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

Poasonova regresija i primene

- Master rad -

Mentor:
Prof. dr. Zagorka
Lozanov-Crvenković

Kandidat:
Sanja Bojović
460m/10

Novi Sad, *Jun 2014*.

Sadržaj

Predgovor

.....

Tema ovog rada je Poasonova regresija za prebrojive podatke, kao specijalni slučaj uopštenih linearnih modela. Široka primena ovog oblika regresije u mnogim drugim oblastima nauke i prakse bila je primarni motiv za detaljnije upoznavanje sa teorijskom podrškom koja je vezana za njih, kao i za sprovođenje istraživanja. U radu je data i osnovna teorija koja se odnosi na uopštene linearne modele, sa akcentom na Poasonovu slučajnu promenljivu.

U prvom poglavlju su uvedeni osnovne oznake i pojmovi koji su neophodni za dalje razumevanje rada. Drugo poglavlje sadrži kratak istorijski pregled razvoja uopštenih linearnih modela. U trećem poglavlju su definisani uopšteni linearni modeli, a zatim je prikazana njihova konstrukcija i objašnjene su tri osnovne komponente. Uopšteni linearni modeli su ograničeni na članove jedne specijalne familije raspodela, eksponencijalne familije, pa zbog toga dajemo detaljniji pregled osobina važnijih članova ove familije. Eksponencijalna familija raspodela predstavlja bazu za određivanje funkcije raspodele kod uopštenih linearnih modela.

Četvrto poglavlje detaljno opisuje Poasonovu slučajnu promenljivu i modeliranje prebrojivih podataka Poasonovom regresijom, koje se sastoji od četiri osnovna koraka: postavljanje modela, ocenjivanje parametara modela, provera adekvatnosti modela i zaključivanje, u koje spadaju računanje intervala poverenja i testiranje hipoteza, kao i interpretacija rezultata. Ocene parametara modela su izvedene metodom maksimalne verodostojnosti, pomoću algoritma iterativnih težinskih najmanjih kvadrata. Posebna pažnja je posvećena definisanju i rešavanju problema preraspršenosti ili prekoračenja disperzije. Za prevazilaženje ovog problema predloženi su alternativni modeli, kvazi-Poasonov i negativni binomni model.

Peto poglavlje uvodi postavljanje Poasonove regresije za stope, tj. kada podatke posmatramo u procentima.

Šesto poglavlje je rezervisano za primenu modeliranja Poasonovom regresijom na primeru konzumiranja neoporezovanih duvanskih proizvoda. Podaci su obrađeni u statističkom paketu SPSS, a zatim je data analiza promenljivih u modelu, kao i zaključak o statističkim značajnostima parametara modela i interpretacija rezultata.

.....

Posebno se zahvaljujem svom mentoru, *prof. dr. Zagorki Lozanov-Crvenković* prvenstveno na svom stečenom znanju, zatim na stručnim sugestijama, pomoći prilikom izbora literature i profesionalnom usmeravanju pri izradi ovog rada.
Takođe, neizmerno hvala na ogromnoj nesebičnoj podršci i razumevanju Milanu, Dragici, Goranu i Darku.
.....

Novi Sad, *Jun 2014.* *Sanja Bojović*

Uvod

Standardni linearni modeli imaju široku upotrebu, jer se pomoću njih mogu modelirati mnogi tipovi podataka i postoje razne teorije njihove primene. Međutim, sve više se istražuju metode i modeli koji prevazilaze ograničenja standardnih linearnih modela. Uopšteni linearni modeli, koji predstavljaju generalizaciju standardnih linearnih modela, dopuštaju izbor raspodele podataka, pa se na taj način više ne postavlja uslov da podaci imaju normalnu raspodelu ili primenjuju transformacije podataka tako da imaju normalnu raspodelu. Ovi modeli su ograničeni na članove eksponencijalne familije raspodela koja sadrži specijalne slučajeve kao što su normalna, binomna, Poasonova, gama i inverzna Gausova raspodela. Specijalno, Poasonova raspodela je pogodna za modeliranje prebrojivih podataka. Uopšteni linearni modeli su uvedeni od strane Neldera i Vederburna, kao način za ujedinjenje različitih statističkih modela, uključujući linearnu, logističku i Poasonovu regresiju.

Pre svega, u prvom poglavlju ćemo uvesti oznake i osnovne pojmove, a zatim u dugom poglavlju izložiti motivaciju rada, kao i kratak istorijski osvrt. U trećem poglavlju ćemo najpre definisati uopštene linearne modele, koji se sastoje od tri komponente: komponente slučajnosti, sistematičnosti i funkcije veze i biće objašnjena njihova konstrukcija. Tipovi uopštenih linearnih modela će biti razmotreni u smislu izbora familije raspodele, kao i funkcije veze i biće definisan pojam kanoničke veze.

Četvrto poglavlje će biti posvećeno Poasonovoj regresiji i modeliranju prebrojivih podataka. Najjednostavniji uopšteni linearni model za podatke dobijene prebrojavanjem podrazumeva Poasonovu raspodelu komponente slučajnosti i kanoničku log funkciju veze. Kao i podaci dobijeni prebrojavanjem, Poasonove raspodele uzimaju nenegativne celobrojne vrednosti. Poasonova raspodela, koju predstavljamo kao $Y \sim Pois(\mu)$, potpuno je određena srednjom vrednosti μ, s obzirom da je njena disperzija takođe jednaka μ. Ova osobina Poasonove slučajne promenljive da je njena disperzija jednaka srednjoj vrednosti predstavlja i ograničenje u izvesnom smislu. U praksi se često dešava da je disperzija registrovanih prebrojivih podataka veća od srednje vrednosti i taj slučaj se naziva preraspršenost podataka. Preraspršenost predstavlja prekoračenje disperzije koje potiče iz toga kako je definisana stohastička komponenta modela, pri čemu je sistematička struktura modela tačna. Prisustvo preraspršenosti se ne sme ignorisati, jer

čak i ako je forma fitovanog modela tačna, ne uračunavanje preraspršenosti dovodi do netačnih ocena disperzija, čime nastaju previše uski intervali poverenja i suviše male p-vrednosti značajnosti testova. Zbog toga će biti uvedene metode za identifikovanje i prevazilaženje preraspšenosti, tačnije kvazi-Poasonov i negativni binomni regresioni model.

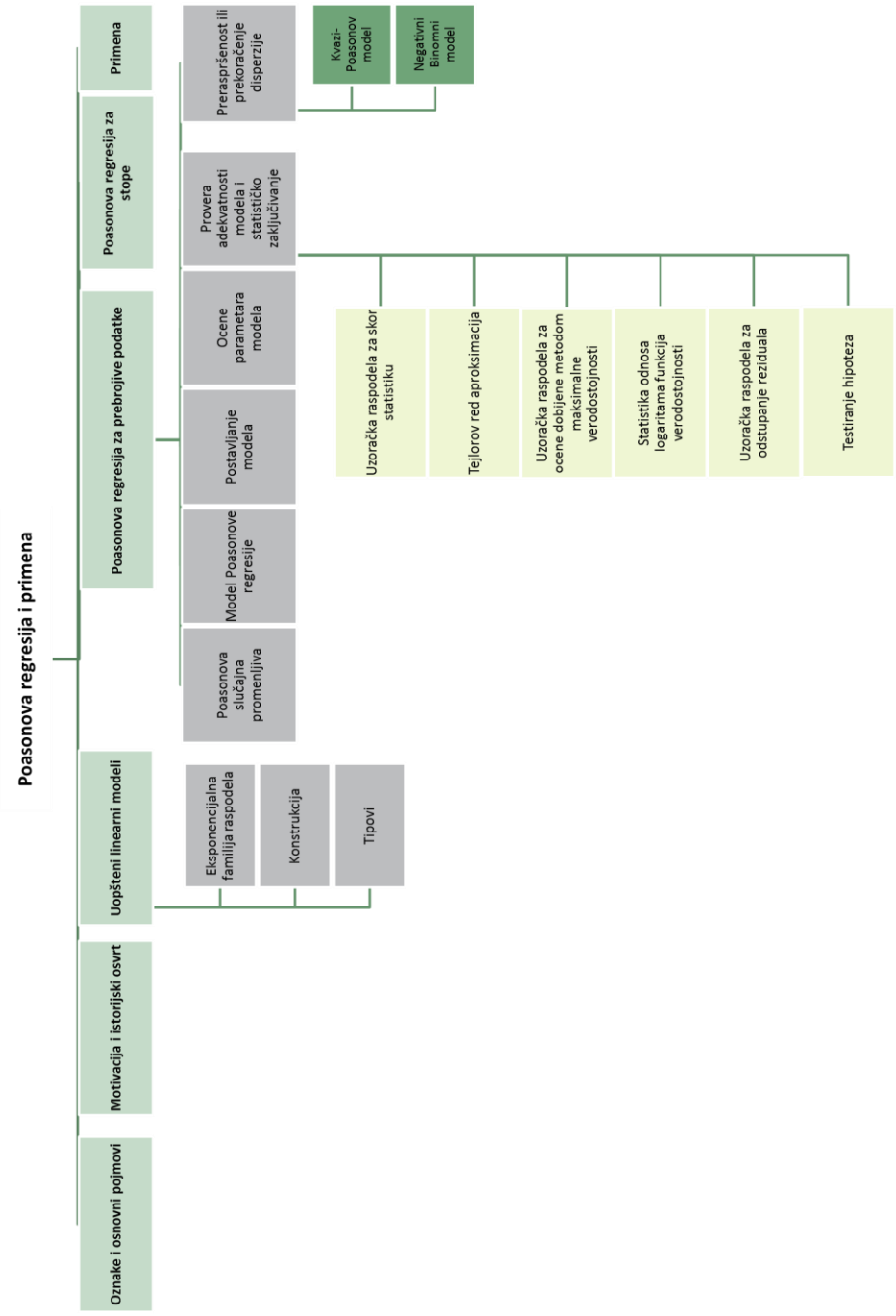
Takođe, u ovom poglavlju će detaljno biti izloženi koraci statističkog modeliranja:

- Određivanje modela – model se određuje iz dva dela: jednačinom koja povezuje obeležje i nezavisne promenljive i raspodelom verovatnoće obeležja.
- Ocena parametara modela, gde će se koristiti algoritam iterativnih težinskih najmanjih kvadrata.
- Provera slaganja modela sa podacima.
- Zaključak – računanje intervala poverenja i testiranje hipoteza o parametrima modela, kao i interpretacija rezultata.

U petom poglavlju pokazujemo da možemo postaviti Poasonovu regresiju tako da posmatramo podatke u procentima. U tom slučaju obeležje predstavljamo kao stopu (ili incidencu).

U šestom poglavlju biće data primena Poasonove regresije na konkretnim podacima i uz pomoć statističkog paketa SPSS. Model će pokazivati kako različiti faktori (na primer, blizina državne granice, raspoloživi prihodi, itd.) utiču na pojavu i obim korišćenja neoporezovanih (ilegalnih) pakovanja duvanskih proizvoda kod potrošača.

Na narednoj stranici dat je kratak pregled sadržaja i ideja rada.


Poasonova regresija i primena
Oznake i osnovni pojmovi
Motivacija i istorijski osvrt
Uopšteni linearni modeli
Eksponencijalna familija raspodela
Konstrukcija
Tipovi
Poasonova regresija za prebrojive podatke
Poasonova slučajna promenljiva
Model Poasonove regresije
Postavljanje modela
Ocene parametara modela
Uzoračka raspodela za skor statistiku
Tejlorov red aproksimacija
Uzoračka raspodela za ocene dobijene metodom maksimalne verodostojnosti
Statistika odnosa logaritama funkcija verodostojnosti
Uzoračka raspodela za odstupanje reziduala
Testiranje hipoteza
Provera adekvatnosti modela i statističko zaključivanje
Prerasprešenost ili prekoračenje disperzije
Kvazi-Poasonov model
Negativni Binomni model
Poasonova regresija za stope
Primena

I. Oznake i osnovni pojmovi

1. Oznake:

Za označavanje slučajnih promenljivih koristimo standardni pristup, pišemo ih velikim slovima latinice, a registrovane vrednosti odgovarajućim malim slovima latinice. Na primer, registrovane vrednosti $y_1, y_2, \dots, y_n$ su realizacije slučajnih promenljivih $Y_1, Y_2, \dots, Y_n$. Grčka slova ćemo koristiti da označimo parametre, a odgovarajuća mala latinična slova za njihove ocene. Simbol ^ ćemo takođe koristiti za ocenjene vrednosti. Na primer, parametar β je ocenjen sa $\hat{\beta}$ ili b. U radu se ponekad nećemo striktno držati ovih pravila, ili da bismo na taj način izbegli suvišne zapise, gde je značenje očigledno iz konteksta, ili ukoliko postoji tradicija alternativnog zapisa (na primer, e ili ε za termine grešaka).

Vektori i matrice, bilo da su stohastički ili ne, se označavaju podebljanim malim i velikim slovima, respektivno. Dakle, $\boldsymbol{y}$ predstavlja vektor realizovanih vrednosti

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

ili vektor slučajnih promenljivih

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

$\boldsymbol{\beta}$ predstavlja vektor parametara, a X je matrica. Oznaka ^{T} se koristi za transponovane matrice ili u slučaju kada vektor kolonu pišemo kao red, na primer, $\boldsymbol{Y} = [Y_1, \dots, Y_n]^T$.

2. Osnovni pojmovi:

Def. 1.1: Preslikavanje $Y: \Omega \to \mathbb{R}$ je *slučajna promenljiva* nad prostorom verovatnoća $(\Omega, \mathcal{F}, \mathrm{P})$ ako $Y^{-1}(S) \in \mathcal{F}$ za svako $S \in \mathcal{B}$, gde je $\mathcal{B} = \mathcal{B}(\mathbb{R})$ Borelovo σ-polje. Ekvivalentno, kažemo da je Y $\mathcal{F}$-merljivo.

Kako je u prostoru verovatnoća $(\Omega, \mathcal{F}, \mathrm{P})$ verovatnoća definisana za svaki skup iz $\mathcal{F}$ i kako $Y^{-1}(S) \in \mathcal{F}$ za svako $S \in \mathcal{B}(\mathbb{R})$, to znači da je za svako $S \in \mathcal{B}(\mathbb{R})$ definisana funkcija

$$P_Y(S) \coloneqq P\{Y \in S\} = P\{\omega | Y(\omega) \in S\} = P\big(Y^{-1}(S)\big).$$

Tako definisana funkcija $P_Y(S)$, $S \in \mathcal{B}(\mathbb{R})$ zove se *raspodela verovatnoća slučajne promenljive* Y.

Def. 1.2: Slučajna promenljiva Y je *diskretna* (diskretnog tipa) ako postoji prebrojiv skup brojeva R_Y takav da je $P\{Y \in \bar{R}_Y\} = 0$, odnosno ako je skup slika od Y najviše prebrojiv skup.

Def. 1.3: Slučajne promenljive $Y_1, Y_2, \ldots$ su *nezavisne* ako su događaji ${Y_1}^{-1}(S_1), {Y_2}^{-1}(S_2), \ldots$ nezavisni za sve Borelove skupove $S_i \in \mathcal{B}(\mathbb{R}), i = 1,2, \ldots.$

Specijalno, za dvodimenzionalnu slučajnu promenljivu diskretnog tipa (X, Y) sa raspodelom $p\big(x_i, y_j\big), i, j = 1,2, \ldots$ lako se proverava potreban i dovoljan uslov za nezavisnost X i Y:

$$P\big(\{X = x_i\} \cap \big\{Y = y_j\big\}\big) = P\{X = x_i\} \cdot P\big\{Y = y_j\big\}, \qquad i, j = 1,2, \ldots$$

ili, kraće

$$p\big(x_i, y_j\big) = p(x_i) \cdot q\big(y_j\big), \qquad i, j = 1,2, \ldots$$

Slučajna promenljiva Y definisana nad prostorom verovatnoća $(\Omega, \mathcal{F}, \mathrm{P})$ je određena svojom raspodelom verovatnoća: $\forall S \in \mathcal{B},\ P_Y(S) = P\{Y \in S\}$. Vidimo da je raspodela verovatnoća $P_Y(\cdot)$ funkcija skupova, a ne tačke. Kako bismo koristili aparat matematičke analize, odgovaralo bi nam da definišemo funkciju tačke koja bi u potpunosti određivala slučajnu promenljivu Y. Zato definišemo funkciju raspodele (verovatnoća) slučajne promenljive Y.

Def. 1.4: Funkcija $F_Y(y)\colon \mathbb{R} \mapsto [0,1]$ definisana sa

$$F_Y(y) = P_Y\big((-\infty, y)\big) = P\{\omega \in \Omega | Y(\omega) < y\}$$

naziva se *funkcija raspodele slučajne promenljive* Y.

Funkcija raspodele F_Y u tački $y \in \mathbb{R}$ predstavlja verovatnoću događaja sastavljenu od onih elementarnih događaja ω čija je slika $Y(\omega)$ manja od y. To kraće pišemo kao

$$F_Y(y) = P\{Y < y\}.$$

Funkcija raspodele postoji i jedinstvena je za svaku slučajnu promenljivu i ona određuje sva bitna svojstva slučajne promenljive. Takođe, treba napomenuti da iako je funkcija raspodele jedinstvena za svaku slučajnu promenljivu, postoji beskonačno mnogo slučajnih promenljivih koje imaju iste raspodele.

Def. 1.5: Preslikavanje $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, $\boldsymbol{Y}: \Omega \to \mathbb{R}^n$ je *n-dimenzionalna slučajna promenljiva* na prostoru verovatnoća $(\Omega, \mathcal{F}, \mathrm{P})$ ako za svako $S \in \mathcal{B}_n$ važi

$$\{\omega | \boldsymbol{Y}(\omega) \in S\} = \{\boldsymbol{Y} \in S\} = \boldsymbol{Y}^{-1}(S) \in \mathcal{F}.$$

Def. 1.6: *Funkcija raspodele n-dimenzionalne slučajne promenljive* $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ je

$$F_{\boldsymbol{Y}}(y_1, \dots y_n) = F_{(Y_1, \dots, Y_n)}(y_1, \dots y_n)$$
$$= P(\{Y_1 < y_1\} \cap \dots \cap \{Y_n < y_n\}), \qquad -\infty < y_1, \dots y_n < \infty.$$

Def. 1.7: *Očekivanje* $E(Y)$ diskretne slučajne promenljive Y sa raspodelom $p(y_k), k = 1,2, \dots$ definiše se sa

$$E(Y) = \sum_{k=1}^{\infty} y_k \, p(y_k),$$

i postoji ako i samo ako

$$\sum_{k=1}^{\infty} |y_k| \; p(y_k) < \infty.$$

Def. 1.8: *Momenat* reda $k, k \in \mathbb{N}$ slučajne promenljive Y je $E(Y^k)$. *Centralni momenat* reda $k, k \in \mathbb{N}$ slučajne promenljive Y je

$$E\left(\left(Y - E(Y)\right)^k\right).$$

Dakle, vidimo da je očekivanje u stvari momenat reda 1.

Def. 1.9: Centralni momenat reda 2 slučajne promenljive Y zove se *disperzija (varijansa)* slučajne promenljive Y i označava se sa $D(Y)$ ili $\sigma^2(Y)$. Dakle,

$$D(Y) = E\left(\left(Y - E(Y)\right)^2\right).$$

Disperzija ili varijansa slučajne promenljive je brojna karakteristika koja predstavlja meru odstupanja od srednje vrednosti.

II. Motivacija i istorijski osvrt

Statističko modeliranje nastalo je kao potreba da se predvidi najverovatnije ponašanje sistema podataka u budućnosti. Osnovna svrha građenja modela je da dobijemo odgovarajuće procene sa malim odstupanjima o tome kako je jedna ili više slučajnih promenljivih povezana sa jednom ili više drugih promenljivih. Standardni linearni modeli imaju široku upotrebu, jer se pomoću njih mogu modelirati mnogi tipovi podataka i postoje razne teorije njihove primene. Međutim, sve više se istražuju metode i modeli koji premašuju ograničenja standardnih linearnih modela. Na primer, postoje brojni tipovi podataka koji nemaju normalnu raspodelu. Da bi se prevazišao ovaj problem mogu da se koriste transformacije u cilju normalizacije podataka. Međutim, diskretna obeležja često znaju da imaju nule za registrovane vrednosti i njihove standardne greške nemaju normalnu raspodelu. Uopšteni linearni modeli, koji predstavljaju ekstenziju standardnih linearnih modela, dopuštaju izbor raspodele podataka, što rešava problem transformacije podataka u normalno raspodeljene. Naravno, da bismo dobili najbolje procene obeležja određenog sistema, vrlo je važno fitovati podatke na odgovarajući način.

Uopšteni linearni modeli se ravijaju u proteklih više od 100 godina. Ukratko, istorija razvoja izgleda ovako:

- Višestruka linearna regresija (Legendre, Gaus – početak XIX veka)
- Eksperimenti na osnovu analize varijanse (ANOVA) – normalna raspodela sa vezom identiteta (Fišer, 1920. – 1935.)
- Funkcija verodostojnosti – uopšteni pristup značajnosti proizvoljnog statističkog modela (Fišer, 1922.)
- Testovi razblaživanja – binomna raspodela sa dodatnom log log vezom (Fišer, 1922.)
- Eksponencijalna familija – klasa raspodela sa dovoljnim statistikama[1] za parametre (Fišer, 1934.)
- Probit analiza – binomna raspodela sa probit vezom (Blis, 1935.)

[1] Statistika je dovoljna u odnosu na statistički model i njegov pridruženi nepoznati parametar, ako nijedna druga statistika koja može biti dobijena iz istog uzorka ne obezbeđuje nijednu dodatnu informaciju.

- Logit za proporcije – binomna raspodela sa logit vezom (Berkson, 1944.; Djuke i Paterson, 1952.)
- Log-linearni modeli za prebrojive podatke – Poasonova raspodela sa log vezom (Birč, 1963.)
- Regresioni modeli za analizu preživljavanja – eksponencijalna raspodela sa recipročnom ili log vezom (Frajgl i Zelen, 1965.; Zipin i Armitage, 1966.; Glaser, 1967.)
- Inverzni polinomi – Gama raspodela sa recipročnom vezom (Nelder, 1966.)

Dakle, poznato je još od vremena Fišera (1934.) da su mnoge od najčešće korišćenih raspodela članovi jedne familije, koju nazivamo eksponencijalna familija raspodela. Do kraja 1960.-ih, bilo je pravo vreme za sintezu ovih različitih modela (Lindsi, 1971.). Nelder i Vederburn su otišli korak dalje i 1972. ujedinili teoriju statističkog modeliranja, naročito regresionih modela, time što su objavili članak o *Uopštenim linearnim modelima*. Oni su pokazali dve stvari. Prvo, da je značajan broj najčešće korišćenih linearnih regresionih modela klasične statistike članova jedne familije, koji se mogu tretirati na isti način. Drugo, da procene maksimalne verodostojnosti kod ovih modela mogu biti dobijene istim algoritmom, *iterativnim težinskim najmanjim kvadratima*. U daljem razvoju, oba elementa su imala podjednaku ulogu.

III. Uopšteni linearni modeli

Kao što smo već napomenuli, uopšteni linearni modeli su uvedeni od strane Neldera i Vederburna, kao način za ujedinjenje različitih statističkih modela, uključujući linearnu, logističku i Poasonovu regresiju. Oni predstavljaju fleksibilnu generalizaciju klasične linearne regresije, koja dozvoljava obeležju da ima standardne greške koje nisu normalno raspodeljene. Uopšteni linearni modeli, dakle, uopštavaju linearnu regresiju tako što dopuštaju linearnom modelu da sadrži obeležja koja imaju raspodelu različitu od normale.

Uopšteni linearni modeli su ograničeni na članove jedne specijalne familije raspodela, *eksponencijalne familije*, koja ima pogodne statističke osobine. Zapravo, ovaj uslov proizilazi iz čisto tehničkih razloga: numerički algoritam, iterativni težinski najmanji kvadrati, koji se koristi za ocene parametara modela, funkcioniše samo unutar ove familije raspodela. Uz pomoć modernih kompjutera, ovo ograničenje se može relativno jednostavno prevazići.

1. Eksponencijalna familija raspodela

Eksponencijalna familija raspodela predstavlja skup raspodela koji sadrži kako neprekidne, tako i na diskretne slučajne promenljive. Članovi ove raspodele imaju mnoge važne osobine, koje se mogu razmatrati uopšteno i važe za sve članove familije. Eksponencijalna familija raspodela predstavlja bazu za određivanje funkcije raspodele kod uopštenih linearnih modela. Posmatrajmo slučajnu promenljivu Y čija raspodela verovatnoća zavisi od parametra θ. Za raspodelu možemo reći da pripada *eksponencijalnoj familiji*, ako ima sledeći oblik

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}, \tag{3.1}$$

gde su a, b, s i t poznate funkcije. Primetimo simetriju između y i parametra θ, koja naročito dolazi do izražaja ako jednačinu (3.1) napišemo u sledećem obliku

$$f(y;\theta) = e^{\left(a(y)b(\theta)+c(\theta)+d(y)\right)},$$

gde je $s(y) = e^{d(y)}$, a $t(\theta) = e^{c(\theta)}$.

Ako je $a(y) = y$, tada kažemo da je raspodela u *kanoničkom* (ili standardnom) *obliku*, a $b(\theta)$ se ponekad naziva *prirodni parametar* raspodele.

Eksponencijalna familija raspodela koju smo upravo definisali sadrži specijalne slučajeve kao što su normalna, binomna, Poasonova, gama i inverzna Gausova raspodela. Sada ćemo razmotriti neke važnije osobine ovih raspodela.

Gausova (normalna) raspodela sa sredinom μ i disperzijom σ^2 ima funkciju gustine

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

Funkciju gustine možemo zapisati u kanoničkom obliku na sledeći način

$$f(y;\mu) = e^{\left(-\frac{y^2}{2\sigma^2}+\frac{y\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right)}.$$

Prirodni parametar je $b(\mu) = \frac{\mu}{\sigma^2}$. U zavisnosti od vrednosti parametara μ i σ, grafici krivih gustina su različiti, ali se mogu uočiti neke zajedničke crte. Sve krive gustine su simetrične u odnosu na pravu $y = \mu$. Promena vrednosti parametra μ dovodi do translacije krive gustine duž apscisne ose. Promena vrednosti parametra σ dovodi do promene spljoštenosti krive gustine (raspršenosti oko tačke $y = \mu$). U slučaju kada su parametri normalne raspodele $\mu = 0$ i $\sigma^2 = 1$ dobijamo normalnu $\mathcal{N}(0,1)$ raspodelu koja se naziva *standardna normalna raspodela*.

Normalna raspodela se koristi za modeliranje neprekidnih podataka koji imaju simetričnu raspodelu. Ona ima široku primenu zbog sledeće tri bitne karakteristike. Prvo, mnoge prirodne pojave mogu dobro da se opišu normalnom raspodelom. Na primer, visina ili krvni pritisak kod ljudi. Drugo, čak i ako slučajne promenljive nemaju normalnu raspodelu (na primer, ako je njihova raspodela asimetrična), raspodela srednjih vrednosti dovoljno velikog broja nezavisnih i jednako raspodeljenih slučajnih promenljivih, pri čemu svaka od njih ima konačnu srednju vrednost i varijansu, približno odgovara normalnoj raspodeli. Ovo je dokazano u Centralnoj graničnoj teoremi, čiju formulaciju i dokaz dajemo u dodatku. Treće, ukoliko neprekidna promenljiva Y nije normalno raspodeljena, često se može identifikovati relativno jednostavna transformacija, kao na primer, $Y' = \log Y$ ili $Y' = \sqrt{Y}$, koja daje podatke Y' sa približno normalnom raspodelom. Zbog toga se veliki deo statističke teorije bavi upravo normalnom raspodelom.

Binomna raspodela je diskretna raspodela koja ima funkciju gustine

$$f(y;p) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Ovde y predstavlja broj uspešnih događaja u n pokušaja, a $n-y$ je broj neuspešnih. Broj $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ se zove *binomni koeficijent*. Binomna raspodela zavisi od dva parametra $n \in \mathbb{N}$ i $p \in (0,1)$. Ako slučajna promenljiva Y ima binomnu raspodelu sa parametrima n i p to zapisujemo $Y: \mathcal{B}(n,p)$. Binomna raspodela, dakle, predstavlja model za izvođenje n istih pokušaja, pri čemu se svaki od njih može realizovati uspešno (sa verovatnoćom p) ili neuspešno (sa verovatnoćom $1-p$), nezavisno od ishoda ostalih pokušaja. Tada slučajna promenljiva $Y: \mathcal{B}(n,p)$ predstavlja broj pokušaja (od n) koji su se uspešno realizovali. Funkciju gustine binomne raspodele možemo zapisati u kanoničkom obliku kao

$$f(y;p) = e^{\left(y\log p - y\log(1-p) + n\log(1-p) + \log\binom{n}{y}\right)}.$$

Binomna raspodela je često prvi izbor kod modeliranja procesa sa binarnim ishodima, kao što su, na primer, broj kandidata koji su položili test (mogući ishod za svakog od kandidata je da je položio ili da je pao), broj pacijenata sa određenom bolesti koji su živi u navedenom vremenskom periodu nakon diagnoze (mogući ishod je da je pacijent živ ili nije).

Poasonova raspodela je diskretna raspodela sa funkcijom gustine koja zavisi od parametra $\mu > 0$:

$$f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!},$$

gde y uzima vrednosti $0,1,2,\ldots$ To možemo drugačije zapisati kao

$$f(y;\mu) = e^{(y\log\mu - \mu - \log y!)},$$

što predstavlja kanonički oblik Poasonove raspodele, s obzirom da je $a(y) = y$. Takođe, vidimo da je prirodni parametar $\log\mu$.

Očekivanje i disperzija Poasonove slučajne promenljive jednaki su μ, tako da nema potrebe ocenjivati posebno svaki od ova dva parametra. Kao što ćemo videti kasnije, Poasonova raspodela je pogodna za modeliranje prebrojivih podataka. Kako se μ povećava, Poasonova raspodela se približava normalnoj. Primeri podataka koji imaju Poasonovu raspodelu su broj slučajnih slovnih grešaka na jednoj stranici časopisa, broj pogrešnih komponenti u kompjuteru, broj čestica pri raspadu radioaktivne materije u određenom vremenskom periodu. Realni podaci koji mogu biti dobro modelirani

pomoću Poasonove raspodele često imaju veću disperziju od srednje vrednosti i tada imamo problem preraspršenosti podataka. U tom slučaju model mora biti prilagođen tako da odražava ovu osobinu. U poglavlju IV. 6. ćemo se detaljnije baviti metodama kojima se modeliranje prilagođava takvim podacima.

Gama raspodela je neprekidna familija sa funkcijom gustine određenom parametrima $\omega, \psi > 0$:

$$p(y) = \left(\frac{y}{\omega}\right)^{\psi-1} \frac{e^{\frac{-y}{\omega}}}{\omega\Gamma(\psi)}, \qquad \text{za } y > 0,$$

gde je $\Gamma(\cdot)$ gama funkcija[2]. Očekivanje i disperzija gama raspodele su, respektivno, $E(Y) = \omega\psi$ i $D(Y) = \omega^2\psi$. Parametar ω utiče na širenje gama raspodele, dok parametar ψ kontroliše nagib raspodele. Što je parametar ψ veći, to je raspodela više simetrična. Gama raspodela je korisna za modeliranje pozitivnih neprekidnih obeležja, kada njihova uslovna disperzija raste zajedno sa njihovom srednjom vrednošću, ali gde je koeficijent varijacije obeležja konstanta.

Inverzna Gausova raspodela je takođe neprekidna familija određena sa dva parametra, μ i λ, sa funkcijom gustine

$$p(y) = \sqrt{\frac{\lambda}{2\pi y^3}}\, e^{-\frac{\lambda(y-\mu)^2}{2y\mu^2}}, \qquad \text{za } y > 0.$$

Očekivanje i disperzija za Y su $E(Y) = \mu$ i $D(Y) = \mu^3/\lambda$. Slično kao i kod gama raspodele, disperzija inverzne Gausove raspodele se povećava sa sredinom, ali mnogo brže. Nagib se takođe povećava sa μ, a smanjuje sa λ.

Primeri raspodela koje ne pripadaju eksponencijalnoj familiji su Košijeva, uniformna, itd.

Sada ćemo prikazati osobine raspodela iz eksponencijalne familije. Pre svega, potrebno je pokazati kako dolazimo do očekivanja i disperzije za $a(Y)$.

Iz definicije gustine raspodele znamo da je površina ispod krive jednaka jedinici, pa važi

$$\int_{-\infty}^{\infty} f(y;\theta)dy = 1, \tag{3.2}$$

[2] Gama funkcija je definisana kao $\Gamma(x) = \int_0^\infty e^{-z}z^{x-1}dz$ i može se smatrati neprekidnim uopštenjem funkcije faktorijala, kada je x nenegativan ceo broj, $x! = \Gamma(x+1)$.

a ukoliko je slučajna promenljiva diskretna, tada umesto integrala koristimo sume.

Ukoliko potražimo prvi izvod po θ, dobijamo

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(y;\theta)dy = \frac{d}{d\theta}1 = 0.$$

Za eksponencijalnu familiju raspodela uvek je dozvoljeno menjati redosled integracije i diferenciranja (što ne mora uvek da važi za raspodele koje ne pripadaju eksponencijalnoj familiji), pa prema tome, dobijamo

$$\int_{-\infty}^{\infty} \frac{df(y;\theta)}{d\theta}dy = 0. \tag{3.3}$$

Analogno, ukoliko dva puta diferenciramo po θ (3.2), važi sledeće

$$\int_{-\infty}^{\infty} \frac{d^2f(y;\theta)}{d\theta^2}dy = 0. \tag{3.4}$$

Dalje, ukoliko jednačinu za raspodelu

$$f(y;\theta) = e^{\left(a(y)b(\theta)+c(\theta)+d(y)\right)}$$

diferenciramo po θ, dobijamo sledeće

$$\frac{df(y;\theta)}{d\theta} = \big(a(y)b'(\theta) + c'(\theta)\big)f(y;\theta).$$

Iz (3.3) sledi

$$\int_{-\infty}^{\infty} \big(a(y)b'(\theta) + c'(\theta)\big)f(y;\theta)\,dy = 0$$

$$\int_{-\infty}^{\infty} a(y)b'(\theta)f(y;\theta)\,dy + \int_{-\infty}^{\infty} c'(\theta)f(y;\theta)\,dy = 0$$

Kako iz definicije očekivanja sledi da je $\int_{-\infty}^{\infty} a(y)f(y;\theta)\,dy = E\big(a(y)\big)$, a na osnovu (3.2) važi da je $\int_{-\infty}^{\infty} c'(\theta)f(y;\theta)\,dy = c'(\theta)$, sledi da je

$$b'(\theta)E\big(a(y)\big) + c'(\theta) = 0.$$

Dakle, važi da je

$$E\big(a(Y)\big) = -\frac{c'(\theta)}{b'(\theta)}. \tag{3.5}$$

Na sličan način dolazimo i do $D\big(a(Y)\big)$.

$$\frac{d^2 f(y;\theta)}{d\theta^2} = \big(a(y)b'(\theta) + c'(\theta)\big)^2 f(y;\theta) + \big(a(y)b''(\theta) + c''(\theta)\big)f(y;\theta). \qquad (3.6)$$

Na osnovu (3.5), prvi sabirak sa desne strane jednakosti (3.6) može biti napisan kao

$$\big(a(y)b'(\theta) + c'^{(\theta)}\big)^2 f(y;\theta) = b'^2(\theta)\Big(a(y) - E\big(a(Y)\big)\Big)^2 f(y;\theta).$$

Tada iz (3.4) sledi

$$\int_{-\infty}^{\infty} \frac{d^2 f(y;\theta)}{d\theta^2} dy = b'^2(\theta)D\big(a(Y)\big) + b''(\theta)E\big(a(Y)\big) + c''(\theta) = 0,$$

jer je po definiciji $\int_{-\infty}^{\infty}\Big(a(y) - E\big(a(Y)\big)\Big)^2 f(y;\theta)dy = D\big(a(Y)\big)$.

Dakle, za disperziju dobijamo da je

$$D\big(a(Y)\big) = \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{b'^3(\theta)}. \qquad (3.7)$$

Dobijene jednakosti za očekivanje i disperziju mogu biti pokazane za sve specijalne slučajeve raspodela iz eksponencijalne familije. Na primer, posmatrajmo kanonički oblik Poasonove raspodele

$$f(y;\mu) = e^{(y\log\mu - \mu - \log y!)},$$

gde imamo da je $a(y) = y$, $b(\theta) = \log\mu$, $c(\theta) = -\mu$ i $d(y) = -\log y!$.

Tada je

$$E\big(a(Y)\big) = -\frac{c'(\theta)}{b'(\theta)} = -\frac{-1}{\frac{1}{\mu}} = \mu,$$

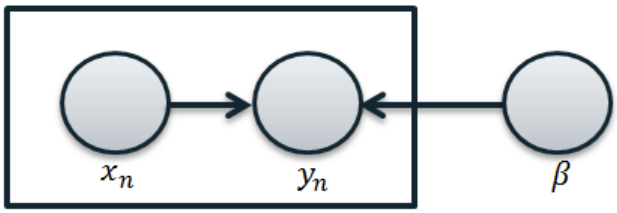
$$D\big(a(Y)\big) = \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{b'^3(\theta)} = \frac{\frac{1}{\mu^2} - 0}{\frac{1}{\mu^3}} = \mu.$$

2. Konstrukcija uopštenih linearnih modela

Uopšteni linearni modeli predstavljaju značajnu generalizaciju linearne regresije u uopšteniju, eksponencijalnu familiju. Na slici 1. možemo videti grafičku reprezentaciju uopštenog linearnog modela, koji je zasnovan na sledećem:

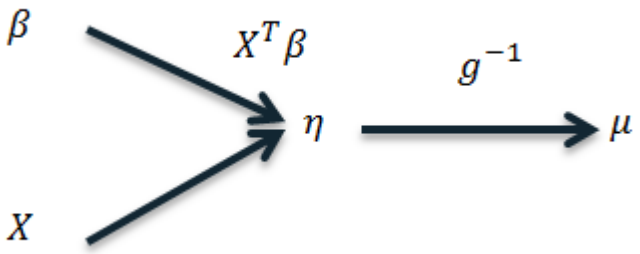
- Registrovane vrednosti se uključuju u model putem linearne funkcije $(X^T\beta)$.
- Uslovno očekivanje zavisne promenljive se predstavlja kao funkcija linearne kombinacije:
$$E(Y|X) = \mu = f(X^T\beta).$$
- Dobijena vrednost se izvodi iz eksponencijalne familije raspodela sa sredinom μ.

Slika 1. Reprezentacija uopštenog linearnog modela



Naredna slika definiše odnose između promenljivih kod uopštenih linearnih modela.

Slika 2. Odnosi između promenljivih kod uopštenih linearnih modela



Dakle, uopšteni linearni modeli se sastoje od tri komponente:

- **Komponenta slučajnosti** definiše uslovnu raspodelu obeležja, Y_i (za i-tu od n nezavisnih vrednosti), za date vrednosti nezavisnih promenljivih u modelu. U originalnoj formulaciji raspodela za Y_i je član eksponencijalne familije raspodela, kao što su normalna, Poasonova, binomna, gama ili inverzna Gausova raspodela.
- **Komponenta sistematičnosti** ili **linearno predviđanje (prediktor)** je linearna funkcija parametara regresije

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

Kao i u linearnom modelu, parametri x_{ij} su prethodno definisane funkcije nezavisnih promenljivih koji ne moraju biti linearno nezavisni, i prema tome, mogu da sadrže kvantitativne nezavisne promenljive, transformacije kvantitativnih nezavisnih promenljivih, polinomne parametre, itd. Zaista, jedna od prednosti uopštenih linearnih modela je to što je struktura linearnog predviđanja poznata.

- Glatka i invertibilna **funkcija veze** $g(\cdot)$ transformiše očekivanje obeležja, $\mu_i \equiv E(Y_i)$, u linearno predviđanje, tj. povezuje komponentu sistematičnosti sa srednjom vrednosti od Y:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} .$$

Kako je funkcija veze invertibilna, možemo takođe da napišemo

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) ,$$

pa se stoga uopšteni linearni modeli mogu posmatrati i kao linearni modeli transformacija očekivanja obeležja ili kao nelinearni regresioni modeli obeležja. Inverzna veza $g^{-1}(\cdot)$ se naziva i funkcija srednje vrednosti. Najčešće korišćene funkcije veze i njihove inverzne vrednosti su date u tabeli 1. Primetimo da veza identiteta naprosto vraća nepromenjen argument, $\eta_i = g(\mu_i) = \mu_i$, a prema tome i $\mu_i = g^{-1}(\eta_i) = \eta_i$ i ona predstavlja najjednostavniju funkciju veze. Druge funkcije veze dozvoljavaju nelinearnost parametra μ u odnosu na predviđanje.

Tabela 1. Najčešće korišćene funkcije veze i njihove inverzne vrednosti

Veza	$\boldsymbol{\eta_i = g(\mu_i)}$	$\boldsymbol{\mu_i = g^{-1}(\eta_i)}$
Identitet	μ_i	η_i
Log	$\ln_e \mu_i$	e^{η_i}
Inverzna	${\mu_i}^{-1}$	${\eta_i}^{-1}$
Inverzno-kvadratna	${\mu_i}^{-2}$	${\eta_i}^{-\frac{1}{2}}$
Kvadratni koren	$\sqrt{\mu_i}$	${\eta_i}^{2}$
Logit	$\ln_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log log	$-\ln_e(-\ln_e \mu_i)$	$e^{-e^{-\eta_i}}$
Komplementarna log log	$\ln_e(-\ln_e(1-\mu_i))$	$1 - e^{-e^{\eta_i}}$

Napomena: μ_i je očekivana vrednost rezultata; η_i je linearno predviđanje; $\Phi(\cdot)$ je kumulativna funkcija raspodele normalne raspodele[3].

Poslednje četiri funkcije veze u tabeli 1. su za binomne podatke, gde Y_i predstavlja udeo uspešnih ishoda od n_i nezavisnih binarnih pokušaja; dakle, Y_i može da primi vrednosti $0, \frac{1}{n_i}, \frac{2}{n_i}, \ldots, \frac{n_i-1}{n_i}, 1$.

Dobar izbor veze će nam otkloniti ograničenja u vezi domena očekivanih rezultata. Na primer, pretpostavimo da je obeležje Y prebrojiva slučajna promenljiva, koja može da primi samo nenegativne celobrojne vrednosti, 0, 1, 2,... Prema tome, i očekivanje μ_i će biti nenegativno (mada ne i obavezno ceo broj), a log veza će preslikati μ_i na celu realnu osu. Međutim, to ne znači da izbor funkcije veze treba da bude u potpunosti određen domenom obeležja.

Pogodna osobina raspodela eksponencijalne familije je to što je uslovna disperzija za Y_i funkcija njene sredine μ_i, recimo $v(\mu_i)$, i parametra disperzije σ_i. U tabeli 2. prikazane su disperzije, kao funkcije od μ_i i σ_i, za najčešće korišćene eksponencijalne familije. Takođe, prikazani su i domeni obeležja i takozvane kanoničke (ili prirodne) funkcije veze u odnosu na svaku familiju. Uopšteni linerani modeli imaju prednost u odnosu na transformacije obeležja kod linearne regresije. To je zbog toga što je izbor transformacije delimično razdvojen od raspodele obeležja. Kanonička veza pojednostavljuje uopšteni linearni model, mada se mogu koristiti i neke druge funkcije veze. Prednost kanoničkih veza je to što minimalna dovoljna statistika[4] za β postoji, tj. sve informacije o β sadržane su u funkciji istih dimenzija kao i β. Konkretno, veze koje se koriste variraju od jedne familije do druge, ali i od jednog do drugog softvera. Tako, na primer, ne bi bilo previše korisno koristiti identitet, log, inverznu, inverzno-kvadratnu ili kvadratni koren vezu za binomne podatke, niti bi imalo smisla uzimati logit, probit, log log ili komplementarnu log log vezu za nebinomne podatke.

Tabela 2. Kanoničke veze, domen rezultata i uslovne funkcije disperzija za raspodele iz eksponencijalne familije

[3] Kumulativna funkcija raspodele normalne raspodele, koja se obično označava grčkim velikim slovom Φ, je integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt.$$

[4] Dovoljna statistika je minimalna dovoljna ako se može predstaviti kao funkcija bilo koje druge dovoljne statistike. Drugim rečima, $S(X)$ je minimalna dovoljna, ako i samo ako

1. $S(X)$ je dovoljna,
2. Ako je $T(X)$ dovoljna, onda postoji funkcija f tako da je $S(X) = f(T(X))$.

Intuitivno, minimalna dovoljna statistika najefikasnije hvata sve moguće informacije o parametru θ.

Familija	Kanonička veza	Domen od Y_i	$D(Y_i\|\eta_i)$
Gausova	Identitet	$(-\infty, +\infty)$	σ_i
Binomna	Logit	$\frac{0, 1, \dots, n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poasonova	Log	0, 1, 2, ...	μ_i
Gama	Inverzna	$(0, \infty)$	$\sigma_i {\mu_i}^2$
Inverzna Gausova	Inverzno-kvadratna	$(0, \infty)$	$\sigma_i {\mu_i}^3$

Napomena: σ_i je parametar disperzije, η_i je linearno predviđanje, a μ_i je očekivanje od obeležja Y_i. Za binomnu familiju, n_i je broj ponavljanja.

3. Tipovi uopštenih linearnih modela

Uopštene linearne modele delimo na standardne i ekstenzije.

Standardni modeli – Uz pomoć softvera za uopštene linearne modele mogu se fitovati standardne raspodele, kao što su Poasonova, binomna, normalna, log-normalna, gama, log-gama, eksponencijalna, Pareto, inverzna Gausova i niz funkcija veze:

- Identitet μ
- Recipročna $\frac{1}{\mu}$
- Kvadratno inverzna $\frac{1}{\mu^2}$
- Kvadratni koren $\sqrt{\mu}$
- Eksponencijalna $(\mu + c_1)^{c_2}$, c_1 i c_2 su poznate
- Log $\log \mu$
- Logit $\log \frac{\mu}{n-\mu}$
- Komplementarna log log $\log\left(-\log \frac{\mu}{n}\right)$
- Probit $\Phi^{-1}\left(\frac{\mu}{n}\right)$

Ekstenzije – Brojne ideje mogu da se koriste za softvere da bi se fitovao model koji nije iz uopštene linearne familije, kao na primer, model čija je raspodela blizu eksponencijalnoj familiji, koji ima parametre unutar funkcije veze, parametre unutar funkcije disperzije, nelinearnu strukturu, itd.

Dalje, prilikom izbora modela, čitav niz regresionih modela se uzima u razmatranje. Sada ćemo uvesti terminologiju, pomoću koje ćemo opisivati zajedničke mogućnosti koje se mogu posmatrati.

- *Kompletan, potpuni ili zasićen model:*
  Model ima onoliko parametara, koliko i registrovanih vrednosti, odnosno, n linearno nezavisnih parametara. Dakle, on reprodukuje podatke tačno, ali bez pojednostavljivanja, i prema tome nije previše pogodan za interpretaciju.
- *Nula-model:*
  Ovaj model ima jedinstvenu srednju vrednost za sve registrovane vrednosti. On je jednostavan, ali obično nema dovoljno reprezentativnu strukturu u odnosu na podatke.
- *Maksimalni model:*
  Predstavlja najveći, najkompleksniji model koji smo spremni da razmotrimo.
- *Minimalni model:*
  Ovaj model sadrži minimalan skup parametara koji moraju biti prisutni.
- *Trenutni model:*
  Ovaj model se nalazi između maksimalnog i minimalnog modela i trenutno je predmet istraživanja.

Zasićeni model opisuje registrovane vrednosti tačno, ali baš zbog toga ima vrlo male šanse da bude pogodan za ponavljanje istraživanja uz korišćenje istih metoda, ali drugih registrovanih vrednosti. On ne naglašava važne osobine podataka. Nasuprot tome, minimalni model ima dobre šanse da odgovara i podacima iz ponovljenih istraživanja. Međutim, bitne karakteristike podataka su kod minimalnog modela obično ispuštene. Dakle, mora se pronaći balans između uspešnosti fitovanja podataka i jednostavnosti.

IV. Poasonova regresija za prebrojive podatke

Poasonova regresija je oblik uopštenih linearnih modela, gde slučajnu promenljivu modeliramo pretpostavljajući da ima Poasonovu raspodelu. Poasonova raspodela podrazumeva slučajne promenljive sa nenegativnim celobrojnim vrednostima, kao što su, na primer, prebrojivi podaci. Takvi podaci se mogu prikazati kao frekvencije, pomoću tabela kontigencije. Takođe, mogu se prikazivati i kao broj ostvarenih događaja, na primer broj saobraćajnih nesreća, koji se analiziraju u odnosu na neke nezavisne promenljive, što u ovom slučaju može biti broj registrovanih motornih vozila ili rastojanje koje prelaze vozači. Dakle, zavisna promenljiva predstavlja broj događaja u određenom vremenskom intervalu.

Kao što smo već napomenuli, kod linearnih modela procene srednjih vrednosti mogu da budu negativne, međutim kada posmatramo prebrojive podatke, sredine moraju biti nenegativne. Prebrojivi podaci mogu uzimati samo (nenegativne) celobrojne vrednosti, što ih čini nekonzistentnim sa Gausovim greškama. Dalje, prebrojivi podaci često ispoljavaju heteroskedastičnost, gde veća disperzija prati veću srednju vrednost. Najjednostavniji uopšteni linearni model za podatke dobijene prebrojavanjem podrazumeva Poasonovu raspodelu komponente slučajnosti. Kao i podaci dobijeni prebrojavanjem, Poasonove slučajne promenljive uzimaju nenegativne celobrojne vrednosti.

1. Poasonova slučajna promenljiva – osnovne osobine i primeri

Poasonova raspodela je diskretna raspodela koja predstavlja verovatnoću da se određeni broj događaja ostvari u zadatom vremenskom intervalu, ako se događaji ostvaruju nezavisno od vremena realizovanja poslednjeg događaja.

Slučajna promenljiva Y ima Poasonovu[5] raspodelu sa parametrom μ, ako za $\mu > 0$ uzima celobrojne vrednosti $y = 0, 1, 2, \ldots$ sa verovatnoćom

[5] Poasonova raspodela je nazvana po francuskom matematičaru *Simonu Denisu Poasonu* (1781.–1840.), koji je prvi uveo ovu raspodelu i objavio je zajedno sa njegovom teorijom verovatnoće 1837. godine u delu pod nazivom "*Istraživanje o verovatnoći presuda u krivičnim i građanskim pitanjima*".

$$P\{Y = y\} = \frac{e^{-\mu}\mu^{y}}{y!}.$$

Očekivanje slučajne promenljive $Y \sim Pois(\mu)$ je

$$E(Y) = \sum_{i=0}^{\infty} i \frac{\mu^{i}}{i!} e^{-\mu} = \mu e^{-\mu} \sum_{i=1}^{\infty} \frac{\mu^{i-1}}{(i-1)!}$$
$$= \mu e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^{i}}{i!} = \mu e^{-\mu} e^{\mu} = \mu.$$

Disperzija slučajne promenljive $Y \sim Pois(\mu)$ je

$$D(Y) = E(Y^2) - E^2(Y) = \sum_{i=0}^{\infty} i^2 \frac{\mu^{i}}{i!} e^{-\mu} - \mu^2$$
$$= \sum_{i=1}^{\infty} (i - 1 + 1) \frac{\mu^{i}}{(i-1)!} e^{-\mu} - \mu^2$$
$$= e^{-\mu} \left(\mu^2 \sum_{i=2}^{\infty} \frac{\mu^{i-2}}{(i-2)!} + \sum_{i=1}^{\infty} \frac{\mu^{i-1}}{(i-1)!} \right) - \mu^2$$
$$= e^{-\mu} \left(\mu^2 \sum_{i=0}^{\infty} \frac{\mu^{i}}{i!} + \mu \sum_{i=0}^{\infty} \frac{\mu^{i}}{i!} \right) - \mu^2$$
$$= e^{-\mu}(\mu^2 e^{\mu} + \mu e^{\mu}) - \mu^2 = \mu.$$

Poasonova raspodela, koju predstavljamo kao $Y \sim Pois(\mu)$, potpuno je određena srednjom vrednosti μ, pošto je njena disperzija takođe jednaka μ. Iz tog razloga, kada su vrednosti u proseku veće, one više i variraju. Kako su očekivanje i disperzija jednaki, faktor koji utiče na jedno, uticaće i na drugo. Dakle, ne možemo pretpostaviti da važi homoskedastičnost za Poasonove podatke.

Primer 1. Pretpostavimo da se na određenoj lokaciji nalazi biljka čiji broj jedinki po m^2 ima raspodelu prema Poasonovom procesu sa srednjom vrednosti 0.2 jedinke po m^2. Hoćemo da odredimo verovatnoću da se na 9 m^2 ne nalazi ni jedna jedinka ove vrste.

Kako broj jedinki ima Poasonovu raspodelu sa sredinom $\mu = 9 \cdot 0.2 = 1.8$, verovatnoća da na 9 m^2 ne živi ova biljka je

$$P\{Y = 0 | \mu = 1.8\} = \frac{e^{-1.8} 1.8^{0}}{0!} = 0.16523.$$

■

Poasonova slučajna promenljiva je zatvorena u odnosu na sabiranje, što znači da je suma nezavisnih Poasonovih slučajnih promenljivih Poasonova slučajna promenljiva sa srednjom vrednosti koja je jednaka sumi odgovarajućih srednjih vrednosti. Specijalno, ako su Y_1 i Y_2 nezavisne, gde $Y_i \sim Pois(\mu_i)$, za $i = 1, 2$, tada

$$Y_1 + Y_2 \sim Pois(\mu_1 + \mu_2).$$

Iz toga sledi da je Poasonova slučajna promenljiva sa sredinom μ jednaka zbiru μ nezavisnih Poasonovih slučajnih promenljivih sa sredinom 1, pa iz Centralne granične teoreme (čiju formulaciju i dokaz dajemo u dodatku) sledi da kako μ raste, Poasonova slučajna promenljiva postaje približno normalna. Sada ćemo dati formalan dokaz osobine zatvorenosti u odosu na sabiranje.

Teorema 1.: Ako su $Y_i \sim Pois(\mu_i)$, $i = 1,2, \dots$ nezavisne slučajne promenljive, gde je $\sum \mu_i < \infty$, tada je

$$S_Y = \sum Y_i \sim Pois\left(\sum \mu_i\right).$$

Dokaz: Daćemo primer koji je specijalni slučaj teoreme za $i = 2$. Generalizacija dokaza se dobija indukcijom.

Neka slučajna promenljiva Y_1 ima Poasonovu $Pois(\mu_1)$ raspodelu, slučajna promenljiva Y_2 ima Poasonovu $Pois(\mu_2)$ raspodelu i neka su Y_1 i Y_2 nezavisne. Odredimo raspodelu zbira $Y_1 + Y_2$.

Najpre, primetimo da slučajne promenljive Y_1, Y_2 i $Y_1 + Y_2$ imaju isti skup mogućih vrednosti. Za proizvoljno $n \in \{0, 1, 2, \dots\}$, imamo

$$P\{Y_1 + Y_2 = n\} = \sum_{k=0}^{n} P(\{Y_1 = k\} \cap \{Y_2 = n - k\}).$$

Kako su slučajne promenljive Y_1 i Y_2 nezavisne, imamo da je

$$\begin{aligned} P\{Y_1 + Y_2 = n\} &= \sum_{k=0}^{n} P\{Y_1 = k\} \cdot P\{Y_2 = n - k\} \\ &= \sum_{k=0}^{n} \frac{{\mu_1}^k}{k!} e^{-\mu_1} \frac{{\mu_2}^{n-k}}{(n-k)!} e^{-\mu_2} \\ &= e^{-(\mu_1+\mu_2)} \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!} {\mu_1}^k {\mu_2}^{n-k} \end{aligned}$$

$$= e^{-(\mu_1+\mu_2)} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k} {\mu_1}^k {\mu_2}^{n-k}$$
$$= \frac{(\mu_1+\mu_2)^n}{n!} e^{-(\mu_1+\mu_2)}, \qquad n = 0, 1, \dots$$

Dakle, $Y_1 + Y_2 : Pois(\mu_1 + \mu_2)$.

■

Korisna posledica ove osobine u praktičnom radu je to što možemo da analiziramo individualne ili grupne podatke, a da dobijemo isti rezultat. Specijalno, neka Y_{ij} označava broj događaja koji su se dogodili u j-toj jedinici i-te grupe i neka Y_i označava ukupan broj događaja u grupi i. Tada, pod uobičajenim pretpostavkama o nezavisnosti, ako $Y_{ij} \sim Pois(\mu_i)$, za $j = 1, 2, \dots, n_i$, tada $Y_i \sim Pois(n_i \mu_i)$. To znači da ako su individualne prebrojive slučajne promenljive Y_{ij} Poasonove sa sredinom μ_i, tada je i ukupna slučajna promenljiva Y_i Poasonova sa sredinom $n_i \mu_i$. Dakle, dobijamo istu funkciju verodostojnosti ako radimo sa pojedinačnim prebrojivim podacima Y_{ij} ili sa ukupnim Y_i.

Poasonova raspodela je povezana sa druge dve diskretne raspodele, binomnom i multinomijalnom. Prvo ćemo dati vezu između binomne i Poasonove raspodele. Ako je broj uspešnih ishoda u n pokušaja binomne raspodele, gde broj pokušaja $n \to \infty$, a verovatnoća uspešnog ishoda $p \to 0$, tako da $np \to \mu$, raspodela uspešnih ishoda je približno Poasonova sa sredinom μ. Odavde sledi da je Poasonova raspodela dobar izbor za modeliranje retkih događaja, tj. događaja koji se najverovatnije neće desiti u bilo kojoj pojedinačnoj situaciji (kako je p malo), ali mogu da se dogode prilikom mnogo nezavisnih pokušaja (odnosno, n je veliko). U praksi, binomnu raspodelu $\mathcal{B}(n, p)$ aproksimiramo Poasonovom ako je n veliko i $np < 10$. Tada uzimamo $\mu = np$ i prelazimo na Poasonovu raspodelu $Pois(\mu)$.

Teorema 2.: Neka je S_n slučajna promenljiva koja predstavlja broj realizacija događaja, tj. Bernulijeva slučajna promenljiva, $S_n : \mathcal{B}(n, p)$. Ako je u Bernulijevoj šemi $np \to \mu > 0$, kada $n \to \infty$, onda

$$P\{S_n = j\} \to \frac{\mu^j}{j!} e^{-\mu}, \qquad j = 0,1, \dots, \qquad n \to \infty.$$

Dokaz: Na osnovu pretpostavki teoreme imamo da je

$$p \sim \frac{\mu}{n}, \qquad \text{kada } n \to \infty.$$

Sada je

$$P\{S_n = j\} = \binom{n}{j} p^j q^{n-j}$$
$$= \frac{n(n-1)\dots(n-j+1)}{j!} p^j (1-p)^{n-j}$$
$$\sim \frac{n(n-1)\dots(n-j-1)}{j!} \left(\frac{\mu}{n}\right)^j \left(1-\frac{\mu}{n}\right)^{n-j}$$
$$= \frac{\mu^j}{j!}\left(1-\frac{\mu}{n}\right)^n \left(1-\frac{\mu}{n}\right)^{-j} \frac{n}{n}\frac{n-1}{n}\dots\frac{n-j+1}{n}, \qquad n \to \infty.$$

Kako je

$$\left(1-\frac{\mu}{n}\right)^n \to e^{-\mu}, \qquad n \to \infty,$$

$$\left(1-\frac{\mu}{n}\right)^{-j} \frac{n}{n}\frac{n-1}{n}\dots\frac{n-j+1}{n} \to 1, \qquad n \to \infty,$$

sledi

$$P\{S_n = j\} \to \frac{\mu^j}{j!} e^{-\mu}, \qquad j = 0,1,\dots, \qquad n \to \infty.$$

■

Poasonova raspodela je usko povezana i sa multinomnom raspodelom, koja predstavlja uopštenje binomne raspodele. Za n nezavisnih pokušaja, gde svaki od njih vodi do realizovanja (uspešnog pokušaja) tačno jedne od k kategorija, pri čemu svaka kategorija ima unapred datu verovatnoću uspeha, multinomna raspodela daje verovatnoću uspešnosti proizvoljne kombinacije brojeva različitih kategorija. Parametri koji određuju multinomnu raspodelu su, dakle, broj događaja $n > 0$ i $p_1, \dots, p_k$ koje predstavljaju verovatnoće realizacije svake kategorije (naravno, $\sum_{i=1}^{k} p_i = 1$). Srednja vrednost je data sa $E(Y_i) = np_i$, dok je disperzija $D(Y_i) = np_i(1-p_i)$. Neka su dalje, sa $Y_1, \dots, Y_k$ označeni mogući ishodi svakog pokušaja i pretpostavimo da je verovatnoća realizacije Y_i u svakom pokušaju jednaka p_i, $i = 1, \dots k$. Verovatnoća da se u n pokušaja Y_1 realizovalo tačno y_1 puta, Y_2 realizovalo tačno y_2 puta, itd. data je sledećom funkcijom

$$f(y_1, \dots, y_k; n, p_1, \dots, p_k) =$$
$$= P(\{Y_1 = y_1\} \cup \dots \cup \{Y_k = y_k\}) =$$
$$= \begin{cases} \dfrac{n!}{y_1! \dots y_k!} {p_1}^{y_1} \dots {p_k}^{y_k}, & \text{kada } \sum_{i=1}^{k} y_i = n \\ 0, & \text{inače} \end{cases}$$

za nenegativne celobrojne vrednosti $y_1, \dots, y_k$. Za $k = 2$ dobijamo binomnu raspodelu, koja je dakle, specijalan slučaj multinomne.

Multinomna raspodela se najčešće koristi za uzorkovanje sa vraćanjem, kada imamo više od dve kategorije. Na primer, neka je populacija od n elemenata podeljena u kategorije $E_1, \dots, E_r$ veličine $n_1, \dots, n_r$. Multinomna raspodela daje verovatnoće za nekoliko mogućih kombinacija slučajnog uzorka sa vraćanjem veličine n, koji je uzet iz ovako date populacije.

Kao drugi primer, posmatrajmo bacanje dvanaest kockica. Kolika je verovatnoća da se svaki broj dobije dva puta? Označimo sa $E_1, \dots, E_6$ šest mogućih brojeva, gde za svaki od njih postoji dva moguća ishoda, a verovatnoća svakog ishoda je $\frac{1}{6}$. Dakle, odgovor je $12!\, 2^{-6} 6^{-12} = 0.0034$.

Veza između Poasonove i multinomne raspodele je data na sledeći način. Ako posmatramo K nezavisnih Poasonovih slučajnih promenljivih $\{Y_1, \dots, Y_K\}$ sa sredinama μ_i, njihova zajednička raspodela, koja zavisi od ukupnog broja prebrojivih podataka $\sum_j Y_j$, je multinomna sa verovatnoćom $\pi_i = \frac{\mu_i}{\sum_j Y_j}$. Ova veza se pokazala veoma bitnom u analizi tabela kontigencije.

Primer 2. U klasičnom tekstu o teoriji verovatnoće Feler (1957.)[6] je uključio brojne primere registrovanih vrednosti koje imaju Poasonovu raspodelu, kao što su na primer podaci o broju avionskih bombi koje su pale na južni deo Londona tokom II svetskog rata. Grad je bio podeljen na 576 malih oblasti, svaka veličine četvrtine kvadratnog kilometra, a zatim su prebrojavane oblasti koje su pogođene tačno k puta. Ukupno je bilo 537 pogodaka, pa je prosečan broj pogodaka po oblasti 0.9323. Kako normalna raspodela nije pogodna za prebrojive podatke, Poasonova raspodela predstavlja standardni izbor. Registrovane vrednosti u tabeli 3. su veoma blizu Poasonove raspodele sa sredinom $\mu = 0.9323$. Dalje, u ovom primeru svaki dan možemo posmatrati kao veliki broj pokušaja, gde svaka od oblasti ima malu verovatnoću da bude pogođena. Ako pretpostavimo da su dani međusobno nezavisni, onda nas to dovodi do binomne raspodele koja je veoma dobro aproksimirana Poasonovom. Drugi primeri događaja koji odgovaraju ovoj raspodeli su radioaktivna dezintegracija, razmena hromozoma unutar ćelija, broj telefonskih poziva pogrešnog broja, broj bakterija u različitim delovima Petrijeve šolje.

[6] Feller, William (1957) *'An Introduction to Probability Theory and Its Applications'*, second edition, John Wiley & Sons, Inc.

Tabela 3. Broj avionskih bombi koje su pale na južni London tokom II svetskog rata

Pogoci	**0**	**1**	**2**	**3**	**4**	**5+**
Registrovani	229	211	93	35	7	1
Očekivani	226.7	211.4	98.6	30.6	7.1	1.6

■

Sada ćemo pogledati neformalno alternativno izvođenje Poasonove raspodele u smislu stohastičkih procesa. Pretpostavimo da se događaji ostvaruju slučajno u vremenu tako da su ispunjeni sledeći uslovi:

- Verovatnoća da se događaj ostvari barem jednom u datom vremenskom periodu proporcionalna je dužini tog vremenskom intervala.
- Verovatnoća da se događaj ostvari dva ili više puta u malo vremenskom periodu je zanemarljiva.
- Broj događaja koji se desio u jednom vremenskom intervalu nezavisan je od broja događaja koji se desio u drugom vremenskom intervalu, ukoliko su intervali disjunktni.

Tada je raspodela verovatnoće broja ostvarenih događaja u određenom vremenskom intervalu Poasonova sa sredinom $\mu = \lambda t$, gde je λ stopa ostvarivanja događaja po jedinici vremena, a t je dužina vremenskog intervala. Proces koji zadovoljava tri gornja uslova se naziva Poasonov proces. Poasonova raspodela je često asimetrična na desnu stranu, pa sledi da je dobro da se koristi za retke događaje.

U primeru avionskih bombi ovi uslovi mogu biti ispunjeni. Što duže traje rat, to je veća verovatnoća da će određena oblast biti pogođena makar jednom. Takođe, verovatnoća da će jedna oblast biti pogođena dva puta u toku istog dana je, na sreću, veoma mala. I na kraju, to što je oblast pogođena u bilo kojem danu je nezavisno od onoga što se događa u susednim oblastima.

2. Model Poasonove regresije

Statističko modeliranje se odvija u četiri koraka:

- *Postavljanje modela* – model se određuje iz dva dela: jednačinom koja povezuje obeležje i nezavisne promenljive i raspodelom verovatnoće obeležja
- *Ocenjivanje parametara modela*
- *Provera adekvatnosti modela* – koliko model dobro fituje podatke

- *Zaključak* – računanje intervala poverenja i testiranje hipoteza o parametrima modela, kao i interpretacija rezultata

3. Postavljanje modela

Pretpostavimo da imamo uzorak obima n, dat sa $y_1, y_2, \dots, y_n$, koji može da se posmatra kao realizacija nezavisnih Poasonovih slučajnih promenljivih, gde je $Y_i \sim Pois(\mu_i)$ i pretpostavimo da hoćemo da pustimo da srednja vrednost μ_i (a samim tim i disperzija) zavise od vektora nezavisnih promenljivih $\boldsymbol{x}_i$. Efekat nezavisnih promenljivih na slučajne promenljive Y_i se modelira kroz parametre μ_i.

Mogli bismo da postavimo jednostavan linearni model oblika

$$\mu_i = E(y_i|x_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

ali ovaj model dopušta da linearno predviđanje sa desne strane jednakosti ima bilo koju realnu vrednost, dok Poasonova srednja vrednost sa leve strane, koja predstavlja očekivanje prebrojive slučajne promenljive, mora da bude nenegativna.

Jednostavno rešenje ovog problema jeste da umesto toga modeliramo logaritam srednje vrednosti koristeći linearni model. Dakle, možemo računati logaritam $\eta_i = \log \mu_i$ i pretpostaviti da se transformisana srednja vrednost ponaša po linearnom modelu $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. To znači da ćemo koristiti uopšteni linearni model sa log vezom. Na osnovu toga možemo zapisati model u sledećem obliku

$$\log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}. \tag{4.1}$$

Iz jednačine(4.1) jednostavno dobijamo model za srednju vrednost

$$\mu_i = E(y_i|\boldsymbol{x}_i) = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}.$$

Dalje,

$$\frac{\partial \mu_i}{\partial \boldsymbol{x}_i} = \frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial \boldsymbol{x}_i} = \frac{\partial E\left(e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}\right)}{\partial \boldsymbol{x}_i} = e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} \beta_i = \mu_i \beta_i.$$

Vidimo da u ovom modelu parametar regresije β_j predstavlja očekivanu promenu logaritma srednje vrednosti po jedinici promene za $\boldsymbol{x}_j$. Povećavanje $\boldsymbol{x}_j$ za jednu jedinicu množi srednju vrednost od Y_j faktorom e^{β_j}, tj.

- Ako je $\beta_j = 0$, tada je $e^{\beta_j} = 1$, pa Y_j i X nisu povezani
- Ako je $\beta_j < 0$, tada je $e^{\beta_j} < 1$ i $\mu = E(Y)$ je e^{β_j} puta manje nego kada je $X = 0$

- Ako je $\beta_j > 0$, tada je $e^{\beta_j} > 1$ i $\mu = E(Y)$ je e^{β_j} puta veće nego kada je $X = 0$.

Glavna pretpostavka Poasonovog modela je da su sredina i disperzija jednake, tj.

$$E(y_i|\boldsymbol{x}_i) = \mu_i = e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} = D(y_i|\boldsymbol{x}_i).$$

Ukoliko imamo slučaj da je $E(y_i|\boldsymbol{x}_i) < D(y_i|\boldsymbol{x}_i)$, podaci su preraspršeni i Poasonov model mora biti modifikovan da bismo dobili dobro slaganje modela sa podacima.

Nezavisne promenljive $X = (X_1, X_2, \dots, X_k)$ u Poasonovim regresionim modelima mogu biti:

1. sve kategoričke; tada za modeliranje prebrojivih podataka koristimo tabele kontigencije i ovi modeli se konvencijom zovu log-linearni modeli;
2. numeričke ili kombinacija numeričkih i kategoričkih promenljivih; ove modele nazivamo Poasonovim regresijama;
3. Ukoliko je Y/t promenljiva koju modeliramo, čak iako su sve nezavisne promenljive kategoričke, regresioni model ćemo nazivati Poasonov, a ne log-linearni.

4. Ocene parametara modela

Metoda maksimalne verodostojnosti i algoritam iterativnih težinskih najmanjih kvadrata

Posmatrajmo uopšteni linearni model koji sadrži nezavisne slučajne promenljive $Y_1, \dots, Y_n$ i neka su $y_1, \dots, y_n$ njihove realizovane vrednosti. Za početak ćemo definisati potrebne funkcije, koje koristimo za *metodu maksimalne verodostojnosti*.

Funkcija maksimalne verodostojnosti za Y_i predstavlja verovatnoću da dati uzorak bude izabran, dakle,

$$l_i(\theta_i, y_i) = P\{Y_i = y_i\} = e^{\left(y_i b(\theta_i) + c(\theta_i) + d(y_i)\right)},$$

gde $\theta_i, i = 1, \dots, n$ predstavlja parametar raspodele.

Kako funkcije $l_i(\theta_i)$ i $\ln l_i(\theta_i)$ postižu maksimum za istu vrednost θ_i, često je lakše naći maksimum prirodnog logaritma funkcije verodostojnosti. Tada je

$$\ln l_i(\theta_i, y_i) = y_i b(\theta_i) + c(\theta_i) + d(y_i).$$

Dalje, izvod funkcije $\ln l_i(\theta_i, y_i)$ po θ_i je

$$U_i(\theta_i, y_i) = \frac{\partial \ln l_i(\theta_i, y_i)}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i).$$

Funkcija U_i se naziva *skor statistika* i ona predstavlja ocenu nepoznatog parametra β_i. Kako U_i zavisi od y_i, možemo je posmatrati kao slučajnu promenljivu

$$U_i = Y_i b'(\theta_i) + c'(\theta_i). \tag{4.2}$$

Očekivana vrednost za U_i je

$$E(U_i) = b'(\theta_i)E(Y_i) + c'(\theta_i),$$

pa iz (3.5) dobijamo da je

$$E(U_i) = b'(\theta_i)\left(-\frac{c'(\theta_i)}{b'(\theta_i)}\right) + c'(\theta_i) = 0.$$

Disperzija od U_i se naziva *matrica informacija* i nju ćemo označavati sa $\mathcal{J}_i$. Na osnovu osobina disperzije o linearnim transformacijama slučajne promenljive i (4.2), dobijamo

$$\mathcal{J}_i = D(U_i) = b'^2(\theta_i)D(Y_i).$$

Dalje, iz (3.7) sledi

$$D(U_i) = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta).$$

Skor statistika U ima primenu kod statističkog zaključivanja o parametrima uopštenih linearnih modela, kao što ćemo videti u poglavlju IV. 5.

Za statistiku U važi da je

$$D(U) = E(U^2) = -E(U').$$

Prva jednakost sledi iz osobine disperzije koja važi za sve slučajne promenljive, da je

$$D(X) = E(X^2) - E^2(X),$$

pa kako je $E(U) = 0$, dobijamo da je $D(U) = E(U^2)$. Da bismo izveli drugu jednakost, prvo ćemo da diferenciramo U po θ. Dakle, iz (4.2) dobijamo da je

$$U' = \frac{\partial U}{\partial \theta} = a(Y)b''(\theta) + c''(\theta).$$

Tada je očekivana vrednost od U data na sledeći način

$$E(U') = b''(\theta)E\big(a(Y)\big) + c''(\theta) =$$
$$= b''(\theta)\left(-\frac{c'(\theta)}{b'(\theta)}\right) + c''(\theta) = -D(U).$$

Dakle, pokazali smo da važi i druga jednakost.

Nakon što smo uveli potrebne definicije, metodom maksimalne verodostojnosti ćemo izvesti ocene parametara $\boldsymbol{\beta}$, koje su povezane sa Y_i, $i = 1, \dots, n$, kroz $E(Y_i) = \mu_i$ i $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Iako se u nekim specijalnim slučajevima ocene mogu dobiti konkretnim matematičkim izrazima, uglavnom u te svrhe koristimo numeričke metode. Ove metode su naravno iterativne i bazirane su na Njutnovom algoritmu.

Za svako Y_i, $i = 1, \dots, n$ važi

$$E(Y_i) = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)} \tag{4.3}$$

$$D(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - b'(\theta_i)c''(\theta_i)}{b'^3(\theta_i)} \tag{4.4}$$

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i. \tag{4.5}$$

Funkcija maksimalne verodostojnosti za sve Y_i je

$$L = \sum_{i=1}^{n} \ln l_i = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i).$$

Da bismo dobili ocenu parametra β_j, potrebno je da izračunamo

$$\frac{\partial L}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \frac{\partial \ln l_i}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{\partial \ln l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right), \tag{4.6}$$

koristeći pravilo lanca za date diferencijale. Razmotrićemo svaki činilac iz (4.6) pojedinačno. Prvo,

$$\frac{\partial \ln l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i).$$

Drugo,

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}.$$

Iz (4.3) i (4.4) dobijamo da je

$$\frac{\partial \mu_i}{\partial \theta_i} = -\frac{c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{b'^2(\theta_i)} = b'(\theta_i)D(Y_i).$$

I na kraju, iz (4.5) sledi da je

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}x_{ij}.$$

Dakle, statistika U_j je

$$U_j = \sum_{i=1}^{n}\left(\frac{\partial \ln l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right) = \sum_{i=1}^{n}\left(\frac{(y_i - \mu_i)}{D(Y_i)}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right). \qquad (4.7)$$

Matrica varijanse i kovarijanse za U_j ima oblik

$$\mathcal{J}_{jk} = E\big(U_j U_k\big),$$

koji predstavlja elemente *matrice informacija* $\mathcal{J}$. Iz (4.7) sledi

$$\mathcal{J}_{jk} = E\left(\sum_{i=1}^{n}\left(\frac{(Y_i - \mu_i)}{D(Y_i)}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right)\sum_{l=1}^{n}\left(\frac{(Y_l - \mu_l)}{D(Y_l)}x_{lk}\left(\frac{\partial \mu_l}{\partial \eta_l}\right)\right)\right)$$
$$= \sum_{i=1}^{n}\frac{E((Y_i - \mu_i)^2)x_{ij}x_{ik}}{\big(D(Y_i)\big)^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \qquad (4.8)$$

jer je $E\big((Y_i - \mu_i)(Y_l - \mu_l)\big) = 0$, za sve $i \neq l$, kako su svi Y_i međusobno nezavisni. Ako iskoristimo da je $E((Y_i - \mu_i)^2) = D(Y_i)$, (4.8) može da se napiše kao

$$\mathcal{J}_{jk} = \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{D(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \qquad (4.9)$$

Tada je

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + \big(\mathcal{J}^{(m-1)}\big)^{-1}U^{(m-1)} \qquad (4.10)$$

gde je $\boldsymbol{b}^{(m)}$ vektor ocena parametara $\beta_1, \dots, \beta_p$ u m-toj iteraciji. U jednačini (4.10), $\big(\mathcal{J}^{(m-1)}\big)^{-1}$ je inverzna matrica matrice informacija sa elementima $\mathcal{J}_{jk}$ datim sa (4.9), a

$U^{(m-1)}$ je vektor sa elementima datim u (4.7), pri čemu su sve ocene dobijene u $\boldsymbol{b}^{(m-1)}$. Ako sada pomnožimo obe strane jednakosti (4.10) sa $\mathcal{J}^{(m-1)}$, dobijamo

$$\mathcal{J}^{(m-1)}\boldsymbol{b}^{(m)} = \mathcal{J}^{(m-1)}\boldsymbol{b}^{(m-1)} + U^{(m-1)}. \tag{4.11}$$

Iz (4.9) $\mathcal{J}$ možemo zapisati kao

$$\mathcal{J} = X^T W X,$$

gde je W dijagonalna matrica dimenzija $n \times n$, sa elementima

$$\omega_{ii} = \frac{1}{D(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{4.12}$$

Izraz sa desne strane jednakosti (4.11) je vektor sa elementima

$$\sum_{k=1}^{p}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{D(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{D(Y_i)}x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

ocenjenim u $\boldsymbol{b}^{(m-1)}$. Ovo sledi iz jednakosti (4.9) i (4.7). Dakle, desna strana jednakosti (4.11) može biti napisana kao

$$X^T W z,$$

gde $\boldsymbol{z}$ ima elemente

$$z_i = \sum_{k=1}^{p} x_{ik}b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \mu_i}{\partial \eta_i}\right), \tag{4.13}$$

pri čemu su μ_i i $\frac{\partial \mu_i}{\partial \eta_i}$ dobijeni za $\boldsymbol{b}^{(m-1)}$.

Prema tome, iterativna jednačina (4.11) može biti zapisana kao

$$X^T W X b^{(m)} = X^T W z. \tag{4.14}$$

Ovaj oblik je analogan normalnim jednačinama za linearne modele dobijene težinskim najmanjim kvadratima, pri čemu je razlika u tome što se kod uopštenih linearnih modela ocene računaju iterativno, jer u opštem slučaju $\boldsymbol{z}$ i W zavise od $\boldsymbol{b}$. Dakle, za uopštene linearne modele ocene dobijene metodom maksimalne verodostojnosti podrazumevaju algoritam *iterativnih težinskih najmanjih kvadrata*.

Većina statističkih softvera, koja sadrži pakete sa procedurama za fitovanje uopštenih linearnih modela, bazirana je na efikasnom algoritmu (4.14). Algoritam je napravljen

tako da uzima neku početnu aproksimaciju $\boldsymbol{b}^{(0)}$ za ocenjivanje $\boldsymbol{z}$ i W, a zatim se rešava (4.14) da bismo dobili $\boldsymbol{b}^{(1)}$, koje se dalje koristi za dobijanje bolje aproksimacije za $\boldsymbol{z}$ i W, i to se nastavlja dok ne dostignemo željenu konvergenciju. Kada je razlika između $\boldsymbol{b}^{(m-1)}$ i $\boldsymbol{b}^{(m)}$ dovoljno mala, $\boldsymbol{b}^{(m)}$ se uzima kao ocena dobijena metodom maksimalne verodostojnosti.

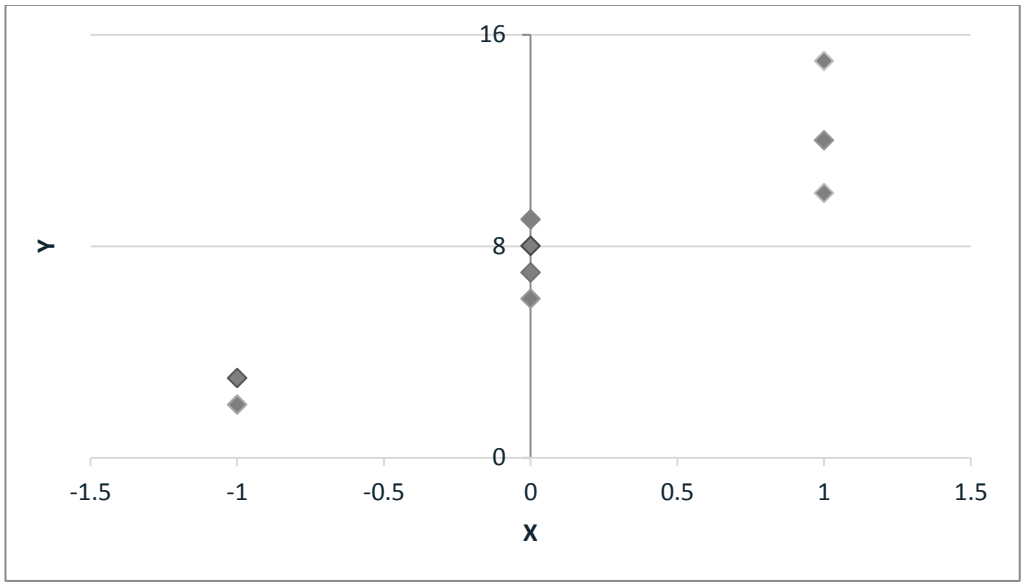
Naredni primer prikazuje primenu algoritma iterativnih težinskih najmanjih kvadrata.

Primer 3. Podaci dati u tabeli 4. su veštački generisani realizovani prebrojivi podaci za Y posmatrani za različite vrednosti nezavisne promenljive X.

Tabela 4. Podaci za primer Poasonove raspodele

$\boldsymbol{y_i}$	2	3	6	7	8	9	10	12	15
$\boldsymbol{x_i}$	-1	-1	0	0	0	0	1	1	1

Slika 3. Grafički prikaz podataka iz primera 3.



Pretpostavimo da su Y_i Poasonove slučajne promenljive. U praksi, pretpostavke o raspodeli podataka bismo doneli ili na osnovu numeričke provere ili na osnovu vizuelnih zaključaka o srednjim vrednostima i varijansama. Za date podatke možemo da primetimo da se disperzija povećava sa Y, što potvrđuje pretpostavku da podaci imaju Poasonovu raspodelu. Tada znamo da je

$$E(Y_i) = D(Y_i). \tag{4.15}$$

Model definišemo tako što pretpostavimo da su Y_i i x_i u linearnom odnosu

$$E(Y_i) = \mu_i = \beta_1 + \beta_2 x_i = \mathbf{x}_i^T \boldsymbol{\beta},$$

gde je

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \text{ i } \mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$

za $i = 1, \dots, n$. Dakle, uzimamo da je funkcija $g(\mu_i)$ funkcija identiteta

$$g(\mu_i) = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i.$$

Tada je $\frac{\partial \mu_i}{\partial \eta_i} = 1$, što pojednostavljuje jednačine (4.12) i (4.13). Iz (4.12) i (4.15) sledi

$$\omega_{ii} = \frac{1}{D(Y_i)} = \frac{1}{\beta_1 + \beta_2 x_i}.$$

Koristeći ocenu $\boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ za $\boldsymbol{\beta}$, jednačina (4.13) postaje

$$z_i = b_1 + b_2 x_i + (y_i - b_1 - b_2 x_i) = y_i.$$

Takođe

$$\mathcal{J} = X^T W X = \begin{bmatrix} \sum_{i=1}^{n} \frac{1}{b_1 + b_2 x_i} & \sum_{i=1}^{n} \frac{x_i}{b_1 + b_2 x_i} \\ \sum_{i=1}^{n} \frac{x_i}{b_1 + b_2 x_i} & \sum_{i=1}^{n} \frac{x_i^2}{b_1 + b_2 x_i} \end{bmatrix}$$

i

$$X^T W \boldsymbol{z} = \begin{bmatrix} \sum_{i=1}^{n} \frac{y_i}{b_1 + b_2 x_i} \\ \sum_{i=1}^{n} \frac{x_i y_i}{b_1 + b_2 x_i} \end{bmatrix}.$$

Ocene metodom maksimalne verodostojnosti su dobijene iterativno iz jednačina

$$(X^T W X)^{(m-1)} \boldsymbol{b}^m = (X^T W \boldsymbol{z})^{(m-1)},$$

gde $(m-1)$ označava ocenu u $\boldsymbol{b}^{(m-1)}$.

Za podatke koje posmatramo $n = 9$

$$y = z = \begin{bmatrix} 2 \\ 3 \\ \vdots \\ 15 \end{bmatrix} \text{ i X} = \begin{bmatrix} \mathrm{x}_1 \\ \mathrm{x}_2 \\ \vdots \\ \mathrm{x}_9 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}.$$

Sa slike 3. dobijamo početne ocene $b_1^{(1)} = 7$ i $b_2^{(1)} = 5$. Tada je

$$(X^T WX)^{(1)} = \begin{bmatrix} 1.821429 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}, \ (X^T W\boldsymbol{z})^{(1)} = \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix},$$

pa sledi,

$$\boldsymbol{b}^{(2)} = \left((X^T WX)^{(1)}\right)^{-1}(X^T W\boldsymbol{z})^{(1)} =$$
$$= \begin{bmatrix} 0.729167 & 0.4375 \\ 0.4375 & 1.0625 \end{bmatrix}\begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix} = \begin{bmatrix} 7.4514 \\ 4.9375 \end{bmatrix}.$$

Iterativni proces se nastavlje dok niz ne konvergira za datu veličinu. Ocene dobijene metodom maksimalne verodostojnosti su $\hat{\beta}_1 = 7.45163$ i $\hat{\beta}_2 = 4.93530$. Za ove vrednosti inverzna matrica matrice informacija $\mathcal{J} = X^T WX$ je

$$\mathcal{J}^{-1} = \begin{bmatrix} 0.7817 & 0.4166 \\ 0.4166 & 1.1863 \end{bmatrix}.$$

Tada je, na primer, 95% interval poverenja za β_2

$$4.9353 \pm 1.96\sqrt{1.1863} \text{ ili } (2.80, 7.07).$$

■

5. Provera adekvatnosti modela i statističko zaključivanje

Dva osnovna alata statističkog zaključivanja su intervali poverenja i testiranje hipoteza. *Intervali poverenja*, koje nazivamo još i intervalima ocena, se sve više koriste od testiranja hipoteza, jer širina intervala poverenja daje i meru preciznosti sa kojom će zaključak biti donesen. Oni su konceptualno mnogo jednostavniji nego određivanje moći statističkih testova. *Testiranje hipoteza* se izvodi tako što se poredi koliko dobro dva povezana modela fituju podatke. Za uopštene linearne modele, dva modela bi trebala da imaju istu raspodelu verovatnoća i istu funkciju veze, ali linearni prediktor jednog modela treba da sadrži više parametara od drugog modela. Jednostavniji model, koji odgovara nultoj hipotezi H_0, mora biti specijalan slučaj drugog, opštijeg modela. Ukoliko jednostavniji model fituje podatke podjednako kao i opštiji model, tada ćemo koristiti, naravno, jednostavniji model i hipoteza H_0 se ne odbacuje. Ako opštiji model fituje

podatke značajnije bolje, tada odbacujemo hipotezu H_0 u korist alternativne hipoteze H_1, koja odgovara opštijem modelu. Da bismo uporedili dva modela, postavljamo statistike koje opisuju koliko dobro model fituje podatke, tj. koliko se model slaže sa podacima. Takve statistike mogu biti bazirane na maksimalnoj vrednosti funkcije verodostojnosti, maksimalnoj vrednosti logaritma funkcije verodostojnosti, kriterijumu minimalne vrednosti sume kvadrata ili razlici statistika za odstupanje reziduala. Proces i logika mogu biti sumirani na sledeći način:

1. Definišemo model M_0 koji odgovara nultoj hipotezi H_0, a zatim definišemo uopšteniji model M_1 (pri čemu je M_0 specijalan slučaj modela M_1).
2. Fitujemo model M_0 i izračunamo statistiku G_0 koja pokazuje koliko se model dobro slaže sa podacima. Zatim, fitujemo model M_1 i izračunamo statistiku G_1 koja pokazuje koliko se taj model dobro slaže sa podacima.
3. Izračunamo poboljšanje u fitovanju, obično $G_1 - G_0$, ali možemo da posmatramo i $\frac{G_1}{G_0}$.
4. Koristimo uzoračku raspodelu za $G_1 - G_0$ (ili neku analognu statistiku) da bismo testirali nultu hipotezu da je $G_1 = G_0$, protiv alternativne hipoteze $G_1 \neq G_0$.
5. Ukoliko nulta hipoteza da je $G_1 = G_0$ nije odbačena, tada H_0 nije odbačena i jednostavnosti radi, koristićemo model M_0. Ukoliko je hipoteza da je $G_1 = G_0$ odbačena, tada je odbačena i hipoteza H_0 i smatramo da je model M_1 bolji.

Za oba tipa statističkog zaključivanja, i intervale poverenja i testiranje hipoteza, potrebna je uzoračka raspodela. Za intervale poverenja potrebna je uzoračka raspodela ocena. Kod testiranja hopoteza potrebna je uzoračka raspodela statistike koja pokazuje koliko se model dobro slaže sa podacima.

Ukoliko je S statistika koju posmatramo, tada je osnovna ideja da je pod određenim uslovima aproksimacija

$$\frac{S - E(S)}{\sqrt{D(S)}} \sim \mathcal{N}(0,1)$$

ili, ekvivalentno[7]

$$\frac{\big(S - E(S)\big)^2}{D(S)} \sim \chi^2(1),$$

gde su $E(S)$ i $D(S)$ očekivanje i disperzija od S, respektivno.

[7] Ako su $X_1, \dots, X_n$ nezavisne slučajne promenljive sa $\mathcal{N}(0,1)$ raspodelom, tada $X_1^2 + \cdots + X_n^2 \sim \chi^2(n)$.

Ako imamo vektor statistika koje posmatramo $\boldsymbol{s} = \begin{bmatrix} S_1 \\ \vdots \\ S_p \end{bmatrix}$, sa asimptotskim očekivanjem $E(\boldsymbol{s})$ i asimptotskom matricom varijanse i kovarijanse V, tada približno važi da je

$$\big(\boldsymbol{s} - E(\boldsymbol{s})\big)^T V^{-1} \big(\boldsymbol{s} - E(\boldsymbol{s})\big) \sim \chi^2(p) \tag{4.16}$$

što obezbeđuje da je matrica V nesingularna, pa postoji jedinstvena inverzna matrica V^{-1}.

Uzoračka raspodela za skor statistiku

Pretpostavimo da su $Y_1, \dots, Y_n$ nezavisne slučajne promenljive iz uopštenog linearnog modela sa parametrima $\boldsymbol{\beta}$, gde je $E(Y_i) = \mu_i$ i $g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i$. Iz jednačine (4.7) skor statistike imaju sledeći oblik

$$U_j = \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{(Y_i - \mu_i)}{D(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right), \qquad j = 1, \dots, p.$$

Kako je $E(Y_i) = \mu_i$, za sve i, sledi da je

$$E\big(U_j\big) = 0, \qquad \text{za } j = 1, \dots, p,$$

što je konzistentno sa opštim rezultatom da je očekivanje od skor statistike jednako 0. Matrica varijanse i kovarijanse za skor statistiku je matrica informacija $\mathcal{J}$ sa elementima matrice

$$\mathcal{J}_{jk} = E\big(U_j U_k\big),$$

koji su dati jednačinom (4.9).

Ukoliko postoji samo jedan parametar β, skor statistika ima asimptotsku uzoračku raspodelu

$$\frac{U}{\sqrt{\mathcal{J}}} \sim \mathcal{N}(0,1)$$

ili ekvivalentno,

$$\frac{U^2}{\mathcal{J}} \sim \mathcal{X}^2(1),$$

jer je $E(U) = 0$ i $D(U) = \mathcal{J}$.

Ukoliko imamo vektor parametara

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

tada je skor statistika vektor

$$\boldsymbol{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix},$$

koji ima multivarijantnu normalnu raspodelu $\boldsymbol{U} \sim \mathcal{N}(0, \mathcal{J})$, makar asimptotski, pa sledi da za veće uzorke važi da je

$$\boldsymbol{U}^T \mathcal{J}^{-1} \boldsymbol{U} \sim \mathcal{X}^2(p).$$

Tejlorov red aproksimacija

Za dobijanje asimptotskih uzoračkih raspodela za različite statistike korisno je koristiti *Tejlorov red aproksimacija*. Tejlorov red aproksimacija za funkciju $f(x)$, sa jednom nezavisnom promenljivom x, u tački t je

$$f(x) = f(t) + (x - t)\left(\frac{\partial f}{\partial x}\right)_{x=t} + \frac{1}{2}(x - t)^2 \left(\frac{\partial^2 f}{\partial x^2}\right)_{x=t} + \cdots$$

Za logaritam funkcije verodostojnosti koja ima samo jedan parametar β prva tri člana razvoja Tejlorovog reda aproksimacija u tački ocene b su

$$\ln l(\beta) = \ln l(b) + (\beta - b)U(b) + \frac{1}{2}(\beta - b)^2 U'(b),$$

pri čemu je $U(b) = \partial \ln l / \partial \beta$ statistika koja predstavlja ocenu parametra β, za $\beta = b$. Ako $U'(b) = \partial^2 \ln l / \partial \beta^2$ aproksimiramo njegovim očekivanjem $E(U') = -\mathcal{J}$, aproksimacija postaje

$$\ln l(\beta) = \ln l(b) + (\beta - b)U(b) - \frac{1}{2}(\beta - b)^2 \mathcal{J}(b),$$

gde je $\mathcal{J}(b)$ informacija za $\beta = b$. Odgovarajuća aproksimacija za logaritam funkcije verodostojnosti za vektor parametara $\boldsymbol{\beta}$ je

$$ln\, l(\boldsymbol{\beta}) = ln\, l(\boldsymbol{b}) + (\boldsymbol{\beta} - \boldsymbol{b})^T \boldsymbol{U}(\boldsymbol{b}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{b})^T \mathcal{J}(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b}),$$

gde je $\boldsymbol{U}$ vektor, a $\mathcal{J}$ matrica informacija.

Za funkciju U sa jednim parametrom β prva dva člana Tejlorovog niza aproksimacija u tački b daju

$$U(\beta) = U(b) + (\beta - b)U'(b).$$

Ako U' aproksimiramo sa $E(U') = -\mathcal{J}$, dobijamo

$$U(\beta) = U(b) - (\beta - b)\mathcal{J}(b).$$

Analogno, za vektor parametara $\boldsymbol{\beta}$ dobijamo

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{U}(\boldsymbol{b}) - \mathcal{J}(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b}). \qquad (4.17)$$

Uzoračka raspodela za ocene dobijene metodom maksimalne verodostojnosti

Jednačina (4.17) se može iskoristiti za dobijanje uzoračke raspodele ocene dobijene metodom maksimalne verodostojnosti $\boldsymbol{b} = \widehat{\boldsymbol{\beta}}$. Po definiciji, $\boldsymbol{b}$ je ocena koja maksimizira $\ln l(\boldsymbol{b})$ (kao i $l(\boldsymbol{b})$), pa je $\boldsymbol{U}(\boldsymbol{b}) = 0$. Tada je

$$\boldsymbol{U}(\boldsymbol{\beta}) = -\mathcal{J}(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b}),$$

ili, ekvivalentno

$$(\boldsymbol{b} - \boldsymbol{\beta}) = \mathcal{J}^{-1}\boldsymbol{U},$$

čime je obezbeđeno da je $\mathcal{J}$ nesingularna matrica. Ako je $\mathcal{J}$ konstantna, tada je $E(\boldsymbol{b} - \boldsymbol{\beta}) = 0$, jer je $E(\boldsymbol{U}) = 0$. Dakle, $E(\boldsymbol{b}) = \boldsymbol{\beta}$, barem asimptotski, pa je $\boldsymbol{b}$ konzistentna ocena za $\boldsymbol{\beta}$. Dovoljan uslov za konzistentnost je da je

$$\lim_{n\to\infty} E((\boldsymbol{b} - \boldsymbol{\beta})^2) = 0.$$

Matrica varijanse i kovarijanse za $\boldsymbol{b}$ je

$$E\big((\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})^T\big) = \mathcal{J}^{-1}E(\boldsymbol{U}\boldsymbol{U}^T)\mathcal{J} = \mathcal{J}^{-1},$$

jer je $\mathcal{J}^{-1} = E(\boldsymbol{U}\boldsymbol{U}^T)$, a $(\mathcal{J}^{-1})^T = \mathcal{J}^{-1}$, kako je $\mathcal{J}$ simetrična matrica.

Asimptotska uzoračka raspodela za $\boldsymbol{b}$ je, na osnovu (4.16)

$$(\boldsymbol{b}-\boldsymbol{\beta})^T \mathcal{J}(\boldsymbol{b})(\boldsymbol{b}-\boldsymbol{\beta}) \sim \chi^2(p). \tag{4.18}$$

Ova statistika se naziva *Valdova statistika*.

Statistika odnosa logaritama funkcija verodostojnosti

Jedan od načina da procenimo adekvatnost modela jeste da ga uporedimo sa opštijim modelom, koji sadrži maksimalan broj parametara koji se mogu oceniti. Takav model se zove kompletan (potpuni ili zasićen) model. To je uopšteni linearni model, koji ima istu raspodelu i funkciju veze kao i model koji posmatramo.

Pretpostavimo da ima n promenljivih $Y_i, i = 1, \dots, n$ koje posmatramo, pri čemu sve u opštem slučaju imaju drugačije vrednosti za linearnu komponentu $\boldsymbol{x}_i^T\boldsymbol{\beta}$. Tada se potpuni model definiše sa n parametara. U ovom slučaju, maksimalan broj parametara koji mogu biti ocenjeni za potpuni model jednak je broju potencijalno različitih linearnih komponenti, što može biti manje od n.

Opštije, označimo sa m maksimalan broj parametara koji mogu biti ocenjeni. Neka $\boldsymbol{\beta}_{max}$ označava vektor parametara potpunog modela, a $\boldsymbol{b}_{max}$ ocenu za $\boldsymbol{\beta}_{max}$ dobijenu metodom maksimalne verodostojnosti. Funkcija verodostojnosti za potpuni model u tački $\boldsymbol{b}_{max}$, $l(\boldsymbol{b}_{max};\, y)$, biće veća od bilo koje druge funkcije verodostojnosti za date registrovane vrednosti, sa pretpostavkama o istoj raspodeli i funkciji veze, jer ona daje najkompletniji opis podataka. Označimo sa $l(\boldsymbol{b};\, y)$ maksimalnu vrednost funkcije verodostojnosti za posmatrani model. Tada pomoću odnosa

$$\lambda = \frac{l(\boldsymbol{b}_{max};\, y)}{l(\boldsymbol{b};\, y)}$$

možemo da ocenimo koliko se dobro model slaže sa podacima. U praksi se koristi logaritam gornjeg razlomka, što zapravo predstvlja razliku izmedju logaritama funkcija verodostojnosti

$$\log\lambda = \log l(\boldsymbol{b}_{max};\, y) - \log l(\boldsymbol{b};\, y). \tag{4.19}$$

Velike vrednosti dobijene za $\log\lambda$ ukazuju na to da posmatrani model slabo opisuje podatke u odnosu na potpuni model. Da bismo odredili kritičnu oblast za $\log\lambda$, potrebno je da znamo njegovu uzoračku raspodelu.

U narednom poglavlju videćemo da $2\log\lambda$ ima hi-kvadrat raspodelu. Prema tome $2\log\lambda$ je statistika koju češće koristimo umesto $\log\lambda$.

Uzoračka raspodela za odstupanje reziduala

Odstupanje reziduala, koje nazivamo još i statistika logaritama funkcija verodostojnosti, je

$$D_r = 2(ln\, l(\boldsymbol{b}_{max}, y) - ln\, l(\boldsymbol{b}, y)).$$

Iz jednačine

$$ln\, l(\boldsymbol{\beta}) = ln\, l(\boldsymbol{b}) + (\boldsymbol{\beta} - \boldsymbol{b})^T \boldsymbol{U}(\boldsymbol{b}) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{b})^T \mathcal{J}(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b}),$$

ako je $\boldsymbol{b}$ ocena dobijena metodom maksimalne verodostojnosti za parametar $\boldsymbol{\beta}$, tako da je $\boldsymbol{U}(\boldsymbol{b}) = 0$, sledi

$$ln\, l(\boldsymbol{\beta}) - ln\, l(\boldsymbol{b}) = -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{b})^T \mathcal{J}(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b}).$$

Prema tome, statistika

$$2(ln\, l(\boldsymbol{b}, y) - ln\, l(\boldsymbol{\beta}, y)) = (\boldsymbol{\beta} - \boldsymbol{b})^T \mathcal{J}(\boldsymbol{b})(\boldsymbol{\beta} - \boldsymbol{b})$$

ima hi-kvadrat raspodelu $\chi^2(p)$, gde je p broj parametara, iz (4.18).

Odavde možemo izvesti uzoračku raspodelu za odstupanje reziduala

$$\begin{aligned} D_r &= 2(ln\, l(\boldsymbol{b}_{max}, y) - ln\, l(\boldsymbol{b}, y)) = \\ &= 2(ln\, l(\boldsymbol{b}_{max}, y) - ln\, l(\boldsymbol{\beta}_{max}, y)) - 2(ln\, l(\boldsymbol{b}, y) - ln\, l(\boldsymbol{\beta}, y)) \\ &+2(ln\, l(\boldsymbol{\beta}_{max}, y) - ln\, l(\boldsymbol{\beta}, y)). \end{aligned} \qquad (4.20)$$

Za $2(ln\, l(\boldsymbol{b}_{max}, y) - ln\, l(\boldsymbol{\beta}_{max}, y))$ znamo da ima $\chi^2(m)$ raspodelu, gde je m broj parametara potpunog modela. Dalje, $2(ln\, l(\boldsymbol{b}, y) - ln\, l(\boldsymbol{\beta}, y))$ ima $\chi^2(p)$ raspodelu, gde je p broj parametara u modelu koji posmatramo. Na kraju, $\upsilon = 2(ln\, l(\boldsymbol{\beta}_{max}, y) - ln\, l(\boldsymbol{\beta}, y))$, je pozitivna konstanta koje će biti blizu nule ukoliko posmatrani model fituje podatke približno dobro kao i potpuni model. Dakle, tada je uzoračka raspodela za odstupanje reziduala, približno,

$$D_r \sim \chi^2(m - p, \upsilon),$$

gde υ predstavlja parametar necentralnosti raspodele χ^2. Odstupanje reziduala postavlja bazu za većinu testova hipoteza kod uopštenih linearnih modela.

***Primer 4**. Odstupanje reziduala za Poasonov model*

Pretpostavimo da su $Y_1, \dots, Y_n$ nezavisne slučajne promenljive i $Y_i \sim Pois(\mu_i)$. Tada je logaritam funkcije verodostojnosti

$$\ln l(\boldsymbol{\beta}, \mathrm{y}) = \sum y_i \log \mu_i - \sum \mu_i - \sum \log y_i!.$$

Za zasićen model, μ_i su različite za sve $i = 1, \dots, n$, tako da je $\boldsymbol{\beta} = [\mu_1, \dots, \mu_n]^T$.Ocene dobijene metodom maksimalne verodostojnosti su $\hat{\mu}_i = y_i$, pa je maksimalna vrednost logaritma funckije verodostojnosti

$$\ln l(\boldsymbol{\beta}_{max}, \mathrm{y}) = \sum y_i \log y_i - \sum y_i - \sum \log y_i!.$$

Pretpostavimo da model koji želimo da koristimo ima $p < n$ parametara. Ocena dobijena metodom maksimalne verodostojnosti $\boldsymbol{b}$ se može koristiti da bismo izračunali ocene $\hat{\mu}_i$, pa su tada fitovane vrednosti $\hat{y}_i = \hat{\mu}_i$, jer je $E(Y_i) = \mu_i$. Maksimalna vrednost logaritma funckije verodostojnosti je u ovom slučaju

$$ln\, l(\boldsymbol{b}, y) = \sum y_i \, log\, \hat{y}_i - \sum \hat{y}_i - \sum log\, y_i!.$$

Tada je D_r

$$\begin{aligned} D_r &= 2(ln\, l(\boldsymbol{\beta}_{max}, y) - ln\, l(\boldsymbol{b}, y)) \\ &= 2\left(\sum y_i \, log \frac{y_i}{\hat{y}_i} - \sum (y_i - \hat{y}_i)\right). \end{aligned}$$

Za većinu modela se može pokazati da je $\sum y_i = \sum \hat{y}_i$. Dakle, D_r se može napisati u sledećem obliku

$$D_r = 2 \sum o_i \log \frac{o_i}{e_i},$$

gde je o_i oznaka za registrovanu vrednost y_i, a e_i označava ocenu očekivane vrednosti $\hat{y}_i$.

Vrednost za D_r se u ovom slučaju može izračunati. Ta vrednost se može uporediti sa raspodelom $\chi^2(n - p)$. Sledeći primer ilustruje ovu ideju.

Podaci u tabeli 5. odgovaraju primeru 3. gde su podaci sa Poasonovom raspodelom modelirani linearno (pravom linijom). Fitovane vrednosti su

$$\hat{y}_i = b_1 + b_2 x_i$$

gde je $b_1 = 7.45163$, a $b_2 = 4.9353$. Tada je $D_r = 2(0.94735 - 0) = 1.8947$, što je u slaboj vezi sa stepenima slobode, $n - p = 9 - 2 = 7$. U stvari, D_r je ispod 5% repa raspodele $\chi^2(7)$, prema čemu se model dobro slaže sa podacima (što je i logično za mali skup veštački generisanih podataka).

Tabela 5. Rezultati Poasonove regresije iz primera 3.

x_i	y_i	$\hat{y}_i$	$y_i \log y_i/\hat{y}_i$
-1	2	2.51633	-0.45931
-1	3	2.51633	0.52743
0	6	7.45163	-1.30004
0	7	7.45163	-0.43766
0	8	7.45163	0.56807
0	9	7.45163	1.69913
1	10	12.38693	-2.14057
1	12	12.38693	-0.38082
1	15	12.38693	2.87112
Ukupno	72	72	0.94735

■

Testiranje hipoteza

Hipoteze o vektoru parametara $\boldsymbol{\beta}$ dužine p mogu da se testiraju pomoću uzoračke raspodele Valdove statistike

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \mathcal{J} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2(p).$$

Alternativni metod koji se koristi je poređenje dva modela i koliko se oni dobro slažu sa podacima. Modeli moraju biti ugnježdeni ili u hijerarhijskom odnosu, tj. moraju imati istu raspodelu verovatnoća i istu funkciju veze, gde je linearna komponenta jednostavnijeg modela M_0 specijalni slučaj linearne komponente uopštenijeg modela M_1.

Neka nulta hipoteza

$$H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

odgovara modelu M_0, a uopštenija hipoteza

$$H_1: \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

odgovara modelu M_1, pri čemu je $q < p < n$.

Testiramo hipotezu H_0 protiv H_1 koristeći razliku između statistika za odstupanje reziduala

$$\begin{aligned} &\Delta D_r = D_{r0} - D_{r1} = \\ &= 2(ln\, l(\boldsymbol{b}_{max}, y) - ln\, l(\boldsymbol{b}_0, y)) - 2(ln\, l(\boldsymbol{b}_{max}, y) - ln\, l(\boldsymbol{b}_1, y)) \\ &= 2(ln\, l(\boldsymbol{b}_1, y) - ln\, l(\boldsymbol{b}_0, y)). \end{aligned}$$

Ukoliko oba modela dobro opisuju podatke, tada $D_{r0} \sim \chi^2(n-q)$ i $D_{r1} \sim \chi^2(n-p)$, pa $\Delta D_r \sim \chi^2(p-q)$, uz pretpostavku da važi potrebna nezavisnost promenljivih. Ako je ΔD_r konzistentna sa $\chi^2(p-q)$ raspodelom, obično biramo model M_0 koji odgovara hipotezi H_0, jer je on jednostavniji.

Ukoliko vrednost za ΔD_r upada u kritičnu oblast (odnosno, vrednost je veća od gornjeg repa raspodele $\chi^2(p-q)$ za $100 \text{ x } \alpha\%$), tada odbacujemo hipotezu H_0 u korist hipoteze H_1, zbog toga što model M_1 značajno bolje opisuje podatke od modela M_0 (iako to i dalje ne znači da se model M_1 naročito dobro slaže sa podacima).

Kako se odstupanje reziduala može izračunati na osnovu registrovanih podataka, ΔD_r predstavlja dobar metod za testiranje hipoteza.

6. Preraspršenost ili prekoračenje disperzije

Iako Poasonova slučajna promenljiva obezbeđuje slučajnost u strukturi prilikom modeliranja prebrojivih podataka, ona nije dovoljno fleksibilna da izdrži sve probleme ovakve regresije. Poasonova slučajna promenljiva je ograničena u smislu da je njena disperzija jednaka srednjoj vrednosti. Zato se uvode razna uopštenja Poasonove regresije koja mogu biti vrlo korisna za neke skupove podataka, jer pomoću njih, na primer, objašnjavamo veću disperziju nego što je očekivana (preraspršenost) i više ili manje registrovanih vrednosti prebrojivih podataka (često više ili manje nula nego što je očekivano).

Postoje najmanje četiri razloga zašto dolazi do većih varijacija oko uslovnog očekivanja Poasonovog regresionog modela. Pre svega, može doći do izostavljanja bitnih parametara. Drugo, mogu biti netačni oblici korišćenih funkcija. Treće, može da postoji slučajna varijacija uslovnih očekivanja. Četvrto, može postojati zavisnost između

događaja koji čine prebrojive podatke. Preraspršenost ne predstavlja tek bilo koju veću varijaciju uslovnih raspodela prebrojivih podataka. Prekoračenje usled izostavljanja bitnih parametara ili druge greške u sistematičnom delu modela ne predstavljaju preraspršenost. Ukratko, ukoliko postoje greške u sistematičkom delu Poasonovog modela, ne postoji drugi način popravljanja osim postavljanja ovog dela kako treba. Ukoliko je sistematički deo modela tačan, što znači da ni jedan važan parametar nije izostavljen i da su funkcije dobro definisane, a ipak postoje povećane varijacije oko fitovanih vrednosti, uzrok može biti stohastičko uslovno očekivanje. Preraspršenost predstavlja prekoračenje koje potiče iz toga kako je definisana stohastička komponenta modela, pri čemu je sistematička struktura modela tačna. Potencijalno rešenje može biti zamena Poasonove raspodele negativnom binomnom raspodelom.

Najčešći slučaj zbog čega dolazi do preraspršenosti je nemodeliranje heterogenosti, gde razlike u srednjim vrednostima među registrovanim vrednostima nisu uzete u obzir u modelu. Primetimo da se ovo takođe može desiti i za binomne podatke (a prema tome i u logističkom regresionom modelu), jer binomna slučajna promenljiva takođe ima osobinu da je njena disperzija tačno determinisana sredinom. Postoje specifični testovi pravljeni tako da identifikuju preraspršenost, ali obično su dovoljne standardne statistike za procene slaganja modela sa podacima, X^2 i G^2. Prisustvo preraspršenosti se ne sme ignorisati, jer čak i ako je forma fitovanog Poasonovog modela tačna, ne uračunavanje preraspršenosti dovodi do ocena disperzija procenjenih koeficijenata koje su previše male, čime nastaju previše uski intervali poverenja i suviše male p-vrednosti značajnosti testova. Specijalno, ocene standardnih greški procenjenih koeficijenata su previše male za faktor koji predstavlja odnos između prave standardne devijacije i procenjene devijacije na osnovu Poasonove regresije. Na primer, ako je prava standardna devijacija od y za 20% veća od devijacije na osnovu Poasonove regresije, procenjene standardne greške bi morale biti za 20% veće da bi uspele da reflektuju situaciju.

Kako je preraspršenost prebrojivih podataka vrlo čest slučaj, postoji nekoliko modela koji su razvijeni za takve podatke. Kvazi-Poasonova i negativna binomna regresija su najčešće korišćene i dostupne su u najvećem broju softvera.

Kvazi-Poasonov i negativni binomni model imaju isti broj parametara i oba mogu da se koriste za rešavanje problema preraspršenosti prebrojivih podataka. U velikom broju slučajeva, oba metoda će dati slične rezultate, međutim postoje bitne razlike između ova dva modela. Disperzija kod kvazi-Poasonovog modela je linearna funkcija srednje vrednosti, dok je kod negativnog binomnog modela disperzija kvadratna funkcija sredine. Ova razlika u obliku disperzije utiče na težinske koeficijente u algoritmu iterativnih težinskih najmanjih kvadrata prilikom fitovanja modela prema podacima. Kako je

disperzija funkcija srednje vrednosti, veliki i mali prebrojivi podaci će imati drugačije težinske koeficijente kod kvazi-Poasonove i negativne binomne regresije.

Kvazi-Poasonov model

U slučaju kada je disperzija prebrojivih podataka veća nego što je modelirana sa Poasonovim modelom, jedan od načina da prevaziđemo ograničenje da je srednja vrednost jednaka disperziji jeste da uvedemo *parametar disperzije*, koji će dozvoljavati prekoračenje disperzije u ovom smislu.

Neka su $Y_1, \dots, Y_n$ nezavisne slučajne promenljive i neka je $E(Y_i) = \mu_i$. Sada ćemo uvesti parametar disperzije θ, takav da je

$$D(Y_i) = \theta\mu_i.$$

Kada je $\theta > 1$, tada je disperzija veća nego što je srednja vrednost, a za $\theta < 1$ imamo slučaj disperzije koja je manja u odnosu na očekivanu po Poasonovom modelu. Prilagođavanje Poasonovog regresionog modela pomoću parametra disperzije koji je linearno zavisan od funkcije sredine, naziva se *kvazi-verodostojan metod* (ili *kvazi-Poasonov metod*).

Naziv kvazi-verodostojna funkcija je prvi uveo Vederburn 1974. godine da bi opisao funkciju koja ima slične osobine kao i funkcija verodostojnosti, osim što kvazi-verodostojna funkcija zapravo ne uzima u obzir ni jednu raspodelu verovatnoća. Umesto da uključuje raspodelu verovatnoća podataka, ovaj metod definiše samo odnos između funkcije srednje vrednosti i disperzije. Dakle, disperzija je u stvari prikazana kao funkcija srednje vrednosti.

Kao posledicu uvođenja parametra disperzije za preraspršene podatke dobićemo ocene standardnih grešaka, koje su sve pomnožene sa $\sqrt{\theta}$ u odnosu na Poasonov regresioni model. Prema tome, ukoliko zanemarimo prekoračenje disperzije, možemo doći do pogrešnih zaključaka.

Negativni Binomni model

Za Poasonov model kod koga prepoznajemo šum kod merenja prebrojivih podataka, možemo definisati i drugu modifikaciju kod koje na standardni model dodajemo stohastički deo ε_i, tj.

$$\log \lambda_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

Očekivanje i disperzija za svako ε_i su jednaki nuli i sva ε_i su međusobno nezavisno generisana. Kao posledica uticaja ε_i, λ_i predstavlja modifikovanu verziju od μ_i za šum ε_i. Čak i ako posmatramo jedan slučaj, λ_i može da varira za različite registrovane podatke, tako da slučajevi sa istim skupom pretpostavki u opštem slučaju neće imati istu vrednost λ. Ovako posmatran model za prebrojive podatke može da se shvati kao Poasonov model sa dvostrukom slučajnosti, jer pored slučajnosti koja je uključena u formulaciju Poasonovog modela, postoji i drugi izvor slučajnosti koji je generisan u ε_i.

U ovakvoj formulaciji bitno je napomenuti da je $\boldsymbol{x}_i^T \boldsymbol{\beta}$ dobro definisano. Nijedna promenljiva nije izostavljena i funkcije su dobro definisane. Drugim rečima, sistematički deo modela je tačan.

Pre nego što pređemo na procese za ocenjivanje parametara regresije, potrebno je da postavimo određene pretpostavke o osobinama za ε_i. Poasonova formulacija može biti izmenjena, tako da je

$$f(y_i|\boldsymbol{x}_i, \lambda_i) = \frac{e^{-\mu_i \lambda_i} (\mu_i \lambda_i)^{y_i}}{y_i!}$$

što znači da uslovna raspodela za y_i koja zavisi od $\boldsymbol{x}_i$ i λ_i, ipak ostaje i dalje Poasonova. Međutim, sada se postavlja pitanje kako da odredimo raspodelu za y_i koji zavise samo od $\boldsymbol{x}_i$, jer su $\boldsymbol{x}_i$ zapravo nezavisne promenljive koje posmatramo.

Funkcija raspodele za y_i koja zavisi samo od posmatranih $\boldsymbol{x}_i$ je data sa

$$f(y_i|\boldsymbol{x}_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} {r_i}^{y_i} (1 - r_i)^{\theta},$$

gde je

$$r_i = \frac{\mu_i}{\mu_i + \theta}.$$

Iz matematički praktičnih razloga koristimo gama raspodelu, a parametar θ je određen *a priori* ili ocenjen.

Gornja formulacija nam daje negativnu binomnu raspodelu. Negativna binomna raspodela je diskretna raspodela verovatnoća koja pokazuje broj uspešnih pokušaja u nizu nezavisnih i jednako raspodeljenih Bernulijevih pokušaja, pre nego što se određeni broj neuspešnih pokušaja dogodi. Ova raspodela se bavi nenegativnim celim brojevima, ali sa manjim ograničenjima u odnosu na Poasonovu raspodelu. Negativna binomna

raspodela ima dodatni parametar koji dozvoljava da disperzija bude veća od očekivanja. Očekivanje je jednako μ_i, što odgovara Poasonovoj raspodeli. Ovo je veoma važan odnos između dve raspodele, jer to implicira da je funkcija očekivane srednje vrednosti ista, bilo da koristimo Poasonovu ili negativnu binomnu raspodelu. Obe raspodele, u suštini, procenjuju istu stvar. Zbog toga, u praksi se često dešava da ocenjeni koeficijenti regresije pomoću ove dve procedure nemaju velike razlike. Dakle, ukoliko postoje problemi sa funkcijom srednje vrednosti kada koristimo Poasonovu raspodelu, isti problemi će ostati i ako pređemo na negativnu binomnu raspodelu.

Disperzija za uslovnu srednju vrednost μ_i nije μ_i, već

$$\mu_i(1 + (1/\theta)\mu_i) = \mu_i + (1/\theta)\mu_i{}^2.$$

Za $\theta > 0$, disperzija je modifikovana tako da rešava preraspršenost. Što je manja vrednost parametra $\theta > 0$, to je veća preraspršenost i raspodela se sve više razlikuje od Poasonove. Ukoliko $\theta \to \infty$ možemo da se vratimo na Poasonovu raspodelu, jer tada negativna binomna raspodela teži Poasonovoj. Ukoliko je $\theta < 0$, tada imamo slučaj da su disperzije manje nego što je to po Poasonovom modelu očekivano. Međutim, kakva god da je vrednost parametra θ, svako μ_i je pomnoženo istim faktorom.

Vrednosti parametara $\boldsymbol{\beta}$ i θ mogu biti ocenjene metodom maksimalne verodostojnosti. Takođe, možemo dobiti i ocene standardnih grešaka za oba parametra. Dakle, možemo da zaključimo da ukoliko je sistematički deo Poasonovog modela tačan, negativna binomna raspodela može rešiti određene probleme vezane za prekoračenje disperzije.

Jedan od načina da proverimo da li postoji preraspršenost podataka je da to uradimo pomoću ocena iz negativnog binomnog modela. Kako nam ovaj model daje ocenu parametra disperzije θ, potrebno je da testiramo da li je θ značajno različito od 0. Dakle, postavljamo hipotezu $H_0: \theta = 0$, protiv alternativne hipoteze $H_0: \theta \neq 0$. U slučaju kada je:

1. $\theta = 0$, koristimo Poasonov model;
2. $\theta > 0$, postoji preraspršenost;
3. $\theta < 0$, disperzija je manja od srednje vrednosti (što je redak slučaj).

V. Poasonova regresija za stope

Kao što smo videli, kod Poasonovog modela obeležje Y_i je prebrojiva slučajna promenljiva. Međutim, možemo posmatrati i Y_i/t, stopu (ili incidencu) kao obeležje, pri čemu t predstavlja vreme, prostor ili neki drugi skup. Tada imamo sledeću uopšteni linearni model:

$$g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

- **Komponenta slučajnosti**: Slučajna promenljiva Y_i ima Poasonovu raspodelu, a t predstavlja prostor ili vreme. Očekivanje za stopu Y_i/t je $E(Y_i/t) = \mu_i$, dakle važi $E(Y_i) = \mu_i t$;
- **Komponenta sistematičnosti** ili **linearno predviđanje** za Poasonovu regresiju je linearna funkcija parametara regresije iz skupa nezavisnih promenljivih $X = (X_1, X_2, \dots, X_k)$;
- **Funkcija veze** je logaritam stope, $\log(Y_i/t)$.

Poasonov regresioni model za očekivanu stopu ostvarivanja događaja je

$$\log(\mu_i/t) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Ovo možemo zapisati kao

$$\log \mu_i - \log t = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

$$\log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \log t.$$

Član $\log t$ služi za podešavanje. Grupa posmatranja može imati istu vrednost t ili svako pojedinačno posmatranje može imati drugačiju vrednost. $\log t$ takođe utiče na ocenu srednje vrednosti prebrojivih podataka

$$\mu_i = e^{\boldsymbol{x}_i^T \boldsymbol{\beta} + \log t} = t e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}.$$

Odavde vidimo da su prebrojivi podaci proporcionalni u odnosu na t. Primetimo da tumačenje ocene parametara $\boldsymbol{\beta}$ ostaje isto; jedino što moramo da pomnožimo prebrojive podatke sa t.

VI. Konstrukcija i analiza modela Poasonove regresije na primeru konzumiranja neoporezovanih duvanskih proizvoda

Decenijskim istraživanjima naučno je provereno da cigarete predstavljaju proizvod koji određenom upotrebom izaziva zavisnost. Međutim, potrošači su u najvećem broju zemalja prilično osetljivi na promene cena ovog proizvoda. Definišimo *dostupnost* kao odnos tržišne cene paklice cigareta na najpopularnijoj ceni, sa jedne strane, i prosečnog raspoloživog mesečnog prihoda, sa druge strane, gde prosečni raspoloživi mesečni prihod podrazumeva ostatak od prosečnog mesečnog prihoda, nakon plaćanja svih redovnih fiksnih mesečnih troškova. *Prag dostupnosti* predstavlja najveći procenat prosečnog raspoloživog mesečnog prihoda, koji je potrošač spreman da da za jednu jedinicu proizvoda. Zbog intenzivne akcizne politike, u većini zemalja potrošači su dovedeni do praga dostupnosti kod cigareta, što znači da svako sledeće povećanje cena uzrokuje prelazak dela potrošača legalnih cigareta na, jeftinije, ilegalne.

Udeo državnih prihoda od akciza na cigarete ima tendenciju smanjivanja sa razvojem ekonomije. U zemljama koje nemaju dobro razvijenu i zdravu ekonomiju, procenat državnog budžeta koji dolazi od akciza na cigarete dostiže i 10%. Zbog toga je pravilna dinamika akcizne politike ključna za planiranje razvoja zemalja u tranziciji.

U ovom radu ćemo pokazati kakva je zavisnost broja paklica na koje nije plaćen porez, a koje su prodate u radnji u odnosu na različite faktore, kao što su, na primer, udaljenost radnje od najbliže granice, pol, starost i stepen obrazovanja potrošača, dostupnost cigareta potrošaču, itd. Podaci su veštački generisani, a populacija je veličine $N = 300$.

Metodologija istraživanja se zasniva na anketiranju potrošača na mestu prodaje, licem u lice, sledećim upitnikom:

Upitnik:

1. *Koliko imate godina?* _____
2. *Pol* *m/ž*
3. *Stepen obrazovanja (1-8)* _____

4. *Koliko tačno cigareta popušite dnevno u proseku?* _____
5. *Koliki je Vaš prosečan mesečni prihod?* _____
6. *Koliko tačno mesečno trošite na cigarete u proseku?* _____
7. *Da li primate neki oblik socijalne pomoći?* *da/ne*

Popunjava anketar:

8. *Da li se radnja se nalazi u mestu koje ima više ili manje od 5,000 stanovnika?* *više/manje*
9. *Koja je udaljenost radnje od najbliže granice (u km)?* _____
10. *Paklica koju ima potrošač ima:*
    i. *akciznu markicu Republike Srbije*
    ii. *akciznu markicu druge zemlje*
    iii. *nema akciznu markicu*

U tabeli 6. prikazujemo kratak pregled svih nezavisnih promenljivih modela, kao i njihove osnovne karakteristike. Za nezvisnu promenljivu AkcMarkica uzimamo vrednosti 0=potrošač je kupio paklicu cigareta sa akciznom markicom Republike Srbije i n=potrošač je kupio paklicu cigareta sa akciznom markicom neke druge zemlje ili bez akcizne markice, gde je $n > 0$ broj paklica koje je kupio potrošač.

Tabela 6. Nezavisne promenljive modela, njihove potencijalne vrednosti i SPSS naziv

Promenljiva	**Vrednosti**	**SPSS naziv**
Godine	18,19,20,...	God
Pol	0=muški 1=ženski	Pol
Stepen obrazovanja	1=I stepen ⋮ 8=VIII stepen	StObraz
Dnevna potrošnja cigareta	$(0, +\infty)$	ADC
Prosečan mesečni prihod	$(0, +\infty)$	PrMesPr
Mesečna potrošnja na cigarete	$(0, +\infty)$	PrMesCig
Primanje nekog vida socijalne pomoći	0=da 1=ne	SocPom
Urbanost naselja u kome se nalazi objekat	0=urban 1=rural	UrbRur
Udaljenost objekta od najbliže granice	$(0, +\infty)$	distKM

Nakon što smo uneli podatke u softverski paket za obradu podataka SPSS, pozivamo analizu za Poasonovu regresiju. U modelu ćemo razmatrati kakav je uticaj svih nezavisnih promenljivih pojedinačno na zavisnu promenljivu, kao i uticaj nekih kombinovanih faktora, kao što su interakcije između broja godina i prosečnog broja konzumiranih cigareta, tipa naselja u kome je posmatrani objekat i prosečnog broja konzumiranih cigareta, prosečnih mesečnih prihoda i prosečne mesečne potrošnje na cigarete. Za Hi-kvadrat test i intervale poverenja koristićemo Wald-ovu statistiku, pri čemu je nivo intervala poverenja 95%.

Prvo, primetimo na osnovu tabele 7. da su sve ankete uzete u obzir od strane SPSS-a, prilikom analize (što je i logično, s obzirom da su podaci veštački generisani), a to znači da u podacima ne postoje outlier-i, niti nedostaju informacije unutar unesenih podataka. U slučaju da postoje prazne ćelije u tabeli sa podacima, SPSS će jednostavno izostaviti ceo red podataka.

Tabela 7. SPSS pregled nakon procesiranja unetih podataka

Case Processing Summary		
	N	Percent
Included	300	100.0%
Excluded	0	0.0%
Total	300	100.0%

U tabeli 8. prikazujemo kako izgleda pregled kategoričkih nezavisnih promenljivih u modelu. Možemo da primetimo da je populacija skoro ravnomerno raspodeljena prema polu (muški/ženski), da oko 12% anketirane populacije prima neki vid socijalne pomoći, kao i da je odnos urban/rural $55.3\%/44.7\%$.

U tabeli 9. dajemo pregled informacija o zavisnoj promenljivoj, kao i o neprekidnim nezavisnim promenljivama u modelu, gde možemo da vidimo koje su njihove minimalne i maksimalne vrednosti, sredina i standardna devijacija.

Sada ćemo pogledati rezultate koji govore o ukupnoj značajnosti i valjanosti samog modela. Ako pogledamo meru za odstupanje reziduala i vrednost za Pirsonovu Hi-kvadrat statistiku u tabeli 10. videćemo da one iznose 0.656 i 1.743. Za Poasonovu regresiju ove vrednosti treba da budu blizu jedinice, jer ukoliko su veće od 2 imamo indikaciju da su podaci preraspršeni. Dakle, u našem slučaju možemo da zaključimo da se model dobro slaže sa podacima, prema ovom indikatoru.

Dalje, posmatrajmo omnibus test, koji uzima u obzir statistiku odnosa logaritama funkcija verodostojnosti, koja ima Hi-kvadrat raspodelu. Omnibus test predstavlja testiranje hipoteza pri čemu se porede dva modela, trenutni model i model u kome su svi ocenjeni parametri jednaki nuli. Ovaj test pokazuje koliko puta je verovatnije da će se registrovani podaci bolje slagati sa jednim, nego sa drugim modelom. Na osnovu p-vrednosti koju smo dobili, možemo da zaključimo da se model značajno dobro slaže sa podacima.

Ukoliko želimo da poredimo naš model sa nekim drugim modelima, to možemo da uradimo pomoću pokazatelja kao što su AIC, AICC (koji prepravlja model za manje uzorke), BIC i CAIC. Dakle, ovi kriterijumi su uporedivi sa drugim, neugnježdenim modelima. U slučaju poređenja više modela, bolji će biti onaj model koji ima manje vrednosti za ove kriterijume.

Tabela 8. SPSS pregled informacija o kategoričkim nezavisnim promenljivim u modelu

Categorical Variable Information				
			N	Percent
Factor	Pol	0	149	49.7%
		1	151	50.3%
		Total	300	100.0%
	StObraz	1	37	12.3%
		2	48	16.0%
		3	47	15.7%
		4	33	11.0%
		5	30	10.0%
		6	40	13.3%
		7	33	11.0%
		8	32	10.7%
		Total	300	100.0%
	SocPom	0	36	12.0%
		1	264	88.0%
		Total	300	100.0%
	UrbRur	0	166	55.3%
		1	134	44.7%
		Total	300	100.0%

Tabela 9. SPSS pregled informacija o neprekidnim promenljivama u modelu

Continuous Variable Information						
		N	Min	Max	Mean	Std. Deviation
Dependent Variable	AkcMarkica	300	0	4	.14	.518
Covariate	God	300	18	64	41.45	13.607
	ADC	300	17.1	21.9	19.444	1.4366
	PrMesPr	300	13732	69285	41075.03	14694.946
	PrMesCig	300	2562	4669	3626.24	603.701
	distKM	300	3	120	59.84	34.663

Tabela 10a. SPSS pregled informacija o ukupnoj značajnosti modela

Goodness of Fit[a]			
	Value	df	Value/df
Deviance	184.262	281	.656
Scaled Deviance	184.262	281	
Pearson Chi-Square	489.845	281	1.743
Scaled Pearson Chi-Square	489.845	281	
Log Likelihood[b]	-121.824		
Akaike's Information Criterion (AIC)	281.648		
Finite Sample Corrected AIC (AICC)	284.363		
Bayesian Information Criterion (BIC)	352.020		
Consistent AIC (CAIC)	371.020		

Dependent Variable: AkcMarkica
Model: (Intercept), Pol, StObraz, SocPom, UrbRur, God, ADC, PrMesPr, PrMesCig, distKM, PrMesPr * PrMesCig, God * ADC, UrbRur * ADC

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

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

Tabela 10b. SPSS pregled informacija o ukupnoj značajnosti modela

Omnibus Test[a]		
Likelihood Ratio Chi-Square	df	Sig.
32.891	18	.017

Dependent Variable: AkcMarkica
Model: (Intercept), Pol, StObraz, SocPom, UrbRur, God, ADC, PrMesPr, PrMesCig, distKM, PrMesPr * PrMesCig, God * ADC, UrbRur * ADC

Compares the fitted model against the intercept-only model.

U tabeli 11. imamo pregled uticaja svih promenljivih modela. Izrazi koji imaju značajnosti manje od 0.05, imaju primetan i značajan efekat na model i na zavisnu promenljivu. Dakle, nezavisne promenljive koje imaju efekta na model su prosečni mesečni prihod, prosečna mesečna potrošnja na cigarete, udaljenost najbliže granice, kao i promenljiva koja predstavlja interakciju između prosečnih mesečnih prihoda i prosečne mesečne potrošnje na cigarete (odnosno, promenljiva koja predstavlja dostupnost cigareta potrošaču).

Tabela 12. sa ocenama parametara modela pokazuje efekte svakog faktora na model. Pored toga što su prikazani nestandardizovani koeficijenti regresije, njihove standardne greške i intervali poverenja, vidimo takođe i intervale poverenja za eksponencijalne nestandardizovane koeficijente. Eksponencijalni koeficijenti su prikazani u koloni *Exp(B)* i njih posmatramo kada prikazujemo rezultate regresije u obliku stope (ili incidence). Ove vrednosti su jednostavno izračunate kao eksponencijalne vrednosti koeficijenata regresije. U slučaju kada je vrednost eksponencijalnog nestandardizovanog koeficijenta jednaka 1, tada taj koeficijent regresije nema uticaja na model. U slučaju kada je njegova vrednost u intervalu $(0,1)$, tada su parametar i zavisna promenljiva u inverznom odnosu, a kada je > 1, tada koeficijent ima pozitivan uticaj na model.

Tabela 11. SPSS pregled značajnosti pojedinačnih izraza u modelu

Tests of Model Effects			
	Type III		
Source	Wald Chi-Square	df	Sig.
(Intercept)	.157	1	.692
Pol	.485	1	.486
StObraz	2.769	7	.905
SocPom	2.698	1	.100
UrbRur	.078	1	.780
God	.499	1	.480
ADC	.346	1	.557
PrMesPr	**4.921**	**1**	**.027**
PrMesCig	**4.625**	**1**	**.032**
distKM	**5.954**	**1**	**.015**
PrMesPr * PrMesCig	**6.161**	**1**	**.013**
God * ADC	.391	1	.532
UrbRur * ADC	.065	1	.799

Dependent Variable: AkcMarkica
Model: (Intercept), Pol, StObraz, SocPom, UrbRur, God, ADC, PrMesPr, PrMesCig, distKM, PrMesPr * PrMesCig, God * ADC, UrbRur * ADC

Tabela 12a. SPSS pregled svih parametara modela

Parameter Estimates							
			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-2.371	6.7221	-15.546	10.804	.124	1	.724
[Pol=0]	-.231	.3313	-.880	.419	.485	1	.486
[Pol=1]	0[a]						
[StObraz=1]	.052	.7846	-1.486	1.590	.004	1	.947
[StObraz=2]	.661	.7015	-.714	2.036	.887	1	.346
[StObraz=3]	.288	.7040	-1.092	1.668	.167	1	.682
[StObraz=4]	.163	.7761	-1.358	1.685	.044	1	.833
[StObraz=5]	.728	.7224	-.688	2.144	1.016	1	.313
[StObraz=6]	.113	.7784	-1.413	1.639	.021	1	.885
[StObraz=7]	.022	.8277	-1.601	1.644	.001	1	.979
[StObraz=8]	0[a]						
[SocPom=0]	.682	.4155	-.132	1.497	2.698	1	.100
[SocPom=1]	0[a]						
[UrbRur=0]	-1.188	4.2536	-9.525	7.149	.078	1	.780
[UrbRur=1]	0[a]						
God	-.102	.1443	-.385	.181	.499	1	.480
ADC	-.201	.3109	-.811	.408	.418	1	.518
PrMesPr	**.00016**	**7.2633E-05**	**1.877E-05**	**.000**	**4.921**	**1**	**.027**
PrMesCig	**.002**	**.0008**	**.000**	**.003**	**4.625**	**1**	**.032**
distKM	**-.013**	**.0052**	**-.023**	**-.002**	**5.954**	**1**	**.015**
PrMesPr * PrMesCig	**-5.116E-08**	**2.0613E-08**	**-9.156E-08**	**-1.076E-08**	**6.161**	**1**	**.013**
God * ADC	.005	.0074	-.010	.019	.391	1	.532
[UrbRur=0] * ADC	.055	.2177	-.371	.482	.065	1	.799
[UrbRur=1] * ADC	0[a]						
(Scale)	1[b]						

Tabela 12b. SPSS pregled svih parametara modela (nastavak)

Parameter Estimates			
		95% Wald Confidence Interval for Exp(B)	
Parameter	Exp(B)	Lower	Upper
(Intercept)	.093	1.772E-07	49241.042
[Pol=0]	.794	.415	1.520
[Pol=1]	1		
[StObraz=1]	1.053	.226	4.902
[StObraz=2]	1.936	.490	7.658
[StObraz=3]	1.334	.336	5.301
[StObraz=4]	1.177	.257	5.390
[StObraz=5]	2.071	.503	8.533
[StObraz=6]	1.119	.243	5.147
[StObraz=7]	1.022	.202	5.175
[StObraz=8]	1		
[SocPom=0]	1.979	.876	4.467
[SocPom=1]	1		
[UrbRur=0]	.305	7.298E-05	1272.243
[UrbRur=1]	1		
God	.903	.681	1.198
ADC	.818	.445	1.504
PrMesPr	**1.00016**	**1.000**	**1.000**
PrMesCig	**1.0017**	**1.000**	**1.003**
distKM	**.987**	**.977**	**.998**
PrMesPr * PrMesCig	**.9999999**	**1.000**	**1.000**
God * ADC	1.005	.990	1.019
[UrbRur=0] * ADC	1.057	.690	1.619
[UrbRur=1] * ADC	1		
(Scale)			

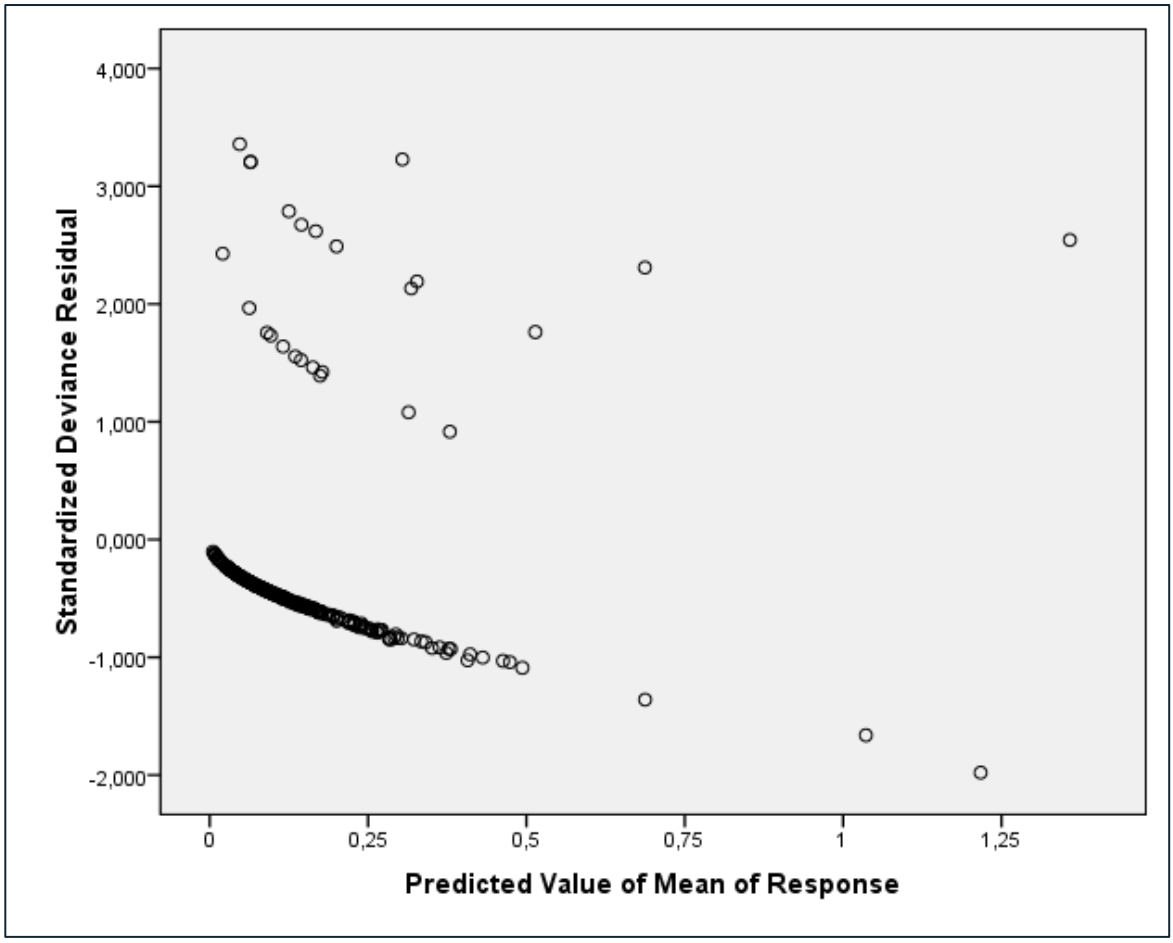
Dependent Variable: AkcMarkica
Model: (Intercept), Pol, StObraz, SocPom, UrbRur, God, ADC, PrMesPr, PrMesCig, distKM, PrMesPr * PrMesCig, God * ADC, UrbRur * ADC

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Takođe, za neformalnu i intuitivnu proveru modela prikazaćemo na grafiku odstupanje reziduala naspram očekivanih linearnih predviđanja.

Slika 4. SPSS grafik odstupanja reziduala prema očekivanim linearnim predviđanjima



Sa grafika na slici 4. možemo da vidimo da su podaci centrirani, jer ne izlaze van intervala $(-3.3, +3.3)$ vrednosti odstupanja reziduala, što znači da su dobro grupisani.

S obzirom da i nakon detaljne analize parametara modela na zavisnu promenljivu značajno utiču prethodno uočene četiri nezavisne promenljive, ostali faktori ne predstavljaju značajne komponente modela.

Dakle, iz modela smo zaključili da je udaljenost posmatranog objekta od granice značajna promenljiva, sa koeficijentom regresije -0.013, pa dobijamo da je eksponencijalni koeficijent $\exp(-0.013) = 0.987$. Kako je za ovaj parametar ocenjena vrednost eksponencijalnog koeficijenta u intervalu $(0,1)$, to znači da su parametar i zavisna promenljiva u inverznom odnosu. Na osnovu toga, možemo da zaključimo da sa

svakim kilometrom bliže granici (tj. sa smanjenjem razdaljine radnje od granice), stopa konzumacije neoporezovanih paklica cigareta raste za $1 - 0.987 = 1.26\%$.

Dalje, sa svakom jedinicom povećanja prosečne mesečne potrošnje na cigarete, stopa konzumacije neoporezovanih paklica cigareta raste za 0.17%, jer su vrednosti koeficijenta regresije i njegove eksponencijalne vrednosti u modelu jednake 0.002 i 1.0017, respektivno. Kako povećanje mesečne potrošnje na cigarete među potrošačima može da bude uzrokovano većom konzumacijom ili kupovinom skupljeg proizvoda, uz modifikaciju upitnika može se proveriti da li ovaj porast konzumacije zapravo predstavlja sliku potrošača koji puši više od proseka populacije, i pri tome kupuje što jeftinije dostupne cigarete.

Parametri modela koji predstavljaju prosečni mesečni prihod i interakciju između prosečnih mesečnih prihoda i prosečne mesečne potrošnje na cigarete (odnosno, dostupnost cigareta potrošaču) imaju koeficijente regresije 0.00016 i -0.0000001, respektivno. Sa povećanjem prosečnih mesečnih prihoda povećava se i konzumacija neoporezovanih paklica cigareta i to za 0.016% po jedinici plaćanja, što je u ovom slučaju dinar, dok povećanje dostupnosti cigareta potrošaču uzrokuje smanjenje vrednosti posmatrane zavisne promenljive -0.00001% .

VII. Zaključak

U ovom radu dat je pregled konstrukcije uopštenih linearnih modela, kao pogodne generalizacije regresionih modela, pri čemu podaci imaju raspodelu iz eksponencijalne familije raspodela. Nakon upoznavanja sa opštim karakteristikama uopštenih linearnih modela i njihovih tipova, posebno je obrađena Poasonova regresija, gde je pored pregleda osobina Poasonove slučajne promenljive data i metodologija modeliranja i analize podataka ovom regresijom.

Poasonova regresija je dobar izbor u slučaju kada su podaci prebrojivi, kao na primer što je broj događaja u nekom ograničenom vremenskom intervalu, pri čemu su događaji međusobno nezavisni. Kako ovaj oblik regresije predstavlja dobar alat za obradu i analizu, modeliranje Poasonovom regresijom dostupno je u većini softverskih paketa za statističku obradu podataka. Prilikom modeliranja podataka Poasonovom regresijom potrebno je obratiti posebnu pažnju na moguću preraspršenost podataka ili prekoračenje disperzije. Kao što smo videli, preraspršenost predstavlja prekoračenje koje potiče iz toga kako je definisana stohastička komponenta modela, pri čemu je sistematička struktura modela tačna. U slučaju preraspršenosti podataka, mogu se koristiti neki od modela koji su razvijeni za ovakve podatke, kao što su na primer Kvazi-Poasonov iili Negativni Binomni model, čiji je teorijski pristup objašnjen u poglavlju IV. 6.

Na kraju rada dat je primer konstrukcije modela Poasonove regresije o konzumiranju neoporezovanih duvanskih proizvoda u zavisnosti od nekoliko promenljivih faktora, pri čemu su analizirani ocenjeni parametri modela, kao i slaganje modela sa podacima u statističkom programu SPSS.

VIII. Dodatak

Klasična centralna granična teorema i dokaz:

Teorema: Ako su $X_1, X_2, \ldots$ nezavisne slučajne promenljive sa istom raspodelom i konačnom disperzijom $D(X_k) = \sigma^2, k = 1, 2, \ldots$ onda važi

$$P\left\{\frac{\sum_{k=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}} < x\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt\,, \qquad n \to \infty.$$

Dokaz: Označimo sa $E(X_k) = a, k = 1, 2, \ldots$ Tada je

$$X_k^* = \frac{X_k - E(X_k)}{\sqrt{D(X_k)}} = \frac{X_k - a}{\sigma}, \qquad k = 1, 2, \ldots.$$

Karakteristična funkcija[8] za X_k^* je

$$f_{X_k^*}(t) = 1 + \frac{tiE(X_k^*)}{1!} + \frac{t^2 i^2 E((X_k^*)^2)}{2!} + o(t^2), \qquad k = 1, 2, \ldots,$$

Znamo da je $E(X_k^*) = 0$, jer X_k^* ima $\mathcal{N}(0,1)$ raspodelu. Tada za $k = 1, 2, \ldots$, važi

$$D(X_k^*) = E((X_k^*)^2) - E^2(X_k^*) = 1,$$

pa sledi da je $E((X_k^*)^2) = 1$. Kada uprostimo jednačinu za karakterističnu funkciju, dobijamo

$$f_{X_k^*}(t) = 1 - \frac{t^2}{2} + o(t^2), \qquad k = 1, 2, \ldots$$

Dalje je

$$\frac{\sum_{k=1}^{n} X_k - E(\sum_{k=1}^{n} X_k)}{\sqrt{D(\sum_{k=1}^{n} X_k)}} = \frac{\sum_{k=1}^{n} X_k - na}{\sqrt{n\sigma^2}}$$

[8] Karakteristična funkcija slučajne promenljive X, u oznaci $f_X(t)$, je funkcija f_X: $\mathbb{R} \mapsto \mathbb{C}$, data sa
$$f_X(t) = E(e^{itX}), \qquad t \in \mathbb{R}, i^2 = -1.$$
Svakoj funkciji raspodele odgovara tačno jedna karakteristična funkcija.

$$= \frac{X_1 + \ldots + X_n - na}{\sqrt{n\sigma^2}} = \frac{1}{\sqrt{n}}\left(\frac{X_1 - a}{\sigma} + \cdots + \frac{X_n - a}{\sigma}\right)$$
$$= \frac{1}{\sqrt{n}}(X_1^* + \cdots + X_n^*).$$

Kako su $X_1, \ldots, X_n$ nezavisne promenljive, karakteristična funkcija slučajne promenljive

$$\frac{\sum_{k=1}^n X_k - E(\sum_{k=1}^n X_k)}{\sqrt{D(\sum_{k=1}^n X_k)}} = \frac{1}{\sqrt{n}}(X_1^* + \cdots + X_n^*)$$

je

$$f_{\frac{1}{\sqrt{n}}\sum_{k=1}^n X_k^*}(t) = f_{\sum_{k=1}^n X_k^*}\left(\frac{t}{\sqrt{n}}\right) = \prod_{k=1}^n f_{X_k^*}\left(\frac{t}{\sqrt{n}}\right) = \left(f_{X_k^*}\left(\frac{t}{\sqrt{n}}\right)\right)^n$$
$$= \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to e^{-\frac{t^2}{2}}, \qquad n \to \infty.$$

Dakle, karakteristična funkcija slučajne promenljive

$$\frac{\sum_{k=1}^n X_k - E(\sum_{k=1}^n X_k)}{\sqrt{D(\sum_{k=1}^n X_k)}}$$

konvergira ka karakterističnoj funkciji slučajne promenljive sa normalnom $\mathcal{N}(0,1)$ raspodelom, pa slučajna promenljiva $\frac{\sum_{k=1}^n X_k - E(\sum_{k=1}^n X_k)}{\sqrt{D(\sum_{k=1}^n X_k)}}$ konvergira u raspodeli[9] ka slučajnoj promenljivoj sa normalnom $\mathcal{N}(0,1)$ raspodelom, kada $n \to \infty$, što je i trebalo pokazati.

■

[9] Niz slučajnih promenljivih $X_1, X_2, \ldots$ konvergira u raspodeli ka slučajnoj promenljivoj X, kada $n \to \infty$, ako niz odgovarajućih funkcija raspodele $F_{X_1}(x), F_{X_2}(x), \ldots$ kompletno konvergira ka funkciji raspodele slučajne promenljive X, $F_X(x)$ (što znači da konvergira za svako $x \in \mathbb{R} \cup \{-\infty, \infty\}$ za koje je $F_X(x)$ neprekidna funkcija).

Literatura:

1) Abedijan, I., Van der Merwe, R., Wilkins, N., Jha, P. (1998) *'The Economics of Tobacco Control – Towards an optimal policy mix'*, Applied Fiscal Research Center (AFReC), University of Cape Town
2) Berk, R. i MacDonald, J. M. (2008) *'Overdispersion and Poisson Regression'*, published online: Springer Science+Business Media, LLC
3) Chatterjee, S. i Simonoff, J. S. (2013) *'Handbook of Regression Analysis'*, Wiley
4) Dobson, A. J. (2002) *'An Introduction to Generalized Linear Models'*, second edition, Chapman & Hall/CRC
5) Feller, W. (1968) *'An Introduction to Probability Theory and Its Applications'*, third edition, John Wiley & Sons, Inc.
6) Fox, J. (2008) *'Applied Regression Analysis and General Linear Models',* second edition, SAGE Publications, Inc
7) Gschlossl, S. i Czado, C. (2006) *'Modelling count data with overdispersion and spatial effects',*Springer-Verlag
8) http://en.wikipedia.org/wiki/Normal_distribution
9) http://en.wikipedia.org/wiki/Poisson_distribution
10) http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/ebooks/html/spm/spmhtmlnode27.html#eq_firstll
11) http://www-01.ibm.com/support/knowledgecenter
12) http://www4.stat.ncsu.edu/~hzhang/st522/08Chap7.pdf
13) http://homepages.math.uic.edu/~rgmartin/Teaching/Stat411/Notes/411notes.pdf
14) Lahiri, S. i Saha, S. *'Regression and Generalized Linear Models'*, Department of Statistics, University of Florida
15) Larget, B. (2008) *'Poisson regression'*, Lecture Notes – Department of Botany and of Statistics, University of Wisconsin – Madison
16) Lindsey, J. K. (2000) *'Applying Generalized Linear Models'*, Springer
17) Lozanov-Crvenković, Z. *'Beleške sa predavanja iz Statistike',* Univerzitet u Novom Sadu, Prirodno-matematički fakultet
18) Mouatiassim, Y. i Ezzahid, E. H. (2012) *'Poisson regression and Zero-inflated Poisson regression: application to private health insurance',* Springer
19) Oelerich, A. i Poddig, T. (2004) *'Modified Wald statistics for generalized linear models',* Physica-Verlag

20) Rajter-Ćirić, D. (2008) *'Verovatnoća',* Univerzitet u Novom Sadu, Prirodno-matematički fakultet
21) Rodríguez, G. (2007) *'Lecture Notes on Generalized Linear Models',* dostupno na sajtu http://data.princeton.edu/wws509/notes/
22) Santos-Silva, J. M. C. i Tenreyro, S. (2009) *'On the Existence of the Maximum Likelihood Estimates for Poisson Regression ',* Centre for Economic Performance, London School of Economics and Political Science
23) Soriano, A. G. *'Excise duties and smuggling – The need of joint solutions to a global threat',* University of Valencia
24) Turner, H. (2008) *'Introduction to Generalized Linear Models',* ESRC National Centar for Research Methods, UK and Department of Statistics, University of Warwick, UK
25) Ver Hoef, Jay M. i Boveng, Peter L. (2007) *'Quasi-Poisson vs. Negative Binomial Regression: How should we model overdispersed count data?',* Publications, Agencies and Staff of the U.S Department of Commerce
26) Zuro, Alain F., Ieno, Elena M. i Smith, Graham M. (2007) *'Analysing Ecological Data',* Springer Science + Business Media, LLC

Biografija

Sanja Bojović je rođena 28. okt 1987. godine u Novom Sadu. Završila je Osnovnu školu “Svetozar Marković Toza” u Novom Sadu i uporedo Osnovnu muzičku školu “Josip Slavenski”. Pohađala je gimnaziju “Svetozar Marković”, takođe u Novom Sadu, a zatim 2006. godine upisala je osnovne studije na Prirodno – matematičkom fakultetu u Novom Sadu, smer *Matematika finansija*. Osnovne studije završava u predviđenom roku sa prosečnom ocenom 9.10. Odmah nakon završenih osnovnih studija upisuje master studije na istom fakultetu, smer *Primenjena matematika*. Od januara 2012. godine je zaposlena u kompaniji Japan Tobacco International u Beogradu.

Položila je sve ispite predviđene nastavnim planom i programom za master studije i time stekla uslov za odbranu master rada.
…….

Novi Sad, *Jun 2014.*

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

AU

Mentor: dr. Zagorka Lozanov-Crvenković

MN

Naslov rada: Poasonova regresija i primene

NR

Jezik publikacije: *Srpski (latinica)*

JP

Jezik izvoda: srpski/engleski

JI

Zemlja publikovanja: Republika Srbija

ZP

Uže geografsko područje: Vojvodina

UGP

Godina: 2014.

GO

Izdavač: Autorski reprint

IZ

Mesto i adresa: Prirodno-matematički fakultet

MA Departman za matematiku i informatiku

Trg Dositeja Obradovića 4, 21000 Novi Sad

Fizički opis rada: (8/63/26/12/4/0/0)

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

FO

Naučna oblast: Matematika

NO

Naučna disciplina: Statistika

ND

Ključne reči: Uopšteni linearni modeli, Poasonova regresija, Eksponencijalna familija raspodela, Metod iterativnih težinskih najmanjih kvadrata

PO/UDK:

Čuva se: Biblioteka departmana za matematiku i informatiku,

ČU Prirodno-matematički fakultet,

Trg Dositeja Obradovića 4, 21000 Novi Sad

Važna napomena: nema

VN

Izvod: U master radu smo prikazali uopštene linearne modele, koji predstavljaju ekstenziju standardnih linearnih modela, jer dopuštaju izbor raspodele podataka iz eksponencijalne familije raspodela, što rešava problem transformacije podataka u normalno raspodeljene. Posebno, teorijski je obrađena Poasonova regresija kroz 4 faze statističkog modeliranja. Ona je pogodna za modeliranje pojava koje rezultuju prebrojivim podacima. Na kraju rada je dat praktični primer istraživanja uticaja različitih faktora na konzumaciju neoporezovanih duvanskih proizvoda.

IZ

Datum prihvatanja teme od strane NN veća: 26.02.2014.

DP

Datum odbrane: 2014.

DO

Članovi komisije:

KO

Predsednik: dr Ljiljana Gajić, redovni profesor

Prirodno-matematički fakultet, Novi Sad

Član: dr Zagorka Lozanov-Crvenković, redovni profesor,

Prirodno-matematički fakultet, Novi Sad

Član: dr Ivana Štajner-Papuga, vanredni profesor,

Prirodno-matematički fakultet, Novi Sad

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
KEY WORDS DOCUMENTATION

Accession number:

ANO

Identification umber:

INO

Document type: Monograph type

DT

Type of record: Printed text

TR

Contents Code: Master thesis

CC

Author: Sanja Bojović

AU

Mentor: Zagorka Lozanov-Crvenković Ph.D.

MN

Title: Poisson regression and applications

XI

Language of text: Serbian (latin)

LT

Language of abstract: English/Serbian

LA

Country of publication: Serbia

CP

Locality of publication: Vojvodina

LP

Publication year: 2014.

PY

Publisher: Author's reprint

PU

Publ. place: Faculty of Natural Sciences and Mathematics

PP Department of Mathematics and Computer Sciences
Trg Dositeja Obradovića 4, 21000 Novi Sad

Physical description: (8/63/26/12/4/0/0)

PD

Scientific field: Mathematics

SF

Scientific discipline: Statistics

Key words: Generalized linear models, Poisson regression, Exponential family of distributions, Iterative weighted least square method

UC:

Holding data: Library of the Department of Mathematics and Computer Sciences, Faculty of Natural Sciences, Trg Dositeja Obradovića 4, 21000 Novi Sad

HD

Note: none

Abstract: Master Thesis consists of overview on Generalized Linear Models (GLM), which are extension of standard linear models. GLMs allow the choice of distribution from the exponential family, which solves the transformation problems of non-normally distributed data into normally distributed. Specially, theoretical background is given for Poisson regression through four phases of statistical modeling. Poisson regression is suitable for modeling of count data. Lastly, application is provided on the example of trends in non-duty paid tobacco products consumption based on several different potentially influencing factors.

AB

Accepted by the Scientific Board on: 26th of February 2014.

Defended:

Thesis defend board: Ljiljana Gajić Ph.D., Full professor, Faculty of Natural Sciences and Mathematics, Novi Sad

Member: Zagorka Lozanov-Crvenković Ph.D., Full professor, Faculty of Natural Sciences and Mathematics, Novi Sad

Member: Ivana Štajner-Papuga Ph.D., Assistant professor, Faculty of Natural Sciences and Mathematics, Novi Sad