Analiza vjerojatnosti pokrivanja Waldovih pouzdanih intervala za regresijske koeficijente u Poissonovoj regresijskoj analizi

Ramljak, Irena

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Irena Ramljak

ANALIZA VJEROJATNOSTI POKRIVANJA WALDOVIH POUZDANIH INTERVALA ZA REGRESIJSKE KOEFICIJENTE U POISSONOVOJ REGRESIJSKOJ ANALIZI

Diplomski rad

Voditelj rada:
prof. dr. sc. Vesna Lužar-Stiffler

Zagreb, veljača 2019.

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. _____

Mojoj obitelji

Sadržaj

Uvod

Već u davnoj prošlosti ljudi su primijetili određene zakonitosti u ponašanju pojava i događaja kao i međudjelovanje različitih faktora. Prikupljanjem i proučavanjem podataka moguće je pobliže opisati opaženu vezu te tako predvidjeti buduće ponašanje. Mogućnost predviđanja budućeg ponašanja može imati vrlo veliki značaj. Omogućuje nam pronalaženje najpovoljnijeg načina djelovanja i optimiziranju rada, možemo predvidjeti neke pojave te se tako pripremiti na dolazeće događaje poput na primjer, obilnih kiša koje bi mogle prouzrokovati poplave na nekom području i slično. Stoga, modeliranje podataka ima veliku važnost u svim područjima ljudskog djelovanja, što potvrđuje razvoj tehnologije koja nam danas omogućuje određivanje sve boljih modela nekog ponašanja.

Mnoge probleme nije uvijek lagano modelirati. Odrediti neki model možemo uvijek, ali isto tako postavljamo pitanje koliko je "dobar" odabrani model te možemo li ga na neki način poboljšati tako da smanjimo pogrešku što je moguće više.
U procesu odabira modela određujemo koje su varijable značajne, odnosno o kojim faktorima ovisi ponašanje neke zavisne varijable koje želimo opisati. Određivanje značajnosti varijabli radimo raznim statističkim testovima. Kao kod svih statističkih tvrdnji, određeno zaključivanje vršimo na nekoj razini značajnosti α. Preciznije, ako je $\alpha = 0.05$, tada je šansa da smo pogriješili prilikom zaključivanja 5%.

U osnovnoj školi se već govori o linearnoj funkciji, odnosno linearnoj vezi između dvije varijable. Kasnije se taj problem proširuje na veće dimenzije tako da jedna varijabla ovisi o linearnoj kombinaciji drugih. Uočeno je da mnoge probleme upravo opisuje upravo linearna veza, a jednom takvom vezom ćemo se baviti i u ovome radu, a to je *Poissonova regresija*.

Poissonova regresija pripada *generaliziranim linearnim modelima* o kojima govorimo u prvom dijelu rada. Definirat ćemo što su to generalizirani linearni modeli, reći neka svojstva te navesti primjere koji pripadaju ovoj skupini modela. Sljedeće ćemo precizno definirati Poissonovu regresiju, opisati kako procjenjujemo koeficijente

regresije te navesti primjere gdje se koristi ova vrsta regresije.

Prilikom procjenjivanja koeficijenata regresije, procjenjujemo i njihove $(1-\alpha)\cdot 100\%$ Waldove pouzdane intervale, za $\alpha = 0.05$ i $\alpha = 0.01$.

Glavni cilj ovoga rada jest ispitati vjerojatnost pokrivanja Waldovih pouzdanih intervala koeficijenta Poissonove regresije, odnosno analizirati je li njihova pouzdanost 95% tj., 99% kako se očekuje.

Ispitivanje vjerojatnosti pokrivanja radimo pomoću Monte Carlo simulacija u računalnom softveru `SAS`. Simulacijom različitih faktora, analiziramo kako se ponaša promatrana vrijednost.
Navedena analiza je sadržaj drugoga dijela rada. Prvo su uvedene potrebne teorijske tvrdnje i definicije te je potom opisana Monte Carlo metoda i analizirana je pogreška prilikom simulacije. Nakon toga, navedeni su podaci koji se koriste, opisan je algoritam simulacije te potom modeliranja.

U zadnjem dijelu rada se nalaze dobiveni rezultati, pripadne tablice i grafovi te njihova analiza.

Poglavlje 1

Generalizirani linearni modeli

1.1 Općenito o generaliziranim linearnim modelima

U mnogim problematikama želimo opisati podatke oblika $(y_i, x_{1i}, \ldots, x_{ki})$ za i = 1,. . . ,n, $k \in \mathbb{N}$, pri čemu je y_i realizacija neke slučajne varijable Y_i čija distribucija ovisi o vrijednostima $x_{1i}, \ldots, x_{ki}$. Drugim riječima, želimo pronaći model koji će nam za danu vrijednost $x_{1i}, \ldots, x_{ki}$ procijeniti vrijednost za y_i, tako da je pogreška što manja.
Odabir najboljeg modela nije jednostavno pitanje, a jedan od najpoznatijih modela koji nam daje rješenje traženog problema je linearna regresija.
Opći oblik linearne regresije je

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \tag{1.1}$$

pri čemu su β_0,. . . , β_k koeficijenti modela koje procjenjujemo, $x_1, \ldots, x_k$ prediktorske varijable ili kovarijate, ε slučajna pogreška za koju se pretpostavlja da dolazi iz N(0, σ^2) distribucije te se pretpostavlja da su komponente od Y nezavisno distribuirane s konstantnom varijancom.
Matrično možemo pisati

$$Y = X\beta + \varepsilon \tag{1.2}$$

gdje su

$$Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$$

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$$

$$\beta = (\beta_0, \ldots, \beta_k)^T \in \mathbb{R}^{k+1}$$

jedinični vektori, a $X = (1, x_1, \ldots, x_k) \in M_{n,k+1}(\mathbb{R})$ matrica čiji su stupci

$$1 = (1, \ldots, 1)^T \in \mathbb{R}^n \qquad x_i = (x_{1i}, \ldots, x_{ni})^T, \qquad i = 1, \ldots, k \tag{1.3}$$

Nadalje, pretpostavljamo da je matrica X punog ranga tj., da su prediktorske varijable međusobne nezavisne. Iz oblika modela dobivamo da vrijedi

$$\mu = E[Y|x] = X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{1.4}$$

odnosno, Y i x su linearno povezane.
Izraz (1.4) možemo zapisati i na sljedeći način:

1. Komponente slučajne varijable Y su međusobno nezavisne i normalno distribuirane s očekivanjem μ i konstantnom varijancom
2. Linearna kombinacija prediktorskih varijabli $x_1, \ldots, x_k$ i parametara $\beta_0, \ldots, \beta_k$ definira parametar η tj.,
$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{1.5}$$
3. Veza između slučajne varijable Y i kovarijata $x_1, \ldots, x_k$ je
$$\mu = \eta$$

Gornji zapis je generalizacija izraza (1.4). Zadnji izraz iz generalizacije možemo zapisati kao

$$\eta = g(\mu)$$

gdje je g funkcija identitete u slučaju linearne regresije. Funkciju g zovemo funkcija veze (eng. *link function*).
Nadalje, generalizaciju još možemo proširiti. Slučajne komponente mogu imati bilo koju razdiobu iz eksponencijalne familije razdioba,a ne samo normalnu razdiobu te vezna funkcija može biti bilo koja monotona diferencijabilna funkcija. Time smo dobili cijelu skupinu generaliziranih linearnih modela (GLM).

Definicija 1.1.1. *Slučajna varijabla Y ima razdiobu iz eksponencijalne familije ako je njena funkcija gustoće oblika*

$$f(y, \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{1.6}$$

za neke funkcije a, b i c. Parametar θ zovemo prirodni parametar, dok je ϕ parametar disperzije ili skaliranja.

Lako se pokaže da vrijedi

$$E[Y] = \mu = b'(\theta) \tag{1.7}$$
$$Var[Y] = a(\phi)b''(\theta) \tag{1.8}$$

gdje su b' i b'' prva, odnosno druga derivacija funkcije b. Vidimo da očekivanje ne ovisi o parametru ϕ dok varijanca općenito ovisi o oba parametra. Može se pokazati da je funkcija b neprekidna i invertibilna, osim u nekim specijalnim slučajevima pa možemo definirati **funkciju varijance**

$$V(\mu) = b''(\theta) \tag{1.9}$$

Sada varijancu od Y možemo pisati i kao

$$Var[Y] = a(\phi)V(\mu) \tag{1.10}$$

Za GLM imamo da vrijedi

$$g(\mu_i) = g(b'(\theta_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Iz gornje jednakosti možemo definirati **prirodnu funkciju veze** kao

$$g = (b')^{-1} \tag{1.11}$$

$$\Rightarrow g(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

Primjer 1.1.2. *Neka je Y slučajna varijabla s normalnom distribucijom s parametrima μ i σ^2. Funkcija gustoće slučajne varijable Y je*

$$f_Y(y, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \tag{1.12}$$
$$= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma)\right)\right] \tag{1.13}$$

Iz (1.6) i (1.13) slijedi da je

$\theta = \mu$

$\phi = \sigma^2$

$a(\phi) = \phi$

$b(\theta) = = \frac{\theta^2}{2}$

$$c(y,\phi) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma)\right)$$

Nadalje, iz izraza (1.7) i (1.8) možemo odrediti očekivanje i varijancu slučajne varijable Y. Slijedi

$$b(\theta) = \frac{\theta^2}{2} \Rightarrow E[Y] = b'(\theta) = \theta = \mu \tag{1.14}$$

$$a(\phi) = \phi \Rightarrow Var[Y] = a(\phi)b''(\theta) = \phi = \sigma^2 \tag{1.15}$$

što smo već i znali od prije.

U sljedećoj tablici su prikazane osnovne karakteristike za neke distribucije eksponencijalne familije. Vidi [14] Table 2.1.

	Normalna	Poisson	Binomna	Gama	Inverzna Gausova
Oznaka	$\mathrm{N}(\mu,\sigma^2)$	$\mathrm{P}(\mu)$	B(m,p)/m	$\mathrm{G}(\mu,\nu)$	$\mathrm{IG}(\mu,\sigma^2)$
Im(y)	$\langle-\infty,\infty\rangle$	$\langle 0,\infty\rangle$	$\langle 0,1]$	$\langle 0,\infty\rangle$	$\langle 0,\infty\rangle$
ϕ	σ^2	1	$\frac{1}{m}$	ν^{-1}	σ^2
$b(\theta)$	$\frac{\theta^2}{2}$	e^θ	$\log(1+e^\theta)$	$-\log(-\theta)$	$-\sqrt{-2\theta}$
c(y,ϕ)	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma)\right)$	$-\log y!$	$\log\binom{m}{my}$	$-\log\Gamma(\nu)$	$-\frac{1}{2}\left(log(2\pi\phi y^3) + \frac{1}{\phi y}\right)$
E[y|θ]	θ	e^θ	$\frac{e^\theta}{1+e^\theta}$	$-\frac{1}{\theta}$	$\frac{1}{\sqrt{-2\theta}}$
$\theta(\mu)$	id_μ	log	logit	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
$\mathrm{V}(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

Tablica 1.1: Karakteristike nekih od distribucija eksponencijalne familije

1.2 Poissonova linearna regresija

U ovome poglavlju ćemo se fokusirati na jedan GLM, a to je Poissonova regresija. Kao što se da naslutiti iz prethodnog poglavlja, komponente Y_i zavisne varijable Y, su međusobno nezavisne i imaju Poissonovu distribuciju, koja pripada eksponencijalnoj familiji. Prema (1.5) moramo još pronaći funkciju vezu i time ćemo točno odrediti model Poissonove regresije.

Definicija 1.2.1. *Kažemo da slučajna varijabla Y ima Poissonovu distribuciju s parametrom $\mu > 0$ i pišemo $Y \sim P(\mu)$ ako je njena funkcija vjerojatnosti*

$$f_Y(y) = \frac{\mu^y}{y!}e^{-\mu}, \qquad y \in ImY = \{0, 1, 2, \dots\} \tag{1.16}$$

Uočimo da izraz (1.16) možemo zapisati i na sljedeći način

$$f_Y(y) = \frac{\mu^y}{y!}e^{-\mu} = \exp(y \log \mu - \mu - \log y!) \tag{1.17}$$

Iz (1.6) i (1.17) slijedi

$\theta = \log \mu$

$\phi = 1 \Rightarrow\ a(\phi) = 1$

$b(\theta) = e^\theta$

$c(y, \phi) = -\log y!$

Dakle vrijedi,

- Prirodni parametar Poissonove distribucije je $\log \mu$
- Iz (1.7) slijedi da je E[Y] = μ
- Iz (1.9) i (1.10) slijedi Var[Y] = μ
- Iz (1.11) slijedi da je prirodna funkcija veze $g(\mu) = \log \mu$, zbog $\mu = g^{-1}(\eta)$

$$\Rightarrow \mu = e^\eta = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

Uzimajući u obzir sve prethodne tvrdnje, model Poissonove regresije možemo opisati kao sljedeće.
Neka su $y_1, \dots, y_n$ gdje je $n \in \mathbb{N}$ opservacije koje su realizacije nezavisnih slučajnih varijabli $Y_1, \dots, Y_n$ takvih da vrijedi $Y_i \sim P(\mu_i)$. Neka su $x_1, \dots, x_k$ prediktorske varijable te $\beta_0, \beta_1, \dots, \beta_k \in \mathbb{R}$, $\forall k \in \mathbb{N}$. Neka je

$$\eta = X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

gdje je X matrica dizajna definirana kao u (1.3).
Slučajna varijabla $Y = (Y_1, \dots, Y_n)$ je s kovarijatama $x_1, \dots, x_k$ povezana na sljedeći način

$$\mu = E[Y|x] = e^\eta = e^{X\beta} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \tag{1.18}$$

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

Preostaje nam još procijeniti koeficijente $\beta_0, \beta_1, \ldots, \beta_k$, a to činimo metodom najveće vjerodostojnosti (eng. *maximum likelihood estimation*).

Iz Definicije 1.2.1 i (1.18) vrijedi

$$\mathbb{P}(Y|x,\beta) = \frac{\mu^y}{y!}e^{-\mu} = \frac{e^{y\beta x}}{y!}e^{-e^{\beta x}} \tag{1.19}$$

$$= \prod_{i=1}^{n} \frac{e^{y_i\beta x_i}}{y_i!}e^{-e^{\beta x_i}} \tag{1.20}$$

Sada želimo pronaći skup koeficijenata $\beta = \{\beta_0, \beta_1, \ldots, \beta_k\}$ tako da je gornja vjerojatnost maksimalna.

Definicija 1.2.2. *Neka je $(x_1, \ldots, x_n)$ opaženi uzorak slučajne varijable X s gustoćom $f(x|\theta)$ gdje je $\theta = (\theta_1, \ldots, \theta_k)$, $n, k \in \mathbb{N}$. Funkcija vjerodostojnosti parametra θ je*

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.21}$$

Iz Definicije 1.2.2 i (1.20) slijedi da je funkcija vjerodostojnosti parametra β jednaka

$$L(\beta|x,Y) = \prod_{i=1}^{n} \frac{e^{y_i\beta x_i}}{y_i!}e^{-e^{\beta x_i}} \tag{1.22}$$

Prethodnu jednakost možemo logaritmirati pa dobivamo log-vjerodostojnost

$$l(\beta|x,Y) = \log L = \sum_{i=1}^{n}(y_i\beta x_i - e^{\beta x_i} - \log(y_i!)) \tag{1.23}$$

Zbog toga što je funkcija $f(x) = \log x$ monotona, maksimum vjerodostojnosti će se postići u istim točkama kao i maksimum log vjerodostojnosti. Uočimo, da možemo izbaciti $-\log(y_i!)$ iz sume u (1.23) jer on neće utjecati na rezultat pošto ne sadrži β. Dakle, možemo promatrati

$$l(\beta|x,Y) = \sum_{i=1}^{n}(y_i\beta x_i - e^{\beta x_i}) \tag{1.24}$$

Napokon, da bismo pronašli maksimum moramo riješiti sustav jednadžbi

$$\frac{\partial l(\beta|x,Y)}{\partial \beta} = 0 \tag{1.25}$$

Time ćemo dobiti procijenjene parametre Poissonove regresije. Dobivene procjene označimo sa $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\}$. Nadalje, željeli bismo znati kolika je preciznost procijenjenih parametara, a tu mjeru ćemo dobiti s pouzdanim intervalima. Dakle, od interesa nam je pronaći pouzdane intervale za parametre $\hat{\beta}$.

Definicija 1.2.3. *Neka je X slučajni uzorak iz populacije s jednodimenzionalnim parametrom θ. $(1-\alpha)\cdot 100\%$ pouzdan interval za θ je slučajni interval $[\hat{\theta}_1, \hat{\theta}_2]$ tako da*

$$\mathbb{P}(\theta_1(X) \leq \theta \leq \theta_2(X)) = 1 - \alpha \tag{1.26}$$

*Kažemo da je **vjerojatnost pokrivanja intervala** $[\hat{\theta}_1, \hat{\theta}_2]$ jednaka $1-\alpha$.*

Uočimo, θ je prava vrijednost parametra dok su granice, θ_1 i θ_2 slučajne varijable. Najčešće određujemo 90%, 95% i 99% pouzdane intervale. Njihova interpretacija je sljedeća. Neka imamo dovoljno veliki broj opažanja tog intervala. Tada će u (1-α) $\cdot$100% slučajeva, prava vrijednost parametra θ biti unutar intervala, dok se u $\alpha\cdot 100\%$ slučajeva to neće dogoditi. Pouzdani intervali nisu jedinstveni te postoji nekoliko metoda njihove konstrukcije, no mi ćemo se koncentrirati na jedan.

Ako imamo veliki uzorak populacije, prema centralnom graničnom teoremu vrijedi sljedeće

$$\frac{\hat{\beta}_j - \beta_j}{\sigma(\hat{\beta}_j)} \sim \mathcal{N}(0,1) \qquad\qquad j = 0, 1, \ldots, k \tag{1.27}$$

pa iz toga slijedi da je $(1-\alpha)\cdot 100\%$ pouzdan interval za β_j, $j = 0, 1, \ldots, k$

$$[\hat{\beta}_j - z_{(1-\frac{\alpha}{2})} \cdot \sigma(\hat{\beta}_j), \hat{\beta}_j + z_{(1-\frac{\alpha}{2})} \cdot \sigma(\hat{\beta}_j)] \tag{1.28}$$

pri čemu je $\sigma(\hat{\beta}_j)$ standardna devijacija procjene za β_j, a $z_{(1-\frac{\alpha}{2})}$ je $(1-\frac{\alpha}{2})$ kvantil standardne normalne distribucije. Time smo definirali **Waldove pouzdane intervale** za koeficijente β_j, $\forall j \in \{0, 1, \ldots, k\}$.

U ovisnosti o veličini β_j mijenja se i očekivana vrijednost za Y. Naime, koeficijente regresije možemo interpretirati na sljedeći način

1. Ako je $\beta_j = 0$, tada prediktorska varijabla x_j nije značajna u regresijskom modelu.

2. Ako je $\beta_j > 0$, tada je $\mu = E[Y|x]$ veće za e^{β_j} nego kada je $x_j = 0$.

3. Ako je $\beta_j < 0$, tada je $\mu = E[Y|x]$ manje za e^{β_j} nego kada je $x_j = 0$.

1.3 Primjeri korištenja Poissonove regresije

Poissonova regresija se koristi u mnogim primjerima gdje želimo modelirati diskretne podatke iz skupa $\mathbb{N}_0$ u ovisnosti o nekim prediktorskim nezavisnim varijablama. Najčešće ti podaci opisuju rijetke događaje u nekom vremenskom intervalu kao na primjer:

- Broj kupovina namjernica kućanstva određenom supermarketu u tjedan dana.
- Broj kvarova nekog stroja u različitim radnim uvjetima.
- Broj kreditnih kartica po osobi.
- Broj bakterija na nekom području.
- Broj poziva u telefonskoj centrali u jednom satu.

Događaji koji su opisani Poissonovom distribucijom često se promatraju kroz neki vremenski period, npr. dan, tjedan, mjesec, godina pa stoga funkciju vjerojatnosti za Poissonovu slučajnu varijablu u Definiciju 1.2.1 možemo pisati

$$f_Y(y) = \frac{(t\mu)^y}{y!} e^{-t\mu}, y \in ImY = \{0, 1, 2, \dots\} \tag{1.29}$$

gdje je t vremensko razdoblje u kojem se promatra neki događaj. Na primjer, ako promatramo vremensko razdoblje od 30 dana, onda nam $t = \frac{7}{30}$ predstavlja jedan tjedan, dok je $t = 1$ mjesec dana.
Navedimo sada jedan poznati primjer korištenja Poissonove regresije.

Primjer 1.3.1. *(Rakovi i sateliti) U istraživanju su prikupljeni podaci o ponašanju rakova bodljaša, pogleda J. Brockmann, Ethology 1996., Agresti (2007) Sec. 3 i Agresti (2013) Sec. 4.3.*
Promatrani su ženski rakovi bodljaši koji su imali muškog raka prikvačenog za njihovo tijelo. Osim muškog raka koji je na tijelu, u blizini ženka može imati još muških rakova, koji se zovu sateliti. Želimo napraviti model koji će predvidjeti broj muških satelita, ovisno o karakteristikama ženskih rakova. Prediktorske varijable su: boja raka, stanje kralježnice, širina oklopa i težina raka.

Poglavlje 2

Analiza

2.1 Opis problema

Kod modeliranja podataka regresijskim modelom pa tako i Poissonovim želimo odgovoriti na neka pitanja. Na primjer, koliko veliki uzorak nam je potreban da bismo mogli dobro opisati podatke, koja kovarijata značajno utječe na zavisnu varijablu, je li dobiveni model dobar model i slično. U ovome radu ćemo se baviti pitanjima vezano za regresijske koeficijente i njihove pouzdane intervale.
U prethodnom poglavlju smo rekli da su zavisna varijabla i kovarijate linearno povezani. Preciznije, zavisna varijabla je linearna kombinacija kovarijata x_i i parametara β_i. Pronaći model koji opisuje zavisnu varijablu, znači pronaći koeficijente β_i. Uz dobivenu procjenu, možemo odrediti njihove 95% i 99% Waldove pouzdane intervale. Ono što nas zanima jest, na koji način različiti faktori utječu na vjerojatnost pokrivanja tih pouzdanih intervala.
Prvi faktor koji promatramo je tip distribucije prediktorskih varijabli. Zanima nas što se događa s vjerojatnosti pokrivanja ako znamo da prediktorska varijabla dolazi iz neke određene distribucije te ima li razlike ako dolazi iz neke druge distribucije.
Drugi faktor koji promatramo je duljina ili veličina uzorka. Promatrat ćemo ponašanje vjerojatnosti pokrivanja s obzirom na različite duljine uzoraka.
Sljedeće što promatramo je broj prediktorskih varijabli. Naime, želimo vidjeti postoji li razlika u slučaju da imamo jednu ili dvije prediktorske varijable te ako imamo više prediktorskih varijabli, jesu li vjerojatnosti pokrivanja pouzdanih intervala za β_i međusobno različite.
Zadnji faktor koji gledamo je korelacija između prediktorskih varijabli. Naime, nije rijetko da među prediktorskim varijablama postoje neke koje su međusobno korelirane. Na primjer, visina i težina zajedno ulaze u mnoge modele,a znamo da postoji korelacija između te dvije vrijednosti. Zbog toga ćemo promatrati kako različita razina

korelacije između kovarijata utječe na vjerojatnost pokrivanja pouzdanih intervala. Naravno, moglo bi se dogoditi da neka određena kombinacije faktora čini razliku, stoga ćemo analizirati vjerojatnost pokrivanja Waldovih pouzdanih intervala za sve kombinacije gore navedenih faktora.
Za ispitivanje utjecaja navedenih faktora provodit ćemo Monte Carlo studiju u programskom softveru `SAS`.

2.2 Korelacija i VIF

U ovome odjeljku ćemo navesti definicije i teoreme koje ćemo koristiti.
Vjerojatnost pokrivanja pouzdanog intervala definirana je u Definiciji 1.2.3. Kod diskretnih populacije poput Poissonove, često ne možemo dobiti da vjerojatnost pokrivanja $[\hat{\theta}_1(X), \hat{\theta}_2(X)]$ bude točno $(1-\alpha)$ stoga u definiciji zahtijevamo $\geq (1-\alpha)$.

Jedan od faktora koji promatramo je korelacija prediktorskih varijabli.

Definicija 2.2.1. ***Kovarijanca*** *slučajnih varijabli X i Z je broj*

$$cov[X, Z] := E[(X - E[X])(Z - E[Z])] \tag{2.1}$$

ako desna strana postoji.(2.1) je ekvivalentno

$$cov[X, Z] = E[XZ] - E[X]E[Z] \tag{2.2}$$

Uočimo, $cov[X, X] = Var[X]$.

Korolar 2.2.2. *Neka su X,Z slučajne varijable za koje postoji kovarijanca. Neka su $\alpha, \beta, \gamma, \delta \in \mathbb{R}$, tada vrijedi*

$$cov[\alpha X + \beta, \gamma Z + \delta] = \alpha\gamma cov[X, Z] \tag{2.3}$$

Dokaz. Vrijedi

$$E[\alpha X + \beta] = \alpha E[X] + \beta \quad i \quad E[\gamma Z + \delta] = \gamma E[Z] + \delta$$

Iz toga slijedi da je

$$\alpha X + \beta - E[\alpha X + \beta] = \alpha(X - E[X])$$
$$\gamma Z + \delta - E[\gamma Z + \delta] = \gamma(Z - E[Z])$$

Dobivamo

$$\begin{aligned} cov[\alpha X + \beta, \gamma Z + \delta] &= E[\alpha(X - E[X])\gamma(Z - E[Z])] \\ &= \alpha\gamma E[(X - E[X])(Z - E[Z])] \\ &= \alpha\gamma cov[X, Z] \end{aligned}$$

□

Kovarijanca slučajnih varijabli mjeri njihovu linearnu povezanost. Istu povezanost možemo mjeriti i **koeficijentom korelacije** što je zapravo kovarijanca između standardiziranih slučajnih varijabli. Vrijedi

$$\rho = corr[X, Z] := E[\frac{X - E[X]}{\sigma(X)} \cdot \frac{Z - E[Z]}{\sigma(Z)}] \tag{2.4}$$

$$\overset{(2.3)}{=} \frac{E[(X - E[X])(Z - E[Z])]}{\sigma(X)\sigma(Z)} \tag{2.5}$$

$$\overset{(2.1)}{=} \frac{cov[X, Z]}{\sigma(X)\sigma(Z)} \tag{2.6}$$

Za koeficijent korelacije vrijedi

$$-1 \leq \rho \leq 1 \tag{2.7}$$

Korolar 2.2.3. *Neka su X,Z slučajne međusobno nezavisne varijable. Tada je korelacija između X,Z jednaka 0 i kažemo da su slučajne varijable X i Z nekorelirane.*

Dokaz. Tvrdnja jednostavno slijedi iz relacije (2.2) i linearnosti očekivanja. Naime vrijedi

$$\rho = \frac{cov[X, Z]}{\sigma(X)\sigma(Z)} \overset{(2.2)}{=} \frac{E[XZ] - E[X]E[Z]}{\sigma(X)\sigma(Z)} \overset{nez.}{=} \frac{E[X]E[Z] - E[X]E[Z]}{\sigma(X)\sigma(Z)} = 0 \tag{2.8}$$

□

Neka sada imamo višestruku linearnu vezu

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.9}$$

Procjenom parametara β_i, $\forall i = 0, \ldots, k$ dobivamo model koji opisuje zavisnu varijablu y u ovisnosti o $x_1, \ldots, x_k$.
Sada bismo se mogli pitati je li dobiveni model dobar model? Jedna od mjera koja nam govori koliko je model dobar je **koeficijent determinacije** R^2, koji nam govori koliko je varijabilnosti podataka objašnjeno modelom. Što je model bliže 1 to je bolje.

Definicija 2.2.4. *Omjer varijabilnosti objašnjene linearnim modelom i ukupne varijabilnosti zove se koeficijent determinacije i definiran je kao*

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.10}$$

gdje je

$\bar{y} = \frac{1}{n}\sum y_i$

$SS_{res} = \sum(y_i - \hat{y}_i)^2$

$SS_{tot} = \sum(y_i - \overline{y})^2$

pri čemu su y_i stvarne vrijednosti, a $\hat{y}_i$ vrijednosti dobivene iz procijenjenog modela.

U linearnim vezama s više prediktorskih varijabli može se dogoditi da su prediktorske varijable međusobno korelirane. Gotovo u svim slučajevima, ne želimo visoku korelaciju između varijabli jer bi to moglo dovesti do pogrešnih rezultata. Pretpostavimo da je varijabla x_t za $t \in 1, \ldots, k$ korelirana s jednom ili više varijabli. Jedan od indikatora multikreliranosti je **faktor inflacije varijance** (VIF). On nam govori koliko puta je varijanca svakog koeficijenta povećana. Na primjer, ako je $VIF_t = 1.6$, onda kažemo da je varijanca koeficijenta β_t veća za 60% od očekivane varijance kada ne bi postojala međusobna korelacija.
Faktor inflacije varijance je definiran kao

$$VIF_t = \frac{1}{1 - R_t^2} \tag{2.11}$$

pri čemu je R_t^2 koeficijent determinacije modela u kojemu nam je x_t zavisna varijabla, a prediktorske varijable su sve druge osim nje. Ako je R_t^2 veliki, to znači da varijablinost varijable x_t možemo objasniti pomoću drugih varijabli, a to povlači veću vrijednost VIF_t. Za više o faktoru inflacije varijance možete pogledati u [1].

2.3 Monte Carlo metoda

Monte Carlo (MC) metode ili simulacije su vrlo bitni i korisni računalni algoritmi koji imaju vrlo veliku primjenu u matematici, fizici, financijama, bankarstvu i mnogim drugim područjima. Ime su dobili po gradu Monte Carlo koji je poznat po mnogim kockarnicama u kojima se igraju mnoge igre poput bacanja kockica, ruleta i sl. koje se temelje na slučajnom ishodu. Ovaj pristup se zasniva na generiranju velikoga broja slučajnih brojeva i slučajnih uzoraka s ciljem dobivanja procjene nekoga

svojstva. Često se koriste kada trebamo napraviti neku procjenu ili donijeti odluku, a pri tome postoji određena nesigurnost. Omogućuje uvid kako se ponašaju određeni parametri s obzirom na različite uvjete koje se mogu dogoditi te tako omogućuju poboljšanje usluga, uštedu resursa, otkrivanja zakonitosti pronalaženju distribucija događaja i slično.

Uobičajeni tijek Monte Carlo simulacije je sljedeći:

1. Generirati N međusobno nezavisnih uzoraka koji zadovoljavaju uvjete koji nas zanimaju.
2. Izračunati vrijednost parametra T za svaki od uzorka $\Rightarrow T_1, \ldots, T_N$.
3. Ako je N dovoljno veliki, koristeći vrijednosti $T_1, \ldots, T_N$ bismo trebali dobiti dobru aproksimaciju stvarne vrijednosti za parametra T.

Kao kod svih procjena, možemo se zapitati koliko je dobra procjena dobivena MC metodom. Stoga, prije provođenja same simulacije potrebno je postaviti neke određene uvjete koje će nam osigurati da je dobivena procjena dovoljno dobra. Za početak, potrebno je definirati

1. Veličina odnosno parametar koji se procjenjuje
2. Dozvoljena pogreška procjene

U našem slučaju, koristit ćemo MC metodu da bismo provjerili je li vjerojatnost pokrivanja pouzdanih intervala uistinu ona koju očekujemo. Dakle, za koeficijente Poissonove regresije smo odredili pouzdane intervale kao u (1.28).

Pitamo se je li vjerojatnost pokrivanja zaista $\geq 1 - \alpha$? Da bismo to provjerili koristimo tako zvanu *Hits and Miss* metodu. Naime, za N realizacija intervala ćemo promatrati je li stvarna vrijednost upala u procijenjeni pouzdani interval. Možemo reći da je problem određivanja vjerojatnosti pokrivanja ekvivalentan pronalaženju vrijednosti parametra p Binomne razdiobe u N nezavisnih pokušaja, pri čemu je p vjerojatnost da će stvarna vrijednost upasti u procijenjeni interval.
Definirajmo povoljan događaj kao onaj u kojoj se stvarna vrijednost nalazi unutar pouzdanog intervala te neka je p vjerojatnost da se povoljni događaj dogodio. Tada imamo

$$\text{Z = broj povoljnih događaja} \Rightarrow Z \sim B(N, p)$$

Procjenu za parametar p ćemo dobiti kao omjer broja povoljnih događaja u N ponavljanja tj.

$$\hat{p} = \frac{broj\ \ povoljnih\ \ događaja}{broj\ \ ukupnih\ \ događaja} = \frac{Z}{N}$$

Sada možemo odrediti očekivanje i standardna pogreška procjene proporcije za $\hat{p}$. Vrijedi

$$E[\hat{p}] = E\left[\frac{Z}{N}\right] = \frac{1}{N}E[Z] = \frac{1}{N} \cdot Np = p$$
$$\sigma_{\hat{p}} = \sqrt{Var\left(\frac{Z}{N}\right)} = \sqrt{\frac{p(1-p)}{N}}$$

pri čemu je p stvarna vrijednost proporcije. U našem slučajevima, očekivane stvarne vrijednosti za p bi trebale biti 0.95 te 0.99.
Na početku smo rekli da zadajemo dozvoljenu pogrešku procjene odnosno, željeli bismo da se procijenjena vrijednost ne razlikuje od prave za više od neke vrijednosti d. Iz gornjih izraza vidimo da će greška biti manja ukoliko je N veći. Odredimo broj ponavljanja pokusa N, tako da nam pogreška ne bude veća od neke vrijednosti d.

$$d = \sqrt{\frac{p(1-p)}{N}} \Rightarrow N = \frac{p(1-p)}{d^2}$$

Ako za stavimo d = 0.01, za $p_1 = 0.95$ i $p_2 = 0.99$ dobivamo

$$N_1 = \frac{0.95 \cdot 0.05}{0.01^2} = 475$$
$$N_2 = \frac{0.99 \cdot 0.01}{0.01^2} = 99$$

Dakle, ako bismo željeli da nam greška procjene parametra bude manja od 0.01, trebali bismo izvršiti barem 475 ponavljanja u slučaju kada nam je p = 0.95 te najmanje 99 ponavljanja u slučaju kada je p = 0.99.
U ovome radu ćemo koristiti N = 1000 kako bismo dobili bolji rezultat. Za N = 1000 dobivamo da nam je greška

$$d_1 = \sqrt{\frac{0.95(1-0.95)}{1000}} \approx 0.007$$
$$d_2 = \sqrt{\frac{0.99(1-0.99)}{1000}} \approx 0.003$$

2.4 Podaci

Ukratko smo opisali glavni cilj ovoga rada, a sada ćemo se posvetiti konkretnim podacima koje koristimo. Kao što smo već rekli, želimo vidjeti na koji način različiti faktori utječu na vjerojatnost pokrivanja 95% i 99% Waldovih pouzdanih intervala za regresijske koeficijente u modelu Poissonove regresije.

Eksperiment ćemo provesti za sve kombinacije faktora:

- veličina uzorka $n = 20, 50, 100, 200$
- tip distribucije prediktorskih varijabli:

  Normalna (0,1)

  Laplaceova

  Gamma (sa shape parametrom 0.5)

  Kontaminirana normalna (90% podataka iz N(0,1) a 10% podataka iz N(0,3.2))
- broj prediktorskih varijabli: 1, 2
- koreliranost prediktorskih varijabli: 0.3, 0.5, 0.7, 0.9

Očito, koreliranost prediktorskih varijabli ćemo promatrati samo u slučaju kada u modelu imamo dvije varijable.

Kako bismo željeli usporediti rezultate, za sve distribucije ćemo staviti da je očekivanje $\mu = 0$ i standardna devijacija $\sigma = 1$. Za neke distribucije koje ćemo koristiti, nije odmah vidljivo na koji način moramo simulirati podatke da bismo dobili da je $\mu = 0$ i $\sigma = 1$. Stoga, moramo odrediti s kojim će parametrima biti simulirani podaci u svakoj distribuciji da bismo ispunili traženi uvjet.

Određivanje parametara

Normalna distribucija

Iz Primjera 1.1.2. znamo da ako slučajna varijabla dolazi iz N(0,1), onda joj je očekivanje $\mu = 0$ i standardna devijacija $\sigma = 1$ pa imamo zadovoljen uvjet simulacije.

Laplaceova distribucija

Definicija 2.4.1. *Slučajna varijabla X dolazi iz Laplace*(λ, b) *distribucije ako joj je funkcija gustoće*

$$f_X(x) = \frac{1}{2b} e^{-\frac{|x-\lambda|}{b}} \tag{2.12}$$

Tada prema [18] vrijedi

$$\mathbb{E}X = \lambda$$
$$VarX = 2b^2$$

Sada vidimo da bismo zadovoljili uvjet za simulaciju, moramo riješiti sustav jednadžbi

$$\lambda = 0$$
$$2b^2 = 1$$

Dakle, simulirat ćemo slučajne vrijednosti iz *Laplace*$(0, \frac{\sqrt{2}}{2})$.

Gama distribucija

Zbog traženog uvjeta, promatrat ćemo generaliziranu gama distribuciju koja se još u literaturi naziva i treći tip Pearsonove distribucije.

Definicija 2.4.2. *Neka je* $X \sim \Gamma(\alpha, \beta, \theta)$ *tada je njegova funkcija gustoće*

$$f_X(x) = \frac{1}{\beta\Gamma(\alpha)} \Big(\frac{x-\theta}{\beta}\Big)^{\alpha-1} e^{-\frac{x-\theta}{\beta}} \tag{2.13}$$

Prema [19] vrijedi,

$$\mathbb{E}X = \alpha\beta + \theta$$
$$VarX = \alpha\beta^2$$

U zadatku imamo zadano da nam je $\alpha = 0.5$ pa onda moramo riješiti sustav

$$0.5\beta + \theta = 0$$
$$0.5\beta^2 = 1$$

Rješavanjem dobivamo da moramo simulirati podatke iz $\Gamma(\frac{1}{2}, \sqrt{2}, -\frac{\sqrt{2}}{2})$.

Kontaminirana normalna distribucija

Za kontaminiranu normalnu distribuciju imamo određene sve parametre za simulaciju. Dakle, 90% uzorka simuliramo iz N(0,1) dok preostalih 10% simuliramo iz N(0,3.2). Nakon što smo dobili uzorak, moramo ga standardizirati kako bismo dobili da je očekivanje $\mu = 0$ i standardna devijacija $\sigma = 1$ tj., koristimo formulu

$$x_{stand} = \frac{x - \mu_u}{\sigma_u}$$

gdje su μ_u i σ_u očekivanje i standardna devijacija uzorka.

2.5 Simulacija i model

Kako smo odredili parametre za sve distribucije, možemo simulirati podatke. Ono što ćemo raditi je procjena koeficijenta i njihovih pouzdanih intervala u Poissonovoj regresiji u kojoj su nam poznate stvarne vrijednosti koeficijenata.
Koristit ćemo Monte Carlo simulacije, odnosno svaku kombinaciju faktora ćemo ponavljati veliki broj puta, u našem slučaju 1000, kako bismo dobili što bolje procijene parametara.
Vrijednosti stvarnih regresijskih koeficijenata su:

- u modelu s jednim prediktorom $\beta_0 = 2$, $\beta_1 = -4$
- u modelu s dva prediktora $\beta_0 = 2$, $\beta_1 = -4$, $\beta_2 = 2$

(2.14)

Sada ćemo opisati simulaciju podataka na kojima ćemo temeljiti rezultate. U slučaju jedne prediktorske varijable imamo sljedeći algoritam. Uzmimo jednu distribuciju iz koje će nam dolaziti prediktorska varijable. Neka je n duljina uzorka.

1. Na slučajan način odredimo vrijednost za varijablu x iz odabrane distribucije.
2. Računamo $\eta = 2 - 4x$
3. Za vrijednost y uzimamo slučajan broj iz Poissonove distribucije s parametrom e^{η}

Ako u modelu imamo dvije prediktorske varijable, tada koristimo sljedeći algoritam. Fiksirajmo distribuciju iz koje će nam dolaziti vrijednosti za prediktorske varijable, neka je n duljina uzorka te odredimo jesu li varijable međusobno nezavisne ili korelirane s koeficijentom ρ.

1. Na slučajan način odredimo vrijednost za varijable x_1 i x_2 iz odabrane distribucije tako da varijable budu nezavisne ili međusobno korelirane s koeficijentom ρ.

2. Računamo $\eta = 2 - 4x_1 + 2x_2$

3. Za vrijednost y uzimamo slučajan broj iz Poissonove distribucije s parametrom e^{η}

U oba algoritma smo dobili jedan simulirani broj iz Poissonove distribucije. Gornji račun ponavljamo n puta da bismo dobili jedan uzorak,a potrebno nam je 1000 takvih uzoraka.
Ponavljanjem gornjih algoritama dobili smo skup podataka kojemu ćemo sada pristupiti na *"obrnuti"* način. Naime, za te simulirane podatke ćemo tražiti odgovarajući model Poissonove regresije. Dakle, znamo stvaran model, ali sad nas zanima koliko će biti dobre procijenjene vrijednosti.
Za svaki od 1000 uzoraka duljine n računamo sljedeće:

Neka je $i = 1, \ldots, 1000$ redni broj uzorka. Vrijednosti zavisne varijable $Y^{(i)} = (y_1, \ldots, y_n)$ su dobivene u i-toj simulaciji kao i vrijednosti prediktorskih varijabli $x_1^{(i)} = (x_{11}^{(i)}, \ldots, x_{1n}^{(i)})$ i $x_2^{(i)} = (x_{21}^{(i)}, \ldots, x_{2n}^{(i)})$.
Sada za taj uzorak možemo odredimo koji Poissonov regresijski model $y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ najbolje opisuje $Y^{(i)}$ u ovisnosti o $x_1^{(i)}$ i $x_2^{(i)}$.
Određivanjem modela, za svaki uzorak dobivamo neke procjene za koeficijente β i njihove pouzdane intervale. Sada promatramo jesu li stvarni pripadni koeficijenti, oni definirani u (2.14), sadržani u pouzdanim intervalima njihovih procjena. Potom, računamo omjer broja onih koji su sadržani i onih koji nisu. Dobivena proporcija nam govori kolika je stvarna vjerojatnost pokrivanja Waldovih pouzdanih intervala za pripadne koeficijente.

Dakle, za sve moguće kombinacije faktora imat ćemo po 1000 replikacija. Za svaku od tih replikacija ponavljamo gore opisanu simulaciju kako bismo dobili skup podataka na kojem ćemo primijeniti Poissonovu regresiju te naposljetku odrediti stvarne vjerojatnosti pokrivanja Waldovih pouzdanih intervala za koeficijente Poissonove regresije.

Poglavlje 3

Rezultati

U ovom poglavlju ćemo iznijeti dobivene rezultate za sve kombinacije faktora:

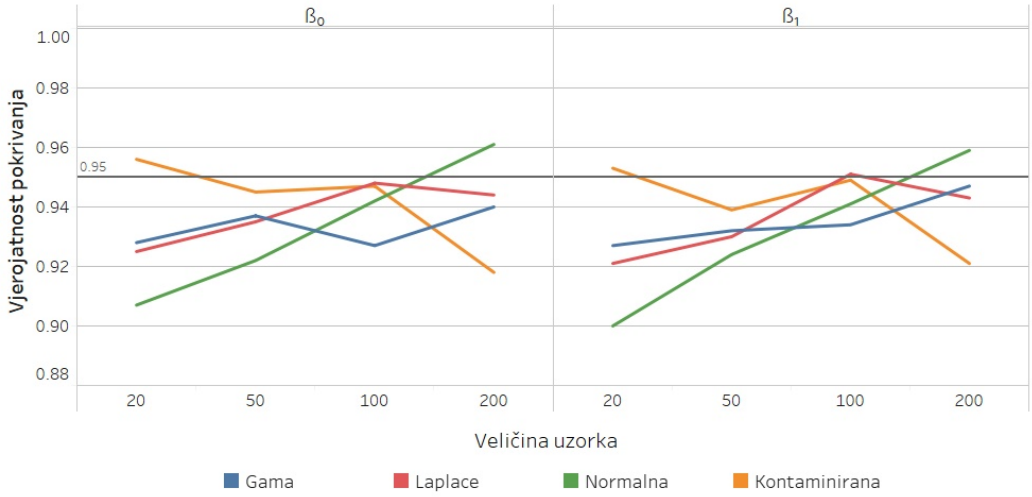
- tip distribucije
- duljina uzoraka (n = 20, 50, 100, 200)
- vrsta modela (jedna ili dvije prediktorske varijable)
- koreliranost varijabli
- vrsta pouzdanog intervala (95% i 99%)

3.1 Jednoparametarski modeli

U tablicama 3.1 i 3.2 se nalaze vjerojatnosti pokrivanja za 95% i 99% pouzdane intervale u slučaju samo jedne prediktorske varijable. Grafički prikaz vrijednosti vidljiv je na slikama 3.1 i 3.2. Vjerojatnosti promatramo po parametrima β_0, β_1 za svaku različitu distribuciju uzimajući u obzir duljinu uzorka.
Ako pogledamo dobivene rezultate, možemo primijetiti da je u većini slučajeva vjerojatnost pokrivanja pouzdanih intervala manja od očekivane. Nadalje, vidljivo je da kada kovarijate dolaze iz gama, Laplace i normalne distribucije, vjerojatnost pokrivanja raste kako raste duljina uzorka, ali kada je kovarijata iz kontaminirane normalne distribucije, tada je vjerojatnost pokrivanja najveća za n = 20, a najmanja za n = 200. Uočimo da se za normalnu distribuciju postiže očekivana vrijednost kada je duljina uzorka n = 200.

Parametar	n	Normalna	Laplace	Gama	Kontaminirana
β_0	20	0.907	0.925	0.928	0.956
β_0	50	0.922	0.935	0.937	0.945
β_0	100	0.942	0.948	0.927	0.947
β_0	200	0.961	0.944	0.940	0.918
β_1	20	0.900	0.921	0.927	0.953
β_1	50	0.924	0.930	0.932	0.939
β_1	100	0.941	0.951	0.934	0.949
β_1	200	0.959	0.943	0.947	0.921

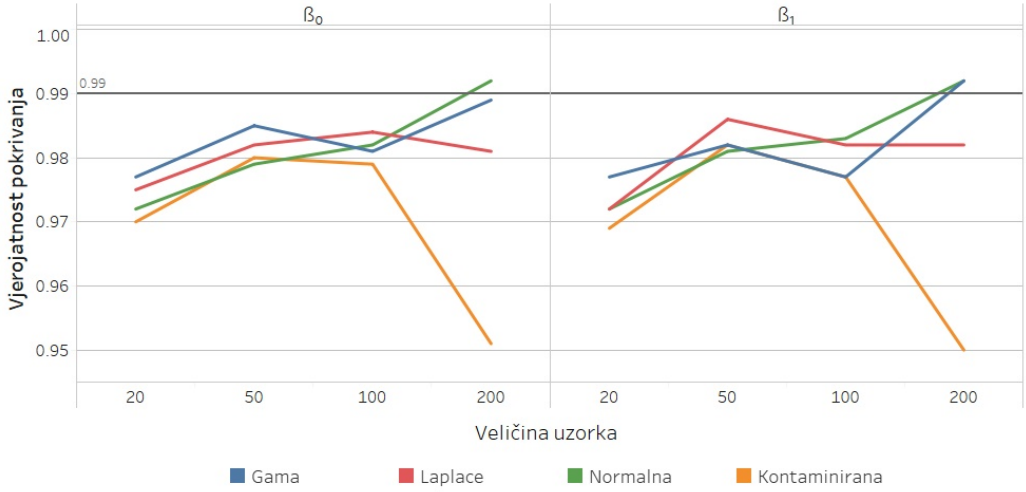
Tablica 3.1: Vjerojatnost pokrivanja 95% pouzdanih intervala za jednoparametarske modele



Slika 3.1: Vjerojatnost pokrivanja 95% pouzdanog intervala u jednoparametarskom modelu

Parametar	n	Normalna	Laplace	Gama	Kontaminirana
β_0	20	0.972	0.975	0.977	0.970
β_0	50	0.979	0.982	0.985	0.980
β_0	100	0.982	0.984	0.981	0.979
β_0	200	0.992	0.981	0.989	0.951
β_1	20	0.972	0.972	0.977	0.969
β_1	50	0.981	0.986	0.982	0.982
β_1	100	0.983	0.982	0.977	0.977
β_1	200	0.992	0.982	0.992	0.950

Tablica 3.2: Vjerojatnost pokrivanja 99% pouzdanih intervala za jednoparametarske modele



Slika 3.2: Vjerojatnost pokrivanja 99% pouzdanog intervala u jednoparametarskom modelu

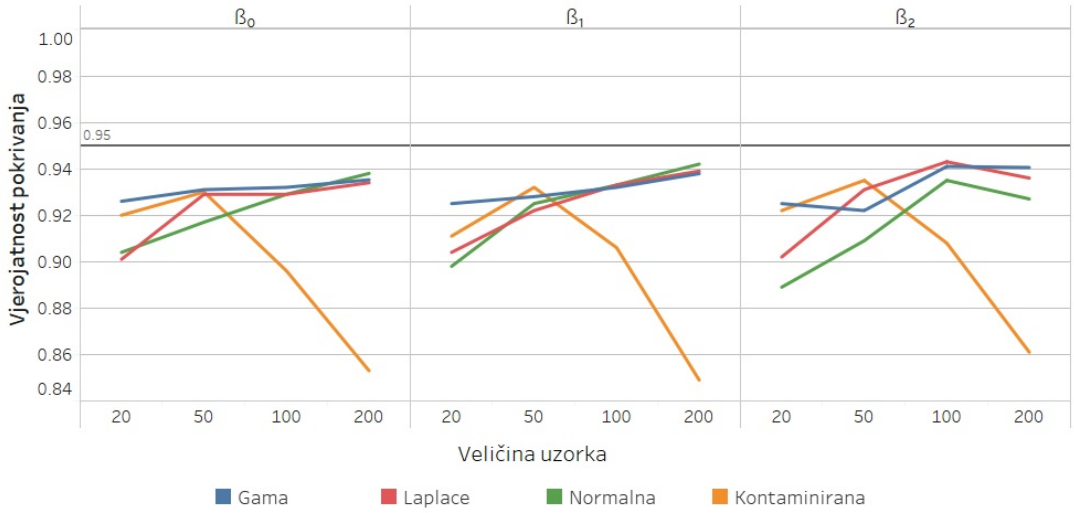
3.2 Dvoparametarski modeli

U tablicama 3.3 i 3.4 se nalaze vjerojatnosti pokrivanja za 95% i 99% pouzdane intervale u slučaju dvije prediktorske varijable. Grafički prikaz vjerojatnosti vidljiv je na slikama 3.3 i 3.4. Vjerojatnosti promatramo po parametrima $\beta_0, \beta_1, \beta_2$ za svaku različitu distribuciju uzimajući u obzir duljinu uzorka.
Slično kao i kod jednoparametarskih modela, zaključujemo da je vjerojatnost pokrivanja manja od očekivane, no ovdje je to slučaj za sve distribucije i za sve duljine

uzorka. Možemo reći da prilikom povećavanja duljina uzorka, vjerojatnost pokrivanja pouzdanih intervala ima tendenciju rasta kod gama, Laplace i normalne distribucije, dok kod kontaminirane normalne distribucije značajno opada.

Parametar	n	Normalna	Laplace	Gama	Kontaminirana
β_0	20	0.904	0.901	0.926	0.920
β_0	50	0.917	0.929	0.931	0.930
β_0	100	0.929	0.929	0.932	0.896
β_0	200	0.938	0.934	0.935	0.853
β_1	20	0.898	0.904	0.925	0.911
β_1	50	0.925	0.922	0.928	0.932
β_1	100	0.933	0.933	0.932	0.906
β_2	200	0.942	0.939	0.938	0.849
β_2	20	0.889	0.902	0.925	0.922
β_2	50	0.909	0.931	0.922	0.935
β_2	100	0.935	0.943	0.941	0.908
β_2	200	0.927	0.936	0.940	0.861

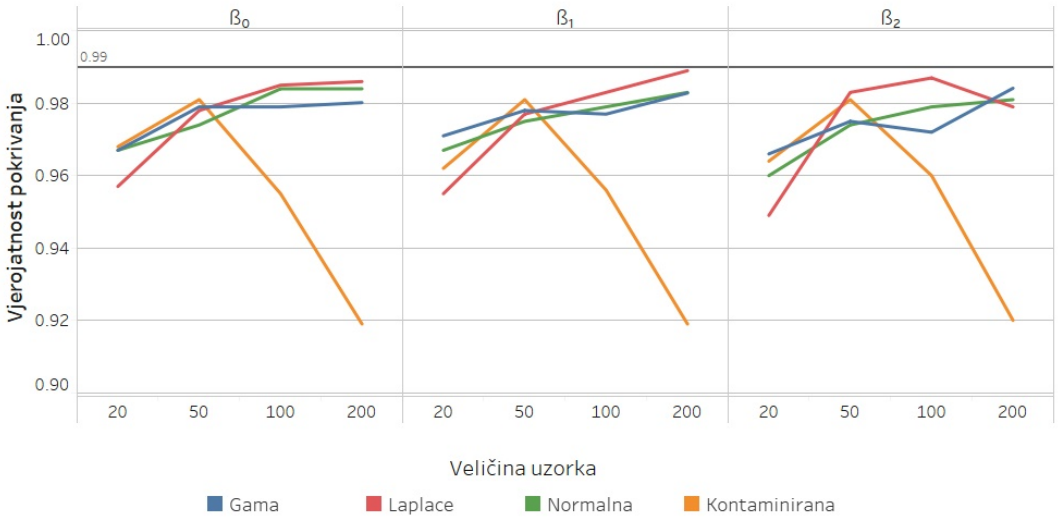
Tablica 3.3: Vjerojatnost pokrivanja 95% pouzdanih intervala za dvoparametarske modele



Slika 3.3: Vjerojatnost pokrivanja 95% pouzdanog intervala u dvoparametarskom modelu

Parametar	n	Normalna	Laplace	Gama	Kontaminirana
β_0	20	0.967	0.957	0.967	0.968
β_0	50	0.974	0.978	0.979	0.981
β_0	100	0.984	0.985	0.979	0.955
β_0	200	0.984	0.986	0.980	0.919
β_1	20	0.967	0.955	0.971	0.962
β_1	50	0.975	0.977	0.978	0.981
β_1	100	0.979	0.983	0.977	0.956
β_2	200	0.983	0.989	0.983	0.919
β_2	20	0.960	0.949	0.966	0.964
β_2	50	0.974	0.983	0.975	0.981
β_2	100	0.979	0.987	0.972	0.960
β_2	200	0.981	0.979	0.984	0.920

Tablica 3.4: Vjerojatnost pokrivanja 99% pouzdanih intervala za dvoparametarske modele



Slika 3.4: Vjerojatnost pokrivanja 99% pouzdanog intervala u dvoparametarskom modelu

Sljedeće tablice i slike će prikazivati vjerojatnosti pokrivanja pouzdanih intervala po distribucijama. Usporedit ćemo ponašanje vjerojatnosti po parametrima $\beta_0, \beta_1, \beta_2$ s obzirom na korelacije prediktorskih varijabli. Kao što smo već rekli, promatrali smo pet različitih slučajeva za $\rho = 0.0, 0.3, 0.5, 0.7, 0.9$. Vrijednosti za vjerojatnosti pokrivanja pouzdanih intervala u slučaju kada je korelacija 0.0 su iste one kao u

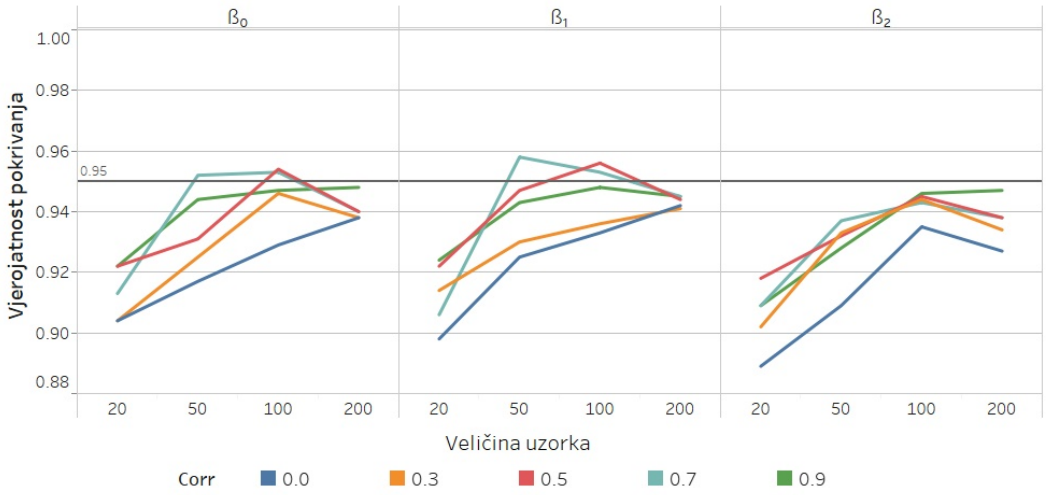
tablicama 3.3 i 3.4 jer su u tom slučaju prediktorske varijable nezavise pa su prema Korolaru 2.2.3 i nekorelirane.

Normalna distribucija

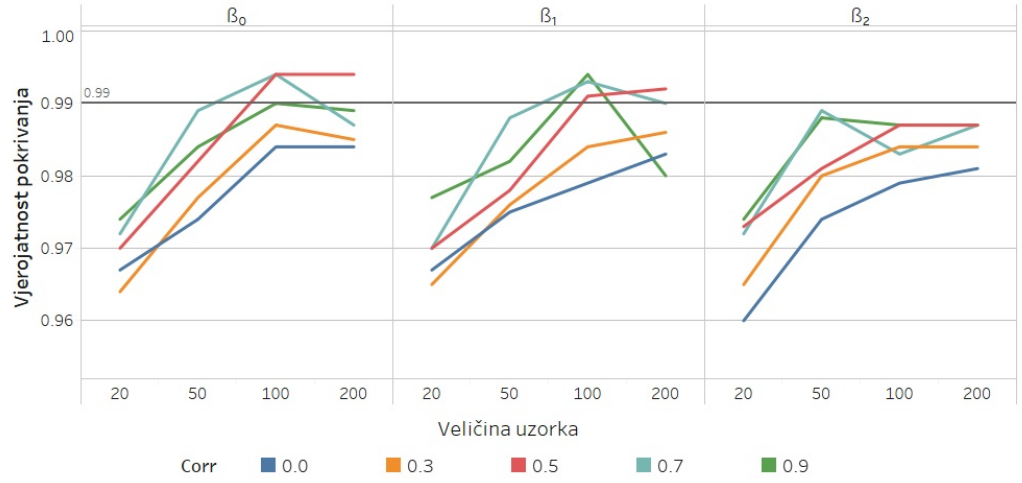
U tablicama 3.5 i 3.6 se nalaze vjerojatnosti pokrivanja za 95% i 99% Waldove pouzdane intervale. Grafički prikaz vjerojatnosti vidljiv je na slikama 3.5 i 3.6.

n	korelacija	β_0	β_1	β_2
20	0.0	0.904	0.898	0.889
50	0.0	0.917	0.925	0.909
100	0.0	0.929	0.933	0.935
200	0.0	0.938	0.942	0.927
20	0.3	0.904	0.914	0.902
50	0.3	0.925	0.930	0.933
100	0.3	0.946	0.936	0.944
200	0.3	0.938	0.941	0.934
20	0.5	0.922	0.922	0.918
50	0.5	0.931	0.947	0.932
100	0.5	0.954	0.956	0.945
200	0.5	0.940	0.944	0.938
20	0.7	0.913	0.906	0.909
50	0.7	0.952	0.958	0.937
100	0.7	0.953	0.953	0.943
200	0.7	0.940	0.945	0.938
20	0.9	0.922	0.924	0.909
50	0.9	0.944	0.943	0.928
100	0.9	0.947	0.948	0.946
200	0.9	0.948	0.945	0.947

Tablica 3.5: Vjerojatnost pokrivanja 95% pouzdanih intervala za normalno distribuirane kovarijate



Slika 3.5: Vjerojatnost pokrivanja 95% pouzdanog intervala u dvoparametarskom modelu za normalnu distribuciju



Slika 3.6: Vjerojatnost pokrivanja 99% pouzdanog intervala u dvoparametarskom modelu za normalnu distribuciju

Iz dobivenih rezultata, možemo primijetiti da je većina vrijednosti manja od očekivanih te je uočljiv porast vrijednosti s povećanjem duljine pojedinog uzorka. Nadalje, vidljivo je da su vjerojatnosti pokrivanja gotovo u svim slučajevima, najmanje kada su kovarijate nekorelirane.

n	korelacija	β_0	β_1	β_2
20	0.0	0.967	0.967	0.960
50	0.0	0.974	0.975	0.974
100	0.0	0.984	0.979	0.979
200	0.0	0.984	0.983	0.981
20	0.3	0.964	0.965	0.965
50	0.3	0.977	0.976	0.980
100	0.3	0.987	0.984	0.984
200	0.3	0.985	0.986	0.984
20	0.5	0.970	0.970	0.973
50	0.5	0.982	0.978	0.981
100	0.5	0.994	0.991	0.987
200	0.5	0.994	0.992	0.987
20	0.7	0.972	0.970	0.972
50	0.7	0.989	0.988	0.989
100	0.7	0.994	0.993	0.983
200	0.7	0.987	0.990	0.987
20	0.9	0.974	0.977	0.974
50	0.9	0.984	0.982	0.988
100	0.9	0.990	0.994	0.987
200	0.9	0.989	0.980	0.987

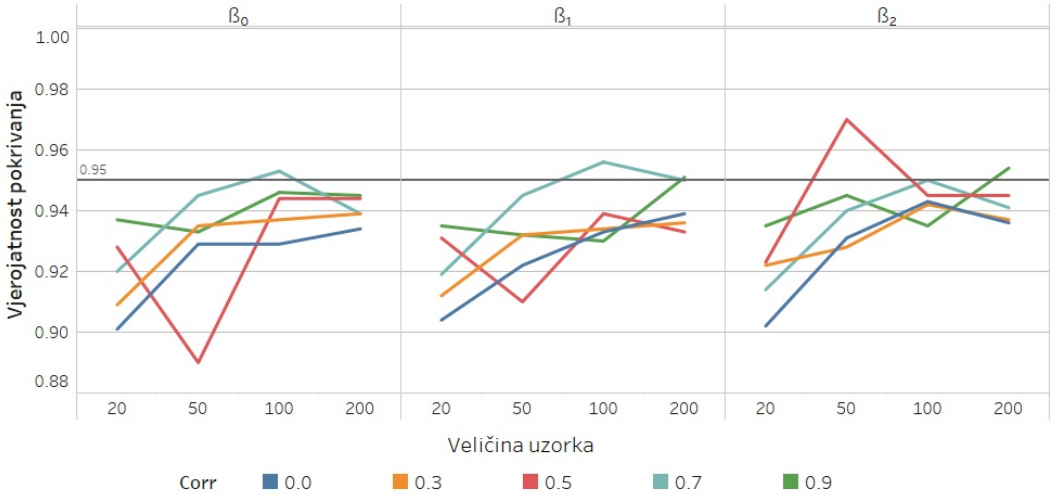
Tablica 3.6: Vjerojatnost pokrivanja 99% pouzdanih intervala za normalno distribuirane kovarijate

Laplace distribucija

U tablicama 3.7 i 3.8 se nalaze vjerojatnosti pokrivanja za 95% i 99% Waldove pouzdane intervale. Grafički prikaz vjerojatnosti vidljiv je na slikama 3.7 i 3.8. Kao i kod ostalih rezultata i u ovome slučaju je većina vrijednosti ispod očekivane. Usporedimo li vrijednosti dobivene s manjom i većom duljinom uzorka, vidimo da se vjerojatnost povećava. Nadalje, ne možemo reći da postoji uočljivo ponašanje vjerojatnosti s obzirom na razinu korelacije između prediktorskih varijabli.

n	korelacija	β_0	β_1	β_2
20	0.0	0.901	0.904	0.902
50	0.0	0.929	0.922	0.931
100	0.0	0.929	0.933	0.943
200	0.0	0.934	0.939	0.936
20	0.3	0.909	0.912	0.922
50	0.3	0.935	0.932	0.928
100	0.3	0.937	0.934	0.942
200	0.3	0.939	0.936	0.937
20	0.5	0.928	0.931	0.923
50	0.5	0.890	0.910	0.970
100	0.5	0.944	0.939	0.945
200	0.5	0.944	0.933	0.945
20	0.7	0.920	0.919	0.914
50	0.7	0.945	0.945	0.940
100	0.7	0.953	0.956	0.950
200	0.7	0.939	0.950	0.941
20	0.9	0.937	0.935	0.935
50	0.9	0.933	0.932	0.945
100	0.9	0.946	0.930	0.935
200	0.9	0.945	0.951	0.954

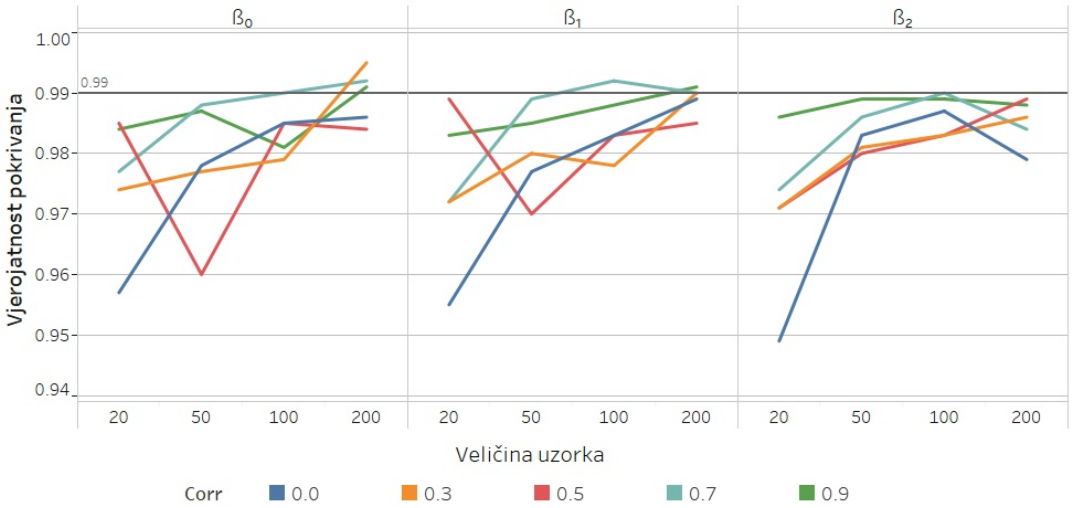
Tablica 3.7: Vjerojatnost pokrivanja 95% pouzdanih intervala za kovarijate iz Laplaceove distribucije



Slika 3.7: Vjerojatnost pokrivanja 95% pouzdanog intervala u dvoparametarskom modelu za Laplace distribuciju

n	korelacija	β_0	β_1	β_2
20	0.0	0.957	0.955	0.949
50	0.0	0.978	0.977	0.983
100	0.0	0.985	0.983	0.987
200	0.0	0.986	0.989	0.979
20	0.3	0.974	0.972	0.971
50	0.3	0.977	0.980	0.981
100	0.3	0.979	0.978	0.983
200	0.3	0.995	0.990	0.986
20	0.5	0.985	0.989	0.971
50	0.5	0.960	0.970	0.980
100	0.5	0.985	0.983	0.983
200	0.5	0.984	0.985	0.989
20	0.7	0.977	0.972	0.974
50	0.7	0.988	0.989	0.986
100	0.7	0.990	0.992	0.990
200	0.7	0.992	0.990	0.984
20	0.9	0.984	0.983	0.986
50	0.9	0.987	0.985	0.989
100	0.9	0.981	0.988	0.989
200	0.9	0.991	0.991	0.988

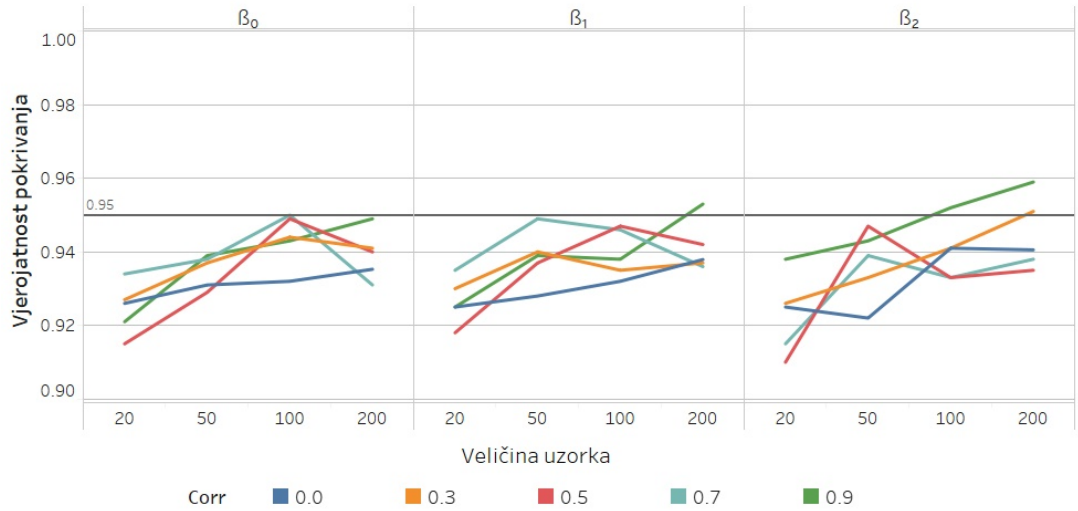
Tablica 3.8: Vjerojatnost pokrivanja 99% pouzdanih intervala za kovarijate iz Laplaceove distribucije



Slika 3.8: Vjerojatnost pokrivanja 99% pouzdanog intervala u dvoparametarskom modelu za Laplace distribuciju

Gama distribucija

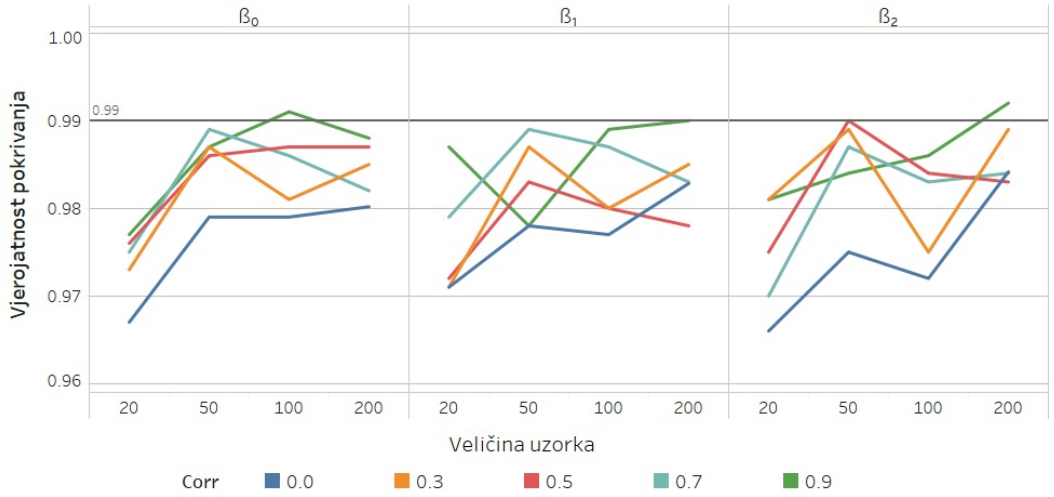
U tablicama 3.9 i 3.10 se nalaze vjerojatnosti pokrivanja za 95% i 99% Waldove pouzdane intervale. Grafički prikaz vjerojatnosti vidljiv je na slikama 3.9 i 3.10.



Slika 3.9: Vjerojatnost pokrivanja 95% pouzdanog intervala u dvoparametarskom modelu za gama distribuciju

n	korelacija	β_0	β_1	β_2
20	0.0	0.926	0.925	0.925
50	0.0	0.931	0.928	0.922
100	0.0	0.932	0.932	0.941
200	0.0	0.935	0.938	0.941
20	0.3	0.927	0.930	0.926
50	0.3	0.937	0.940	0.933
100	0.3	0.944	0.935	0.941
200	0.3	0.941	0.937	0.951
20	0.5	0.915	0.918	0.910
50	0.5	0.929	0.937	0.947
100	0.5	0.949	0.947	0.933
200	0.5	0.940	0.942	0.935
20	0.7	0.934	0.935	0.915
50	0.7	0.938	0.949	0.939
100	0.7	0.950	0.946	0.933
200	0.7	0.931	0.936	0.938
20	0.9	0.921	0.925	0.938
50	0.9	0.939	0.939	0.943
100	0.9	0.943	0.938	0.952
200	0.9	0.949	0.953	0.959

Tablica 3.9: Vjerojatnost pokrivanja 95% pouzdanih intervala za kovarijate iz gama distribucije



Slika 3.10: Vjerojatnost pokrivanja 99% pouzdanog intervala u dvoparametarskom modelu za gama distribuciju

n	korelacija	β_0	β_1	β_2
20	0.0	0.967	0.971	0.966
50	0.0	0.979	0.978	0.975
100	0.0	0.979	0.977	0.972
200	0.0	0.980	0.983	0.984
20	0.3	0.973	0.971	0.981
50	0.3	0.987	0.987	0.989
100	0.3	0.981	0.980	0.975
200	0.3	0.985	0.985	0.989
20	0.5	0.976	0.972	0.975
50	0.5	0.986	0.983	0.990
100	0.5	0.987	0.980	0.984
200	0.5	0.987	0.978	0.983
20	0.7	0.975	0.979	0.970
50	0.7	0.989	0.989	0.987
100	0.7	0.986	0.987	0.983
200	0.7	0.982	0.983	0.984
20	0.9	0.977	0.987	0.981
50	0.9	0.987	0.978	0.984
100	0.9	0.991	0.989	0.986
200	0.9	0.988	0.990	0.992

Tablica 3.10: Vjerojatnost pokrivanja 99% pouzdanih intervala za kovarijate iz gama distribucije

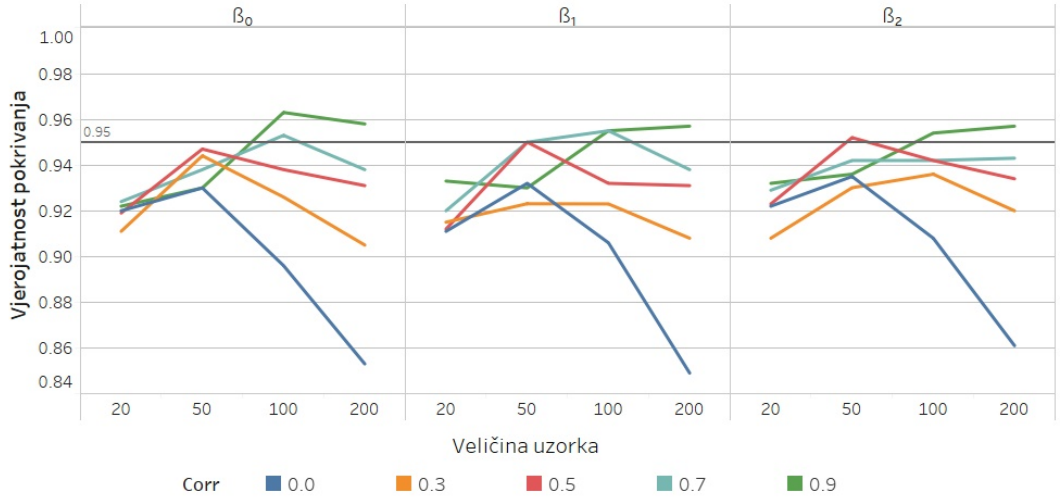
Iz dobivenih rezultata možemo vidjeti, da su u većini slučajeva vjerojatnosti pokrivanja pouzdanih intervala manje od očekivanih. Uočljiv je trend rasta s povećanjem duljine uzorka, dok s povećanjem korelacije između kovarijata ne možemo vidjeti očito ponašanje. Možemo primijetiti da su vrijednosti u slučaju kada je korelacija jednaka 0, gotovo uvijek najmanje što je posebno vidljivo za 99% pouzdane intervale.

Kontaminirana normalna distribucija

U tablicama 3.11 i 3.12 se nalaze vjerojatnosti pokrivanja za 95% i 99% Waldove pouzdane intervale. Grafički prikaz vjerojatnosti vidljiv je na slikama 3.11 i 3.12.

n	korelacija	β_0	β_1	β_2
20	0.0	0.920	0.911	0.922
50	0.0	0.930	0.932	0.935
100	0.0	0.896	0.906	0.908
200	0.0	0.853	0.849	0.861
20	0.3	0.911	0.915	0.908
50	0.3	0.944	0.923	0.930
100	0.3	0.926	0.923	0.936
200	0.3	0.905	0.908	0.920
20	0.5	0.919	0.912	0.923
50	0.5	0.947	0.950	0.952
100	0.5	0.938	0.932	0.942
200	0.5	0.931	0.931	0.934
20	0.7	0.924	0.920	0.929
50	0.7	0.938	0.950	0.942
100	0.7	0.953	0.955	0.942
200	0.7	0.938	0.938	0.943
20	0.9	0.922	0.933	0.932
50	0.9	0.930	0.930	0.936
100	0.9	0.963	0.955	0.954
200	0.9	0.958	0.957	0.957

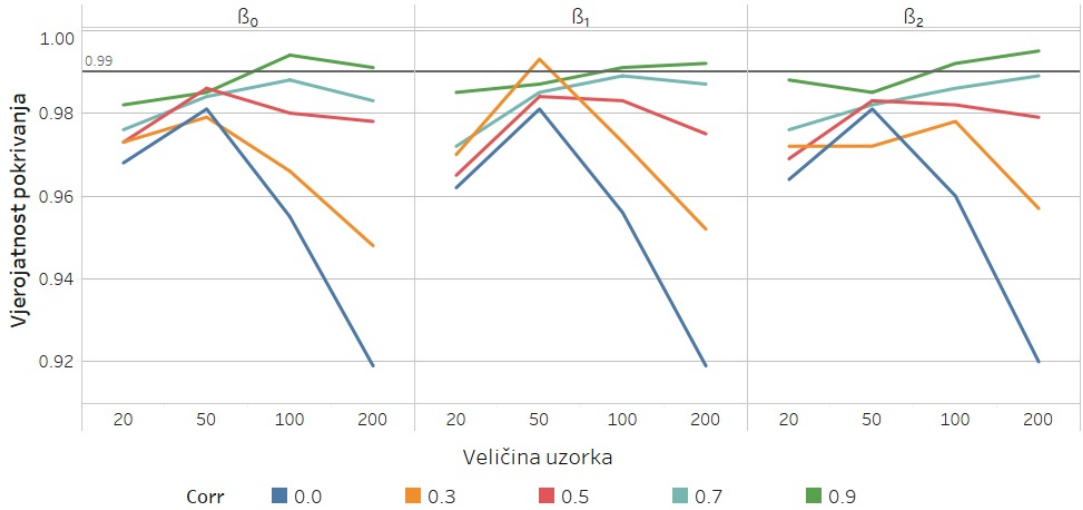
Tablica 3.11: Vjerojatnost pokrivanja 95% pouzdanih intervala za kovarijate iz kontaminirane normalne distribucije



Slika 3.11: Vjerojatnost pokrivanja 95% pouzdanog intervala u dvoparametarskom modelu za kontaminiranu normalnu distribuciju

n	korelacija	β_0	β_1	β_2
20	0.0	0.968	0.962	0.964
50	0.0	0.981	0.981	0.981
100	0.0	0.955	0.956	0.960
200	0.0	0.919	0.919	0.920
20	0.3	0.973	0.970	0.972
50	0.3	0.979	0.993	0.972
100	0.3	0.966	0.973	0.978
200	0.3	0.948	0.952	0.957
20	0.5	0.973	0.965	0.969
50	0.5	0.986	0.984	0.983
100	0.5	0.980	0.983	0.982
200	0.5	0.978	0.975	0.979
20	0.7	0.976	0.972	0.976
50	0.7	0.984	0.985	0.982
100	0.7	0.988	0.989	0.986
200	0.7	0.983	0.987	0.989
20	0.9	0.982	0.985	0.988
50	0.9	0.985	0.987	0.985
100	0.9	0.994	0.991	0.992
200	0.9	0.991	0.992	0.995

Tablica 3.12: Vjerojatnost pokrivanja 99% pouzdanih intervala za kovarijate iz kontaminirane normalne distribucije



Slika 3.12: Vjerojatnost pokrivanja 99% pouzdanog intervala u dvoparametarskom modelu za kontaminiranu normalnu distribuciju

U ovome dijelu,prediktorske varijable dolaze iz kontaminirane normalne distribucije, no kao i u prethodnim distribucijama, vjerojatnosti pokrivanja su manje od očekivanih u gotovo svim situacijama. Očito se vidi da su vrijednosti najmanje kada su kovarijate nekorelirane. Nadalje, ako promatramo različite veličine uzoraka, uočljiva je razlika u ponašanju s obzirom na jačinu koreliranosti. Naime, kada između kovarijata nema korelacije ili je korelacija mala, tada vrijednosti opadaju kako se duljina uzorka povećava, dok kod jakih korelacija možemo reći vrijednosti rastu pri povećanju veličine uzorka.

Generalno, iz dobivenih rezultata je vidljivo

1. Vjerojatnosti pokrivanja Waldovih pouzdanih intervala su manje od očekivanih.
2. S porastom veličine uzorka, vjerojatnost pokrivanja raste kada prediktorske varijable dolaze iz gama, Laplace i normalne distribucije.
3. Kada su kovarijate iz kontaminirane normalne distribucije, tada je vjerojatnost pokrivanja veća za manje duljine uzorka.
4. U slučaju kada su prediktorske varijable iz gama i Laplace distribucije, koreliranost između varijabli nema utjecaja na vjerojatnost pokrivanja.
5. Kada su kovarijate iz normalne i kontaminirane normalne distribucije, tada postoji rast vjerojatnosti pokrivanja s obzirom na jačinu korelacije.

Porast vjerojatnosti pokrivanja u gama, Laplace i normalnoj distribuciji s rastom veličine uzorka se mogla očekivati. Naime, prema centralnom graničnom teoremu za veće n, dobiti ćemo bolju aproksimaciju parametara β_i, $\forall i$.
U slučaju kontaminirane normalne distribucije imamo drugu situaciju. Naime, funkcija gustoće kontaminirane normalne distribucije ima "teže repove" od funkcije gustoće normalne razdiobe što znači da sporije pada prema nuli nego normalna razdioba. Zbog toga, povećanjem duljine uzorka smo povećali vjerojatnost dobivanja puno većih vrijednosti što onda utječe na procjene parametara β_i, $\forall i$. S druge strane, povećanjem korelacije između kovarijata, povećava se standardna pogreška za β_i, $\forall i$, a time povećavamo duljinu Waldovih pouzdanih intervala, što omogućuje da stvarna vrijednost parametra β_i upadne u taj pouzdan interval.

Bibliografija

[1] STAT 501, *12.4 - Detecting Multicollinearity Using Variance Inflation Factors*, The Pennsylvania State University, https://newonlinecourses.science.psu.edu/stat501/node/347/, posjećeno 4. 2. 2019.

[2] STAT 504, *SAS - Poisson Regression Model for Count Data*, The Pennsylvania State University, https://newonlinecourses.science.psu.edu/stat504/node/223/, posjećeno 4. 2. 2019.

[3] SAS/STAT 9.2, *Generalized Linear Models Theory*, SAS Institute Inc., September 2009, https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_genmod_sect030.htm#statug.genmod.genmodoverdisp, posjećeno 4. 2. 2019., second edition.

[4] B. Basrak, *Aktuarska matematika II, 4.dio*, Sveučilište u Zagrebu PMF-Matematički odjel, 2016, https://web.math.pmf.unizg.hr/~bbasrak/pdf_files/AM2slides4dio.pdf, posjećeno 4. 2. 2019.

[5] M. Davidian, *Simulation studies in statistics*, North Carolina State University, 2005, https://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf, posjećeno 4. 2. 2019.

[6] R. Godwin, *Poisson Regression*, University of Manitoba, http://home.cc.umanitoba.ca/~godwinrt/7010/poissonregression.pdf, posjećeno 4. 2. 2019.

[7] J. Goodman, *Accuracy and efficiency of Monte Carlo method*, Bechtel Power Corporation, https://inis.iaea.org/collection/NCLCollectionStore/_Public/19/047/19047359.pdf, posjećeno 4. 2. 2019.

[8] G.J. HAHN, *Sample Sizes for Monte Carlo Simulation*, (1972), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4309200.

[9] M. Huzak, *Vjerojatnost i matematička statistika*, Sveučilište u Zagrebu PMF-Matematički odjel, travanj 2006, `http://aktuari.math.pmf.unizg.hr/docs/vms.pdf`, posjećeno 4. 2. 2019.

[10] J. Neter William Li M.H. Kutner, C. J. Nachtsheim, McGraw-Hill/Irwin, 2005, ISBN 0-07-238688-6, Fifth Edition.

[11] NCSS, *Poisson Regression*, `https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Poisson_Regression.pdf`, posjećeno 4. 2. 2019.

[12] V. Nicosia, *Lecture Notes – Week 7: Monte Carlo methods*, Queen Mary University of London, 2016, `https://qmplus.qmul.ac.uk/pluginfile.php/985596/mod_resource/content/4/MTH739_2016-17_lecture_notes_week7.pdf`, posjećeno 4. 2. 2019.

[13] W. Oberle, *Monte Carlo Simulations: Number of Iterations and Accuracy*, US Army Research Laboratory, July 2015, `https://www.arl.army.mil/arlreports/2015/ARL-TN-0684.pdf`, posjećeno 4. 2. 2019.

[14] J.A. Nelder frs P. McCullagh, *Generalized Linear Models*, Chapman and Hall, 1983, second edition.

[15] Germán Rodríguez, *Lecture Notes on Generalized Linear Models: Poisson Models for Count Data*, Princeton University, 2007, `http://home.cc.umanitoba.ca/~godwinrt/7010/poissonregression.pdf`, posjećeno 4. 2. 2019.

[16] Statistic How To, *Variance Inflation Factor*, September 2015, `https://www.statisticshowto.datasciencecentral.com/variance-inflation-factor/`, posjećeno 4. 2. 2019.

[17] ———, *Poisson Regression*, September 2016, `https://www.statisticshowto.datasciencecentral.com/poisson-regression/`, posjećeno 4. 2. 2019.

[18] Eric W. Weisstein, *Laplace Distribution, From MathWorld–A Wolfram Web Resource*, `http://mathworld.wolfram.com/LaplaceDistribution.html`, posjećeno 27. 12. 2018.

[19] ———, *Pearson Type III Distribution, From MathWorld–A Wolfram Web Resource*, `http://mathworld.wolfram.com/PearsonTypeIIIDistribution.html`, posjećeno 5. 1. 2019.

[20] R. Wicklin, *Simulating Data with SAS*, SAS Institute Inc., September 2013, ISBN 978-1-61290-332-3, second edition.

Sažetak

Poissonova regresija je jedan od generaliziranih linearnih modela (GLM) koji se najčešće koristi za modeliranje podataka koji imaju cjelobrojnu, nenegativnu vrijednost (eng.*count data*). Određivanjem modela, procjenjujem koeficijente regresije i njihove $(1-\alpha)\cdot 100\%$ pouzdane intervale. Simuliranjem različitih uvjeta, analizira se je li stvarna vjerojatnost pokrivanja pouzdanih intervala zaista ona koja se očekuje.

Prvi dio rada govori općenito o generaliziranim linearnim modelima, na koji način se došlo do generalizacije, svojstvima modela te su navedeni neki primjeri ove vrste modela.

U drugom dijelu se uvode definicije i teorijske tvrdnje, navedene su sve kombinacije faktora za simulaciju, opisuje se sama simulacija i modeliranje podataka.

U zadnjem dijelu se analiziraju i uspoređuju dobiveni rezultati. Za provedbu rada korišten je programski softver `SAS`.

Summary

Poisson regression is one of generalized linear model (GLM) which is usually used to model count data. With determination of model, the regression coefficients have been estimated along with their $(1 - \alpha) \cdot 100\%$ confidence intervals. By simulating different conditions, it is analysed if the coverage probability of confidence intervals is the what is expected.

The first part of the paper deals generally with generalized linear models, how generalizations are made, model properties and some examples of this type of model are listed.

In the second part, definitions and theoretical statements have been introduced, all combinations of simulation factors are listed and simulations and modeling of data are described.

In the last part, the results are analysed and compared. The software `SAS` was used for the implementation.

Životopis

Rođena sam 29. rujna 1993. godine u Zagrebu, a živim u Kutini. Tamo sam pohađala Osnovnu školu Zvonimira Franka koju završavam 2008. godine te potom upisujem Srednju školu Tina Ujevića, smjer matematička gimnazija. Po završetku srednje škole, 2012. godine upisujem preddiplomski studij Matematika na Matematičkom odsjeku Prirodoslovno–matematičkog fakulteta Sveučilišta u Zagrebu. 2016. godine upisujem diplomski studij Matematička statistika na istom fakultetu.