Regresijska analiza za brojeće podatke

Matošin, Matea

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Matea Matošin

REGRESIJSKA ANALIZA ZA BROJEĆE PODATKE

Diplomski rad

Voditelj rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, veljača, 2025.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Ovaj rad posvećujem svojim roditeljima koji su mi pružili bezuvjetnu podršku i ljubav u svim trenutcima studiranja. Iznimno sam zahvalna seki i teti za svaki skuhani ručak. Hvala babi i didi za svaku gurnutu kunu u džep.
Zahvaljujem se svojim prijateljima na svakom zajedničkom učenju, zajedničkom osmijehu i zajedničkim suzama. Posebno se zahvaljujem svom dečku za svaku riječ ohrabrenja.

Sadržaj

Uvod

God made the integers, all the rest is the work of man.
- Leopold Kronecker

Regresijska analiza za brojeće podatke predstavlja ključni alat u statističkom modeliranju i analizi podataka u kojima je zavisna varijabla oblika broja događaja ili učestalosti pojavljivanja. Takvi podaci često se susreću u različitim područjima poput biostatistike, ekonomije, sociologije i epidemiologije. Glavna svrha regresijske analize u ovakvim situacijama je modelirati zavisnost između brojeće zavisne varijable i jedne ili više nezavisnih varijabli, koristeći statističke modele koji adekvatno opisuju specifičnu prirodu podataka.

Tradicionalni linearni regresijski modeli nisu prikladni za brojeće podatke zbog svojih pretpostavki o normalnosti reziduala i konstantnoj varijanci, koje nisu zadovoljene kada je zavisna varijabla cijeli broj i diskretna. Umjesto toga, koriste se generalizirani linearni modeli (GLM), koji omogućuju fleksibilnost u odabiru distribucije zavisne varijable i funckija veze. Među najvažnijim GLM-ovima za brojeće podatke ističu se Poissonov i negativni binomni model.

Prvi dio rada usmjeren je na predstavljanje teorijskog okvira generaliziranih linearnih modela. Fokus je na uvodu u regresijske modele, definiciji GLM-a, eksponencijalnoj familiji distribucija, konceptu funkcija veze te osnovnim metodama procjene parametara. Ovaj teorijski okvir pruža temelj za razumijevanje specifičnih modela prilagođenih za brojeće podatke.

Drugo poglavlje bavi se Poissonovim generaliziranim linearnim modelom. Poissonova regresija je standardni pristup za modeliranje brojećih podataka kada su zadovoljene pretpostavke o ekvidisperziji (jednakosti varijance i srednje vrijednosti). U poglavlju se razmatraju ključne komponente modela, uključujući procjenu parametara, dijagnostika modela te identifikaciju potencijalnih problema poput prekomjerne disperzije - najava idućeg poglavlja. Treće poglavlje posvećeno je negativnoj binomnoj regresiji, koja predstavlja prirodno proširenje Poissonovog modela za slučajeve kada varijanca podataka premašuje očekivanu vrijednost (prekomjerna disperzija). U poglavlju se obrađuju osnovne karakteristike negativne binomne distribucije i metode procjene parametara.

Četvrto poglavlje fokusira se na problem prekomjerne disperzije koja se javlja kada varijanca premašuje srednju vrijednost, što može narušiti validnost zaključaka Poissonovih modela. Naposljetku zadnje poglavlje obrađuje praktične primjere sa stvarnim podatcima na razmatranim modelima.

Cilj ovog rada je predstaviti teorijski okvir i praktične alate za regresijsku analizu brojećih podataka, uz isticanje ključnih problema i rješenja kroz dijagnostiku i prilagodbu modela. Na kraju, rad nudi uvid u prednosti i izazove korištenja GLM-ova u analizi brojećih podataka.

Poglavlje 1

Generalizirani linearni modeli

U ovom poglavlju će biti objašnjene regresijske metode i njihovi ciljevi. Najjednostavniji oblik regresije je linearna regresija, no zbog određenih prepostavki na linarnu regresiju i mogućnost fleksibilnijeg pristupa uvest će se korištenje generaliziranih linearnih modela koji omogućuju rješavanje mnogih ograničenja linearne regresije. Uvođenjem GLM-a omogućuje se analiza složenijih podataka i problema koji nadilaze mogućnosti klasične linearne regresije, a pritom se zadržava interpretabilnost i jednostavnost primjene modela.

1.1 Uvod u regresijske modele

Regresijski modeli predstavljaju skup statističkih metoda namijenjenih modeliranju odnosa između zavisne varijable i jedne ili više nezavisnih varijabli. Cilj ovih metoda je opisati odnos između zavisnih i nezavisnih varijabli, procijeniti parametre modela, provjeriti koliko je model adekvatan, odnosno koliko dobro opisuje podatke, te interpretirati rezultate i donijeti zaključke. Osnovni tipovi regresije uključuju linearnu regresiju, višestruku linearnu regresiju, logističku regresiju, polinomijalnu regresiju, Poissonovu regresiju i mnoge druge. Regresijski modeli su često jednostavni, mogu se proširiti kako bi opisali složenije odnose među varijablama, što ih čini izrazito korisnima u praktičnoj primjeni. Međutim, u stvarnim situacijama podaci često ne zadovoljavaju idealne pretpostavke — greške nisu normalno distribuirane, a varijable nisu uvijek nezavisne, zbog čega postoji potreba za razvojem modela koji se mogu prilagoditi takvim izazovima i omogućiti točnije i pouzdanije zaključke.

1.2 Definicija i osnovna ideja

Linearni model (LM) je najjednostavniji model gdje se pretpostavlja da je odnos između zavisne varijable Y i nezavisne varijable X linearan.
Generalizirani linearni model (GLM) proširuje sposobnost linearnih modela kako bi se prilagodio podatcima.

Kod LM-a pretpostavljamo da zavisna varijabla slijedi normalnu distribuciju s konstantnom varijancom, dok je kod GLM-a dopušteno da zavisna varijabla slijedi bilo koju distribuciju iz eksponencijalne familije - Normalnu, eksponencijalnu, Poissonovu, binomnu, beta, gamma i mnoge druge. U ovom radu će biti značajna Poissonova i negativna binomna distribucija.
Kod LM-a očekivana vrijednost odziva (zavisne varijable) se modelira kao linearna kombinacija prediktora. GLM koristi *funkciju veze* za modeliranje nelinearnih odnosa između prediktora i srednje vrijednosti distribucije odziva.

Prednosti GLM-ova u usporedbi s transformiranjem podataka

Tradicionalni način modeliranja podataka transformira Y tako da približno slijedi normalnu raspodjelu s konstantnom varijancom; zatim se primjenjuje obična linearna regresija. Nasuprot tome, kod GLM-ova, izbor funkcije veze odvojen je od izbora slučajne komponente. Ako je funkcija veze korisna u smislu da je linearni model za prediktore izvediv za tu funkciju, nije potrebno da stabilizira varijancu ili osigurava normalnost. To je zato što postupak prilagodbe maksimizira vjerojatnost za izbor distribucije za Y, a taj izbor nije ograničen na normalnu raspodjelu.

Neka g označava funkciju, poput logaritamske funkcije, koja može biti funkcija veze u GLM pristupu ili transformacijska funkcija u transformiranom podatkovnom pristupu. Prednost GLM formulacije je u tome što parametri modela opisuju $g[\mathbb{E}(Y)]$, umjesto $\mathbb{E}[g(Y)]$ kao u transformiranom pristupu podacima. Kod GLM pristupa, ti parametri također opisuju učinke objašnjavajućih varijabli na $\mathbb{E}(Y)$, nakon primjene inverzne funkcije za g. Takvi su učinci relevantniji od učinaka objašnjavajućih varijabli na $\mathbb{E}[g(Y)]$.

GLM

Pretpostavka o postojanju linearne veze između očekivanja zavisne varijable i kovarijata je jedna od osnovnih ideja linearnih modela. Zapišimo to ovako:

$$E(Y_i) = \sum_{j=1}^{k} \beta_j x_{ij},$$

gdje je k broj kovarijata. Pretpostavljamo i normalnost

$$Y_i \sim N\left(\sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2\right).$$

Kod generaliziranih linearnih modela pretpostavljamo

$$g\left(E(Y_i)\right) = \sum_{j=1}^{k} \beta_j x_{ij}$$

i

$$E(Y_i) = g^{-1}\left(\sum_{j=1}^{k} \beta_j x_{ij}\right),$$

gdje je:

- g^{-1} inverz funkcije veze g,
- $\sum_{j=1}^{k} \beta_j x_{ij}$ linearni prediktor (Detaljnije u [2]).

1.3 Eksponencijalna familija distribucija

Teorija generaliziranih linearnih modela omogućuje istovremeno modeliranje ovisnosti zavisne varijable na numeričke i kategorijalne nezavisne varijable. Teorija pretpostavlja da slučajna zavisna varijabla slijedi razdiobu iz eksponencijalne familije, no u praksi je ova restrikcija prihvatljiva jer eksponencijalna familija obuhvaća najčešće korištene razdiobe, poput normalne, Poissonove i binomne razdiobe [6].

Funkcija gustoće eksponencijalnih

Kažemo da slučajna varijabla Y pripada nekoj eksponencijalnoj familiji ako joj gustoća ima sljedeći oblik:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{1.1}$$

gdje je:

- $a(\phi)$ funkcija disperzije, gdje je ϕ parametar disperzije ili raspršenja,
- b funkcija koja je dvaput neprekidno diferencijabilna i b' je invertibilna,
- c funkcija koju možemo ignorirati jer nema utjecaja u procesu promjene parametra GLMa i
- θ prirodni parametar.

Poissonova distribucija kao eksponencijalna familija

Uz pomoć funkcije gustoće pokažimo da je Poissonova distribucija dio eksponencijalne familije.

Funkcija gustoće Poisonove distribucije je:

$$f_Y(y;\mu) = \frac{\mu^y e^{-\mu}}{y!} = \exp\left[y\log\mu - \mu - \log y!\right] \quad y \in \mathbb{R}.$$

Želimo oblik eksponencijalne familije 1.1.

Odmah vidimo da je:

$$\begin{aligned} \theta &= \log\mu, \\ a(\phi) &= \phi = 1, \\ b(\theta) &= e^\theta = \mu, \\ c(y,\phi) &= -\log y!. \end{aligned}$$

Dakle, Poissonova distribucija pripada eksponencijalnoj familiji.

Negativna binomna distribucija kao eksponencijalna familija

Uz pomoć funkcije gustoće pokažimo da je negativna binomna distribucija dio eksponencijalne familije. Funkcija gustoće negativne binomne distribucije je:

$$f_Y(y; r, p) = \binom{y + r - 1}{y} p^r (1 - p)^y, \quad y = 0, 1, 2, \ldots \tag{1.2}$$

Želimo oblik eksponencijalne familije 1.1.
Uzmimo prirodni logaritam:

$$\begin{aligned} \ln f_Y(y; r, p) &= \ln \binom{y + r - 1}{y} + r \ln p + y \ln(1 - p) \\ &= y \ln(1 - p) + r \ln p + \ln \binom{y + r - 1}{y}. \end{aligned}$$

Definirajmo prirodni parametar θ:

$$\theta = \ln (1 - p).$$

Tada vrijedi:

$$e^\theta = 1 - p, \quad p = 1 - e^\theta \quad i \quad \ln(p) = \ln(1 - e^\theta).$$

Supstitucijom dobijemo:

$$\ln f_Y(y; r, \theta) = y\theta - (-r \ln(1 - e^\theta)) + \ln \binom{y + r - 1}{y},$$

gdje je:

$$\begin{aligned} \theta &= \ln (1 - p), \\ a(\phi) &= \phi = 1, \\ b(\theta) &= -r \ln(1 - e^\theta), \\ c(y, \phi) &= \ln \binom{y + r - 1}{y} \\ &= \ln \frac{\Gamma(r)\, y!}{\Gamma(y + r)} = \ln \Gamma(y + r) - \ln y! - \ln \Gamma(r)^1. \end{aligned}$$

Dakle, negativna binomna distribucija pripada eksponencijalnoj familiji.

[1] Za pozitivne cijele brojeve n vrijedi $\Gamma(n) = (n - 1)!$.

Očekivanje i varijanca eksponencijalne familije

Pokažimo da općenito očekivanje od Y ovisi isključivo o parametru θ i da varijanca ovisi o parametru θ i ϕ.
Funkcija log-vjerodostojnosti unutar neke eksponencijalne familije je $l(y;\theta,\phi) = \log(f(y;\theta,\phi))$.
Dokaz iduća dva rezultata se nalazi u 6.

$$\mathbb{E}\left[\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right] = 0 \quad \text{i} \quad \mathbb{E}\left[\frac{\partial^2 l(Y;\theta,\phi)}{\partial\theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right)^2\right] = 0. \tag{1.3}$$

Za log-vjerodostojnost eksponencijalnih familija iz 1.1vrijedi da je prva derivacija jednaka:

$$\frac{\partial l}{\partial\theta} = \frac{y - b'(\theta)}{a(\phi)}.$$

Stoga iz prve jednakosti u 1.3 slijedi

$$\mathbb{E}[Y] = b'(\theta).$$

Ponovnim diferenciranjem log - vjerodostojnosti dobijemo:

$$\frac{\partial^2 l}{\partial\theta^2} = -\frac{b''(\theta)}{a(\phi)},$$

također iz 1.3 imamo

$$\mathbb{E}\left[\frac{\partial^2 l(Y;\theta,\phi)}{\partial\theta^2}\right] = -\mathbb{E}\left[\left(\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right)^2\right].$$

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] \\
&= \mathbb{E}\left[(Y - b'(\theta))^2\right] \\
&= a(\phi)^2\mathbb{E}\left[\frac{(Y - b'(\theta))^2}{a(\phi)^2}\right] \\
&= a(\phi)^2\mathbb{E}\left[\left(\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right)^2\right] \\
&= -a(\phi)^2\mathbb{E}\left[\frac{\partial^2 l(Y;\theta,\phi)}{\partial\theta^2}\right] \\
&= a(\phi)b''(\theta).
\end{aligned}$$

Slijedi da je:

$$\mathbb{E}[Y] = b'(\theta), \tag{1.4}$$

$$\mathrm{Var}(Y) = a(\phi)b''(\theta). \tag{1.5}$$

Pokazali smo da očekivanje od Y ovisi samo o parametru θ i ne ovisi o parametru ϕ, dok varijanca ovisi o oba parametra θ i ϕ.

Poissonova distribucija

Očekivanje i varijanca od Y

$$\mathbb{E}[Y] = b'(\theta) = e^{\theta} = \mu, \tag{1.6}$$

$$\text{Var[Y]} = a(\phi)\, b''(\theta) = e^{\theta} = \mu. \tag{1.7}$$

Negativna binomna distribucija

$$\begin{aligned} \mathbb{E}[Y] &= \frac{\partial}{\partial\theta}\left(-r\log(1-e^{\theta})\right) \\ &= -r\frac{-e^{\theta}}{1-e^{\theta}} \\ &= \frac{re^{\theta}}{1-e^{\theta}} \\ &= \frac{r(1-p)}{p}. \end{aligned} \tag{1.8}$$

$$\begin{aligned} \text{Var}[Y] &= a(\phi)\frac{\partial^2}{\partial\theta^2}b(\theta) \\ &= \frac{\partial}{\partial\theta}\left(r\frac{e^{\theta}}{1-e^{\theta}}\right) \\ &= r\frac{e^{\theta}}{(1-e^{\theta})^2} \\ &= \frac{r(1-p)}{p^2}. \end{aligned} \tag{1.9}$$

1.4 Funkcija veze

U ovom odjeljku objašnjavamo važnost funkcija veze u povezivanju očekivane vrijednosti odgovora s prediktorskom funkcijom. Funkcija veze ima ključnu ulogu u generaliziranim linearnim modelima (GLM-ovima), omogućujući prilagodbu različitim vrstama podataka korištenjem različitih distribucija i odgovarajućih funkcija veze.

Uloga funkcije veze

Funkcija veze $g(\cdot)$ povezuje očekivanu vrijednost odgovora $\mathbb{E}[Y]$ s linearnom prediktorskom funkcijom η, koja je linearna kombinacija prediktora:

$$g(\mathbb{E}[Y]) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

gdje je:

- $\mathbb{E}[Y]$: očekivana vrijednost zavisne varijable Y,
- $g(\cdot)$: funkcija veze,
- $\boldsymbol{\eta}$: linearna prediktorska funkcija,
- $\mathbf{X}$: matrica prediktora,
- $\boldsymbol{\beta}$: vektor parametara.

Jedna od pretpostavki GLM-a je da je očekivana vrijednost zavisne varijable Y glatka i invertibilna funkcija prediktora η. Ta veza se može izraziti kao:

$$\theta = h(\eta) \quad \text{i} \quad \mu = \mathbb{E}Y = b'(\theta) = b'(h(\eta)).$$

Iz toga slijedi da je

$$\mu = g^{-1}(\eta) \quad \text{odnosno} \quad \eta = g(\mu).$$

Funkcija veze g je definirana kao $g = h^{-1} \circ b'^{-1}$.

Prirodni parametar θ je također glatka funkcija od η:

$$\theta = b'^{-1}(\mu) = h(\eta).$$

Ako je funkcija h identiteta, tj $h \equiv \text{id}$, tada vrijedi $\theta \equiv \eta$, pa je funkcija veze

$$g = b'^{-1}.$$

Ovisno o prirodi zavisne varijable i specifičnostima problema, biramo odgovarajuću funkciju veze.

Najčešće korištene funkcije veze su funkcija identiteta, logaritamska funkcija, logistička (logit) funkcija, probit funkcija, recipročna funkcija, kao i mnoge druge koje se primjenjuju ovisno o prirodi podataka i zahtjevima modela.

Poissonova funkcija veze

Logaritamska funkcija ključna je za Poissonovu regresiju i druge slučajeve kada je zavisna varijabla diskretna, poput broja događaja.

Pokažimo da je logaritamska funkcija zaista funkcija veze Poissonove regresije:

$$\begin{aligned} g(\mathbb{E}[Y]) &\longrightarrow \theta, \\ \mathbb{E}[Y] &= \mu = e^{\theta}, \\ \theta &= \log(\mathbb{E}[Y]). \end{aligned}$$

Logaritamska funkcija veze je:

$$g(\mathbb{E}[Y]) = \log(\mathbb{E}[Y]). \tag{1.10}$$

Dakle, logaritamska funkcija $\log(\mu_i)$ linearizira odnos između srednje vrijednosti μ_i i prediktorskih varijabli x_i. Ovo omogućava jednostavnu interpretaciju koeficijenata $\boldsymbol{\beta}$, gdje svaki koeficijent β_j predstavlja promjenu log-srednje vrijednosti za jednu jedinicu promjene u x_j, uz pretpostavku da su ostale varijable konstantne.

Ova karakteristika čini logaritamsku funkciju prirodnom funkcijom veze za Poissonovu regresiju.

Negativna binomna funkcija veze

Pronađimo funkciju veze negativne binomne regresije

$$g(\mathbb{E}[Y]) \to \theta$$

Neka je:

$$\gamma = \mathbb{E}[Y] = \frac{re^{\theta}}{1 - e^{\theta}} = \frac{r}{e^{-\theta} - 1}.$$

Inverz od γ *je* :

$$\gamma^{-1} = \frac{e^{-\theta} - 1}{r},$$

$$\begin{aligned} r\gamma^{-1} &= e^{-\theta} - 1, \\ r\gamma^{-1} + 1 &= e^{-\theta}, \\ \log(r\gamma^{-1} + 1) &= -\theta. \end{aligned}$$

$$\begin{aligned}\theta &= -\log\left(r\gamma^{-1}+1\right)\\ &= log\left(\frac{1}{r\gamma^{-1}+1}\right)\\ &= log\left(\frac{1}{\frac{r}{\mathbb{E}[Y]}+1}\right)\\ &= log\left(\frac{\mathbb{E}[Y]}{r+\mathbb{E}[Y]}\right).\end{aligned}$$

Negativna binomna funkcija veze je:

$$g(\mathbb{E}[Y]) = log\left(\frac{\mathbb{E}[Y]}{r+\mathbb{E}[Y]}\right). \tag{1.11}$$

Funkcije veze u okviru generaliziranih linearnih modela pružaju fleksibilnost pri modeliranju odnosa između zavisne varijable i prediktora. Odabirom odgovarajuće funkcije veze moguće je prilagoditi model specifičnostima problema, čime se osigurava njegova preciznost, konzistentnost i interpretabilnost rezultata.

1.5 Osnovne procjene parametara

Procjena parametara u generaliziranim linearnim modelima može biti zahtjevan zadatak. Najčešće se koristi postupak **maksimalne vjerodostojnosti** (MLE) za procjenu parametara modela. U nastavku je sažet pregled ključnih koraka koji čine postupak procjene parametara.

Maksimalna vjerodostojnost (MLE)

MLE

Cilj metode maksimalne vjerodostojnosti je pronaći vrijednosti parametara koji maksimiziraju funkciju vjerodostojnosti. Funkcija vjerodostojnosti opisuje vjerojatnost promatranih podataka $y_1, y_2, \ldots, y_n$ s obzirom na pretpostavljenu distribuciju i zadane parametre.

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta} = \boldsymbol{\eta}.$$

Za n nezavisnih opažanja, funkcija vjerodostojnosti $L(\beta)$ je:

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

gdje je $f(y_i; \theta, \phi)$ funkcija gustoće za y_i.

Umjesto funkcije vjerodostojnosti, često koristimo **log-vjerodostojnost** $\ell(\boldsymbol{\beta})$, jer logaritam pretvara umnožak u zbroj i pojednostavljuje izračune:

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i; \theta, \phi).$$

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i; \theta, \phi) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right]. \tag{1.12}$$

Maksimalizacija $\ell(\boldsymbol{\beta})$ zahtijeva računanje derivacija. Koristimo lančano pravilo:

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Vrijedi iz 1.4 i 1.5 da je :

$$\begin{aligned}
&\frac{\partial L_i}{\partial \beta_j} = [y_i - b'(\theta_i)]/a(\phi), \\
&\mu_i = b'(\theta_i), \\
&\text{var}(Y_i) = b''(\theta_i) a(\phi).
\end{aligned}$$

Također vrijedi:

$$\begin{aligned}
&\frac{\partial L_i}{\partial \theta_i} = \frac{(y_i - \mu_i)}{a(\phi)}, \\
&\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \text{var}(Y_i)/a(\phi), \\
&\eta_i = \sum_j \beta_j x_{ij}, \\
&\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}, \\
&\eta_i = g(\mu_i)
\end{aligned}$$

Slijedi :

$$\frac{\partial L_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i) x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

Zbrajanjem preko n opažanja dobivamo jednadžbe vjerodostojnosti,

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, \ldots. \tag{1.13}$$

Iako se $\boldsymbol{\beta}$ ne pojavljuje eksplicitno u ovim jednadžbama, ona je prisutna implicitno kroz μ_i, jer vrijedi:

$$\mu_i = g^{-1}\left(\sum_j \beta_j x_{ij}\right).$$

Različite funkcije veze daju različite skupove jednadžbi.

Izračun Poissonovog MLE-a i MLE-a za negativnu binomnu regresiju bit će prikazan u narednim poglavljima.

Devijanca

Jedan od pristupa procjeni adekvatnosti modela je usporedba s općenitijim modelom koji sadrži maksimalan broj parametara koji se mogu procijeniti. Takav model naziva se **zasićeni model** i predstavlja generalizirani linearni model koji koristi istu distribuciju i funkciju veze kao model od interesa.

Za n opažanja $Y_i, i = 1, \ldots, n$, s različitim vrijednostima linearne komponente $\mathbf{x}_i^T\boldsymbol{\beta}$, zasićeni model je definiran s n parametara.

Općenito, neka m označava maksimalni broj parametara koji se može procijeniti. Neka $\boldsymbol{\beta}_{\max}$ označava vektor parametara za zasićeni model, a $\mathbf{b}_{\max}$ označava procjenu maksimalne vjerodostojnosti za $\boldsymbol{\beta}_{\max}$. Funkcija vjerodostojnosti za zasićeni model, evaluirana u $\mathbf{b}_{\max}$, $L(\mathbf{b}_{\max}; \mathbf{y})$, bit će veća od bilo koje druge funkcije vjerodostojnosti za ova opažanja, s istom distribucijom i funkcijom veze, jer pruža najpotpuniji opis podataka.

Neka $L(\mathbf{b}; \mathbf{y})$ označava maksimalnu vrijednost funkcije vjerodostojnosti za model koji nas zanima. Tada je omjer log-vjerodostojnosti definiran kao:

$$\Lambda = \frac{L(\mathbf{b}_{\max}; \mathbf{y})}{L(\mathbf{b}; \mathbf{y})}. \tag{1.14}$$

Za logaritam omjera vjerodostojnosti koristimo:

$$\log \Lambda = l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y}). \tag{1.15}$$

I provodimo sljedeći statistički test:

H_0 : Model od interesa opisuje podatke jednako dobro kao i zasićeni model.

H_1 : Zasićeni model bolje opisuje podatke.

Velike vrijednosti $\log \Lambda$ idu u korist alternativne hipoteze i sugeriraju da model koji nas zanima lošije opisuje podatke u usporedbi sa zasićenim modelom.

$2 \log \Lambda$ ima χ^2 distribuciju sa $n - k$ stupnjeva slobode (Dokaz se nalazi u [6], poglavlje 7.5). Stoga se $2 \log \Lambda$, a ne $\log \Lambda$, češće koristi kao statistika. Ova mjera naziva se **devijanca**.

Poglavlje 2

Poissonov generalizirani linearni model

Poissonova regresija je posebna vrsta GLM-a namijenjena modeliranju podataka gdje je zavisna varijabla diskretna, predstavlja broj pojavljivanja nekog događaja unutar određenog intervala i slijedi Poissonovu distribuciju. Cilj Poissonove regresije je uspostaviti povezanost između zavisne varijable (broj događaja) i nezavisnih varijabli (prediktora), pri čemu se očekivani broj događaja modelira pomoću odgovarajuće funkcije veze.
Prvobitni razvoj brojećih modela bio je usmjeren na područja poput aktuarstva, biostatistike i demografije. Međutim, s razvojem naprednijih statističkih metoda i računalnih tehnologija, primjena ovih modela proširila se na ekonomiju, politiku, sociologiju i mnoge druge discipline [4].

Prisjetimo se osnovnih pojmova o Poissonovoj distribuciji.

Definicija 2.0.1. Slučajna varijabla X ima **Poissonovu distribuciju** s parametrom $\mu > 0$ s vjerojatnostima:

$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!}, \quad k = 0, 1, 2, \ldots$$

gdje je:

- k broj događaja,
- μ očekivanje.

Pišemo $X \sim \mathcal{P}(\mu)$.

Za slučajnu varijablu X vrijedi:

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot P(X = k) = \sum_{k=0}^{\infty} k \frac{e^{-\mu}\mu^k}{k!} = e^{-\mu} \sum_{k=0}^{\infty} \frac{k \cdot \mu^k}{k!}$$
$$= e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^k}{(k-1)!} = \mu \cdot e^{-\mu} \sum_{k=0}^{\infty} \frac{\mu^{k-1}}{(k-1)!}.$$

Jer vrijedi $\sum_{k=0}^{\infty} \frac{\mu^k}{k!} = e^{\mu}$, slijedi

$$\mathbb{E}[X] = \mu. \tag{2.1}$$

Za varijancu vrijedi:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

Izračunajmo

$$\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 \frac{e^{-\mu}\mu^k}{k!}.$$

Zapišimo $k^2 = k^2 - k + k = k(k-1) + k$ i izračunajmo prethodni izraz za prve dvije vrijednosti $k = 0, 1$.
Tada slijedi,

$$\mathbb{E}[X^2] = e^{-\mu} \left[\mu + \sum_{k=2}^{\infty} \left(k^2 - k + k\right) \cdot \frac{\mu^k}{k!} \right]$$
$$= e^{-\mu} \left[\mu + \sum_{k=2}^{\infty} (k-1) \cdot k \cdot \frac{\mu^k}{k!} + \sum_{k=2}^{\infty} k \cdot \frac{\mu^k}{k!} \right].$$

Uvrstimo μ u drugu sumu i dobijemo

$$\mathbb{E}[X^2] = e^{-\mu} \left[\mu \cdot \sum_{k=1}^{\infty} \frac{\mu^{k-1}}{(k-1)!} + \mu^2 \cdot \sum_{k=2}^{\infty} \frac{\mu^{k-2}}{(k-2)!} \right]$$
$$= e^{-\mu} \left[\mu \cdot e^{\mu} + \mu^2 \cdot e^{\mu} \right]$$
$$= \mu + \mu^2.$$

Iz navedene formule za varijancu vrijedi:

$$\mathrm{Var}[X] = \mu + \mu^2 - \mu^2.$$
$$\mathrm{Var}[X] = \mu. \tag{2.2}$$

2.1 Poissonova regresija

Neka je $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ zavisna slučajna varijabla, gdje su kovarijate Y_i međusobno nezavisne i označavaju broj događaja n_i za i-tu kovarijatu i slijede Poissonovu distibuciju. Neka je $\mathbf{X} = (1, x_1, \ldots, x_k)^T \in M_{n,k+1}$ matrica prediktora čiji su vektori stupci $\mathbf{1} = (1, \ldots, 1)^T \in M_{n,1}$ i $\mathbf{x_i} = (x_{1i}, \ldots, x_{ni})^T, i = 1, \ldots, k$, te neka je $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T \in M_{k+1,1}$ vektor koeficijenata.
Iz 1.1, 1.4 i 1.5 smo pokazali da

$$\begin{aligned} &\mathbb{E}[Y_i] = \mu_i, \\ &\mathrm{Var}[Y_i] = \mu_i, \\ &\theta = \log \mu_i. \end{aligned}$$

Zavisna varijabla Y i nezavisne varijable $x_1, \ldots, x_k$ su povezane putem sljedeće relacije:

$$\mu = e^{\eta} = e^{\mathbf{X}\boldsymbol{\beta}} = \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right),$$

gdje je vektor očekivanih vrijednosti $\mu = (\mu_1, \ldots, \mu_n)$.

Stoga, očekivanje za generalizirani linearni model je:

$$\mathbb{E}[Y_i] = \mu_i = e^{x_i^T \boldsymbol{\beta}}, \quad Y_i \sim \mathrm{P}(\mu_i). \tag{2.3}$$

Funkcija veze g je:

$$g(\mu_i) = \log(e^{x_i^T \boldsymbol{\beta}}) = x_i^T \boldsymbol{\beta}. \tag{2.4}$$

2.2 Procjena parametara

Primjenom metode maksimalne vjerodostojnosti [1] želimo procijeniti koeficijente $\beta_0, \beta_1, \ldots, \beta_k$.

Vrijedi da je:

$$f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}) = \frac{\mu^y}{y!} e^{-\mu} = \frac{e^{x\beta y}}{y!} e^{-e^{x\beta}} = \prod_{i=1}^{n} \frac{e^{x_i \beta y_i}}{y_i!} e^{-e^{x_i \beta}}.$$

[1] Maximum likelihood estimation

Iz 1.12 slijedi

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left[y_i \log \mu_i - \mu_i - \log y_i! \right] \\ &= \sum_{i=1}^{n} \left[y_i x_i^T \boldsymbol{\beta} - e^{x_i^T \beta} - \log y_i! \right] \ ^2 \\ &= \sum_{i=1}^{n} \left[y_i x_i^T \boldsymbol{\beta} - e^{x_i^T \beta} \right]. \end{aligned} \tag{2.5}$$

Da bismo pronašli maksimum moramo riješiti sustav jednadžbi

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_i} = 0, \quad i = 0, \ldots, k \\ \sum_{i=1}^{n} x_i \left[y_i - e^{x_i^T \beta} \right] = 0. \end{aligned} \tag{2.6}$$

Prethodne jednadžbe definiraju sustav od $k + 1$ nelinearnih jednadžbi i $k + 1$ nepoznanica β. Budući da za $\hat{\boldsymbol{\beta}}$ ne postoji analitičko rješenje, potrebne su iterativne metode za rješavanje ovog problema. Najčešće korištene metode su gradijentne metode, poput Newton-Raphsonove metode, koje omogućuju numeričko rješavanje $\hat{\boldsymbol{\beta}}$. Konvergencija je garantirana jer je funckija log-vjerodostojnosti globalno konkavna. U praksi je potrebno manje od 10 iteracija (Poglavlje 3.8 u [4].).

Distribucijski rezultati za procjenu $\hat{\boldsymbol{\beta}}$ nisu dostupni u egzaktnoj formi stoga se zaključivanje temelji na asimptotskim rezultatima. Postoje tri glavna pristupa rješavanju ovog problema ([4], poglavlje 2.):

1. Procjenu $\hat{\boldsymbol{\beta}}$ možemo promatrati kao procjenitelj koji maksimizira funkciju 2.5 log-vjerodostojnosti . U tom slučaju primjenjuje se teorija maksimalne vjerodostojnosti za izvođenje zaključaka o parametrima.

2. Procjenu $\hat{\boldsymbol{\beta}}$ možemo definirati pomoću sustava jednadžbi 2.6, koje imaju sličnu interpretaciju kao i jednadžbe za metodu najmanjih kvadrata (OLS)[3]. Ovo omogućuje zaključivanje pod pretpostavkama koje uključuju samo srednju vrijednost i, po potrebi, varijancu. Generalizirani linearni modeli koriste ovu strukturu kako bi proširili klasične regresijske metode na širi raspon problema.

[2]Možemo izbaciti $-\log y_i!$ s obzirom da izraz ne sadrži β i ne utječe na rezultat.

[3]OLS - Ordinary least squares

3. Pristup temeljen na momentima: Budući da jednadžba ($\mathbb{E}[Y_i] = e^{x_i^T \boldsymbol{\beta}}$) implicira momentni uvjet $E[(y_i - e^{x_i^T \boldsymbol{\beta}})x_i] = 0$, moguće je definirati procjenitelj $\hat{\boldsymbol{\beta}}$ kao rješenje odgovarajućeg momentnog uvjeta u uzorku. Ovo je pristup temeljen na modelima momenta, koji koristi empirijske momente podataka za procjenu parametara.

Svaki od navedenih pristupa ima svoje prednosti i ograničenja, ovisno o karakteristikama podataka i ciljevima analize. Procjena maksimalne vjerodostojnosti je općenito nekonzistentna ako je gustoća pogrešno definirana, za određene gustoće konzistentnost se može postići i uz djelomično pogrešno definiran model. Jedan od primjera je procjena maksimalne vjerostostojnosti u linearnom regresijkom modelu pod pretpostavkom da je y_i nezavisno distribuiran kao $N(x_i^T\beta_0, \sigma^2)$. Tada $\hat{\beta}_{\mathrm{ML}}$ odgovara OLS procjenitelju, koji može biti konzistentan čak i uz nenormalnost i heteroskedastičnost, budući da je osnovni uvjet za konzistentnost točno definirano uvjetno očekivanje: $\mathbb{E}[y_i \mid x_i] = x_i^T\beta_0$.
Iz 2.6 vrijedi:

$$\mathbb{E}[(y_i - \exp(x_i^T\beta_0))x_i] = 0.$$

Prethodni uvjet je zadovoljen ako vrijedi $\mathbb{E}[y_i \mid x_i] = \exp(x_i^T\beta_0)$. Općenitije, ovakvi rezultati vrijede za procjenu maksimalne vjerojatnosti modela sa definiranom gustoćom koja pripada linearnoj eksponencijalnoj obitelji i za procjenu usko povezanih klasa generaliziranih linearnih modela.

Iako konzistentnost u ovim modelima zahtijeva samo točnu definiranost očekivanja, pogrešna definiranost varijance dovodi do neispravnih statističkih zaključaka zbog netočno prikazanih t-statistika i standardnih pogrešaka. Na primjer, u linearnom regresijskom modelu uobičajene standardne pogreške za OLS su netočne ako je pogreška heteroskedastična, a ne homoskedastična. Potrebno je prilagoditi standardne pogreške kako bi se osigurale ispravne vrijednosti standardnih pogrešaka.

2.3 Dijagnostika modela

Analiza reziduala

Nakon prilagodbe modela podacima, ključno je utvrditi koje prediktorske varijable imaju značajan utjecaj. Hipoteze o parametrima β_k mogu se testirati korištenjem Waldove statistike ili statistike omjera vjerodostojnosti. Prema [1] test omjera vjerodostojnosti preferira se u slučajevima malih uzoraka[4] ili kada su vrijednosti parametara velike, jer pruža pouzdanije rezultate u takvim situacijama.

[4]"Mali" uzorak je onaj s manje od otprilike 30 opažanja.

Za testiranje značajnosti jednostavnog modela, postavljamo hipoteze:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

Primjerice, za parametar β_k statistika testa može se izraziti kao:

$$W = \frac{\hat{\beta}_k - \beta_k}{\epsilon(\hat{\beta}_k)} \sim N(0, 1),$$

gdje je $\hat{\beta}_k$ procijenjena vrijednost parametra, a $\epsilon(\hat{\beta}_k)$standardna pogreška procjene.

Prilagođene vrijednosti modela definirane su sljedećim izrazom:

$$\hat{Y}_i = \hat{\mu}_i = \exp(x_i^T \hat{\boldsymbol{\beta}}), \quad i = 1, \ldots, n.$$

Ove vrijednosti često se označavaju s e_i, budući da su procjene očekivanih vrijednosti $E(Y_i) = \mu_i$. Budući da za Poissonovu razdiobu vrijedi $\text{Var}[Y_i] = E[Y_i]$, standardna pogreška Y_i procjenjuje se pomoću $\sqrt{e_i}$. Na temelju toga, reziduali su definirani kao:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}, \tag{2.7}$$

gdje o_i označava promatranu vrijednost Y_i.

Devijanca

Ako su varijable odgovora $Y_1, \ldots, Y_n$ nezavisne i $Y_i \sim \text{Poisson}(\mu_i)$, funkcija log-vjerodostojnosti je:

$$l(\boldsymbol{\beta}; y) = \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log y_i!.$$

Za zasićeni model, sve vrijednosti μ_i su različite, pa je $\boldsymbol{\beta} = [\mu_1, \ldots, \mu_k]^T$. Maksimalne procjene vjerodostojnosti su $\hat{\mu}_i = y_i$, pa je maksimalna vrijednost funkcije log-vjerojatnosti iz:

$$l(\mathbf{b}_{\max}; y) = \sum_{i=1}^{n} y_i \log y_i - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log y_i!.$$

Pretpostavimo da model od interesa ima $k < n$ parametara. Maksimalna procjena vjerodostojnosti $\mathbf{b}$ može se koristiti za izračunavanje procjena $\hat{\mu}_i$ i odgovarajućih prilagođenih vrijednosti $\hat{y}_i = \hat{\mu}_i$, jer $\mathbb{E}[Y_i] = \mu_i$. Maksimalna vrijednost log-vjerodostojnosti u ovom slučaju je:

$$l(\mathbf{b}; y) = \sum_{i=1}^{n} y_i \log \hat{y}_i - \sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} \log y_i!.$$

Stoga je devijanca iz 1.15 definirana kao:

$$D = 2\left[l(\mathbf{b}_{\max}; y) - l(\mathbf{b}; y)\right], \tag{2.8}$$

što se može zapisati kao:

$$D = 2\left[\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - \sum_{i=1}^{n}(y_i - \hat{y}_i)\right].$$

Za većinu modela vrijedi relacija $\sum y_i = \sum \hat{y}_i$. Kao rezultat toga, devijanca se može izraziti kao:

$$D = 2\sum_{i=1}^{n} o_i \log\left(\frac{o_i}{e_i}\right), \tag{2.9}$$

gdje su o_i opažene vrijednosti zavisne varijable y_i, a e_i procijenjene očekivane vrijednosti $\hat{y}_i$.

Vrijednost devijance može se usporediti s χ^2-distribucijom sa stupnjevima slobode $n-k$, gdje je k broj procijenjenih parametara. Detaljan dokaz ovog rezultata može se pronaći u [6].

Prekomjerna disperzija

Postoje situacije u kojima podaci o brojanju ne zadovoljavaju pretpostavke standardnog Poissonovog modela. Dva najčešća problema koja se pritom javljaju su:

- Prekomjerna disperzija:
  - Varijabilnost podataka premašuje očekivanu razinu prema Poissonovoj pretpostavci o ekvidisperziji ($\mathrm{Var}(Y) = \mu$).
  - Prekomjerna disperzija nastaje kada je varijanca veća od srednje vrijednosti, što može dovesti do podcijenjenih standardnih pogrešaka, precijenjenih test statistika, prekomjerne značajnosti i preusko postavljenih intervala pouzdanosti.

- Problemi sa ”nulama”:
  - Odsustvo nula: Pojavljuju se kada su iz uzorka isključeni članovi populacije s vrijednošću $Y = 0$ (npr., isključenje sudionika koji nisu konzumirali alkohol tijekom promatranja).
  - Višak nula: Javlja se kada uzorak uključuje pojedince koji nikada ne bi pokazali promatrano ponašanje (npr., osobe koje ne piju alkohol iz zdravstvenih, religijskih ili drugih razloga) [5].

U ovom diplomskom radu će biti objašnjena prekomjerna disperzija, a detaljnije o odsutnim nulama se može pronaći u [4].

Prekomjerna disperzija može nastati zbog:

- Nepoznate heterogenosti:

  Razlike među pojedincima koje nisu objašnjene regresijskim modelom, često uzrokovane izostavljenim prediktorima (npr., ako se u modelu za predikciju konzumacije alkohola ne uključi spol, neobjašnjena heterogenost povezana sa spolom uzrokuje dodatnu varijabilnost.).

- Zavisnosti među brojanjem događaja:

  Kada se pojedini događaji ne javljaju neovisno jedni o drugima (npr., konzumacija drugog alkoholnog pića ovisi o tome je li osoba prethodno popila prvo piće.).

Kako bi se problemi prekomjerne disperzije adekvatno riješili, koriste se prilagođeni modeli poput prekomjerne disperzirane Poissonove regresije (Quasi-Poissonova regresija) i negativne binomne regresije.

1. Negativna binomna regresija: Ovaj model uvodi dodatni parametar disperzije (α), čime se omogućava modeliranje varijance koja raste brže od srednje vrijednosti, tj. $\text{Var}(Y_i) = \mu_i + \alpha\mu_i^2$.

2. Quasi-Poisson model: Ova metoda proširuje standardni Poissonov model dopuštajući da varijanca bude proporcionalna srednjoj vrijednosti, tj. $\text{Var}(Y_i) = \phi\mu_i$, gdje je ϕ dodatni parametar disperzije koji se procjenjuje iz podataka.

Obje metode omogućuju bolje modeliranje podataka s prekomjernom disperzijom, a odabir između njih ovisi o strukturi podataka i ciljevima analize. Prekomjerna disperzija i problemi s viškom nula često se pojavljuju u analizi podataka o brojanju te mogu značajno utjecati na točnost Poissonovih modela. Ovi modeli omogućuju precizniju analizu i pružaju pouzdanije statističke zaključke [5].

Poglavlje 3

Negativan binomni generalizirani linearni model

Negativna binomna regresija je proširenje Poissonove regresije koje se koristi za modeliranje podataka gdje je zavisna varijabla diskretna i predstavlja broj pojavljivanja nekog događaja u određenom intervalu, ali varijanca podataka **premašuje** srednju vrijednost, što je poznato kao već definirana prekomjerna disperzija. Ovaj model je posebno koristan kada podaci ne zadovoljavaju pretpostavku ekvidisperzije, koja je osnovna karakteristika Poissonove regresije. Negativna binomna regresija uvodi **dodatni parametar** disperzije koji omogućuje modelu da prilagodi povećanu varijabilnost podataka.

Primarna svrha negativne binomne regresije je uspostavljanje veze između zavisne varijable, broja događaja, i nezavisnih varijabli, prediktora, pri čemu se očekivani broj događaja modelira pomoću funkcije veze. Ovaj model je posebno prikladan za brojeće podatke koji imaju značajnu heterogenost, odnosno kada različiti pojedinci ili skupine u uzorku pokazuju različite stope pojavljivanja događaja, što Poissonov model ne može adekvatno obuhvatiti.

Negativna binomna regresija se primjenjuje u širokom spektru područja. Rani razvoj ovog modela bio je usmjeren na biostatistiku i ekologiju, gdje se često susreću podaci sa značajnom prekomjernom disperzijom. S vremenom, zahvaljujući razvoju statističkih metoda i računalne tehnologije, negativna binomna regresija našla je primjenu u ekonomiji, društvenim znanostima, epidemiologiji, inženjerskim znanostima i drugim disciplinama [4].

Za bolje razumijevanje negativne binomne regresije potrebno je podsjetiti se osnovnih pojmova o negativnoj binomnoj distribuciji i njenoj vezi s Poissonovim modelom, kao i njenoj sposobnosti da modelira povećanu varijabilnost kroz dodatni parametar disperzije.

Definicija 3.0.1. Slučajna varijabla X ima **negativnu binomnu distribuciju** s parametrima $r > 0$ i $p \in (0, 1)$ i prima vrijednosti iz skupa $\mathbb{N}_0$ s vjerojatnostima:

$$p_i = P(X = k) = \binom{k + r - 1}{k} p^r (1 - p)^k, \quad k = 0, 1, 2, \ldots$$

gdje je:

- k broj neuspjeha,
- r broj uspjeha,
- p vjerojatnost uspjeha pri svakom pokušaju.

Pišemo $X \sim \text{NB}(r, p)$.

Za slučajnu varijablu X vrijedi:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot P(X = k) = \sum_{k=1}^{\infty} k \binom{k + r - 1}{k} p^r (1 - p)^k.$$

Vrijedi da je $k! = k(k - 1)!$ i $(r - 1)! = \frac{r!}{r}$ pa slijedi: $k\binom{k+r-1}{k} = r\binom{k+r-1}{k-1}$. Sada zapišemo očekivanje kao

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} r \binom{k + r - 1}{k - 1} p^r (1 - p)^k. \tag{3.1}$$

Također, znamo da vrijedi:

$$\sum_{k=0}^{\infty} \binom{k + r - 1}{k} p^r (1 - p)^k = 1. \tag{3.2}$$

Neka su $y = x - 1, \quad q = r + 1$ pa jednadžbu 3.1 možemo zapisati

$$\begin{aligned} \mathbb{E}[X] &= r \sum_{y=0}^{\infty} \binom{q + y - 1}{y} p^{q-1} (1 - p)^{y+1} \\ &= \frac{r(1 - p)}{p} \sum_{y=0}^{\infty} \binom{q + y - 1}{y} p^q (1 - p)^y \\ &\overset{(3.2)}{=} \frac{r(1 - p)}{p}. \end{aligned} \tag{3.3}$$

Za slučajnu varijablu X vrijedi: $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2$. Izračunajmo na sličan način kao i očekivanje:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=2}^{\infty} k(k-1)\binom{k+r-1}{k} p^r (1-p)^k \\
&= \sum_{k=2}^{\infty} r(r+1)\binom{k+r-1}{k-2} p^r (1-p)^k \\
&= r(r+1)\sum_{y=0}^{\infty}\binom{q+y-1}{y} p^{q-2}(1-p)^{y+2} \\
&= \frac{r(r+1)(1-p)^2}{p^2}.
\end{aligned} \tag{3.4}$$

$$\begin{aligned}
\mathrm{Var}[X] &= \frac{r(r+1)(1-p)^2}{p^2} + \frac{r(1-p)}{p} - \frac{r^2(1-p)^2}{p^2} \\
&= \frac{r(1-p)}{p}\left[\frac{(r+1)(1-p)}{p} + \frac{p}{p} - \frac{r(1-p)}{p}\right] \\
&= \frac{r(1-p)}{p^2}.
\end{aligned} \tag{3.5}$$

3.1 Negativna binomna regresija

Negativna binomna regresija je proširenje Poissonove regresije. Negativna binomna regresija uvodi dodatni parametar disperzije.

Prisjetimo se iz 1.1, 1.8 i 1.9 vrijedi:

$$\begin{aligned}
\mathbb{E}[Y] &= \frac{r(1-p)}{p}, \\
\mathrm{Var}[Y] &= \frac{r(1-p)}{p^2}, \\
\theta &= \ln(1-p).
\end{aligned}$$

Iz 1.11:

$$g(\mathbb{E}[Y]) = log\left(\frac{\mathbb{E}[Y]}{r + \mathbb{E}[Y]}\right).$$

No, zbog jednostavnosti parametriziramo u smislu $\dot{}\mu$ i $\alpha = \frac{1}{r}$ te zapišemo očekivanje i varijancu kao:

$$\mathbb{E}(Y) = \mu, \tag{3.6}$$

$$\text{Var}(Y) = \mu + \alpha\mu^2. \tag{3.7}$$

Indeks $\alpha > 0$ je vrsta već spomenutog parametra disperzije. Kako $\alpha \to 0$ tako $\text{Var}(Y) \to \mu$ što implicira da negativna binomna distribucija konvergira ka Poissonovoj. Parametar α je obično nepoznat, a njegova procjena pruža uvid u razinu prekomjerne disperzije pristutne u podatcima. Ako fiksiramo $\alpha = r$, tada možemo gustoću negativne binomne distribucije danu formulom 1.2 izraziti u obliku prirodne eksponencijalne familije 1.1. GLM modeli za negativnu binomnu regresiju pretpostavljaju da je α konstanta za sva opažanja, no njezina vrijednost ostaje nepoznata i mora se procijeniti ([1], 4.3.4).

Također vrijedi i:

$$\text{Var}(Y) = \phi \cdot \mathbb{E}(Y), \tag{3.8}$$

gdje je $\phi > 1$ paramater koji se procjenjuje.
Ove dvije varijance, (3.7) i (3.8), označavat će se kao NB2 i NB1, redom.

NB1 i NB2 funkcije varijance

U Poissonovom regresijskom modelu, y_i ima srednju vrijednost $\mu_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$ i varijancu μ_i. Za negativan binomni regresijski model zadržavamo pretpostavku da je srednja vrijednost $\exp(\mathbf{x}_i^T\boldsymbol{\beta})$, dok u podatcima kada varijanca nije jednaka srednjoj vrijednosti uvodimo opću notaciju za uvjetnu varijancu y_i:

$$\omega_i = \text{Var}[Y_i \mid \mathbf{x}_i].$$

Modeliramo varijancu kao funkciju srednje vrijednosti, koristeći:

$$\omega_i = \omega(\mu_i, \alpha),$$

za neku funkciju $\omega(\cdot)$, gdje je α skalarni parametar prekomjerne disperzije. Većini modela opća funkcija varijance za negativnu binomnu regresiju je:

$$\omega_i = \mu_i + \alpha\mu_i^p,$$

gdje je konstanta p određena. Analiza se obično ograničava na dva posebna slučaja, uz Poissonov slučaj gdje je $\alpha = 0$.

1. **NB1** funkcija varijance gdje je $p = 1$. Tada je varijanca:

$$\omega_i = (1 + \alpha)\mu_i,$$

što je višekratnik srednje vrijednosti. U GLM-u ovo obično pišemo:

$$\omega_i = \phi\mu_i, \tag{3.9}$$

gdje je $\phi = 1 + \alpha$ pa varijancu možemo zapisati i kao $\text{Var}[Y_i] = \phi \cdot \mathbb{E}[Y_i]$.

2. **NB2** funkcija varijance gdje je $p = 2$. Tada je varijanca kvadratna u srednjoj vrijednosti:

$$\omega_i = \mu_i + \alpha\mu_i^2. \tag{3.10}$$

NB1 se također naziva Quasi-Poisson modelom i može se koristiti kada su podaci prekomjerno ili nedovoljno raspršeni. S druge strane, NB2 se primjenjuje isključivo za modeliranje podataka koji pokazuju prekomjernu raspršenost.

3.2 Procjena parametara

Procjenitelji za ϕ_{NB1}, ϕ_{NB2}

Standardni procjenitelj za ϕ_{NB1} je jednak:

$$\hat{\phi}_{NB1} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \tag{3.11}$$

Funkcija varijance $\omega_i = \phi\mu_i$ implicira $\mathbb{E}[(y_i - \mu_i)^2] = \phi\mu_i$ i stoga $\phi = \mathbb{E}\left[\frac{(y_i-\mu_i)^2}{\mu_i}\right]$. Dijeljenje s $(n - k)$ umjesto s n se koristi kao korekcija za stupnjeve slobode (Detaljnije u [4], poglavlja 2.3.2 i 3.2.3).

Standardni procjenitelj za ϕ_{NB2} je jednak:

$$\hat{\phi}_{NB2} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2 - \hat{\mu}_i}{\hat{\mu}_i^2}. \tag{3.12}$$

Funkcija varijance $\omega_i = \mu_i + \alpha\mu_i^2$ podrazumijeva $\mathbb{E}[(y_i - \mu_i)^2 - \mu_i] = \alpha\mu_i^2$ i stoga

$$\alpha = \mathbb{E}\left[\frac{(y_i - \mu_i)^2 - \mu_i}{\mu_i^2}\right].$$

Detaljnije u [4], poglavlja 2.3.2 i 3.2.3.

NB2 i MLE

NB2 funkcija log-vjerodostojnosti je:

$$\ln L(\alpha,\beta) = \sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\ln\left(j+\alpha^{-1}\right)\right) - \ln y_i! - \left(y_i+\alpha^{-1}\right)\ln\left(1+\alpha\exp(\mathbf{x}_i^T\beta)\right) + y_i\ln\alpha + y_i\mathbf{x}_i^T\beta\right\}. \tag{3.13}$$

NB2 MLE $(\hat{\beta}_{NB2}, \hat{\alpha}_{NB2})$ je rješenje uvjeta prvog reda:

$$\sum_{i=1}^{n}\frac{y_i-\mu_i}{1+\alpha\mu_i}x_i = 0,$$

$$\sum_{i=1}^{n}\left\{\frac{1}{\alpha^2}\left(\ln(1+\alpha\mu_i) - \sum_{j=0}^{y_i-1}\frac{1}{j+\alpha^{-1}}\right) + \frac{y_i-\mu_i}{\alpha(1+\alpha\mu_i)}\right\} = 0. \tag{3.14}$$

Pod uvjetom da je očekivanje ispravno definirano, NB2 MLE je konzistentan za β. Iz $\mathbb{E}[y_i - \mu_i|x_i] = 0$ slijedi da je očekivana vrijednost lijeve strane u 3.14 jednaka 0 ako je srednja vrijednost ispravno definirana (Detaljnije u [4], 3.3.1).

NB1 i MLE

NB1 funkcija log-vjerodostojnosti je:

$$\ln L(\alpha,\beta) = \sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\ln\left(j+\alpha^{-1}\exp(\mathbf{x}_i^T\beta)\right)\right) - \ln y_i! - \left(y_i+\alpha^{-1}\exp(\mathbf{x}_i^T\beta)\right)\ln(1+\alpha) + y_i\ln\alpha\right\}. \tag{3.15}$$

NB1 MLE je rješenje uvjeta prvog reda:

$$\sum_{i=1}^{n}\left\{\left(\sum_{j=0}^{y_i-1}\frac{\alpha^{-1}\mu_i}{j+\alpha^{-1}\mu_i}\right)x_i + \alpha^{-1}\mu_i x_i\right\} = 0,$$

$$\sum_{i=1}^{n}\left\{\frac{1}{\alpha^2}\left(\sum_{j=0}^{y_i-1}\frac{\mu_i}{j+\alpha^{-1}\mu_i}\right) - \alpha^{-2}\mu_i\ln(1+\alpha) - \frac{\alpha}{1+\alpha} + y_i\alpha\right\} = 0 \tag{3.16}$$

Procjena temeljena na prva dva momenta NB1 gustoće daje Poissonov GLM procjenitelj, koji se također naziva NB1 GLM procjeniteljem (Detaljnije u [4], 3.3.3).

Jasno je da možemo razmotriti i druge modele NBp, koji se razlikuju od NB1 i NB2, a zadovoljavaju uvjete $\mathbb{E}(Y_i) = \mu_i$ i $\text{Var}(Y_i) = \mu_i + \alpha\mu_i^p$, gdje p nije fiksiran na vrijednosti 1 ili 2, već se procjenjuje.

Međutim, u ovom diplomskom radu bit će obrađeni isključivo modeli NB1 i NB2.

3.3 Dijagnostika modela

Analiza reziduala

Suma kvadrata Pearsonovih reziduala, danih izrazom (2.7), daje Pearsonovu χ^2 statistiku sa $n - p$ stupnjeva slobode:

$$\chi^2 = \sum_{i=1}^{n} r_i^2 = \sum \frac{(o_i - e_i)^2}{e_i}. \tag{3.17}$$

Pearsonov χ^2 test ili jednostavno χ^2 test provjerava odstupanja između promatranih vrijednosti i očekivanih vrijednosti. Postavljene hipoteze su:

H_0 : Model dobro odgovara podacima.

H_1 : Model ne odgovara dobro podacima.

χ^2 test je statistički test za ocjenu prihvatljivosti pretpostavljenog modela. Pearsonov statistički test za model Y_i, sa očekivanjem μ_i i varijancom ω_i, definiran je kao::

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\omega}_i}, \tag{3.18}$$

gdje su $\hat{\mu}_i$ i $\hat{\omega}_i$ procijenjene vrijednosti za μ_i i ω_i.

Sada ćemo prikazati Poissonov regresijski model uz prethodnu formulu gdje je:

$$\omega_i = \mu_i,$$

slijedi:

$$P_p = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \tag{3.19}$$

Vrijednost statistike P_p može se interpretirati na sljedeći način:

- Ako $P_p > n - k$ - prekomjerna disperzija, [1]
- Ako $P_p < n - k$ - nedovoljna disperzija,
- Ako $P_p = n - k$ - greška u definiranju uvjetne srednje vrijednosti.

Za NB1 model vrijedi 3.9 i 3.11. Međutim, to implicira ([4], 5.3) da P uvijek približno iznosi $(n - k)$, pa P više nije koristan dijagnostički alat.
Ako umjesto toga koristimo NB2 model, gdje vrijedi 3.10 i 3.12, tada je P koristan dijagnostički alat i vrijedi:

$$P_{NB} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \alpha\hat{\mu}_i^2}}. \tag{3.20}$$

Pearsonova statistika se ne može koristi za testiranje je li prekomjerna disperzije adekvatno modelirana u modelu. Međutim, može se koristiti ako je primijenjen NB2 MLE. U tom slučaju, odstupanje P od $(n-k)$ može zapravo ukazivati na grešku u definiranju uvjetne srednje vrijednosti.

Devijanca

Devijanca iz 1.15 i 2.8 za NB2 model je definirana:

$$D_{NB2} = 2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i + \alpha^{-1})\ln\left(\frac{y_i + \alpha^{-1}}{\hat{\mu}_i + \alpha^{-1}}\right)\right\}. \tag{3.21}$$

Devijanca se razlikuje u dva slučaja. U prvom slučaju saturirani model ima poseban parametar za svaku razinu prediktora, dok u drugom slučaju postoji jedan parametar za svaki podskup. Te slučajeve nazivamo grupiranim i negrupiranim podacima.

- Grupirani podatci - asimptotski prate približno χ^2- distribuciju, sa $n - k$ stupnjeva slobode.
- Negrupirani podatci - nije poznato koju distribuciju prate.

[1] Ako su očekivanje i varijanca dobro definirani vrijedi $\mathbb{E}\left[\sum_{i=1}^{n}(y_i - \mu_i)^2/\omega_i\right] = n$ jer $\mathbb{E}\left[(y_i - \mu_i)^2/\omega_i\right] = 1$. U praksi se P uspoređuje s $(n - k)$, zbog stupnjeva slobode procjene μ_i.

Poglavlje 4

Prekomjerna disperzija

Iako se u Poissonovoj regresiji pretpostavlja jednaka raspršenost, stvarni podaci mogu biti prekomjerno raspršeni. Ignoriranje prekomjerne raspršenosti podataka može rezultirati:

- premalim procjenama standardnih pogrešaka (SEE),
- precijenjenim testnim statistikama za procjene parametara,
- precijenjenom značajnošću,
- suženim pouzdanim intervalima.

Za rješavanje prekomjerne disperzije koriste se korekcija standardnih pogrešaka i negativna binomna regresija.

Poissonov regresijski model s prekomjernom disperzijom

U ovaj model uključujemo drugi parametar ϕ koji se koristi u procjeni uvjetne varijance. Model procijenjen s ovom korekcijom pretpostavlja distribuciju greške sa očekivanjem μ i varijancom $\mu\phi$.
Jedan od najjednostavnijih pristupa za rješavanje prekomjerne disperzije je skaliranje pomoću parametra ϕ.

Računanje parametra ϕ dano je idućom formulom:

$$\phi = \frac{\chi^2}{df}, \tag{4.1}$$

gdje je Pearsonov χ^2 test definiran u 3.19, sa $df = n - k$ stupnjeva slobode.

Vrijednost parametra ϕ može se interpretirati na sljedeći način:

- $\phi < 1$ - nedovoljna disperzija,[1]
- $\phi = 1$ - ekvidisperzija,
- $\phi > 1$ - prekomjerna disperzija.

Ako je parametar $\phi > 1$, procijenjene standardne pogreške se mogu skalirati faktorom $\sqrt{\phi}$. Također, prekomjerna disperzija se može riješiti korištenjem negativne binomne regresije - NB1 i NB2 modela.

Negativni binomni regresijski model

Procjena za α za NB1 model se lako može interpretirati. Ako je vrijednost parametra α:

- $\alpha < 0$ - nedovoljna disperzija,
- $0 < \alpha < 1$ - umjerena prekomjerna disperzija,
- $\alpha > 1$ - značajna prekomjerna disperzija.

Procjena za α za NB2 model se lako može interpretirati ako ju zapišemo u sljedećem obliku: $(1 + \alpha\mu_i)\mu_i$. Ako je vrijednost parametra α:

- $\alpha < 0$ - nedovoljna disperzija[2],
- ako je $\alpha\mu_i > 1$ - prekomjerna disperzija ovisi o vrijednosti zavisne varijable.

Za $\alpha = 0$, NB1 i NB2 modeli svode se na Poissonov model. Negativni binomni model može opisati samo slučajeve s prekomjernom disperzijom.

Ako je procijenjeni parametar α nizak, razlike između NB1 i NB2 modela bit će minimalne. U takvim slučajevima NB1 model obično je prikladniji zbog svoje jednostavnosti. Ako je α značajno velik, NB2 model bolje će odgovarati podacima, jer omogućuje kvadratni rast varijabilnosti i bolje modelira prekomjernu disperziju. ([4], 3.4)

[1] Rijetka među stvarnim podatcima, no teorijski moguća.

[2] Procijenjena varijanca je negativna za $\alpha < \frac{-1}{\mu_i}$.

Također, i devijanca se može koristiti za određivanje prekomjerne disperzije.

Srednja vrijednost χ^2-distribucije odgovara broju stupnjeva slobode, odnosno:

$$E\left(\chi^2_{n-k}\right) = n - k. \tag{4.2}$$

Stoga, ako model dobro odgovara podacima i ako devijanca prati χ^2 distribuciju, očekujemo da je skalirana devijanca modela blizu svoje srednje vrijednosti $n - k$, tj.

$$\frac{D_M}{\phi} \approx n - k. \tag{4.3}$$

Ako je $\phi = 1$ (kao što je to slučaj za Poissonov i negativni binomni model) vrijedi da su devijanca i skalirana devijanca jednaka. Za ove modele vrijedi:

$$D_M \approx n - k \tag{4.4}$$

ili ekvivalentno

$$\frac{D_M}{n - k} \approx 1. \tag{4.5}$$

Dakle, brzi test adekvatnosti Poissonovog ili negativnog binomnog modela može se provesti dijeljenjem rezidualne devijance sa stupnjevima slobode kako bi se provjerilo je li rezultat blizu jedan. Ako je omjer:

- $\frac{D_M}{n-k} < 1$ - nedovoljna disperzija,
- $\frac{D_M}{n-k} > 1$ - prekomjerna disperzija.

Kod malih skupova podataka potrebno je biti oprezan. Kako bi devijanca pratila χ^2-distribuciju, procijenjene vrijednosti predviđene modelom trebaju biti dovoljno velike (Detaljnije u [1], 4).[3]

[3]Trebaju biti veće od 5 iako [1] napominje da je očekivana frekvencija ćelije od 1 prihvatljiva sve dok manje od 20% očekivanih frekvencija ima vrijednosti manje od 5.

Poglavlje 5

Praktična primjena modela

U praktičnom dijelu pokušavamo metode iz teorijskog dijela primjeniti na stvarnim skupovima brojećih podataka. Glavni ciljevi su prepoznati adekvatan model za dani skup podataka i provesti razne statističke testove o kvaliteti modela i odabiru nezavisnih varijabli.

5.1 Primjena metoda na *Crab satellites* skupu podataka

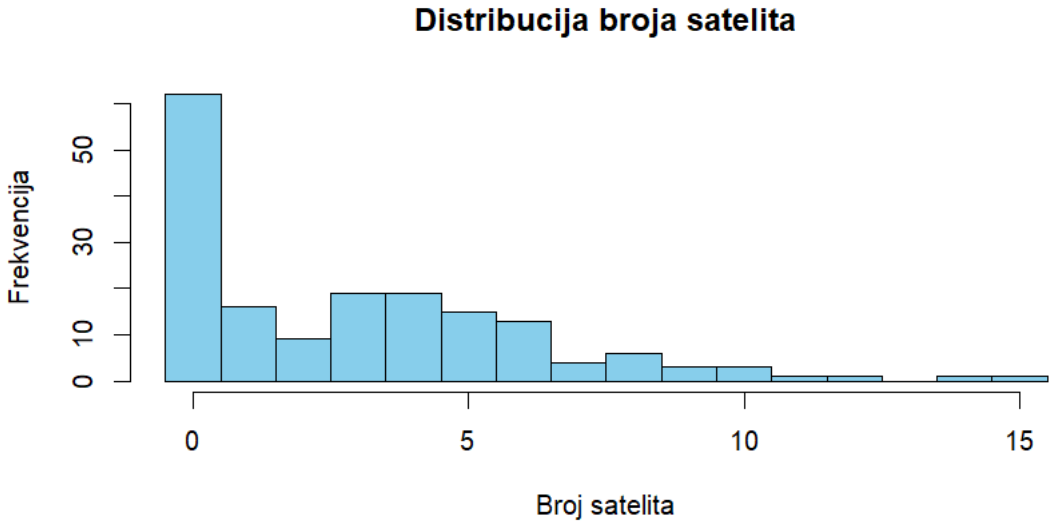
U prvom primjeru koristimo javno dostupni *Crab Satellites* skup podataka dostupan na `https://rdrr.io/rforge/countreg/man/CrabSatellites.html`.

Skup podataka je baziran na istraživanju ženskih rakova potkovičara na otoku u Meksičkom zaljevu. Tijekom sezone mriještenja, ženke migriraju na obalu radi parenja, u paru s mužjakom pričvršćenim na njihov stražnji dio. One kopaju u pijesku i polažu nakupine jaja. Tijekom mriještenja, drugi mužjaci rakova mogu se grupirati oko para i također oploditi jaja. Ovi mužjaci rakova koji se grupiraju oko ženskog raka nazivaju se *sateliti*. Satelitski mužjaci formiraju velike skupine oko nekih parova, dok druge potpuno ignoriraju. U eksperimentalnim manipulacijama, nakon što su vraćeni uklonjeni sateliti, opet su ih izvorni parovi privukli dok su se parovi bez satelita nastavili gnijezditi sami. Manipulacije su također otkrile da sateliti nisu samo kopirali ponašanje drugih mužjaka. Sateliti se nisu nasumično okupljali oko parova, [3] pokazuje da okupljanje nije uzrokovano okolišnim čimbenicima, već svojstvima gnijezdećih ženki rakova. Veće ženke, koje su u boljem stanju, privlače više satelita. Ženke s mnogo satelita bile su veće i u boljem stanju, ali nisu polagale više jaja od ženki s malo ili bez satelita.

Skup podataka sadrži 173 opažanja sa 5 varijabli:

- *color* (boja) → od 1 do 4, gdje je 1 najsvjetlija boja oklopa, a 4 najtamnija boja oklopa,
- *weight* (težina oklopa u kg) → od 1.20 do 5.20,
- *width* (širina oklopa u cm) → od 21.0 do 33.5,
- *spine* (stanje oklopa) → vrijednost 1, 2 ili 3, gdje je 1 najgore, a 3 najbolje,
- *satellites* (broj satelita) → od 0 do 15

Nezavisne varijable uključuju dvije numeričke varijable (težina i širina oklopa) i dvije kategoričke varijable (boja ženke rakova i stanje oklopa). Zavisna varijabla je broj satelita. Pogledajmo distribuciju zavisne varijable:



Slika 5.1: Distribucija zavisne varijable

Vidimo da veliki broj primjeraka ima zavisnu varijablu s vrijednosti 0 što može predstavljati problem. S druge strane, vidimo da su i ostale vrijednosti poprilično zastupljene pa ima smisla ovom skupu podataka pristupiti Poissonovom regresijom.

Poissonov model

Prvi pokušaj je Poissonova regresija kojom dobijemo sljedeće rezultate:

Varijabla	**Koeficijent**	**p-vrijednost**
Intercept	-0.52381	0.58101
width	0.02728	0.56954
color	-0.18503	0.00541
spine	0.04007	0.48061
weight	0.47319	0.00412
Devijanca	551.83 (168 stupnjeva slobode)	

Tablica 5.1: Rezultati regresije: koeficijenti, p-vrijednosti i devijanca

Prvo zaključujemo da su jedino koeficijenti uz varijable color i weight statistički značajno različiti od 0, što sugerira da je u model dovoljno uključiti te dvije varijable. No, za početak treba odrediti je li Poissonov model adekvatan za dane podatke. Vrijednost devijance nam ne govori ništa dok ne primjenimo odgovarajući statistički test. Kao što je opisano u dijelu 2.3, uspoređujemo vrijednost devijance s χ^2 distribucijom sa 168 stupnjeva slobode:

	Vrijednost
Devijanca	551.8301
Stupnjevi slobode (df)	168
P-vrijednost testa	0

Tablica 5.2: Rezultati testa devijance

Dakle, odbacujemo nultu hipotezu koja govori da Poissonov model dobro opisuje podatke. Možemo također i provesti brzi test naveden u 4.5:

$$\frac{D}{n-k} = \frac{551.8301}{168} \approx 3.3 \tag{5.1}$$

Vidimo da Poissonov model svakako nije adekvatan za podatke, a vrijednost 3.3 nam sugerira da je u pitanju prekomjerna disperzija. Isti rezultat dobijemo koristeći i Pearsonov test iz 3.3. Sljedeći korak je pokušati prilagoditi negativni binomni model s NB1 funkcijom varijance.

Negativni binomni (NB1) model

Koristeći negativni binomni model s NB1 funkcijom varijance 3.9, dobijemo sljedeće rezultate:

Varijabla	**Koeficijent**	**p-vrijednost**
intercept	-0.52381	0.760
width	0.02728	0.752
color	-0.18503	0.124
spine	0.04007	0.696
weight	0.47319	0.113
Devijanca	551.83 (168 stupnjeva slobode)	
Parametar raspršenosti (α)	2.241527	

Tablica 5.3: Rezultati regresije.

Odmah vidimo da sada nijedan koeficijent nije statistički značajan, ali koeficijenti za weight i color imaju daleko značajnije vrijednosti nego ostali. Također, iz vrijednosti devijance, kao i prije zaključujemo da model ne opisuje dobro podatke. Isto tako, vrijednost koeficijenta α iz (4) sugerira prekomjernu raspršenost, toliku da je čak i NB1 model ne može uhvatiti. Sve ovo sugerira korištenje negativnog binomnog modela s NB2 funkcijom varijance 3.10.

Negativni binomni (NB2) model

Rezultati NB2 modela su sljedeći:

Varijabla	**Koeficijent**	**p-vrijednost**
intercept	-0.73139	0.7008
width	0.02098	0.8298
color	-0.17920	0.1646
spine	0.01633	0.8906
weight	0.63783	0.0719
Devijanca	196.70 (168 stupnjeva slobode)	
Parametar raspršenosti (α)	1.045	

Tablica 5.4: Rezultati regresije

Slično kao i prije, nijedan koeficijent nije statistički značajan ali su opet color i weight značajniji nego ostali. Vidimo dosta manju vrijednost devijance pa se nadamo dobiti bolje rezultate po pitanju adekvatnosti modela:

	Vrijednost
Devijanca	196.70
Stupnjevi slobode (df)	168
P-vrijednost testa	0.064

Tablica 5.5: Rezultati testa devijance

Vidimo da ne možemo odbaciti nultu hipotezu da model dobro opisuje podatke, pa zaključujemo da je ovaj model adekvatan.

Odabir nezavisnih varijabli

Sada pokušajmo reducirati model i usporediti ga s ovim modelom. Prirodno je zadržati samo varijable color i weight s obzirom da su one kroz sva 3 navedena modela pokazale najveću značajnost koeficijenata. Slijede rezultati za NB2 model samo s varijablama color i weight:

Varijabla	**Koeficijent**	**p-vrijednost**
intercept	-0.3209	0.562
weight	0.7068	0.0000115
color	-0.1735	0.148
Devijanca	196.65 (170 stupnjeva slobode)	
Parametar raspršenosti (α)	1.046	

Tablica 5.6: Rezultati regresije

Varijabla weight je statistički značajna, pa pokušajmo sada samo nju zadržati:

Varijabla	**Koeficijent**	**p-vrijednost**
intercept	-0.8637	0.0328
weight	0.7599	0.00000146
Devijanca	196.16 (171 stupanj slobode)	
Parametar raspršenosti (α)	1.074	

Tablica 5.7: Rezultati regresije

Uz pomoć testa omjera vjerodostojnosti, sličan onome u 1.14, želimo vidjeti je li model sa samo jednom varijablom (weight) značajno lošije opisuje podatke nego model sa sve 4 varijable. Model s 4 varijable ima vrijednost 2 puta negativne log-vjerodostojnosti od -746.388, dok model s jednom varijablom ima vrijednost -748.643. Koristeći χ^2 test s 3 stupnja slobode, dobijemo p-vrijednost od 0.52, što znači da ne možemo odbaciti nultu hipotezu da modeli jednako dobro opisuju podatke.
Zaključak, kao najadekvatniji model biramo negativni binomni model s NB2 funkcijom varijance i jednom varijablom (weight) iz 2 glavna razloga:

- Između Poissonovog modela, NB1 i NB2 modela je pokazano da jedino NB2 model može ispravno modelirati varijancu.
- Od svih NB2 modela je najjednostavniji (najmanji broj nezavisnih varijabli), a pritom ne žrtvuje adekvatnost nad skupom podataka.

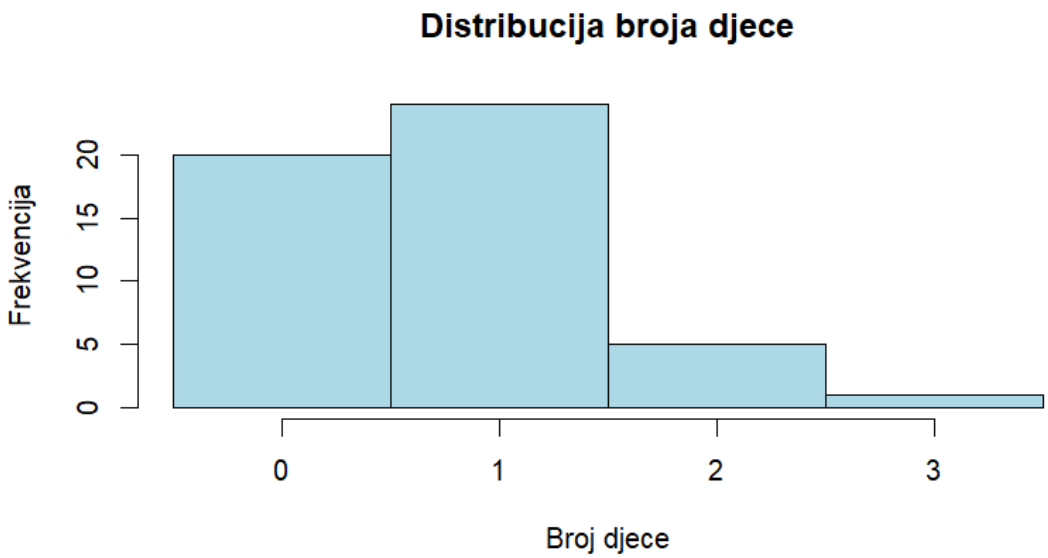
5.2 Primjena metoda na *Fertilitet* skupu podataka

Na podatcima u skupu podataka *Fertilitet* koji se može pronaći na `http://www.efzg.hr/UserDocsImages/sta/jarneric/Fertilitet.xls` se istražuje pod kojim uvjetima žena odlučuje imati više djece, koji od faktora plaće, obrazovanje i godina utječu na fertilitet. Skup podataka sadrži 50 opažanja sa 3 varijable:

- *godine* → od 18 do 58,
- *plaća* (godišnja neto plaća u 000 kn) → od 39 - 90,
- *obrazovanje* → 1 - fakultet, 0 - bez fakulteta,
- *djeca* (broj djece) → od 0 do 3.

Nezavisne varijable su godine, plaća i fakultet, gdje su godine i plaća numeričke varijable, a obrazovanje kategorička. Zavisna varijabla je broj djece.

Prikažimo histogram zavisne varijable broja djece:



Slika 5.2: Distribucija zavisne varijable

S obzirom na prirodu podataka, nule u ovim podatcima ne bi trebale predstavljati problem. Ima smisla ovom skupu podataka pristupiti Poissonovom regresijom.

Poissonov model

Prvi pokušaj je Poissonova regresija kojom dobijemo sljedeće rezultate:

Tablica 5.8: Rezultati regresije: koeficijenti, p-vrijednosti i devijanca

Varijabla	**Koeficijent**	**p-vrijednost**
Intercept	-4.81673	0.000012
Starost	0.04153	0.004121
Plaća	0.04368	0.000314
Obrazovanje	-0.43472	0.206262
Devijanca	20.684	46 stupnjeva slobode

Tablica 5.9: Rezultati regresije: koeficijenti, p-vrijednosti i devijanca

Model koji se procjenjuje je:

$$\log \mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, \tag{5.2}$$

gdje je x_1 starost, x_2 plaća te x_3 obrazovanje. Prvo zaključujemo da su koeficijenti uz varijable starost i plaća statistički značajno različite od 0, što sugerira da je u model dovoljno uključiti samo te varijable.

Odredimo je li Poissonov model adekvatan za podatke, to jest jesu li podatci prekomjerno raspršeni. Kao što smo napravili u prvom primjeru, usporedit ćemo vrijednost devijance sa χ^2 distribucijom sa 46 stupnjeva slobode:

	Vrijednost
Devijanca	20.684
Stupnjevi slobode (df)	46
P-vrijednost testa	0.9995344

Tablica 5.10: Rezultati testa devijance

Dakle, ne odbacujemo nultu hipotezu koja govori da Poissonov model dobro opisuje podatke.

Također, iz brzog testa navedenog u 4.5 vidimo da je omjer manji od 1 te slijedi da podatci nisu prekomjerno raspršeni.

$$\frac{D}{n-k} = \frac{20.684}{46} \approx 0.45 \tag{5.3}$$

Vrijednost Pearsonove statistike je 21.19329 što je manje od 46 stupnjeva slobode, pa prema 3.19 imamo iste zaključke.

S obzirom da Poissonov model dobro opisuje podatke, nije potrebno provjeravati modele negativne binomne regresije.

Odabir nezavisnih varijabli

Sada reducirani model želimo usporediti s početnim modelom i vidjeti koji model bolje opisuje podatke:

Varijabla	**Koeficijent**	**p-vrijednost**
Intercept	-4.63706	0.00002
Starost	0.03892	0.006499
Plaća	0.03960	0.000781
Devijanca	22.274	47 stupnjeva slobode

Tablica 5.11: Rezultati regresije: koeficijenti, p-vrijednosti i devijanca

Kao kod punog modela, usporedbom vrijednosti devijance sa χ^2 distribucijom dobijemo:

	Vrijednost
Devijanca	22.274
Stupnjevi slobode (df)	47
P-vrijednost testa	0.9991776

Tablica 5.12: Rezultati testa devijance

Ne odbacujemo nultu hipotezu, reducirani model sa dvije varijable starosti i plaće dobro opisuje podatke.
Uz pomoć testa vjerodostojnosti želimo provjeriti opisuje li model sa dvije varijable značajno lošije podatke nego puni model sa sve tri varijable.

Model	**Stupnjevi slobode**	**Devijanca**
Reducirani	47	22.274
Puni	46	20.684
LR statistika = 1.5901, p-vrijednost = 0.2073		

Tablica 5.13: Usporedba Poissonovih modela

S obzirom na p vrijednost od 0.2073, ne možemo odbaciti nultu hipotezu da modeli jednako dobro opisuju podatke Između ova dva modela, bolji model je reducirani model sa dvije varijable - godine i plaća zbog njegove jednostavnosti.

Interpretacija koeficijenata: Nema smisla interpretirati slobodni koeficijent kada su sve nezavisne varijable jednake 0. Svaka dodana godina žene povećava očekivani broj djece $e^{0.04153} \approx 1.042$ puta, to jest 4.2%. Za veću plaću od 1000 kuna, očekivani broj djece se poveća $e^{0.04368} \approx 1.045$ puta, to jest 4.5%. S obzirom da je koeficijent kategoričke varijable obrazovanje negativan, to znači da će žene koje imaju fakultet imati manji broj očekivane djece za razliku od onih koje nemaju fakultet i to $e^{-0.43472} \approx 0.648$ puta, to jest osobe sa fakultetom imaju 35.2% manje djece od žena bez visokog obrazovanja.

Očekivani broj djece za ženu sa 26 godina, visokim obrazovanjem i netom godišnjom plaćom od 90000 kuna je 0.7859511, a za istu tu ženu povećanje plaće od 22000 kuna, povećava očekivani broj djece na 2.054466.

Zaključak

Prilagodili smo dva različita modela dvama skupovima podataka, s različitim brojem opažanja i različitim brojem nezavisnih varijabli. Poissonova regresija je jednostavnija i omogućuje direktnu interpretaciju, no prikladna je samo kada podaci nemaju prekomjernu disperziju kao što je to bio drugi primjer. Međutim, kada je prisutna prekomjerna disperzija, što je bio slučaj u prvom primjeru, pokazalo se da negativna binomna regresija bolje modelira podatke. Prvo smo pokušali koristiti NB1 model, no podaci su i dalje pokazivali preveliku raspršenost. Stoga je NB2 model bio bolji izbor, jer omogućuje fleksibilnije modeliranje varijance i bolje opisuje podatke.

Poglavlje 6

Dodatak A

Dokažimo 1.3.

[2] Unutar eksponencijalne familije promatramo funkciju log-vjerodostojnosti $l(y;\theta,\phi) = \log(f(y;\theta,\phi))$. Pokazat ćemo dva rezultata iz statističke teorije:

$$\mathbb{E}\left[\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right] = 0 \quad \text{i} \quad \mathbb{E}\left[\frac{\partial^2 l(Y;\theta,\phi)}{\partial\theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right)^2\right] = 0. \tag{6.1}$$

Kako bismo dokazali prvu od navedenih jednakosti, pretpostavljamo da je moguće diferencirati izraz

$$\int f(y;\theta,\phi)dx$$

po parametru θ tako da jednostavno uvedemo znaka diferenciranja pod integral[1]. Budući da je ovaj integral jednak 1 za sve θ, diferenciranjem dobijemo:

$$0 = \int \frac{\partial}{\partial\theta} f(y;\theta,\phi)dx.$$

Nadalje, možemo napisati:

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(y;\theta,\phi)\,\frac{f(y;\theta,\phi)}{f(y;\theta,\phi)}dx \\ &= \int \frac{\partial}{\partial\theta} l(y;\theta,\phi) f(y;\theta,\phi)dx \\ &= \mathbb{E}\left[\frac{\partial l(Y;\theta,\phi)}{\partial\theta}\right]. \end{aligned}$$

[1] Unutar eksponencijalnih familija distribucija ovo je uvijek moguće.

Kako bismo pokazali drugu od navedenih jednakosti, diferenciramo iduću jednakost po θ

$$\int \frac{\partial}{\partial \theta} l(y;\theta,\phi) f(y;\theta,\phi) dx = 0,$$

dobijemo sljedeće:

$$\int \frac{\partial^2 l(y;\theta,\phi)}{\partial \theta^2} f(y;\theta,\phi) dx + \frac{\partial l(y;\theta,\phi)}{\partial \theta} \frac{\partial f(y;\theta,\phi)}{\partial \theta} dx = 0.$$

Diferenciranjem $l(y;\theta,\phi) = \log(f(y;\theta,\phi))$ po θ vrijedi

$$\frac{\partial l(y;\theta,\phi)}{\partial \theta} = \frac{1}{f(y;\theta,\phi)} \frac{\partial f(y;\theta,\phi)}{\partial \theta},$$

stoga

$$\int \frac{\partial^2 l(y;\theta,\phi)}{\partial \theta^2} f(y;\theta,\phi) dx = -\int \left(\frac{\partial l(y;\theta,\phi)}{\partial \theta} \right)^2 f(y;\theta,\phi) dx.$$

Vrijedi

$$\mathbb{E}\left[\frac{\partial^2 l(Y;\theta,\phi)}{\partial \theta^2} \right] = -\mathbb{E}\left[\left(\frac{\partial l(Y;\theta,\phi)}{\partial \theta} \right)^2 \right] = 0.$$

Pokazali smo i drugu jednakost u 6.1.

Poglavlje 7

Kodovi

Crab Satellites

```
hist(CrabSatellites$satellites,
     breaks = seq(-0.5, max(CrabSatellites$satellites) + 0.5, by
        = 1),
     col = "skyblue",
     main = "Distribucija broja satelita",
     xlab = "Broj satelita",
     ylab = "Frekvencija")
```

```
# Poissonova regresija
poisson_model <- glm(satellites ~ width + color + spine + weight,
                     family = poisson,
                     data = CrabSatellites)

summary(poisson_model)
```

```
#Reducirana Poissonova regresija
poisson_model_reduced <- glm(satellites ~ color + weight,
                     family = poisson,
                     data = CrabSatellites)

summary(poisson_model_reduced)
```

```
# Devijanca i stupnjevi slobode
model_deviance <- poisson_model$deviance
df_residual <- poisson_model$df.residual

# p vrijednost hi kvadrat test
p_value_deviance <- 1 - pchisq(model_deviance, df_residual)
```

```
# Pearsonova hi-kvadrat statistika
pearson_residuals <- residuals(poisson_model, type = "pearson")
pearson_chisq <- sum(pearson_residuals^2)

# p vrijednost Pearsonova testa
p_value_pearson <- 1 - pchisq(pearson_chisq, df_residual)
```

```
#NB1 model
nb1_model <- glm(satellites ~ width + color + spine + weight,
                         family = quasipoisson,
                         data = CrabSatellites)
summary(nb1_model)
```

```
#NB1 model reducirani
nb1_model_reduced <- glm(satellites ~ color + weight,
                 family = quasipoisson,
                 data = CrabSatellites)
summary(nb1_model_reduced)
```

```
#NB2 model
nb2_model <- glm.nb( satellites ~ width + color + spine +
                         weight,
                    data = CrabSatellites)

summary(nb2_model)
```

```
# Devijanca i stupnjevi slobode reziduala
model_deviance <- nb2_model$deviance
df_residual <- nb2_model$df.residual

# p vrijednost hi kvadrat test
p_value_deviance <- 1 - pchisq(model_deviance, df_residual)
```

```
# NB2 model reducirani
nb2_model_reduced <- glm.nb(satellites ~  weight + color,
                            data = CrabSatellites)
summary(nb2_model_reduced)
```

```
#NB2 potpuno reducirani
nb2_model_reduced_total <- glm.nb(satellites ~  weight,
                          data = CrabSatellites)

summary(nb2_model_reduced_total)
```

```
Usporedba dva modela
anova(nb2_model, nb2_model_reduced_total, test = "Chisq")
```

Fertilitet

```
hist(Fertilitet$djeca,
     breaks = seq(-0.5, max(Fertilitet$djeca) + 0.5, by = 1),
     col = "lightblue",
     main = "Distribucija broja djece",
     xlab = "Broj djece",
     ylab = "Frekvencija",
     right = FALSE)
```

```
#Poissonova regresija
poisson_model_pr2 <- glm(djeca ~  starost + placa + factor(
   obrazovanje)       ,
                    family = poisson,
                    data = Fertilitet)
summary(poisson_model_pr2)
```

```
# Devijanca i stupnjevi slobode
model_deviance_2 <- poisson_model_pr2$deviance
df_residual_2 <- poisson_model_pr2$df.residual

# p vrijednost hi kvadrat test
p_value_deviance_2 <- 1 - pchisq(model_deviance_2, df_residual_2)
```

```
# Pearsonova hi - kvadrat statistika
pearson_residuals_2 <- residuals(poisson_model_pr2, type = "
    pearson")
pearson_chisq_2 <- sum(pearson_residuals_2^2)

# p vrijednost Pearsonova testa
p_value_pearson_2 <- 1 - pchisq(pearson_chisq_2, df_residual_2)
```

```
poisson_model_pr2_reduced <- glm(djeca ~  starost + placa
    ,
                          family = poisson,
                          data = Fertilitet)
summary(poisson_model_pr2_reduced)
```

```
Devijanca i stupnjevi slobode
model_deviance_2_reduced <- poisson_model_pr2_reduced$deviance
df_residual_2_reduced <- poisson_model_pr2_reduced$df.residual

#p vrijednost hi kvadrat testa
p_value_deviance_2_reduced <- 1 - pchisq(model_deviance_2_reduced
    , df_residual_2_reduced)
```

```
#Usporedba dva modela
anova(poisson_model_pr2, poisson_model_pr2_reduced, test = "Chisq
    ")
```

```
#Predikcija
novi_podaci <- data.frame(starost = 26, placa = 90, obrazovanje =
     1)
predvidjeni_broj_djece <- predict(poisson_model_pr2, newdata =
    novi_podaci, type = "response")

novi_podaci_2 <- data.frame(starost = 26, placa = 112,
    obrazovanje = 1)
predvidjeni_broj_djece_2 <- predict(poisson_model_pr2, newdata =
    novi_podaci_2, type = "response")
```

Bibliografija

[1] A. Agresti, *Categorical Data Analysis*, Wiley-Interscience, Hoboken, NJ, 2002.

[2] B. Basrak, *Generalizirani linearni modeli*, https://web.math.pmf.unizg.hr/~bbasrak/pdf_files/FinPrak/FPchap7.pdf, Pristupljeno: 05. prosinca 2024.

[3] H. Jane Brockmann, *Satellite Male Groups in Horseshoe Crabs, Limulus polyphemus*, Ethology **102** (1996), br. 1, 1–21, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0310.1996.tb01099.x.

[4] A. C. Cameron i P. K. Trivedi, *Regression Analysis of Count Data*, Cambridge University Press, Cambridge, 1998.

[5] S. Coxe, S. G. West i L. S. Aiken, *The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives*, Journal of Personality Assessment **91** (2009), 121–136.

[6] A. J. Dobson, *An Introduction to Generalized Linear Models*, Chapman and Hall/CRC, London, 2002.

Sažetak

U ovom diplomskom radu obrađena je regresijska analiza za brojeće podatke kroz pristup generaliziranih linearnih modela, s naglaskom na Poissonovu i negativu binomnu regresiju. Generalizirani linearni modeli pružaju fleksibilno modeliranja podataka koristeći eksponencijalnu familju distribucija i odgovarajuće funkcije veze. U radu su detaljno analizirani Poissonova regresija koja je osnovni pristup za modeliranje brojećih podataka i negativna binomna regresija koja uključuje dodatni parametar disperzije za rješavanje problema prekomjerne raspršenosti.

Praktična primjena ovih modela na stvarne skupove podataka omogućuje usporedbu njihovih performansi. Rezultati teorijskog i praktičnog dijela potvrđuju da negativna binomna regresija bolje odgovara podatcima sa prekomjernom raspršenosti, dok je Poissonova regresija primjerenija kada vrijedi pretpostavka o jednakoj raspršenosti.

Odabir odgovarajućeg modela za brojeće podatke ključan je za točnost statističkih procjena i interpretaciju rezultata. Pravilna dijagnostika modela je ključna u donošenju ispravnih zaključaka i osigurava da modeli adekvatno reprezentiraju stvarne podatke.

Summary

In this thesis, regression analysis for count data is explored through the framework of generalized linear models, with a focus on Poisson and negative binomial regression. Generalized linear models provide a flexible approach to data modeling using the exponential family of distributions and appropriate link functions. The study presents a detailed analysis of Poisson regression, which serves as the fundamental model for count data, and negative binomial regression, which includes an additional dispersion parameter to deal with overdispersion.
The practical application of these models to real datasets allows for a comparison of their performance. The results from both theoretical and empirical analyses confirm that negative binomial regression is more suitable for data with overdispersion, while Poisson regression is more appropriate when the assumption of equidispersion holds.

The selection of an appropriate model for count data is crucial for the accuracy of statistical estimates and the interpretation of results. Proper model diagnostics play a key role in making correct conclusions and ensuring that models adequately represent real-world data.

Životopis

Rođena sam 30. srpnja 1998. godine u Šibeniku. Obrazovanje sam započela u Osnovnoj školi Tina Ujevića nakon čega upisujem Prirodoslovno-matematičku gimnaziju Antuna Vrančića koju završavam 2017. godine. Iste godine upisujem Matematički odsjek, nastavnički smjer na Prirodoslovno- matematičkom fakultetu u Zagrebu. 2021. godine kao prvostupnica edukacije matematike upisujem diplomski studij, smjer Matematička statistika. Tijekom studiranja sam držala instrukcije i započela rad u osnovnoj školi tijekom završnih godina studija.