,089	,548	7,393	1	,007	1,375	1,093	1,730
[agecat=2]	,160	,0999	-,036	,356	2,571	1	,109	1,174	,965	1,428
[agecat=3]	,068	,0960	-,121	,256	,495	1	,482	1,070	,886	1,291
[agecat=4]	,062	,0958	-,126	,250	,422	1	,516	1,064	,882	1,284
[agecat=5]	-,054	,1029	-,256	,148	,272	1	,602	,948	,775	1,160
[agecat=6]	0[a]	.	.	.	.	.	.	1	.	.
[gender=F]	-,153	,0497	-,250	-,055	9,465	1	,002	,858	,779	,946
[gender=M]	0[a]	.	.	.	.	.	.	1	.	.
[area=A]	-,355	,1266	-,604	-,107	7,875	1	,005	,701	,547	,898
[area=B]	-,385	,1274	-,635	-,135	9,133	1	,003	,680	,530	,873
[area=C]	-,282	,1250	-,528	-,037	5,104	1	,024	,754	,590	,963
[area=D]	-,380	,1367	-,648	-,112	7,738	1	,005	,684	,523	,894
[area=E]	-,213	,1462	-,500	,074	2,122	1	,145	,808	,607	1,076
[area=F]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	,001[b]	2,8680E-005	,001	,001						

Dependent Variable: claimcost
Model: (Intercept), agecat, gender, area

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Rezultat statistične analize za primer-logistički model na strani 45 odeljak 4.2

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Intercept)	165,520	1	,000
katvehvalue	22,790	5	,000

Dependent Variable: clm
Model: (Intercept), katvehvalue

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
(Intercept)	-2,647	,0172	-2,681	-2,614	23799,728	1	,000	,071	,068	,073
[katvehvalue=6]	-,818	1,0156	-2,809	1,172	,649	1	,420	,441	,060	3,230
[katvehvalue=5]	-,397	,7240	-1,816	1,022	,301	1	,583	,672	,163	2,778
[katvehvalue=4]	-,571	,5102	-1,571	,429	1,254	1	,263	,565	,208	1,535
[katvehvalue=3]	,102	,1096	-,113	,317	,865	1	,352	1,107	,893	1,373
[katvehvalue=2]	,174	,0389	,097	,250	19,928	1	,000	1,190	1,102	1,284
[katvehvalue=1]	0[a]	.	.	.	.	.	.	1	.	.
(Scale)	1[b]									

Dependent Variable: clm
Model: (Intercept), katvehvalue

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Rezultat statistične analize za primer dat na strani 48 odeljak 4.4 ordinalna regresija

Parameter Estimates

Parameter		B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
Threshold	[degree=1]	,470	,0212	,429	,512	493,968	1	,000	1,601	1,536	1,669
	[degree=2]	5,049	,0450	4,961	5,137	12569,314	1	,000	155,895	142,724	170,281
[roaduserclass=2]		,151	,0270	,098	,204	31,061	1	,000	1,163	1,103	1,226
[roaduserclass=4]		,297	,0368	,224	,369	64,833	1	,000	1,345	1,252	1,446
[roaduserclass=6]		2,449	,0557	2,340	2,558	1931,804	1	,000	11,577	10,379	12,912
[roaduserclass=10]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=1]		-,179	,0323	-,242	-,116	30,697	1	,000	,836	,785	,891
[agecat=2]		-,112	,0321	-,175	-,049	12,222	1	,000	,894	,839	,952
[agecat=3]		-,058	,0364	-,129	,013	2,543	1	,111	,944	,879	1,013
[agecat=5]		,055	,0300	-,004	,114	3,367	1	,067	1,057	,996	1,120
[agecat=6]		,070	,0331	,005	,135	4,457	1	,035	1,072	1,005	1,144
[agecat=7]		,150	,0345	,083	,218	18,996	1	,000	1,162	1,086	1,244
[agecat=10]		0[a]	.	.	.	.	.	.	1	.	.
[sex=F]		,172	,0334	,106	,237	26,496	1	,000	1,188	1,112	1,268
[sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=1] * [sex=F]		,129	,0531	,025	,233	5,903	1	,015	1,138	1,025	1,263
[agecat=1] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=2] * [sex=F]		,118	,0521	,016	,220	5,115	1	,024	1,125	1,016	1,246
[agecat=2] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=3] * [sex=F]		,042	,0600	-,076	,160	,488	1	,485	1,043	,927	1,173
[agecat=3] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=5] * [sex=F]		,028	,0486	-,067	,124	,340	1	,560	1,029	,935	1,132
[agecat=5] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=6] * [sex=F]		,018	,0556	-,091	,127	,109	1	,741	1,019	,913	1,136
[agecat=6] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=7] * [sex=F]		-,148	,0595	-,265	-,032	6,209	1	,013	,862	,767	,969
[agecat=7] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=10] * [sex=F]		0[a]	.	.	.	.	.	.	1	.	.
[agecat=10] * [sex=M]		0[a]	.	.	.	.	.	.	1	.	.
(Scale)		1[b]									

Dependent Variable: degree
Model: (Threshold), roaduserclass, agecat, sex, agecat * sex

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
43650,245	16	,000

Dependent Variable: degree
Model: (Threshold), roaduserclass, agecat, sex, agecat * sex

a. Compares the fitted model against the thresholds-only model.

Tests of Model Effects

Source	Type III Wald Chi-Square	df	Sig.
(Threshold)			
roaduserclass	1968,910	3	,000
agecat	75,886	6	,000
sex	143,205	1	,000
agecat * sex	24,333	6	,000

Dependent Variable: degree
Model: (Threshold), roaduserclass, agecat, sex, agecat * sex

Literatura

[1] P. de Jong, G. Z. Heller, *Generalized Linear Models for Insurance data*, Cambridge University Press, New York, 2008.

[2] E. Ohlsson, B. Johansson, *Non-Life Insurance Pricing with Generalized Linear Models*, Springer, 2010.

[3] James K. Lindsey , *Applying Generalized Linear Models*, Springer, 1997.

[4] Edward W. Frees, *Regression Modelling with Actuarial and Financial Application*, Cambridge University Press, 2010.

[5] P. McCullagh, J. A. Nelder, *Generalized Linear Models-second edition*, Chapman and Hall, 1989.

[6] Annete J. Dobson, *An Introduction to Generalized Linear Models- second edition*, Chapman & Hall, 2002.

[7] Alan Agresti, *Foundation of Linear and Generalized Linear Models*, John Wiley & Sons, 2015.

[8] J. Kočović, M. Mitrašević, V. Rajić, *Aktuarska matematika-prvo izdanje*, Univerzitet u Beogradu, Ekonomski fakultet, 2014.

[9] Z. Lozanov-Crvenković , *Statistika*, Univerzitet u Novom Sadu, Prirodno Matematički fakultet, 2012.

[10] D. Rajter-Ćirić, *Verovatnoća -drugo dopunjeno izdanje* , Univerzitet u Novom Sadu, Prirodno Matematički fakultet, 2009.

[11] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi, *A Praktitioer's Guide to Generalized Linear Models*, Februar 2007.

[12] P. J. Green, B. W. Silverman, *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall, 1994.

[13] R. Williams, *Brief Introduction to Generalized Linear Models*, University of Notre Dame, http://www3.nd.edu/ rwilliam

[14] H. Turner, *Introduction to Generalized Linear Models*, ESRC National Centre for Research Methods, UK and Department of Statistics, University of Warwick, UK, 2008.

[15] G. Gvozdić , *Primenjena logistička regresija*, master rad, Univerzitet u Novom Sadu, Prirodno Matematički fakultet, 2011.

[16] Z. Lužanin, *'Beleške sa predavanja iz Ekonometrije'*, Univerzitet u Novom Sadu, Prirodno Matematički fakultet, 2014/2015.

Biografija

Tatjana Vučenović je rođena 1. avgusta 1991. u Novom Sadu. Završila je osnovnu školu ,,Petar Kočić" u Inđiji 2006 godine. Iste godine upisala je prirodno-matematički smer Gimnazije u Inđiji. Nakon odbranjenog maturskog rada iz matematike, oblast **Matrice** i završene srednje škole 2010. godine sa odličnim uspehom, upisala je Prirodno-Matematički fakultet u Novom Sadu, smer primenjena matematika modul matematika finansija. Osnovne studije matematike završila je 7. septembra 2013. godine. U julu 2013. godine položila je prijemni ispit i upisala master studije na istom fakultetu nastavljajući smer primenjena matematika. Položila je sve ispite na master studijama koji su predviđeni planom i programom zaključno sa septembrom 2015. godine, i time stekla uslov za odbranu master rada. Za vreme studija bila je stipendista opštine Inđija, Republike Srbije kao i dobitnica nagrade fakulteta za ostvaren uspeh u toku osnovnih studija.
U periodu april-jun 2016. godine radila je kao nastavnik matematike u osnovnoj školi ,,Jovan Popović" u Inđiji.
Njena oblast interesovanja su statistika, finansijska i aktuarska matematika.

U Novom Sadu, 2016. Tatjana Vučenović

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: Tatjana Vučenović
AU
Mentor: dr Dora Seleši
MN

Naslov rada: Uopšteni linearni modeli sa primenama u aktuarstvu
NR

Jezik publikacije: srpski(latinica)
JP
Jezik izvoda: srpski/engleski
JI
Zemlja publikovanja: Srbija
ZP
Uže geografsko područje: Vojvodina
UGP
Godina: 2016.
GO
Izdavač: Autorski reprint
IZ
Mesto i adresa: Novi Sad, Prirodno-matematički fakultet, Trg Dositeja Obradovića 4
MA
Fizički opis rada: 7/80/0/22/16/0/1
(broj poglavlja/strana/lit. citata/tabela/slika/grafika/priloga)
FO

Naučna oblast: Matematika
NO
Naučna disciplina: Aktuarska matematika
ND
Predmetna odrednica/Ključne reči: Uopšteno linearno modeliranje, eksponencijalna familija raspodela, varijansna funkcija, metoda maksimalne verodostojnosti, devijansa, analiza reziduala, modeliranje neprekidnih promenljivih u aktuarstvu, gama regresija, inverzna Gausova regresija, modeliranje kategorijalnih promenljivih, logistički model, Poasonova regresija, pojam ,,overdispersion", uopšteni aditivni modeli
PO

UDK:
Čuva se: u biblioteci Departmana za matematiku i informatiku, Novi Sad
ČU
Važna napomena:
VN
Izvod: Tema master rada su uopšteni linearni modeli. Prvo je definisan matematički model i objašnjena je teorijska osnova rada koja uključuje eksponencijalnu familiju raspodela, varijansnu funkciju, metode za ocene nepoznatih parametara modela, testiranje hipoteza, statistike za procenu adekvatnosti fitovanja (devijansa i Pirsonova statistika), reziduali i još neke dijagnostike modela. Zatim je kroz primere pokazano da se iznosi zahteva za odštetu mogu modelirati pomoću gama i inverzne Gausove raspodele. Ukratko su objašnjene kategorijalne promenljive i dat je primer u kome binarne podatke modeliramo pomoću logističke regresije. Zatim je objašnjena Poasonova regresija koja je pogodna za modeliranje prirodnih brojeva kao na primer broj zahteva za odštetu unutar odgovarajuće grupe polisa osiguranja. Definisan je i pojam ,,overdispersion" i obrađen je jedan primer u kome za modeliranje koristimo negativnu binomnu raspodelu. Na kraju rada je kratak teorijski osvrt na neke naprednije modele koji predstavljaju proširenja uopštenog linearnog modela , bez datih primera. Svi rezultati su dobijeni u statističkom programu SPSS.
IZ

Datum prihvatanja teme od strane NN Veća: 8. 7. 2016.
DP
Datum odbrane:
DO
Članovi komisije:
Predsednik: dr Sanja Rapajić, vanredni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Dora Seleši, vanredni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, mentor

Član: dr Nataša Krklec Jerinkić, docent, Prirodno-matematički fakultet, Univerzitet u Novom Sadu
KO

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
KEY WORDS DOCUMENTATION

Accession number:
ANO
Identification number:
INO
Document type: Monograph type
DT
Type of record: Printed text
TR
Contents code: Master's thesis
CC
Author: Tatjana Vučenović
AU
Mentor: Dora Seleši, PhD
MN

Title: Generalized linear models with application in actuarial science
TI

Language of text: Serbian
LT
Language of abstract: English/Serbian
LA
Country of publication: Serbia
CP
Locality of publication: Vojvodina
LP
Publication year: 2016.
PY
Publisher: Author's reprint
PU

Publication place: Novi Sad, Faculty of Science, Dositeja Obradovića 4
PP
Physical description: 7/80/0/22/16/0/1
(chapters/pages/literature/tables/pictures/graphics/appendices)
PD
Scientific field: Mathematics
SF
Scientific discipline: Actuarial science
SD
Subject / Key words: Generalized Linear modelling, exponential family of distributions, variance function, method of maximum likelihood estimation, deviance, analysis of residuals, modelling continuous variables in actuarial science, gamma regression, inverse Gaussian regression, modelling categorical variables, logistic model, Poisson regression, concept ,,overdispersion", generalized additive models
SKW

UC:
Holding data: Library of the Department of Mathematics and Informatics, Novi Sad
HD
Note:
N
Abstract: The subject of this master's thesis are generalized linear models. Firstly, the mathematical model was defined and the theoretical basis of the thesis explained, which included exponential family of distributions, variance function, methods for assessing unknown parameters of a model, hypothesis testing, goodness of fit (deviance and Pearson statistics), residuals and some other model diagnostics. Subsequently, examples were given to show how claim amounts can be modelled using gamma and inverse Gaussian distribution. Categorial variables were briefly explained and an example was given showing use of logistical regression for binary data modelling. Poisson regression was later described, suitable for modelling of natural numbers such as the number of claims within a certain group of claims. Overdispersion was defined and an example was given showing use of negative binomial distribution. Finally, a brief theoretical review of some advanced models which are expansions of generalized linear model was given, without examples. Statistical software used for all results was SPSS.
AB

Accepted by Scientific Board on: July 8, 2016
ASB
Defended:
DE
Thesis defend board:

President: Sanja Rapajić, PhD, associate professor, Faculty of Science, University of Novi Sad

Member: Dora Seleši, PhD, associate professor, Faculty of Science, University of Novi Sad, mentor

Member: Nataša Krklec Jerinkić, PhD, assistant professor, Faculty of Science, University of Novi Sad

DB