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

Špehar, Marija

SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO–MATEMATIČKI FAKULTET
MATEMATIČKI ODSJEK

Marija Špehar

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

Diplomski rad

Voditelj rada: Doc. dr. sc. Vesna Lužar-Stifer

Zagreb, ožujak, 2019. godina

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. _____

Sadržaj

Uvod

U današnje vrijeme se u mnogim područjima znanosti često javlja potreba za proučavanjem prisutstva ili odsutstva nekog svojstva ili pojave. Podatke s kojima se susrećemo u takvim situacijama nazivamo binarnim podacima. Prisutnost svojstva koje promatramo zovemo uspjehom, a odsutnost neuspjehom te označujemo jedinicom i nulom. Binarnim podacima možemo pristupiti na dva načina. Prvi način je da promatramo svaku obzervaciju zasebno te nam tada uspjeh i neuspjeh odgovaraju realizacijama slučajne varijable s Bernoullijevom razdiobom. Drugi način je da grupiramo obzervacije koje su jednake po ostalim karakteristikama mjerenim u eksperimentu te po grupama promatramo broj obzervacija koje imaju promatrano svojstvo (ukupan broj jedinica), što odgovara realizaciji binomne slučajne varijable.

U povijesti je prvi model za takve podatke razvio Ronald Fisher 1922. godine kada je u svojim eksperimentima promatrao otopine i smjese te ispitivao prisutnost kontaminanta. Takav model uključivao je transformiranje podataka funkcijom $g(x) = \log(-\log(1-x))$, danas poznatom pod nazivom $log - log$. Logistički model prvotno je predstavio Joseph Berkson 1944. godine. Na temelju biološkog eksperimenta i *probit* regresije, koju je razvio Chester Bliss desetak godina ranije, utvrdio je novi, jednostavniji model. Berkson je dao alternativu inverzu funkcije distribucije jedinične normalne razdiobe te pokazao da je logistička funkcija također pogodna za modeliranje takvih podataka. Model je skraćeno prozvao *logit* modelom, po uzoru na Blissov *probit* model. Tijekom 60-ih i 70-ih godina razvijali su se prvi algoritmi za procjenu parametara takvih modela metodom maksimalne vjerodostojnosti, a prijelomnom godinom smatra se 1972. kada su John Nelder i Robert Wedderburn razvili metodologiju statističkog modeliranja i pripadne modele jednim imenom prozvali generaliziranim linearnim modelima. Na taj način su modeli za binarne podatke postali dio veće klase modela te su se u tom okviru nastavili dalje razvijati. Danas se za modeliranje binarnih podataka najčešće koristi logistički model upravo zbog svoje jednostavnosti i lakše interpretacije u odnosu na *probit* funkciju. Svoju je primjenu našao u brojnim znanostima, medicini, ekonomiji, aktuarstvu...

Ovaj rad ćemo započeti definiranjem osnovnih pojmova i kratkom motivacijom koja će nas uvesti u klasu generaliziranih linearnih modela kojoj, kako smo naveli, pripada lo-

gistički model. Navest ćemo glavne razlike multinomne i dihotomne logističke regresije te kratko predstaviti vrste multinomne logističke regresije - nominalnu i ordinalnu, a detaljnije ćemo se posvetiti dihotomnoj logističkoj regresiji. Opisat ćemo način postavljanja modela i objasniti interpretaciju parametara te predstaviti jednu metodu dobivanja procjena parametara modela. Razmatrat ćemo koliko dobro model opisuje stvarne podatke, a na kraju teorijskog dijela pokazati spomenuta svojstva na primjeru.

U drugom dijelu rada provest ćemo istraživanje kojemu je primarni cilj analiza vjerojatnosti pokrivanja Waldovih pouzdanih intervala za regresijske koeficijente u dihotomnoj logističkoj regresiji. U tu svrhu koristit ćemo jednostavne modele s jednom i dvije prediktorske varijable koje su generirane uzorcima iz različitih distribucija i različitih duljina. Opisat ćemo postupak rješavanja problema i promotriti dobivene rezultate.

Poglavlje 1

Osnovni pojmovi

Definicija 1.0.1. *Neka je $(\Omega, \mathcal{F})$ izmjeriv prostor i $\mathcal{P}$ familija vjerojatnosnih mjera na $(\Omega, \mathcal{F})$. Trojku $(\Omega, \mathcal{F}, \mathcal{P})$ nazivamo statistička struktura.*

Familija vjerojatnosti često je parametrizirana:

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}, \qquad \Theta = \text{parametarski prostor}$$

Definicija 1.0.2. *Neka je na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ dan slučajni vektor $\boldsymbol{Y} : \Omega \to \mathbb{R}^n$. Za fiksni $\theta \in \Theta$ označimo s $F(\cdot; \theta)$ funkciju distribucije od $\boldsymbol{Y}$ u odnosu na vjerojatnost $\mathbb{P}_\theta$. Familiju $\mathcal{P}' = \{F(\cdot; \theta) : \theta \in \Theta\}$ nazivamo statističkim modelom, a za vektor $\boldsymbol{Y}$ kažemo da pripada tom statističkom modelu.*

Definicija 1.0.3. *n-dimenzionalni slučajni uzorak na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je niz $X_1, X_2, ..., X_n$ slučajnih varijabli (vektora) na $(\Omega, \mathcal{F})$ takvih da su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost $\mathbb{P} \in \mathcal{P}$.*

Definicija 1.0.4. *Statistika na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$ je svaka slučajna varijabla (vektor) $T : \Omega \to \mathbb{R}^d$ takva da postoji $n \in \mathbb{N}$ i n-dimenzionalni slučajni uzorak $(X_1, X_2, ..., X_n)$ na $(\Omega, \mathcal{F}, \mathcal{P})$ te izmjerivo preslikavanje $t : \mathbb{R}^n \to \mathbb{R}^d$ takvo da je $T = t(X_1, X_2, ..., X_n)$*

Definicija 1.0.5. *Neka su $L_n = l_n(X_1, ..., X_n)$ i $D_n = d_n(X_1, ..., X_n)$ statistike slučajnog uzorka $X_1, ..., X_n$. Za $[L_n, D_n]$ kažemo da je $(1 - \alpha) \cdot 100\%$ pouzdani interval za parametar τ ako vrijedi*

$$\mathbb{P}(L_n \leq \tau \leq D_n) \geq 1 - \alpha, \quad \alpha \in \langle 0, 1 \rangle.$$

Poglavlje 2

Uvod u model

2.1 Motivacija

Generalizirani linearni modeli su svojevrsno proširenje klasičnih linearnih modela, kod kojih se promatra linearna veza varijable odaziva (zavisne) Y i nezavisnih varijabli poticaja ili prediktora $X_1, X_2, ..., X_p$. Pretpostavimo da imamo n međusobno nezavisnih opažanja od Y, koje označimo s $Y_1, Y_2, ..., Y_n$. Tada linearnu vezu opisujemo s

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \quad i = 1, ..., n$$

odnosno matrično

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

gdje su $\boldsymbol{Y}^T = (Y_1, Y_2, ..., Y_n)$, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_p)$ vektor parametara modela, $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}_{\cdot 1}, ..., \boldsymbol{x}_{\cdot p}]$ matrica poticaja, pri čemu je $\boldsymbol{x}_{\cdot j}$ vektor stupac realizacije j-te varijable poticaja. S $\boldsymbol{\epsilon}$ označavamo vektor slučajnih grešaka. Za slučajne varijable koje modeliraju te greške pretpostavljamo da su očekivanja 0, međusobno nekorelirane i jednake varijance. Ukoliko pretpostavimo da je y_i realizacija slučajne varijable Y_i i da je $\mathbb{E}(Y_i) = \mu_i$, vrijedi

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{X\beta} =: \boldsymbol{\mu}.$$

Dodatna pretpostavka je da greške dolaze iz normalne distribucije, što implicira normalnost promatrane varijable odaziva.

2.2 Generalizirani linearni modeli

Kod generalziranih linearnih modela uvodimo nekoliko drugačijih pretpostavki:
Pretpostavljamo da svaka komponenta varijable odaziva $\boldsymbol{Y}$ potječe iz eksponencijalne familije, što znači da ima gustoću oblika

$$f_Y(y;\theta,\rho) = \exp\left(\frac{y\theta - b(\theta)}{a(\rho)} + c(y,\rho)\right) \tag{1}$$

za neke funkcije a, b, c. Parametar θ zove se prirodni parametar, a ρ parametar disperzije ili skaliranja. Funkcija a parametra ρ zove se funkcija disperzije i omogućuje dodatnu fleksibilnost u modelu, tako da ne moraju svi odazivi imati istu varijancu, dok je b dva puta neprekidno diferencijabilna, tako da je b' invertibilna, što će nam biti važno kasnije u računu. Funkciju c obično ignoriramo jer nam nije od velikog značaja.

Uvodimo pojam linearnog prediktora i funkcije povezivanja, čija je zadaća uspostaviti vezu komponenata linearnog prediktora $\boldsymbol{\eta}$ s odgovarajućim komponentama očekivanja $\boldsymbol{\mu}$.

Definicija 2.2.1. *Za parametre β_0, β_1,..., β_p i vrijednosti varijabli poticaja, vektore stupce $\boldsymbol{x}_{\cdot 1}$,...,$\boldsymbol{x}_{\cdot p}$ ($\boldsymbol{x}_{\cdot \mathbf{0}} = \mathbf{1}$) definiramo*

$$\boldsymbol{\eta} = \sum_{i=0}^{p} \boldsymbol{x}_{\cdot i}\beta_i$$

i nazivamo ga linearnim prediktorom.

Definicija 2.2.2. *Monotonu diferencijablinu funkciju $g\colon \mathbb{R} \to \mathbb{R}$ takvu da*

$$\eta_i = g(\mu_i) \quad i = 1, ..., n$$

nazivamo funkcijom povezivanja, kraće poveznicom (engl. link function).

Sada model možemo matrično zapisati u obliku

$$\boldsymbol{Y} = g^{-1}(\boldsymbol{X\beta}) + \boldsymbol{\epsilon}$$

pri čemu je $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{Y}) = g^{-1}(\boldsymbol{X\beta})$.

Promotrimo funkciju log-vjerodostojnosti $l(\theta,\rho;y) = log f_Y(y;\theta,\rho)$ za dane θ, ρ, y unutar neke eksponencijalne familije i prisjetimo se dobro poznatih rezultata iz statističke teorije:

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = 0 \quad \text{i} \quad \mathbb{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathbb{E}\left(\frac{\partial l}{\partial \theta}\right)^2 = 0. \tag{2}$$

Iz (1) slijedi

$$l(\theta; y) = \frac{y\theta - b(\theta)}{a(\rho)} + c(y, \rho)$$

pa imamo da je

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\rho)} \quad \text{i} \quad \frac{\partial^2 l}{\partial \theta^2} = -\frac{b''(\theta)}{a(\rho)}. \tag{3}$$

Uvrštavanjem jednakosti (3) u (2) dobivamo:

$$0 = \mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = \frac{\mu - b'(\theta)}{a(\rho)},$$

stoga je

$$\mu = \mathbb{E}(Y) = b'(\theta) \Leftrightarrow \theta = b'^{-1}(\mu).$$

Na sličan način kao što smo dobili očekivanje, dobivamo i varijancu:

$$0 = -\frac{b''(\theta)}{a(\rho)} + \frac{var(Y)}{a^2(\rho)} \quad \Rightarrow \quad var(Y) = b''(\theta)a(\rho).$$

Dakle, očekivanje ne ovisi o ρ, dok je varijanca produkt dvije funkcije: $a(\rho)$, koja uključuje parametar skaliranja i $b''(\theta)$, koja ovisi o prirodnom parametru i zvat ćemo je funkcija varijance.
Kako je b' neprekidna i invertibilna funkcija i vrijedi da je $\mu = b'(\theta)$, uvodimo novi parametar, tzv. parametar srednje vrijednosti *(engl. mean value parameter)*. Budući da je $\theta = b'^{-1}(\mu)$, kako smo već pokazali, dobro je definirana funkcija varijance relacijom

$$\mu \mapsto V(\mu) = b''(\theta) = b''(b'^{-1}(\mu)).$$

Dana funkcija određuje način na koji varijanca ovisi o očekivanju. Da bismo naglasili taj utjecaj, izrazimo varijancu u obliku:

$$var(Y) = a(\rho)V(\mu).$$

Prije nego krenemo ova obilježja promatrati na konkretnim primjerima, definirajmo pojam kanonske funkcije povezivanja. Stavimo da je

$$\theta = h(\eta),$$

a očekivanje glatka i invertibilna funkcija linearnog predviditelja η oblika $b' \circ h$, za neku funkciju h, odnosno

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

Znamo da je $\mu = g^{-1}(\eta)$, tj. $\eta = g(\mu)$, ali ovog puta promatrajmo funkciju povezivanja g kao kompoziciju dvije funkcije: $g = h^{-1} \circ b'^{-1}$.
Imamo da je

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

U slučaju da je $h \equiv id$, vrijedi $\theta \equiv \eta$, a

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

se naziva kanonska funkcija povezivanja.

Pogledajmo sada sljedeće primjere. Pretpostavimo prvo da je varijabla Y iz normalne, a zatim iz binomne distribucije.

Primjer 2.2.3. *Normalna distribucija*
Fukcija gustoće normalne distribucije je oblika

$$f_Y(y;\theta,\rho) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2}\right) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log 2\pi\sigma^2\right),$$

što je oblika jednadžbe (1), pri čemu je
$\theta = \mu$, $\rho = \sigma^2$, $a(\rho) = \rho$, $b(\theta) = \frac{\theta^2}{2}$, $c(y,\rho) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log 2\pi\sigma^2\right)$.
Dakle, prirodni parametar normalne distribucije je jednak μ, parametar skaliranja je σ^2, a iz (2) i (3) možemo izračunati očekivanje i varijancu:

$$\begin{aligned} b(\theta) = \frac{\theta^2}{2} \quad &\Rightarrow \quad \mathbb{E}(Y) = b'(\theta) = \theta = \mu \\ a(\rho) = \rho \quad &\Rightarrow \quad var(Y) = a(\rho)b''(\theta) = \rho = \sigma^2 \end{aligned}$$

Uočimo da je kod klasičnog linearnog modela kanonska funkcija povezivanja identiteta i da varijanca ne ovisi o očekivanju (zbog $b''(\theta) = 1$), ali ćemo vidjeti da to nije slučaj kod svih distribucija.

Prisjetimo se:
Kažemo da slučajna varijabla X ima Bernoullijevu distribuciju s parametrom π ako je njezina razdioba dana s

$$\mathbb{P}(X = x) = \pi^x(1-\pi)^{1-x}, \qquad x \in \{0,1\}.$$

Pišemo $X \sim B(1,\pi)$.

Primjer 2.2.4. *Binomna distribucija*
Napomena: *Kod binomne distribucije, umjesto oznake μ koristit ćemo oznaku π.*
Pretpostavimo da je $Z \sim B(n, \pi)$ i stavimo da je $Y = \frac{Z}{n}$, tako da je $Z = nY$.
Gustoća diskretne varijable Z za $\theta = \pi$ dana je s

$$f_Z(z; \theta, \rho) = \binom{n}{z} \pi^z (1-\pi)^{n-z},$$

a nakon supstitucije, funkcija gustoće varijable $Y \sim B(1, \pi)$, za $y = \frac{k}{n}$, $k = 0, ..., n$, je oblika:

$$\begin{aligned} f_Y(y; \theta, \rho) &= \binom{n}{ny} \pi^{ny} (1-\pi)^{n-ny} \\ &= \exp\left(n(y \log \pi + (1-y)\log(1-\pi)) + \log \binom{n}{ny} \right) \\ &= \exp\left(n\left(y \log\left(\frac{\pi}{1-\pi} \right) + \log(1-\pi) \right) + \log \binom{n}{ny} \right). \end{aligned}$$

Nakon što smo izraz zapisali u obliku jednadžbe (1), čitamo da je $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, (uočimo da je inverzno preslikavanje $\pi = e^\theta / 1 + e^\theta$), $\rho = n$, $a(\rho) = \frac{1}{\rho}$, $b(\theta) = \log(1 + e^\theta)$, $c(y, \rho) = \log \binom{n}{ny}$.
Prirodni parametar binomne, a specijalno i Bernoullijeve distribucije je $\log\left(\frac{\pi}{1-\pi}\right)$, očekivanje je

$$\mathbb{E}(Y) = b'(\theta) = \frac{e^\theta}{1 + e^\theta},$$

a funkcija varijance

$$V(\pi) = b''(\theta) = \frac{e^\theta}{(1 + e^\theta)^2} = \pi(1-\pi).$$

Ukoliko želimo modelirati varijablu odaziva kao Bernoullijevu slučajnu varijablu, moramo odrediti vjerojatnost $\pi = P(Y = 1) = \mathbb{E}(Y)$ kao funkciju kovarijata. Naravno, ta funkcija mora poprimati vrijednosti u intervalu $(0, 1)$. Jedan često korišten model je

$$\log\left(\frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p = \eta,$$

pri čemu je $h \equiv id$, $\theta \equiv \eta$, a $\pi = e^\theta / 1 + e^\theta$. Funkciju $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ nazivamo logit fukcija i upravo je ona kanonska funkcija povezivanja za binomnu razdiobu.

Sva svojstva koja smo pokazali u primjerima za normalnu i binomnu distribuciju možemo naći u donjoj tablici. Osim tih, navedene su još neke često upotrebljavane distribucije s pripadnim karakteristikama, poput očekivanja, funkcije varijance te kanonske funkcije povezivanja, međutim za njih nećemo provoditi poseban račun.

Tablica 2.1: *Karakteristike nekih univarijatnih distribucija iz eksponencijalne familije*

	Normalna	Poissonova	Binomna	Gamma
Notacija	$\mathrm{N}(\mu,\theta^2)$	$\mathrm{P}(\mu)$	$B(n,\pi)/n$	$G(\mu,\nu)$
ρ	σ^2	1	$1/n$	ν^{-1}
$b(\theta)$	$\theta^2/2$	$\exp(\theta)$	$\log(1+e^\theta)$	$-\log(-\theta)$
$c(y;\theta)$	$\frac{1}{2}\left(\frac{y^2}{\rho}+\log(2\pi\rho)\right)$	$-\log y!$	$\log\binom{n}{ny}$	$\nu\log(\nu y)-\log y-\log\Gamma(\nu)$
$\mathbb{E}(Y;\theta)$	θ	$\exp(\theta)$	$e^\theta/(1+e^\theta)$	$-1/\theta$
$g(\mu)$	μ	$\log\mu$	$\log\left(\frac{\mu}{1-\mu}\right)$	$1/\mu$
$V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2

Od svih distribucija iz eksponencijalne familije, od posebnog interesa u ovom radu bit će Bernoullijeva, tj. binomna distribucija jer ćemo promatrati manifestaciju određenog svojstva, kojeg možemo kodirati nulama i jedinicima.

Poglavlje 3

Logistička regresija

Metoda ispitivanja i analize ovisnosti jedne varijable (zavisne) o jednoj ili više nezavisnih varijabli naziva se regresijska analiza. Linearna i logistička regresija najpopularnije su metode u regresijskoj analizi, a glavna razlika je klasifikacija zavisne varijable, koja je kod logističke regresije kategorijska, a kod linearne kontinuirana, dok nezavisne varijable mogu biti i kategorijske i kontinuirane. Logistička regresija često se koristi ako želimo ispitati učinkovitost određenog lijeka, koji su vanjski utjecaji na bolest, kakva je sposobnost biljnih nametnika da prežive tretiranje insekticidom ili bilo što drugo što možemo razvrstati u kategorije (bračni status, zanimanje, lokacija...). Iz tog razloga, primijenjuje se u raznim granama znanosti poput medicine, farmacije, biologije, kemije, ekonomije, marketinga, demografije, psihologije, sociologije i mnogim drugim.

U praksi se nerijetko događa da je zavisna varijabla binarna ili dihotomna, što znači da može poprimiti vrijednosti koje interpretiramo odgovorima da i ne, tj. jedinicom i nulom pri čemu jedinica označava prisutstvo, a nula odsutstvo određenog svojstva ili pojave.
Pitanja ove vrste možemo opisati i pomoću višestruke linearne regresije na isti način - vrijednosti varijable obilježimo nulom ili jedinicom. Na taj način dobili bismo regresijski model koji bi mogao predvidjeti vrijednost zavisne varijable zajedno s regresijskim koeficijentima, koji bi pokazivali relativan utjecaj svake nezavisne varijable na zavisnu. Međutim, logistička regresija nudi bolje rješenje, budući da nas s pozicije predviđanja zanima kojoj od dvije moguće skupine pripada ispitanik, odnosno jedinica promatranja. Koristeći linearnu regresiju dobili bismo rješenje u kojem zavisna varijabla ima vrijednost između nule i jedinice, a predviđena vrijednost izgleda kao vjerojatnost da jedinica promatranja pripada jednoj ili drugoj skupini. Na primjer, ako je s (0) obilježen slučaj kupovine robe A, a s (1) kupovina robe B i vrijednost zavisne varijable (kupovine) iznosi 0.65 za nekog kupca, onda je vjerojatnost da će kupac kupiti robu B veća jer je vrijednost bliža jedinici. Naime, pretpostavka je da se kod višestruke linearne regresije dobivena vrijednost zavisne varijable u takvim slučajevima može promatrati kao vjerojatnost, tj. proporcija. Još je veći

problem što se kod višestruke linearne regresije često dobiju vrijednosti zavisne varijable manje od nule ili veće od jedinice. S obzirom da se takve vrijednosti ne mogu tumačiti kao vjerojatnosti, postaje jasno da predmetni model nije dobro rješenje pa je potrebno izvršiti određenu vrstu matematičke transformacije zavisne varijable kako bi se dobio logistički regresijski model, a nešto više o tom reći ćemo malo kasnije.

Osim binarne ili dihotomne, postoji i multinomna logistička regresija, no ona se javlja nešto rjeđe. Koristi se kada varijabla odaziva poprima vrijednosti koje se mogu rasporediti u više od dvije kategorije.

3.1 Multinomna logistička regresija

Glavna razlika multinomne ili polinomne logističke regresije u odnosu na dihotomnu je da postoje, kao što smo već spomenuli, barem dvije kategorije $J \geq 2$ kojima pripadaju vrijednosti varijable odaziva. U ovisnosti o kakvim se kategorijama radi, razlikujemo dvije vrste mulitnomne logističke regresije: logistička regresija s nominalnom zavisnom varijablom i logistička regresija s ordinalnom zavisnom varijabom.
Nominalna zavisna varijabla poprima vrijednosti iz skupa u kojem ne postoji ljestvica kvalitete ili redoslijed po kojem je neka kategorija više ili manje vrijedna od druge (npr. plava, zelena, žuta boja ili odgovori na pitanja u nekom upitniku: da, ne, ne znam, nema odgovora), dok kod ordinalne varijable postoji ta ljestvica ili prirodan slijed (npr. dobar, srednji, loš klijent).

Pretpostavimo da je Y slučajna varijabla čije se vrijednosti nalaze u $J \geq 2$ kategorija. Neka su $\pi_j = \mathbb{P}(Y = j)$, $j = 1, ..., J$ odgovarajuće vjerojatnosti tako da vrijedi $\sum_{j=1}^{J} \pi_j = 1$. Promotrimo N nezavisnih realizacija slučajne varijable Y na način da s y_1 označimo broj realizacija prve kategorije, y_2 broj realizacija druge kategorije... Tako dolazimo do vektora $\boldsymbol{y} = [y_1, y_2, ...y_J]^T$, gdje je $y_j \in \{0, 1, ..., N\}$ i $\sum_{j=1}^{J} y_j = 1$.

Sukladno s navedenim oznakama, definirajmo multinomnu distribuciju, koja je početna točka na koju se nastavlja multinomna logistička regresija.

Definicija 3.1.1. *Vektor $\boldsymbol{Y}$ ima multinomnu distribuciju s parametrima $N \in \mathbb{N}$ i $\boldsymbol{\pi} \in \mathbb{R}^J$ ako ima sljedeću funkciju gustoće:*

$$f(\boldsymbol{y}|N) = \frac{N!}{y_1!y_2! \cdots y_J!} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_J^{y_J},$$

a označavamo je s $\boldsymbol{M}(N, \pi_1, ..., \pi_J)$.

Primjetimo, ukoliko $J = 2$ multinomna distribucija svodi se na binomnu.

Multinomni logistički model je specijalan slučaj multivarijatnog generaliziranog linearnog modela. Slično kao kod univarijatnih, multivarijatni generalizirani modeli bazirani su na distribucijskoj i strukturnoj pretpostavci.

Neka su $X_1, X_2, ..., X_K$ nezavisne varijable, zavisna varijabla Y_i je J-dimenzionalni vektor s očekivanjem $\mu_i = \mathbb{E}(Y_i|\boldsymbol{X}_i)$, a s $\boldsymbol{X}$ označavamo matricu dizajna čiji je i-ti redak $\boldsymbol{X}_i^T$.
Distribucijska pretpostavka je da su $\boldsymbol{X}_i^T$ i Y_i nezavisni i da Y_i ima distribuciju koja pripada eksponencijalnoj familiji, tj. ima formu kao u izrazu (1).
Strukturna pretpostavka je da je očekivanje μ_i određeno linearnim prediktorom $\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_i$ u obliku $\mu_i = h(\eta_i)$, odnosno funkcija povezivanja je definirana kao inverz funkcije h, $g(\mu_i) = \eta_i$, na isti način kao što smo definirali u poglavlju *2.2 Generalizirani linearni modeli*, pri čemu je $\boldsymbol{X}$ matrica dizajna koja se sastoji od vektora $\boldsymbol{1}, \boldsymbol{X_1}, ..., \boldsymbol{X_N}$, a $\boldsymbol{\beta}$ matrica nepoznatih parametara koja se sastoji od vektora $\boldsymbol{\beta_1}, ..., \boldsymbol{\beta_K}$.

3.1.1 Nominalna logistička regresija

Napomenimo da na sličan način kao što dolazimo do samog modela dihotomne logističke regresije, dolazimo i do nominalne, međutim detaljniji postupak toga obrađujemo u poglavlju *3.2. Dihotomna logistička regresija*. Ovdje ćemo ukratko prikazati model nominalne logističke regresije.

Promotrimo varijablu Y koja ima multinomnu vjerojatnosnu distribuciju s $J \geq 2$ kategorija. S N označimo broj promatranja slučajne varijable Y. Ako je svako od N promatranja nezavisno, tada svaka varijabla Y_i, $i = 1, ..., N$ ima multinomnu distribuciju.

Obzirom da se pri svakom promatranju realizira jedna od J mogućih vrijednosti zavisne varijable Y, neka je $\boldsymbol{y}$ matrica realizacija s N redaka i $J - 1$ stupaca.

$$\boldsymbol{y} = \begin{bmatrix} y_{11} & \cdots & y_{1(J-1)} \\ \vdots & & \vdots \\ y_{N1} & \cdots & y_{N(J-1)} \end{bmatrix}$$

Svaki y_{ij} predstavlja realizaciju kategorije j pri i-tom promatranju. Ukoliko se kategorija j realizirala, y_{ij} će poprimiti vrijednost 1, a u suprotnom 0. Uočimo da vrijedi $\sum_{j=1}^{J} y_{ij} = 1$.

Neka je $\boldsymbol{\pi}$ matrica dimenzija $N \times (J - 1)$, gdje je svaki π_{ij} vjerojatnost realizacije j-te vrijednosti u i-tom promatranju zavisne varijable, tj. $\pi_{ij} = \mathbb{P}(Y_i = j)$ i vrijedi $\sum_{j=1}^{J} \pi_{ij} = 1$, $\forall i \in \{1, ..., N\}$.

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{11} & \cdots & \pi_{1(J-1)} \\ \vdots & & \vdots \\ \pi_{N1} & \cdots & \pi_{N(J-1)} \end{bmatrix}$$

Matrica dizajna nezavisnih varijabli $\boldsymbol{X}$ je dimenzije $N \times (K+1)$, gdje je K broj nezavisnih varijabli. Prvi stupac sadrži samo jedinice, $x_{i0} = 1$, $\forall i \in \{1, ..., N\}$ jer se veže uz slobodni član *(engl. intercept)*.

$$\boldsymbol{X} = \begin{bmatrix} x_{10} & \cdots & x_{1K} \\ \vdots & & \vdots \\ x_{N0} & \cdots & x_{NK} \end{bmatrix}$$

Matrica parametara $\boldsymbol{\beta}$ je dimenzija $(K + 1) \times (J - 1)$, a parametar β_{kj} veže se uz k-tu nezavisnu varijablu i j-tu vrijednost zavisne varijable, $\forall k \in \{1, ..., K\}$, $\forall j \in \{1, ..., J - 1\}$.

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \cdots & \beta_{0(J-1)} \\ \vdots & & \vdots \\ \beta_{K1} & \cdots & \beta_{K(J-1)} \end{bmatrix}$$

Kako bismo dobili *logit* funkciju kod multinomne logističke regresije služimo se istim transformacijama kao i kod binomne.

Zbog svojih obilježja, nominalna logistička regresija nam omogućuje da za baznu kategoriju odaberemo bilo koju od J kategorija, stoga uzmimo posljednju, J-tu kategoriju kao baznu. Logaritmirana izglednost prvih $J-1$ kategorija bi imala sljedeći oblik:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \log\left(\frac{\pi_{ij}}{1-\sum_{j=1}^{J-1}\pi_{ij}}\right) = \sum_{k=0}^{K} x_{ik}\beta_{kj}, \qquad i = 1, ..., N.$$

odakle slijedi:

$$\pi_{ij} = \frac{e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}{1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}, \qquad j < J,$$

tj. za baznu kategoriju vrijedi:

$$\pi_{iJ} = 1 - \sum_{j=1}^{J-1}\pi_{ij} = \frac{1}{1+\sum_{j=1}^{J-1} e^{\sum_{k=0}^{K} x_{ik}\beta_{kj}}}.$$

Kod ordinalne logističke regresiju uvodimo neke promjene.

3.1.2 Ordinalna logistička regresija

Ukoliko postoji više od dvije kategorije zavisne varijable koje međusobno čine nekakav uređeni slijed, možemo koristiti ordinalnu logističku regresiju. Postoji više modela ordinalne logističke regresije, no onaj koji se u praksi najviše koristi je kumulativni *(engl. proportional odds model)*, stoga ćemo njega objasniti. Ostali su zapravo modifikacije navedenog modela.

Neka je Y_i, $i = 1, ..., N$, zavisna ordinalna varijabla koja pri jednom promatranju poprima jednu od J kategorija. Odgovarajuće vjerojatnosti realizacija svake od kategorija pri i-tom promatranju su $\pi_{i1} = \mathbb{P}(Y_i = 1), ..., \pi_{iJ} = \mathbb{P}(Y_i = J)$. Distribucija slučajne varijable Y_i je multinomna s parametrom $\pi_i = (\pi_{i1}, ..., \pi_{iJ})$.

Za ordinalnu logističku regresiju ključne su kumulativne vjerojatnosti koje definiramo na sljedeći način:

$$\mathbb{P}(Y_i \leq j) = \pi_{i1} + \pi_{i2} + ... + \pi_{ij},$$

gdje je $j = 1, ..., J$ realizirana kategorija zavisne varijable.

Kod ovog modela zanima nas izglednost ili šansa za svaku pojedinu kategoriju zavisne varijable, ali na način da gledamo omjer sume vjerojatnosti realizacije manjih kategorija i sume vjerojatnosti realizacije kategorija većih od njih.
Prethodna definicija nam omogućuje da svaku izglednost zapišemo na sljedeći način:

$$\frac{\mathbb{P}(Y_i \leq j)}{\mathbb{P}(Y_i > j)} = \frac{\mathbb{P}(Y_i \leq j)}{1 - \mathbb{P}(Y_i \leq j)} = \frac{\pi_{i1} + \pi_{i2} + ... + \pi_{ij}}{\pi_{i(j+1)} + ... + \pi_{iJ}}$$

pa možemo izraziti logaritam sklonosti dvije kumulativne vjerojatnosti:

$$\log\left(\frac{\mathbb{P}(Y_i \leq j)}{1 - \mathbb{P}(Y_i \leq j)}\right) = \log\left(\frac{\pi_{i1} + \pi_{i2} + ... + \pi_{ij}}{\pi_{i(j+1)} + ... + \pi_{iJ}}\right).$$

Time mjerimo kolika je šansa realizacije kategorije koja je manja ili jednaka od j-te u odnosu na realizaciju kategorije koja je strogo veća od j-te.

Uključimo li nezavisne varijable u model dobivamo konačni kumulativni model logističke regresije:

$$\log\left(\frac{\mathbb{P}(Y_i \leq j)}{1 - \mathbb{P}(Y_i \leq j)}\right) = \beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + ... + \beta_{Kj}x_{iK}, \qquad j \in \{1, ..., J-1\},$$

gdje je K broj nezavisnih varijabli.

Važna pretpostavka ovog modela je da koeficijenti $\beta_{1j}, \beta_{2j}, ..., \beta_{Kj}$ ne ovise o kategoriji j, nego su jednaki za svaki $j = 1, ..., J$, za razliku od koeficijenata β_{0j} koji ovise o kategoriji j i variraju za svaku od funkcija. Ti koeficijenti su poput slobodnih članova u linearnom regresijskom modelu pa iz tog razloga prethodnu jednakost možemo zapisati tako da slobodni član β_{0j} zamijenimo s α_j:

$$\log\left(\frac{\mathbb{P}(Y_i \leq j)}{1 - \mathbb{P}(Y_i \leq j)}\right) = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK}.$$

Dakle, vektor α je oblika $(\alpha_1, \alpha_2, ..., \alpha_{J-1})$.
Iz prethodne jednadžbe možemo izraziti kumulativne vjerojatnosti u ovisnosti o poznatim vrijednostima nezavisnih varijabli:

$$\mathbb{P}(Y_i \leq j) = \frac{e^{\alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK}}}{1 + e^{\alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK}}}.$$

Budući da se svaka pojedina vjerojatnost može izraziti pomoću kumulativnih jer vrijedi $\mathbb{P}(Y_i = j) = \mathbb{P}(Y_i \leq j) - \mathbb{P}(Y_i \leq j-1)$, pokažimo da za slobodne članove vrijedi $\alpha_j < \alpha_{j+1}$, $\forall j \in \{1, ..., J-1\}$.

Iz nejednakosti $\mathbb{P}(Y_i \leq j) < \mathbb{P}(Y_i \leq j+1)$ i gornje nejednadžbe slijedi:

$$\frac{e^{\alpha_j+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_K x_{iK}}}{1+e^{\alpha_j+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_K x_{iK}}} < \frac{e^{\alpha_{j+1}+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_K x_{iK}}}{1+e^{\alpha_{j+1}+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_K x_{iK}}}.$$

Obzirom da član $\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK}$ ne utječe na nejednakost možemo ga tretirati kao konstantu i izuzeti iz nejednadzbe pa imamo:

$$\frac{e^{\alpha_j}}{1+e^{\alpha_j}} < \frac{e^{\alpha_{j+1}}}{1+e^{\alpha_{j+1}}}.$$

Množenjem dobivamo: $e^{\alpha_j}(1+e^{\alpha_{j+1}}) < (1+e^{\alpha_j})e^{\alpha_{j+1}}$, odnosno

$$e^{\alpha_j} < e^{\alpha_{j+1}},$$

iz čega slijedi $\alpha_j < \alpha_{j+1}$.

Procjenu parametara multinomnog logističkog modela možemo provesti metodom maksimalne vjerodostojnosti *(engl. maximum likelihood estimation)*, a ocjenu kvalitete modela jednim od testova: ANOVA, statistika odstupanja, promatranje AIC i BIC kriterija itd. Detaljnije o tome vidjet ćemo u sljedećem poglavlju.

3.2 Dihotomna logistička regresija

3.2.1 Izglednost i logit funkcija

U konstrukciji modela u centru promatranja je vjerojatnost s kojom slučajna varijabla poprima vrijednost jedan, $\pi_i \in (0, 1)$, tj. vjerojatnost pripadanja jednoj od dviju kategorija. Medutim, zbog ograničenosti na interval kojem ona mora pripadati potrebno je prikladnim transformacijama prijeći na cijelu realnu os. Kako bismo maknuli gornje ograničenje, promatramo izglednost (šansu).

Definicija 3.2.1. *Neka je $A \in \mathcal{F}$ promatrani događaj i $\pi = \mathbb{P}(A)$. Tada broj $\omega = \frac{\pi}{1-\pi}$ nazivamo izglednost (engl. odds) događaja A.*

Izglednost je omjer vjerojatnosti da neka jedinica promatranja pripadne jednoj kategoriji i vjerojatnosti da ne pripadne toj, već drugoj kategoriji $(1 - \pi)$.

Nadalje, logaritmiramo izglednost kako bismo maknuli donje ograničenje. Nakon navedenih transformacija dobivamo logaritmiranu izglednost *(engl. log-odds)*:

$$\text{logit: } (0, 1) \to \mathbb{R} \quad \text{logit } \pi_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right). \tag{4}$$

Primijetimo jedno zanimljivo svojstvo transformacije. Naime, kad vjerojatnost ima vrijednost 1/2 tada izraz (4) ima vrijednost 0. Negativna logaritmirana izglednost reprezentira vjerojatnosti manje od 1/2, a pozitivna veće od 1/2.

Ako promatramo dva događaja A i B, tada definiramo omjer njihovih izglednosti ili omjer šansi *(engl. odds ratio)* kao

$$\frac{\omega(A)}{\omega(B)} = \frac{\frac{\mathbb{P}(A)}{1-\mathbb{P}(A)}}{\frac{\mathbb{P}(B)}{1-\mathbb{P}(B)}}.$$

Njime izražavamo koliko je puta izglednost da se dogodi događaj A veća ili manja od izglednosti da se dogodi B.

3.2.2 Postavljanje modela i interpretacija parametara

Pretpostavimo sada da jedinke koje promatramo možemo klasificirati u k grupa tako da su u pojedinoj grupi jedinke koje imaju jednake kombinacije vrijednosti varijabli poticaja. Takve se grupe nazivaju kovarijatni razredi. Označimo s n_i broj jedinki u i-tom razredu, a s y_i realizacije slučajne varijable Y_i koje označuju broj jedinki i-tog razreda koje imaju promatrano svojstvo, odnosno za koje je $Y_i = Y_{i1} + ... + Y_{in_i}$.
Također, pretpostavimo da su za $i = 1, ..., k$, $Y_i \sim B(n_i, \pi_i)$ međusobno nezavisne slučajne varijable te $y_1, ..., y_k$ njihove realizacije. Neka je $n = n_1 + ... + n_k$.
Definirajmo slučajni vektor $\boldsymbol{Y}^T = (Y_1, ..., Y_k)$ koji pripada eksponencijalnoj familiji i dodatno pretpostavimo da je *logit* vjerojatnosti jednak linearnom prediktoru kao u *Primjeru 1.2.4.*:

$$\begin{aligned} \text{logit}\, \pi_i &= \eta_i \\ &= \boldsymbol{x}_{i\cdot}^T \boldsymbol{\beta} \quad \forall i \in \{1, ..., k\}, \end{aligned}$$

gdje su $\boldsymbol{x}_{i\cdot}^T = (x_{i0}, x_{i1}, ..., x_{ip})$, $x_{i0} = 1$, vektori retci matrice poticaja.
Ovime je dan generalizirani linearni model s funkcijom povezivanja *logit* za modeliranje binarnih podataka.

Funkciju $F : \mathbb{R} \to (0, 1)$ danu s

$$F(x) = \frac{M}{1 + e^{-k(x - x_0)}}$$

nazivamo logistička funkcija.
Logistička funkcija pripada skupini funkcija S-oblika, tzv. sigmoidalnih funkcija. U definiciji M označava maksimum, k nagib funkcije, a x_0 točku u kojoj se događa infleksija. Standardnom logističkom funkcijom nazivamo funkciju kod koje je $k = 1$, $x_0 = 0$ i $M = 1$, odnosno:

$$F(x) = \frac{1}{1 + e^{-x}}.$$

Problemu modeliranja logističkog regresijskog modela možemo pristupiti na način da vjerojtnost $\pi_i(x_{ij})$ shvatimo upravo kao standardnu logističku funkciju j-og prediktora x_{ij}. U tom slučaju nelinearnu vezu imamo imamo za cilj linearizirati, pritom koristeći iste transformacije (izglednost i logaritmiranje) te na poslijetku dobivamo i isti model:

$$\text{logit}\, \pi_i(x_{ij}) = x_{i0} + x_{i1}\beta_1 + ... + x_{ip}\beta_p.$$

Parametri β_i interpretiraju se na isti način kao kod lineranog modela, samo ne u smislu varijable odaziva, nego u smislu logaritmirane izglednosti. Pretpostavimo da promijenimo j-ti prediktor $\boldsymbol{x}_{\cdot j} \to \boldsymbol{x}_{\cdot j} + \mathbf{1}$, a sve ostale držimo fiksnima. Kod linearnog modela ta se

promjena očitovala kao promjena očekivanja varijable odaziva za β_j, a kod logističkog modela imamo:

$$\begin{aligned}\log \frac{\pi_i(x_{ij}+1)}{1-\pi_i(x_{ij}+1)} - \log \frac{\pi_i(x_{ij})}{1-\pi_i(x_{ij})} &= \beta_j \\ \log \frac{\omega(\pi_i(x_{ij}+1))}{\omega(\pi_i(x_{ij}))} &= \beta_j \\ \frac{\omega(\pi_i(x_{ij}+1))}{\omega(\pi_i(x_{ij}))} &= e^{\beta_j},\end{aligned}$$

prilikom čega s $\pi_i(x_{ij})$ označavamo vjerojatnost π_i kao funkciju j-tog prediktora x_{ij}. Prema tome, ukoliko se j-ti prediktor promijeni za jedan, izglednost da jedinka ima promatrano svojstvo promijeni se e^{β_j} puta. Kada mijenjamo neprekidni prediktor koji poprima vrijednosti na nekom intervalu, često nam promjena za jedan nije naročito bitna. Recimo da nas zanima promjena za vrijednost c. Analognim raspisom kao gore dobivamo:

$$\frac{\omega(\pi_i(x_{ij}+c))}{\omega(\pi_i(x_{ij}))} = e^{c\beta_j}$$

Sljedeće što nas zanima je kako takve promjene vrijednosti prediktora utječu na π_i. Donekle zadovoljavajući odgovor nam daje transformirani izraz

$$\pi_i = \frac{e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}}}.$$

Primjećujemo da se s desne strane jednakosti nalazi nelinearna funkcija prediktora i nema jednostavnog načina kako izraziti efekt koji promjena jednog prediktora ima na vjerojatnost slijeva. Možemo promatrati vjerojatnosti π_i kao funkcije s argumentom x_{ij} te računati derivacije:

$$\begin{aligned}\frac{d\pi_i}{dx_{ij}} &= \frac{e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}}\beta_j}{(1+e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}})^2} \\ &= \beta_j \cdot \frac{e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}}} \cdot \frac{1}{1+e^{\boldsymbol{x}_{i\cdot}^T\boldsymbol{\beta}}} \\ &= \beta_j\pi_i(1-\pi_i).\end{aligned}$$

Efekt j-te varijable poticaja na vjerojatnost π_i ovisi o parametru β_j i vrijednosti te vjerojatnosti. Taj broj se najčešće evaluira postavljanjem π_i na vrijednost relativne frekvencije uspjeha (broj jedinki s promatranim svojstvom u odnosu na ukupni broj jedinki).

Nakon što procijenimo vektor parametara $\boldsymbol{\beta}$, uvrštavajući konkretne vrijednosti varijabli poticaja mjerenih kod neke nove jedinke, ovaj model nam daje vjerojatnost s kojom ta jedinka ima promatrano svojstvo.

3.2.3 Procjena parametara

Za razliku od linearne regresije gdje se u svrhu procjene parametara modela koristi metoda najmanjih kvadrata, kod logističkog modela poslužit ćemo se metodom maksimalne vjerodostojnosti.

Pretpostavimo da je $\boldsymbol{Y}^T = (Y_1, ..., Y_n)$ slučajni vektor čije komponente pripadaju modelu $\mathcal{P} = \{f(\cdot; \theta) : \theta \in \Theta\}$ definiranom na statističkoj strukturi $(\Omega, \mathcal{F}, \mathcal{P})$. Ako je $\boldsymbol{y}^T = (y_1, ..., y_n)$ jedna njegova realizacija, tada je vjerodostojnost funkcija $L : \Theta \to \mathbb{R}$ definirana s:

$$L(\theta) \equiv L(\theta; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta) \tag{5}$$

Definicija 3.2.2. *Statistika $\hat{\theta} = \hat{\theta}(\boldsymbol{Y})$ je procjenitelj maksimalne vjerodostojnosti (MLE) za θ ako vrijedi:*

$$L(\hat{\theta}; \boldsymbol{Y}) = \max_{\theta \in \Theta} L(\boldsymbol{Y}; \theta).$$

Maksimizacija funkcije L ekvivalentna je maksimizaciji log-vjerodostojnosti $l = \log(L)$ jer je $\log(\cdot)$ strogo rastuća injekcija pa je to u praksi vrlo često lakše izvesti.
Pogledajmo kako ta funkcija izgleda kada slučajne varijable potječu iz binomne distribucije.

Promatramo k nezavisnih slučajnih varijabli $Y_i \sim B(n_i, \pi_i)$ koje odgovaraju kovarijatnim razredima te generalizirani model $g(\pi_i) = \eta_i$ za $\theta = \boldsymbol{\pi}$:

$$\begin{aligned} L(\boldsymbol{\pi}; \boldsymbol{y}) &= \prod_{i=1}^{k} f(y_i; \boldsymbol{\pi}) \\ &= \prod_{i=1}^{k} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \end{aligned}$$

Logaritmiranjem dobivamo

$$\begin{aligned} l(\boldsymbol{\pi}; \boldsymbol{y}) &= \log L(\boldsymbol{\pi}; \boldsymbol{y}) \\ &= \sum_{i=1}^{k} \left(\log \binom{n_i}{y_i} + y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i) \right) \\ &= C + \sum_{i=1}^{k} \left(y_i \log g^{-1}(\eta_i) + (n_i - y_i) \log(1 - g^{-1}(\eta_i)) \right), \end{aligned}$$

gdje smo s C označili konstantni član $\sum_{i=1}^{k} \log \binom{n_i}{y_i}$ koji nam ne igra nikakvu ulogu u procjeni.
Koristeći definicije linearnog prediktora i funkcije povezivanja, zaključujemo da je log-vjerodostojnost funkcija nepoznatih parametara modela $\beta_0, \beta_1, ..., \beta_p$. Konkretno, uzmemo li u obzir i način na koji smo definirali generalizirani linearni model s funkcijom povezivanja *logit* za modeliranje binarnih podataka:

$$g(\pi_i) = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \eta_i \Rightarrow \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}} \Rightarrow \pi_i = \frac{1}{1 + e^{-\eta_i}},$$

dobivamo:

$$\begin{aligned} l(\boldsymbol{\pi}(\boldsymbol{\beta}); \boldsymbol{y}) &= C + \sum_{i=1}^{k} \left(y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + n_i \log(1 - \pi_i) \right) \\ &= C + \sum_{i=1}^{k} \left(y_i \eta_i + n_i \log \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right) \right) \\ &= C + \sum_{i=1}^{k} \left(y_i \eta_i - n_i \log(1 + e^{\eta_i}) \right) \\ &= C + \sum_{i=1}^{k} \sum_{j=0}^{p} y_i x_{ij} \beta_j - \sum_{i=1}^{k} n_i \log \left(1 + e^{\sum_{j=0}^{p} x_{ij} \beta_j} \right). \end{aligned}$$

Da bismo mogli tražiti procjenitelje maksimalne vjerodostojnosti kao stacionarne točke funkcije log-vjerodostojnosti, trebali bismo uvesti nekoliko novih pojmova i statistički model morao bi zadovoljavati još neke uvjete, međutim u ovom radu nećemo se toliko detaljno baviti time. Dovoljno je znati da bismo ih dobili maksimizacijom gornje funkcije.

3.2.4 Prilagodba modela podacima

Nakon procjene parametara, prirodno pitanje koje se nameće je: "Koliko se dobro model prilagodio podacima?", odnosno: "Kolika je razlika između opaženih realizacija y_i varijabli $Y_i \sim B(n_i, \pi_i)$ i prilagođenih vrijednosti $\hat{y}_i = n_i\hat{\pi}_i$, za $i = 1, ..., k$?". Da bismo odgovorili na to pitanje, mjerit ćemo "udaljenost" modela od stvarnih podataka, što odgovara manjku prilagodbe koju model ima.
Postoje mnoge statistike kojima se opisuje ta razlika, no najčešće se koriste one temeljene na funkciji vjerodostojnosti, stoga ćemo u nastavku predstaviti jednu takvu, a poznata je pod nazivom statistika odstupanja.

Prisjetimo se, za logistički model log-vjerodostojnost je:

$$l(\boldsymbol{\pi}; \boldsymbol{y}) = C + \sum_{i=1}^{k} \left(y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) \right). \tag{6}$$

Za konkretnu realizaciju $\boldsymbol{y}$ funkcija vjerodostojnosti objedinjuje informaciju o nepoznatim parametrima promatranog modela. Vrijednost log-vjerodostojnosti koju dobivamo uvrštavanjem procjenitelja maksimalne vjerodostojnosti u 6 govori nam do koje se mjere promatrani model prilagodio podacima. Budući da ona ovisi o broju opažanja u uzorku, ne možemo je kao takvu koristiti za opisivanje nedostatka prilagodbe. Potrebno ju je usporediti s vrijednošću koju poprima pod pretpostavkom nekog drugog, alternativnog modela. Ovakvim pristupom dobivamo mjeru nedostatka prilagodbe modela koju zovemo odstupanje *(engl. deviance)*. Ona je analogon sume kvadratnih pogrešaka u klasičnom linearnom modelu. U pozadini njezine definicije je test omjera vjerodostojnosti za usporedbu dvaju ugniježđenih modela s pripadnim pretpostavkama:

$$H_0 : \text{Model } M \text{ je točan.}$$
$$H_1 : \text{Model } M \text{ nije točan.}$$

Pri tome je u H_0 model koji se promatra ($p + 1 < k$), a alternativna hipoteza reprezentira tzv. puni ili saturirani model M_f. On je egzaktno prilagođen podacima jer svakom opažanju odgovara jedan parametar ($p + 1 = k$).
Imamo:

$$D = -2 \log \frac{\text{vjerodostojnost modela } M}{\text{vjerodostojnost modela } M_f}. \tag{7}$$

Navedenim izrazom zapravo određujemo koliko je naš model lošiji od perfektne prilagodbe punog modela. Ako s $\tilde{\pi}_i$ označimo procijenjene vrijednosti parametara punog modela, slijedi:

$$
\begin{aligned}
D &= -2l(\hat{\boldsymbol{\pi}}; \boldsymbol{y}) + 2l(\tilde{\boldsymbol{\pi}}; \boldsymbol{y}) \\
&= (2C_2 - 2C_1) + 2\sum_{i=1}^{k}\left(y_i \log\frac{\tilde{\pi}_i}{\hat{\pi}_i} + (n_i - y_i)\log\frac{1-\tilde{\pi}_i}{1-\hat{\pi}_i}\right) \\
&= (2C_2 - 2C_1) + 2\sum_{i=1}^{k}\left(y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right).
\end{aligned}
$$

Zanemarivanjem konstantnog člana dobivamo:

$$
D(\boldsymbol{y}; \hat{\boldsymbol{\pi}}) = 2\sum_{i=1}^{k}\left(y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right).
$$

Kada je model savršeno prilagođen podacima, što u praksi nikad nije slučaj, omjer opaženih i prilagođenih vrijednosti je 1 pa je D jednaka 0. Prema tome, cilj je promatranu statistiku učiniti što manjom. Izraz D poprima velike vrijednosti kada je brojnik u (7) relativno mali u odnosu na nazivnik što ukazuje na nedovoljno dobar model, odnosno na model koji je presiromašan. S druge strane, vrijednosti D su manje kad su vrijednosti brojnika i nazivnika bliske što znači da je promatrani model dovoljno dobar.

Promotrimo na konkretnom primjeru dihotomne logističke regresije s kontinuiranim nezavisnim varijablama pojmove koje smo spominjali u ovom poglavlju.

3.2.5 Primjer: Donnerova karavana

Na proljeće 1846. godine braća Jackob i George Donner sa svojim obiteljima te obitelj Jamesa F. Reeda odlučili su napustiti Springfield (Illinois) i krenuti na selidbu prema Californiji na Zapadu. Organizirali su karavanu s devet velikih zaprežnih kola i mnogo stoke. Žureći se da prijeđu Sierru Nevadu prije nego dođe zima odlučili su krenuti neispitanim putevima i zaglavili u dubokom snijegu.
Od 87 ljudi koji su krenuli na put, do 21. travnja 1847. godine, kada su spašeni, 40 njih je umrlo.

Potaknuti ovim događajem, postavljamo za ciljeve:
• Predvidjeti vjerojatnost preživljavanja ovisno o dobi.
• Uzimajući u obzir spol testirati da li žene u odnosu na muškarce, s većom vjerojatnošću preživljavaju teške uvjete te utječe li starost različito na stopu preživljavanja muškaraca i žena.

Koristit ćemo programski jezik SAS i promatrati 45 opažanja koja sadrže podatke o dobi, spolu (1 = muški spol, 0 = ženski) te je li osoba preživjela (1) ili nije (0).

Uvodimo varijable:
$Y_i = 1$ ako je i-ta osoba preživjela
$Y_i = 0$ ako i-ta osoba nije preživjela
X_{1i} = dob osobe i
$X_{2i} = 1$ ako je osoba i muškog spola
$X_{2i} = 0$ ako je osoba i ženskog spola

3.2.5.1 Predviđanje vjerojatnosti preživljavanja ovisno o dobi

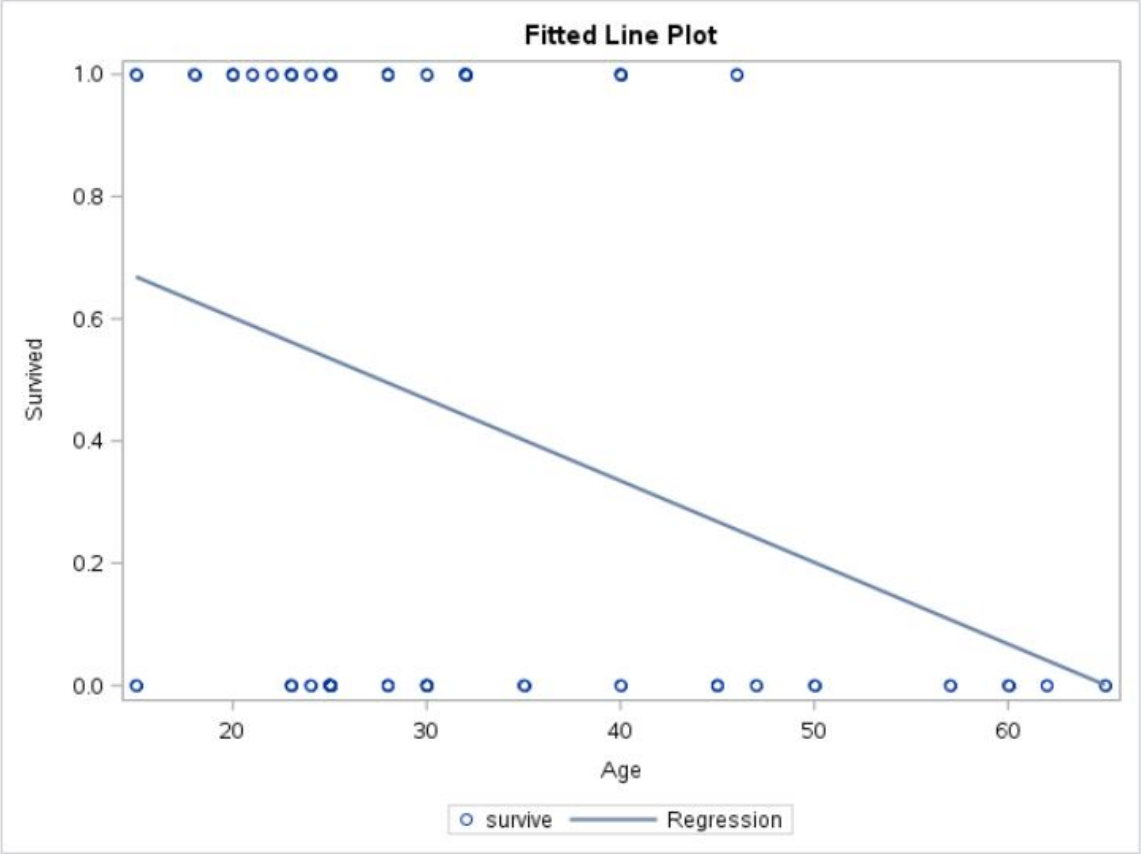
Početni model kojim računamo vjerojatnost preživljavanja uzimajući u obzir dob, bio bi:

$$\mathbb{P}(Y = 1|X_1) = \beta_0 + \beta_1 X_1,$$

a očekivane vrijednosti računali bismo s:

$$\mathbb{E}(Y = 1|X_1) = \beta_0 + \beta_1 X_1.$$

Uočimo da se radi o modelu linearne regresije i u tom slučaju problem bi bila nelinearnost, tj. linearni model mogao bi dati predviđene vrijednosti izvan intervala (0,1). Drugi problem je što varijanca $n\pi(1 - \pi)$ nije konstantna, a on nastaje jer Y slijedi binomnu distribuciju za $n = 1$, odnosno Bernoullijevu.



Slika 3.1: Graf linearnog regresijskog modela

Gornja slika prikazuje graf linearnog regresijskog modela za dane podatke. Procijenjen model je: $\mathbb{E}(Survival|Age) = 0.8692 - 0.01336Age$, pri čemu procjenjene parametre dobivamo iz rezultata u SAS-u. U slučaju da računamo vjerojatnost preživljavanja za 70-godišnju osobu, dobili bismo
$\mathbb{P}(Survival = 1|Age = 70) = 0.8692 - 0.01336 \times 70 = -0.0658$, što je negativna vrijednost, a znamo da vjerojatnost ne može biti negativna.

Dakle, linearni regresijski model nije pogodan za ovakvu vrstu podataka, stoga uvodimo logistički regesijski model (koristeći iste oznake kao u prethodnom poglavlju), u kojem je vjerojatnost preživljavanja $\pi_i = \mathbb{P}(Survival = 1)$ za i-tu osobu dobi X_{1i}, dana s

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i})}{1 + \exp(\beta_0 + \beta_1 X_{1i})}$$

Ako logaritmiramo izglednost da i-ta osoba preživi, logistički regresijski model poprima

oblik:

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_{1i}.$$

Pogledajmo izlaznu tablicu u SAS-u koja daje procjenjene parametre za gornji model:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8183	0.9993	3.3106	0.0688
age	1	-0.0665	0.0322	4.2553	0.0391

Čitamo da je naš model oblika:

$$\log \frac{\pi_i}{1-\pi_i} = 1.8183 - 0.0665 X_{1i}.$$

Očekivana vjerojatnost preživljavanja 70-godišnje osobe (bez obzira na spol, budući da ta varijabla nije uključena u model) sada iznosi

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp(1.8183 - 0.0665 \times 70)}{1 + \exp(1.8183 - 0.0665 \times 70)} = 0.055,$$

a ta vrijednost u kontekstu vjerojatnosti ima smisla.

$$\exp(\beta_0) = \exp(1.8183) = 3.26$$

se smatra procjenom šansom preživljavanja kada je dob = 0, odnosno prije nego se podaci o dobi uzmu u obzir. Tada vjerojatnost preživljavanja iznosi 3,26 / (1 + 3,26) = 0,77.
Za tumačenje ishoda logističke regresije koeficijent β_0 nije bitan, ali je neophodan za model.

Eksponencirani koeficijent β_1

$$\exp(\beta_1) = \exp(-0.0665) = 0.94$$

govori da svake godine povećanje dobi umnožava šanse opstanka za 0.94 u odnosu na šanse koje su bile prošle godine. Dakle, izgledi opstanka smanjuju se s dobi. Ovo je procijenjeni omjer izgleda i u SAS-u ova je vrijednost i njezin interval pouzdanosti od 95% dan u izlaznoj tablici "Odds Ratio Estimates":

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	0.936	0.878	0.997

Ukoliko želimo mjeriti koliko se dobro naš model prilagodio stvarnim podacima, prvo nam na pamet padaju Pearsonova χ^2 statistika ili mjera odstupanja. Međutim, budući da u ovom slučaju imamo samo 45 jedinki, ne možemo koristiti nijednu od spomenutih statistika jer nije zadovoljen uvjet za veličinu uzorka.

Kada su nezavisne varijable kontinuirane, teško je analizirati prilagodbu modela podacima bez grupiranja podataka. Prisjetimo se Hosmer-Lemeshowove statistike koja se uglavnom koristi kod rijetkih podataka i grupira ih po prilagođenim vjerojatnostima tako da se u svakoj grupi nalazi približno jednak broj jedinki.

Partition for the Hosmer and Lemeshow Test					
		survive = 1		survive = 0	
Group	Total	Observed	Expected	Observed	Expected
1	5	0	0.57	5	4.43
2	7	3	1.82	4	5.18
3	4	3	1.65	1	2.35
4	4	1	1.82	3	2.18
5	4	2	1.96	2	2.04
6	8	2	4.31	6	3.69
7	6	3	3.40	3	2.60
8	7	6	4.47	1	2.53

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
8.5165	6	0.2027

Slika 3.2: Rezultat Hosmer-Lemeshowove statistike u SAS-u

Na temelju gornjeg testa zbog velike p-vrijednosti zaključujemo da se model dobro prilagodio podacima.

3.2.5.2 Uzimajući u obzir spol, da li žene s većom vjerojatnošću preživljavaju teške uvjete, u odnosu na muškarce?

Ovo pitanje uključuje modeliranje punog i reduciranog modela logističke regresije te uspoređivanje istih pomoću $-2\log L$ testa.
Općenito, promotrimo:

H_0 : Reducirani model odgovara podacima.
H_a : Puni model odgovara podacima.

Reducirani model sadrži sve varijable osim varijable p i oblika je:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_{p-1} X_{(p-1)i},$$

dok puni model uključuje i varijablu od interesa, p:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_{p-1} X_{(p-1)i} + \beta_p X_{pi}.$$

Prisjetimo se, vrijednost log-vjerodostojnosti govori nam do koje se mjere promatrani model prilagodio podacima. Što je veća ta vrijednost, model bolje odgovara podacima. Dodavanje parametara uvijek rezultira povećanjem log-vjerodostojnosti pa tako puni model uvijek ima veću log-vjerodostojnost od reduciranog modela. Pitanje je zapravo: "Je li dodavanje varijable doista potrebno ili je varijabilnost zavisne varijable dovoljno dobro opisana kraćim modelom?".

Kako bismo testirali utjecaj spola na podatke, uspoređujemo reducirani model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i},$$

gdje je X_{1i} dob i-te osobe, s punimo modelom:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i},$$

gdje je X_{1i} dob i-te osobe, a X_{2i} spol, pri čemu je muški spol = 1, a ženski = 0.
Ako testiramo samo jedan parametar, ovaj test bi trebao dovesti do istog zaključka kao da testiramo hipoteze: $H_0 : \beta_2 = 0$ i $H_a : \beta_2 \neq 0$.

Promatramo li izlazne tablice u SAS-u, možemo vidjeti da za reducirani model $-2 \log L$ iznosi 56.291, dok je za puni model 51.256. Vrijedi da je
56.291 - 51.256 = 5.035 > 5.02 = $\chi^2_1(0.975)$, odakle zaključujemo da na razini značajnosti $\alpha = 0.025$ možemo odbaciti H_0 hipotezu po kojoj spol nema učinka na podatke.

Promotrimo sljedeću tablicu odakle dobivamo drukčije rezultate:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.2304	1.3870	5.4248	0.0199
age	1	-0.0782	0.0373	4.3988	0.0360
sex	1	-1.5973	0.7555	4.4699	0.0345

Čitamo da je p-vrijednost 0.0345 pa ne bismo odbacili H_0 hipotezu u korist alternative. To se događa jer je Waldova kvadratna statistika osjetljiva na rijetkost podataka. Iz tog razloga uzet ćemo u obzir rezultate dobivene $-2\log L$ testom.
Dakle, odlučujemo se za puni model za koji vrijedi:

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp(3.2304 - 0.0782X_1 - 1.5973X_2)}{1 + \exp(3.2304 - 0.0782X_1 - 1.5973X_2)}.$$

Također, iz SAS-a, Hosmer-Lemeshowova statistika daje p-vrijednost = 0.2305, što nam ukazuje da se model umjereno dobro opisuje podatke, štoviše, bolje nego reducirani model.

Kao i ranije, procijenjene omjere izgleda opstanka s njihovim 95% pouzdanim intervalima možemo pronaći u tablici pod nazivom "Odds Ratio Estimates". Za njih ćemo u sljedećem poglavlju raditi simulacijsku analizu.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	0.925	0.860	0.995
sex	0.202	0.046	0.890

Zaključci:
Izgledi da preživi muškarac u odnosu na ženu su $\exp(-1.5973) = 0.202$ puta. Drugim riječima, povećanjem "vrijednosti" na varijabli spol za 1, tj. sa 0 (žene) na 1 (muškarci), šanse preživljavanja povećavaju se 0.202 puta. Odnosno, šanse da muškarac preživi predstavljaju 20.2% odgovarajućih šansi za žene. Ekvivalentno tome možemo reći da su izgledi za preživljavanje žene $1/0.202 \approx 5$ puta veći od izgleda za preživljavanje muškarca.
Omjer izglednosti je konstantan, a to znači da se ne mijenja u ovisnosti o vrijednostima prediktorskih varijabli, što nije slučaj s vjerojatnošću. Vidjeli smo već ranije da se s porastom dobi izgledi za opstanak smanjuju. Isto tako, da su veći izgledi da preživi žena nego muškarac.

Izračunajmo vjerojatnost preživljavanja 24-godišnje žene i muškarca.
Za žene:

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp(3.2304 - 0.0782 \times 24 - 1.5973 \times 0)}{1 + \exp(3.2304 - 0.0782 \times 24 - 1.5973 \times 0} = 0.806.$$

Za muškarce:

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp(3.2304 - 0.0782 \times 24 - 1.5973 \times 1)}{1 + \exp(3.2304 - 0.0782 \times 24 - 1.5973 \times 1} = 0.439.$$

Kako smo i očekivali, vjerojatnost preživljavanja za žene puno je veća nego vjerojatnost preživljavanja za muškarce.

Gore navedene analize pretpostavljaju da nema interakcije između dobi i spola. Da bismo testirali utjecaj interakcije na podatke, usporedimo sljedeće modele:
reducirani:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i},$$

gdje je X_{1i} dob, a X_{2i} spol i-tog putnika,
i puni model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i},$$

uz iste oznake.

Iz izlaznih tablica u SAS-u čitamo da je $-2\log L$ za reducirani model 51.256, a za puni 47.346 pa slijedi 51.256 - 47.346 = 3.91 > 3.84 = $\chi^2_1(0.95)$. Dakle, na razini značajnosti 0.05 možemo odbiti hipotezu H_0 koja izriče da interakcija između dobi i spola nije statistički značajna. Iako slab dokaz, i dalje podupire ono što smo dokazali.

Iz tablice u nastavku primijetimo da ne bismo odbacili hipotezu H_0: $\beta_{age*sex} = 0$ kao što je bio slučaj ranije.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	7.2450	3.2046	5.1114	0.0238
age	1	-0.1940	0.0874	4.9289	0.0264
sex	1	-6.9267	3.3983	4.1546	0.0415
age*sex	1	0.1616	0.0942	2.9385	0.0865

Ponovo zaključujemo da je bolji puni model, odnosno onaj koji uključuje interakciju dobi i spola jer i Hosmer-Lemeshowova statistika daje bolje rezultate nego za reducirani model. Stoga je:

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp(7.2450 - 0.1940X_1 - 6.9267X_2 + 0.1616X_1X_2)}{1 + \exp(7.2450 - 0.1940X_1 - 6.9267X_2 + 0.1616X_1X_2)}.$$

Izračunajmo sada vjerojatnost preživljavanja 24-godišnje žene i muškarca.

Za žene:

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp\left(7.2450 - 0.1940 \times 24 - 6.9267 \times 0 + 0.1616 \times 0\right)}{1 + \exp\left(7.2450 - 0.1940 \times 24 - 6.9267 \times 0 + 0.1616 \times 0\right)} = 0.930.$$

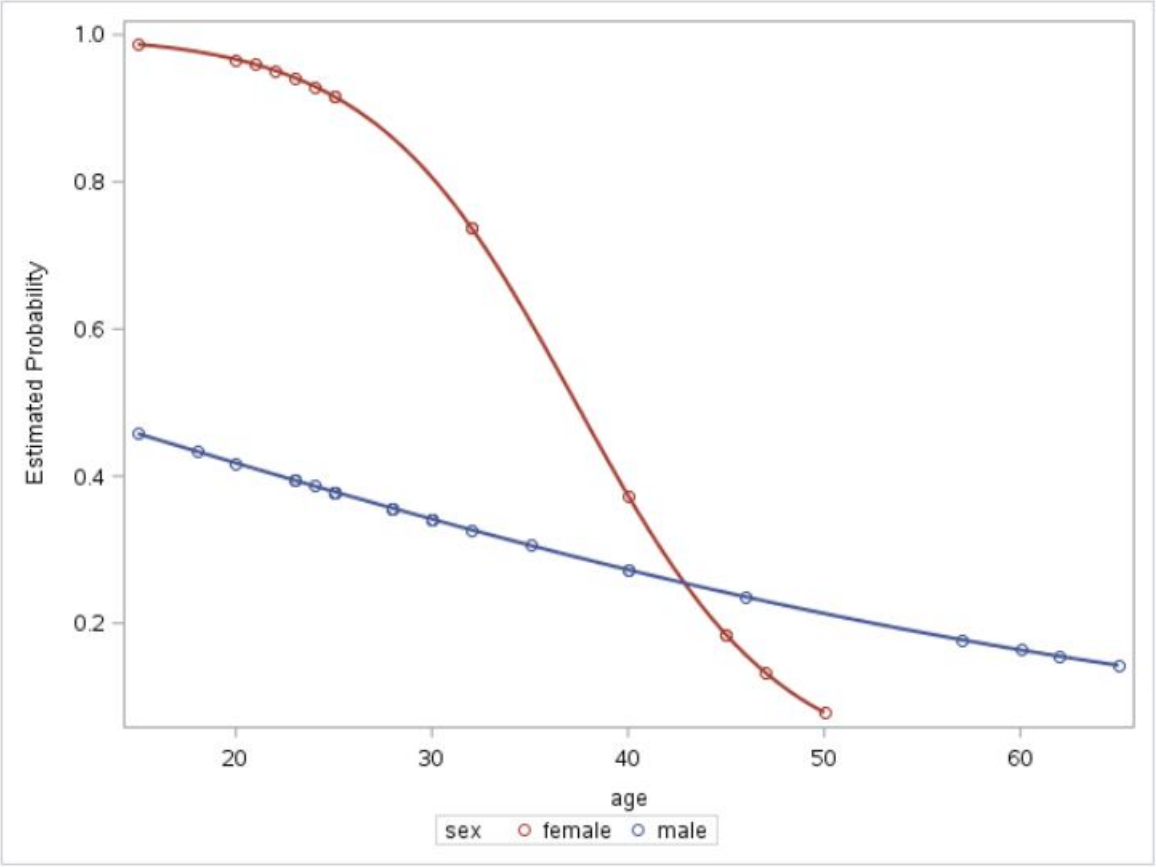
Izgledi za preživljavanje množe se faktorom $\exp(-0,194) = 0,824$ za svaku dodatnu godinu starosti, tj. sa svakom dodatnom godinom starosti izgledi za preživljavanje se smanjuju za otprilike 18%.

Za muškarce:

$$\hat{\pi} = \mathbb{P}(Survival) = \frac{\exp\left(7.2450 - 0.1940 \times 24 - 6.9267 \times 1 + 0.1616 \times (24 \times 1)\right)}{1 + \exp\left(7.2450 - 0.1940 \times 24 - 6.9267 \times 1 + 0.1616 \times (24 \times 1)\right)} = 0.387.$$

Izgledi za preživljavanje množe se faktorom $\exp(-0.0324) = 0.968$ za svaku dodatnu godinu starosti.

Uzmemo li to sve u obzir, jasno možemo vidjeti da se grafički prikaz rezultata modela (Slika 3.3) podudara sa zaključcima do kojih smo došli.



Slika 3.3: Grafički prikaz vjerojatnosti preživljavanja s obzirom na dob i spol

Poglavlje 4

Istraživački problem

4.1 Predstavljanje problema

U ovom dijelu rada provest ćemo Monte Carlo studiju za ispitivanje utjecaja veličine uzorka i tipa distribucije prediktorskih varijabli na vjerojatnosti pokrivanja *(engl. coverage probabilities)* 95% i 99% Waldovih pouzdanih intervala za regresijske koeficijente u modelu dihotomne logističke regresije.

Navedeni eksperiment provest ćemo za kombinacije u kojima je:

- Veličina uzorka n = 20, 50, 100, 200
- Tip distribucije:
  - Normalna N(0,1)
  - Uniformna U(-25,+25)
  - Gamma $\Gamma(0.5, \beta, \theta)$
  - Kontaminirana normalna (90% podataka iz N(0,1), a 10% podataka iz N(0,5))
- Broj prediktorskih varijabli: 1, 2.

Vrijednosti stvarnih (populacijskih) regresijskih koeficijenata bit će:

- U modelu s jednim prediktorom: $\beta_{10} = 2,\ \beta_{11} = 4$
- U modelu s dva prediktora: $\beta_{20} = 2,\ \beta_{21} = 4,\ \beta_{22} = -2$

U modelu s dvije prediktorske varijable promatrat ćemo slučajeve kada su varijable međusobno nekorelirane i kada su korelirane, uzimajući različite koeficijente korelacije: $\rho = 0, 0.3, 0.5, 0.9$.

Monte Carlo simulacija je postupak eksperimentiranja pri kojem se pomoću generiranja slučajnih brojeva te velikog broja izračuna i ponavljanja predviđa ponašanje složenih sustava u matematici, fizici, ekonomiji i drugim granama znanosti. U statistici se često koristi za procjenu određenih svojstava uzorka kao što je npr. procjenjivanje proporcije u našem slučaju.

Od velike važnosti je odrediti broj ponavljanja za svaku kombinaciju faktora. Očito je da se za više ponavljanja dobivaju i precizniji rezultati, tj. manja je pogreška u procjeni. Mi ćemo za svaku od navedenih kombinacija izvesti 500 replikacija zbog vremena izvođenja kao ograničavajućeg faktora, a poseban slučaj će biti analiza rezultata modela s jednom prediktorskom varijablom nakon 2000 ponavljanja uzoraka.

Ideja je za svaku replikaciju generirati n slučajnih brojeva koji slijede određenu distribuciju, a vrijednosti sredine (μ) i standardne devijacije (σ) bit će jednake za sve distribucije i iznositi $\mu = 0, \sigma = 1$. Potom ćemo izračunati 95% i 99% Waldove pouzdane intervale za regresijske koeficijente.

Cilj nam je procijeniti vjerojatnosti pokrivanja, odnosno izračunati proporcije uzoraka (od 500 uzoraka) za koje se populacijski regresijski koeficijenti $\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{22}$ nalaze unutar 95% i 99% intervala pouzdanosti.

4.2 Razrada problema

Budući da ćemo prediktorske varijable simulirati iz različitih distribucija, važno je za sve distribucije odrediti parametre takve da su vrijednosti za očekivanje i standardnu devijaciju jednake ($\mu = 0$, $\sigma = 1$) kako bismo mogli uspoređivati dobivene rezultate.
Provodimo standardizaciju na sljedeći način:

4.2.0.1 Normalna razdioba

U ovom slučaju simuliramo podatke iz normalne razdiobe N(0,1), stoga ne moramo raditi nikakve promjene jer već vrijedi $\mu = 0$ i $\sigma = 1$.

4.2.0.2 Uniformna razdioba

Ukoliko je $X \sim \mathrm{U}(a, b)$, znamo da je $\mathbb{E}(X) = \frac{a+b}{2}$ i $Var(X) = \frac{(b-a)^2}{12}$.
Da bi vrijedilo $\mu = 0$ i $\sigma = 1$, moramo riješiti sustav jednadžbi:

$$0 = \frac{a+b}{2}$$
$$1^2 = \frac{(b-a)^2}{12}.$$

Dobivamo $a = -\sqrt{3}$ i $b = \sqrt{3}$. Dakle, simulirat ćemo podatke iz uniformne razdiobe $\mathrm{U}(-\sqrt{3}, \sqrt{3})$.

4.2.0.3 Gamma razdioba

Općenito, ako je $X \sim \Gamma(\alpha, \beta, \theta)$, vrijedi $\mathbb{E}(X) = \alpha\beta + \theta$ i $Var(X) = \alpha\beta^2$.
Budući da ćemo generirati podatke s parametrom $\alpha = 0.5$, rješavamo sustav jednadžbi:

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

Zaključujemo da moramo generirati podatke koji pripadaju $\Gamma\left(0.5, \sqrt{2}, -0.5\sqrt{2}\right)$ distribuciji.

4.2.0.4 Kontaminirana normalna razdioba

90% podataka bit će iz normalne distribucije N(0,1), a 10% iz N(0,5), stoga radimo standardizaciju i nakon toga kreiramo model.

4.2.0.5 Postupak rješavanja problema

Nakon određivanja parametara, simuliramo podatke tako da za svaku distribuciju simuliramo 500 replikacija uzoraka različitih duljina, $n = 20, 50, 100, 200$.

Na slučajan način odabiremo vrijednost za varijablu x ukoliko radimo model s jednom varijablom (za x i z ako radimo model s dvije varijable) ovisno o distribuciji iz koje dolazi (dolaze). Nakon toga računamo $\eta = 2 + 4x$, odnosno $\eta = 2 + 4x - 2z$. Za vrijednost y uzimamo slučajan broj iz Bernoullijeve razdiobe s parametrom $e^{\eta}/(1+e^{\eta})$. Time dobivamo jedan simulirani broj iz Bernoullijeve razdiobe. Takav postupak ponavljamo n puta kako bismo dobili jedan uzorak, a potrebno nam je 500 replikacija jednog uzorka.

Za svaki od 500 uzoraka duljine n odredimo koji model najbolje opisuje zavisnu varijablu u ovisnosti o nezavisnoj (nezavisnim) te dobivamo procjene za koeficijente, kao i njihove pouzdane intervale. Zatim promatramo da li stvarni pripadni koeficijenti upadaju u pouzdane intervale njihovih procjena te računamo omjer onih koji upadaju i ne upadaju.

Dobivena proporcija nam govori kolika je vjerojatnost pokrivanja Waldovih pouzdanih intervala za pripadne koeficijente.

Za određivanje vjerojatnosti koristit ćemo programski jezik SAS.

4.3 Rezultati

4.3.1 Model jednostavne logističke regresije

U ovom potpoglavlju razmatrat ćemo rezultate dobivene za model s jednom prediktorskom varijablom.

Za početak pogledajmo rezultate dobivene od 500 ponavljanja uzoraka različitih distribucija. Prisjetimo se, vrijednosti stvarnih regresijskih koeficijenata u modelu s jednim prediktorom su $\beta_{10} = 2, \beta_{11} = 4$ i rezultati predstavljaju proporcije uzoraka (od njih 500) za koje se regresijski koeficijenti β_{10}, β_{11} nalaze unutar 95% i 99% pouzdanih intervala njihovih procjena.

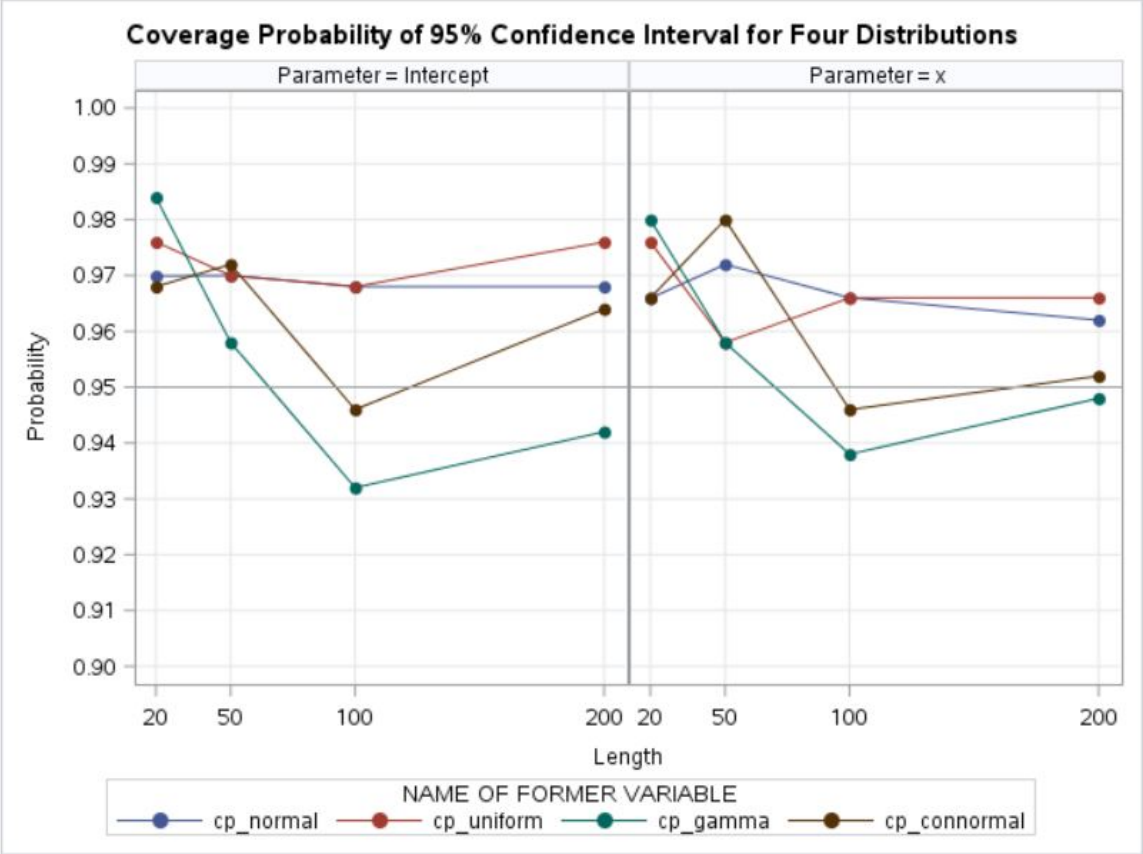
Program u SAS-u nam daje sljedeće rezultate:

Coverage Probability of 95% Confidence Interval for Different Distributions

Parameter	n	cp_normal	cp_uniform	cp_gamma	cp_connormal
Intercept	20	0.96994	0.976	0.984	0.968
Intercept	50	0.97000	0.970	0.958	0.972
Intercept	100	0.96800	0.968	0.932	0.946
Intercept	200	0.96800	0.976	0.942	0.964
x	20	0.96593	0.976	0.980	0.966
x	50	0.97200	0.958	0.958	0.980
x	100	0.96600	0.966	0.938	0.946
x	200	0.96200	0.966	0.948	0.952

Slika 4.1: Tablica vjerojatnosti pokrivanja 95% pouzdanih intervala

Uočavamo da su vrijednosti za uzorke duljina $n = 20, 50$ veće od nominalnih kod svih distribucija. Za uzorke većih duljina najmanje vjerojatnosti javljaju se kod Gamma distribucije i kontaminirane normalne, iako za $n = 20$ Gamma distribucija daje najveće vjerojatnosti. Normalna i uniformna razdioba imaju za velike uzorke najbolje rezultate, kao što možemo vidjeti i iz grafičkog prikaza.



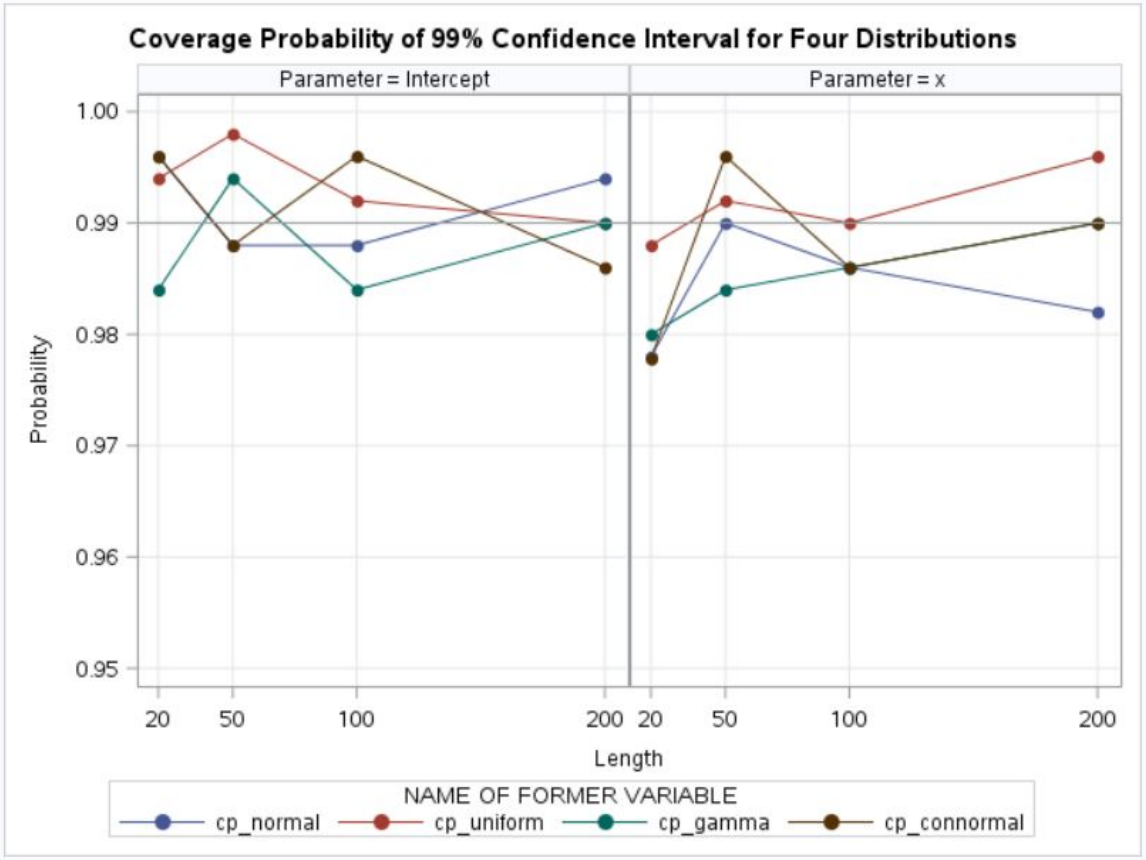
Slika 4.2: Grafički prikaz vjerojatnosti pokrivanja 95% pouzdanih intervala

U slučaju pokrivanja 99% Waldovih pouzdanih intervala, dobivamo drugačije rezultate. Kod Gamma distribucije vjerojatnosti pokrivanja za uzorke duljine $n = 20$ su male neovisno koji parametar promatramo. Ako gledamo rezultate za parametar=*Intercept*, za uzorke najveće duljine uočavamo najveće vrijednosti kod normalne distribucije, a za parametar=x najveću vjerojatnost pogađanja daje uniformna distribucija.

Kao što možemo vidjeti na grafičkom prikazu koji slijedi, linije koje predstavljaju različite distribucije su ispresijecane i teško je uočiti neku pravilnost u ponašanju vrijednosti.

Coverage Probability of 99% Confidence Interval for Different Distributions

Parameter	n	cp_normal	cp_uniform	cp_gamma	cp_connormal
Intercept	20	0.996	0.994	0.984	0.99597
Intercept	50	0.988	0.998	0.994	0.98800
Intercept	100	0.988	0.992	0.984	0.99600
Intercept	200	0.994	0.990	0.990	0.98600
x	20	0.978	0.988	0.980	0.97782
x	50	0.990	0.992	0.984	0.99600
x	100	0.986	0.990	0.986	0.98600
x	200	0.982	0.996	0.990	0.99000



Slika 4.3: Tablica i grafički prikaz vjerojatnosti pokrivanja 99% pouzdanih intervala

4.3.2 Model jednostavne logističke regresije nakon 2000 replikacija

Kako bismo se uvjerili da model s više ponavljanja uzoraka daje preciznije rezultate, provest ćemo analizu vjerojatnosti pokrivanja 95% pozdanih intervala nakon 2000 replikacija.

Mjera nepouzdanosti je srednja kvadratna pogreška aritmetičke sredine (standardna pogreška) i računa se kao:

- $\frac{standardna\ devijacija}{\sqrt{500}}$ za 500 replikacija
- $\frac{standardna\ devijacija}{\sqrt{2000}}$ za 2000 replikacija

Standardna devijacija povećanjem broja mjerenja poprima stalnu vrijednost, tj. ne mijenja se znatno, stoga nju možemo zanemariti. Uspoređujući mjere zaključujemo da je pogreška dvostruko veća u analizi nakon 500 replikacija, što znači da bi rezultati nakon 2000 replikacija bili puno prezicniji u analizi pokrivanja 95% pouzdanih intervala.

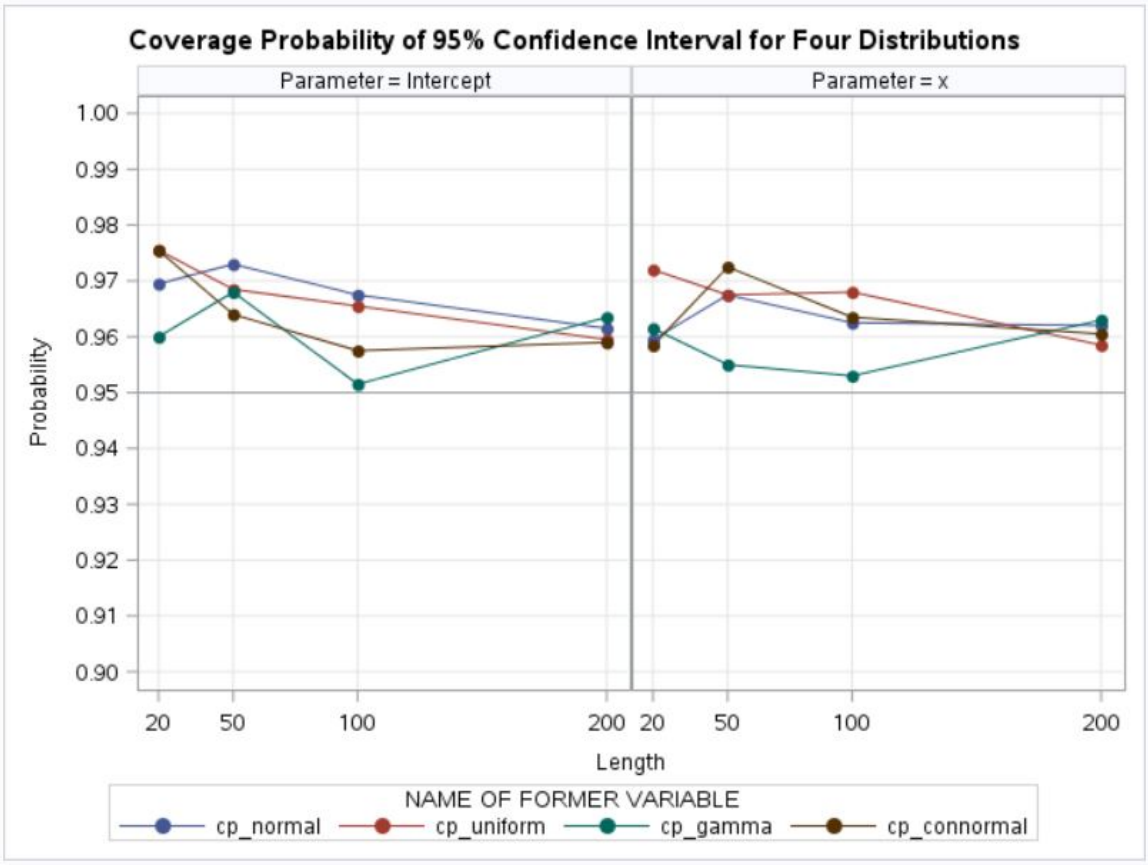
Provodimo postupak dva puta i pri svakom pokretanju generiramo uzorke iznova. Kako je pogreška manja, očekujemo malo preciznije rezultate.

Nakon prvog pokretanja programa dobivamo:

Coverage Probability of 95% Confidence Interval for Different Distributions

Parameter	n	cp_normal	cp_uniform	cp_gamma	cp_connormal
Intercept	20	0.9695	0.9755	0.9600	0.97548
Intercept	50	0.9730	0.9685	0.9680	0.96400
Intercept	100	0.9675	0.9655	0.9515	0.95750
Intercept	200	0.9615	0.9595	0.9635	0.95900
x	20	0.9595	0.9720	0.9615	0.95846
x	50	0.9675	0.9675	0.9550	0.97250
x	100	0.9625	0.9680	0.9530	0.96350
x	200	0.9620	0.9585	0.9630	0.96050

Slika 4.4: Tablica vjerojatnosti pokrivanja 95% pouzdanih intervala



Slika 4.5: Grafički prikaz vjerojatnosti pokrivanja 95% pouzdanih intervala

Ovdje prvo uočavamo da su za uzorke duljine $n = 200$ vjerojatnosti jako blizu, dok su za uzorke manjih duljina udaljenije. Sve su vrijednosti veće od nominalnih. Za parametar=*Intercept* normalna i uniformna distribucija imaju najveće vrijednosti, a za parametar=x uniformna i kontaminirana normalna te vidimo da su zapravo sve linije blizu osim one koja predstavlja Gamma distribuciju.

Pokrenemo li ponovno program, generirajući uzorke iz početka, dobivamo:

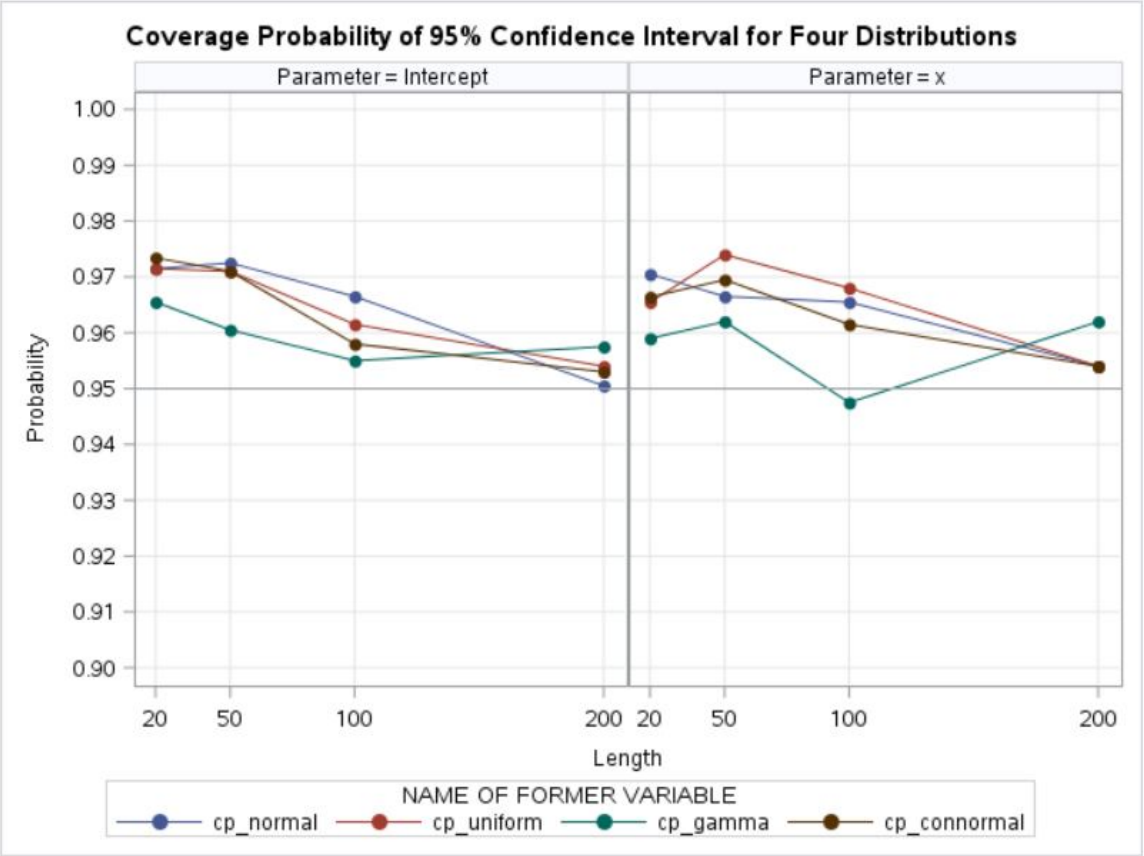
Coverage Probability of 95% Confidence Interval for Different Distributions

Parameter	n	cp_normal	cp_uniform	cp_gamma	cp_connormal
Intercept	20	0.9715	0.97149	0.9655	0.97342
Intercept	50	0.9725	0.97100	0.9605	0.97100
Intercept	100	0.9665	0.96150	0.9550	0.95800
Intercept	200	0.9505	0.95400	0.9575	0.95300
x	20	0.9705	0.96548	0.9590	0.96640
x	50	0.9665	0.97400	0.9620	0.96950
x	100	0.9655	0.96800	0.9475	0.96150
x	200	0.9540	0.95400	0.9620	0.95400

Slika 4.6: Tablica vjerojatnosti pokrivanja 95% pouzdanih intervala

Primjećujemo da su rezultati slični prošlima. Ponovno su vjerojatnosti vrlo blizu ili iste za uzorke najveće duljine. Kod ostalih distribucija su vjerojatnosti malo udaljenije ali prate određeni smjer. Jedino Gamma distribucija malo odskače i u ovom slučaju ima najmanje vrijednosti, a najveće vrijednosti su uglavnom vezane za uniformnu i normalnu distribuciju.

Možemo zaključiti da dobivamo preciznije rezultate koristeći veći broj ponavljanja uzoraka. To se vidi i uspoređivanjem grafičkih prikaza rezultata: Slike 4.2 sa slikama 4.5 i 4.7. Na prvoj slici sve linije su dosta udaljene naspram zadnjih dviju gdje su linije koje predstavljuju vjerojatnosti vezane za različite distribucije puno bliže jedna drugoj.



Slika 4.7: Grafički prikaz vjerojatnosti pokrivanja 95% pouzdanih intervala

4.3.3 Model višestruke logističke regresije

U ovom potpoglavlju bavit ćemo se rezultatima dobivenim u analizi modela s dvije prediktorske varijable.

Vrijednosti stvarnih regresijskih koeficijenata u modelu s dvije prediktorske varijable su $\beta_{20} = 2, \beta_{21} = 4, \beta_{22} = -2$ i rezultati predstavljaju proporcije uzoraka (od njih 500) za koje se regresijski koeficijenti $\beta_{20}, \beta_{21}, \beta_{22}$ nalaze unutar 95% i 99% pouzdanih intervala njihovih procjena.

Razmatrat ćemo dva slučaja: kada su prediktorske varijable nekorelirane i kada su korelirane, uzimajući različite koeficijente korelacije $\rho = 0, 0.3, 0.5, 0.9$. U tu svrhu, prisjetimo se određenih definicija.

Definicija 4.3.1. *Neka su X_i i X_j komponente slučajnog vektora $\mathbb{X} = (X_1, ..., X_n)$ takve da postoji $\mathbb{E}[X_i]$, $\mathbb{E}[X_j]$ i $\mathbb{E}[X_i \cdot X_j]$, $\forall i, j \in 1, ..., n$. Tada kovarijancu slučajnih varijabli X_i i X_j definiramo na sljedeći način:*

$$Cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j].$$

Specijalno vrijedi: $Cov(X_i, X_i) = Var(X_i)$.

Definicija 4.3.2. *Slučajne varijable X_i i X_j su nekorelirane ako je $Cov(X_i, X_j) = 0$. U suprotnom kažemo da su slučajne varijable korelirane.*

Definicija 4.3.3. *Broj*

$$\rho_{X_i,X_j} = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$$

zovemo koeficijent korelacije slučajnih varijabli X_i i X_j. Vrijedi $\rho_{X_i,X_j} \in [-1, 1]$.

4.3.3.1 Međusobno nekorelirane varijable

Kao što smo već spomenuli, podatke smo dobili tako da smo prvo generirali x i z kao slučajne brojeve koji nisu korelirani, a nakon računanja $\eta = 2 + 4x - 2z$, za vrijednost y smo uzeli slučajan broj iz Bernoullijeve razdiobe s parametrom $e^\eta/(1 + e^\eta)$ te postupak ponovili više puta.

Uz pomoć SAS-a došli smo do sljedećih rezultata:

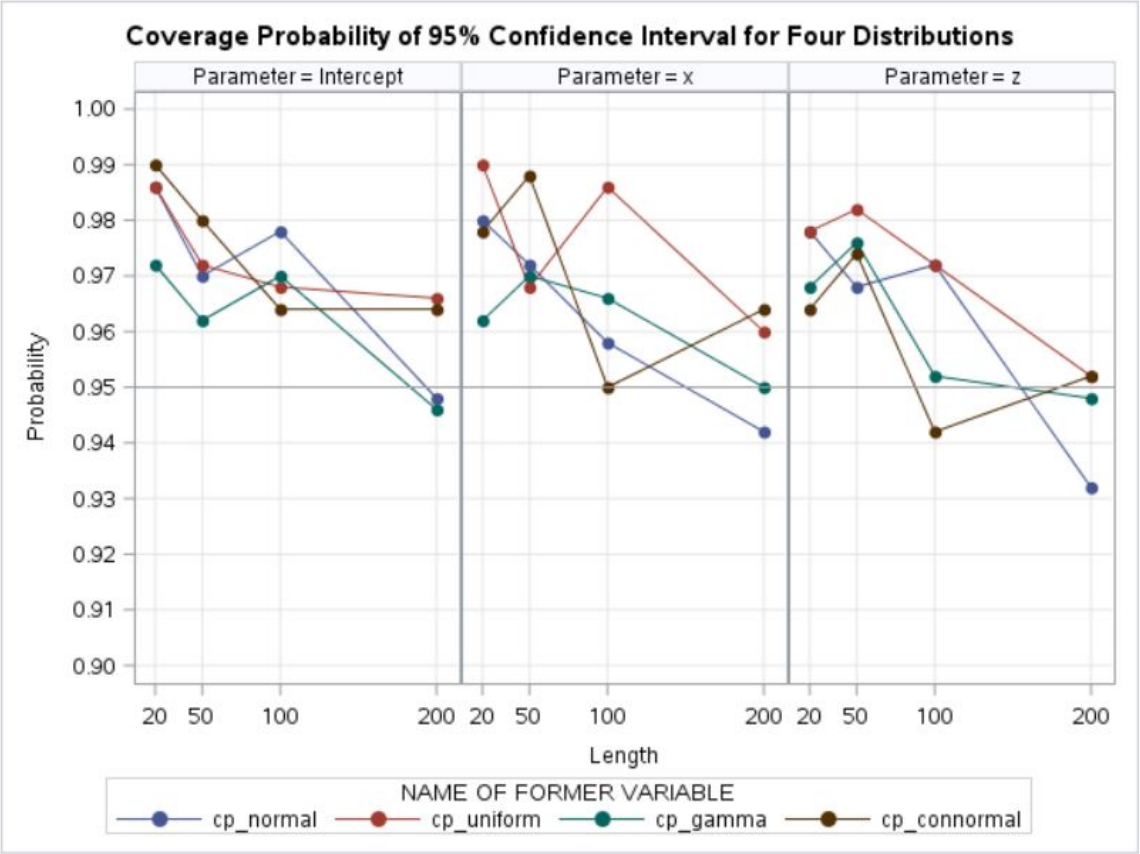
Coverage Probability of 95% Confidence Interval for Different Distributions

n	Parameter	cp_normal	cp_uniform	cp_gamma	cp_connormal
20	Intercept	0.986	0.986	0.972	0.990
20	x	0.980	0.990	0.962	0.978
20	z	0.978	0.978	0.968	0.964
50	Intercept	0.970	0.972	0.962	0.980
50	x	0.972	0.968	0.970	0.988
50	z	0.968	0.982	0.976	0.974
100	Intercept	0.978	0.968	0.970	0.964
100	x	0.958	0.986	0.966	0.950
100	z	0.972	0.972	0.952	0.942
200	Intercept	0.948	0.966	0.946	0.964
200	x	0.942	0.960	0.950	0.964
200	z	0.932	0.952	0.948	0.952

Slika 4.8: Tablica vjerojatnosti pokrivanja 95% pouzdanih intervala

Iz tablice i grafičkog prikaza (Slika 4.9) vidimo da se vjerojatosti smanjuju kako raste veličina uzoraka, n. Za uzorke malih duljina sve su vrijednosti prilično visoke, a za velike uzorke skoro za svaki parametar manje od nominalnih. Neovisno o parametrima, za uniformnu i kontaminiranu normalnu razdiobu su odgovarajuće vjerojatnosti najveće, a za Gamma i normalnu razdiobu najmanje.

Vjerojatnosti pokrivanja 95% pouzdanih intervala za slobodni koeficijent β_{20} su najveće za uzorke duljine $n = 20, 200$, a i za $n = 100$ i najmanja vrijednot je veća od vrijednosti za druga dva koeficijenta. Kod koeficijenta β_{21} (parametar=x) su uglavnom bolji rezultati nego kod koeficijenta β_{22} (parametar=z) ako gledamo najveće vrijednosti za uzorke različitih duljina. Ako gledamo rezultate za $n = 100, 200$ i najmanje vrijednosti su veće nego kod parametra z.



Slika 4.9: Grafički prikaz vjerojatnosti pokrivanja 95% pouzdanih intervala

Pogledajmo sada vjerojatnosti pokrivanja 99% pouzdanih intervala.

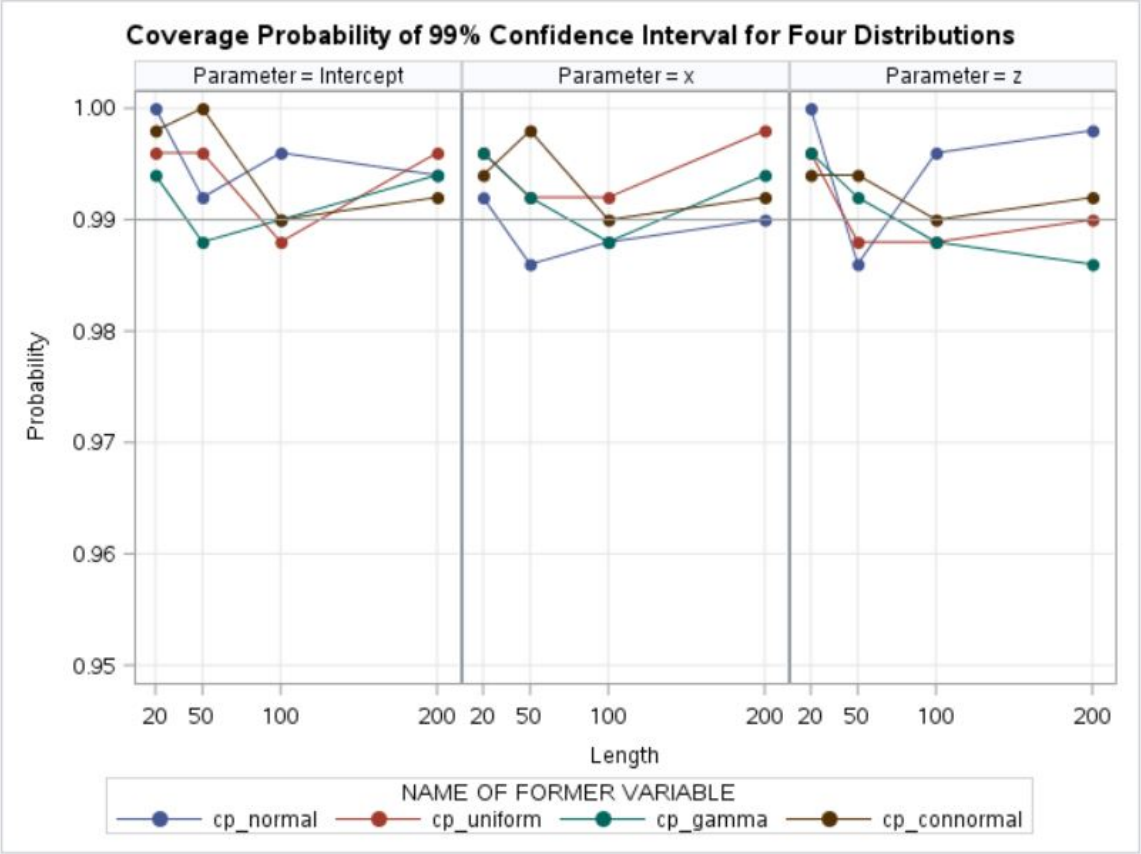
Coverage Probability of 99% Confidence Interval for Different Distributions

n	Parameter	cp_normal	cp_uniform	cp_gamma	cp_connormal
20	Intercept	1.00000	0.996	0.994	0.998
20	x	0.99198	0.996	0.996	0.994
20	z	1.00000	0.996	0.996	0.994
50	Intercept	0.99200	0.996	0.988	1.000
50	x	0.98600	0.992	0.992	0.998
50	z	0.98600	0.988	0.992	0.994
100	Intercept	0.99600	0.988	0.990	0.990
100	x	0.98800	0.992	0.988	0.990
100	z	0.99600	0.988	0.988	0.990
200	Intercept	0.99400	0.996	0.994	0.992
200	x	0.99000	0.998	0.994	0.992
200	z	0.99800	0.990	0.986	0.992

Slika 4.10: Tablica vjerojatnosti pokrivanja 99% pouzdanih intervala

U ovom slučaju ne možemo za sve vrijednosti reći da se smanjuju povećavanjem duljine uzoraka. Na prvoj slici (Slika 4.11) koja prikazuje vjerojatnosti pokrivanja 99% pouzdanih intervala za regresijski koeficijent β_{20} takva konstatacija ne vrijedi jer su neke vrijednosti za uzorke duljine $n = 50, 100$ manje od vrijedosti za $n = 200$, i to za sve distribucije osim normalne. Promatramo li rezultate za koeficijent β_{21} vidimo također da su vrijednosti za $n = 200$ veće od onih za $n = 100$ za sve distribucije. Vrlo slična situacija je i na sličici skroz desno, za parametar=z.

Kao što možemo primijetiti, teško je uočiti neku pravilnost u kretanju dobivenih vjerojatnosti kako za promjenu veličine uzoraka, tako i za vrstu distribucije. Naime, očekivali bismo smanjenje vjerojatnosti s povećanjem veličine uzoraka. Razlog tomu može biti broj ponavljanja pri uzorkovanju ili tzv. problem potpune separacije u logističkoj regresiji koji se javlja pri izvršavanju programa u SAS-u. Nameće nam se pitanje o tome koliko je primjena logističke regresije dobra za male modele i male uzorke.

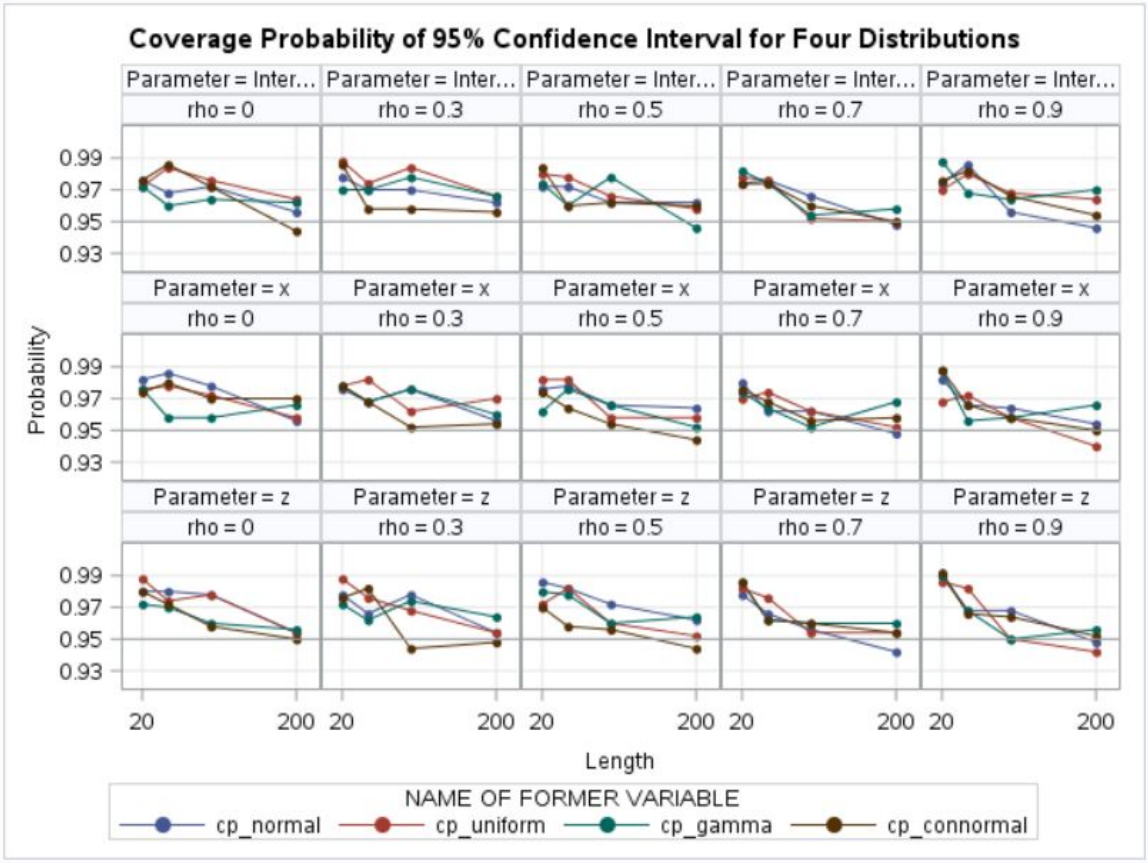


Slika 4.11: Grafički prikaz vjerojatnosti pokrivanja 99% pouzdanih intervala

4.3.3.2 Međusobno korelirane varijable

Kako bismo otkrili što se događa s vjerojatnostima kada se radi o modelu s dvije međusobno korelirane prediktorske varijable, generirali smo varijable x i z iz različitih distribucija koje su nam od interesa tako da vrijedi $z = \rho \cdot x + \sqrt{1-\rho^2} \cdot x$, uzimajući različite vrijednosti za koeficijent korelacije $\rho = 0, 0.3, 0.5, 0.9$.

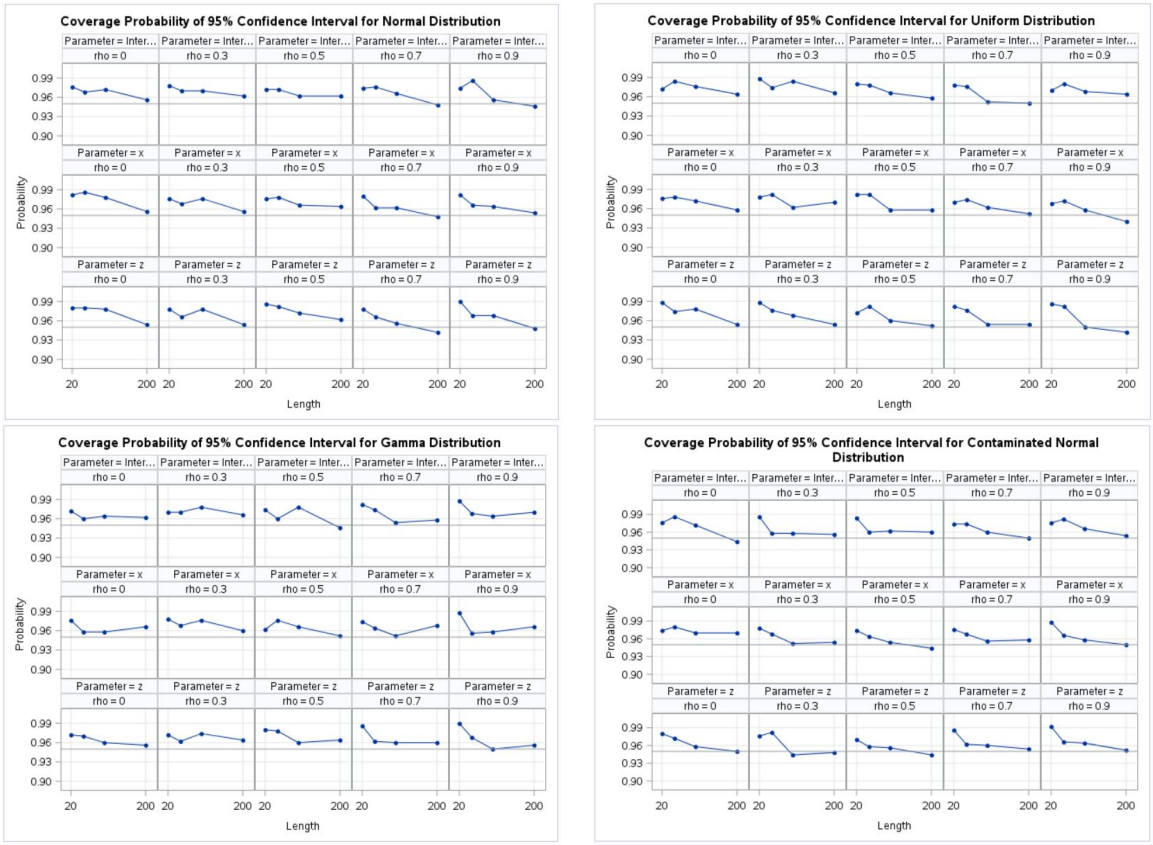
Kao rezultat smo dobili:



Slika 4.12: Grafički prikaz vjerojatnosti pokrivanja 95% pouzdanih intervala

Na temelju gornjeg grafičkog prikaza primjećujemo da su na svakoj sličici vjerojatnosti velike za uzorke malih duljina, a kako se veličine uzoraka povećavaju, vjerojatnosti se smanjuju. Također, uočavamo da su za skoro sve kombinacije vjerojatnosti veće od nominalnih, osim situacija na kojima je koeficijent ρ veći i duljina uzoraka velika ($n = 100, 200$). U slučajevima u kojima je jaka korelacija između varijabli, tj. $\rho = 0.7, 0.9$ možemo vidjeti

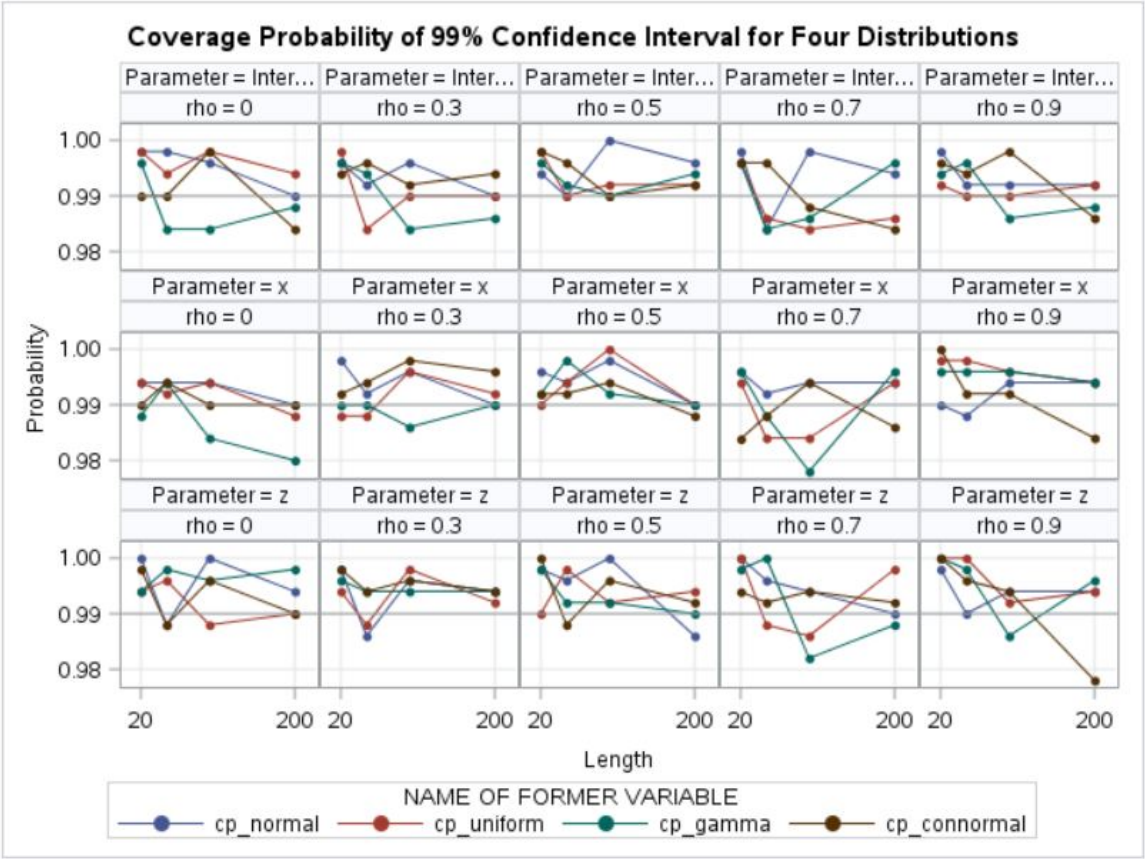
da su linije koje predstavljaju različite distribucije bliže jedna drugoj, odnosno za različite duljine uzoraka imaju rezultate koji su jako blizu jedan drugom, što nije slučaj kod srednje jake korelacije među varijablama.



Slika 4.13: Grafički prikaz vjerojatnosti pokrivanja 95% pouzdanih intervala

Promotrimo li pomnije prikaz svake distribucije zasebno, zaključujemo da kod normalne radiobe povećanjem koeficijenta korelacije vjerojatnosti rezultati počinju ”brže padati” kako se povećava duljina uzoraka. Slično je ponašanje vrijednosti i kod uniformne i kontaminirane normalne razdiobe, za razliku od Gamma distribucije gdje vrijednosti u zadnjem stupcu nakon pada počinju rasti za uzorke velikuh duljina.

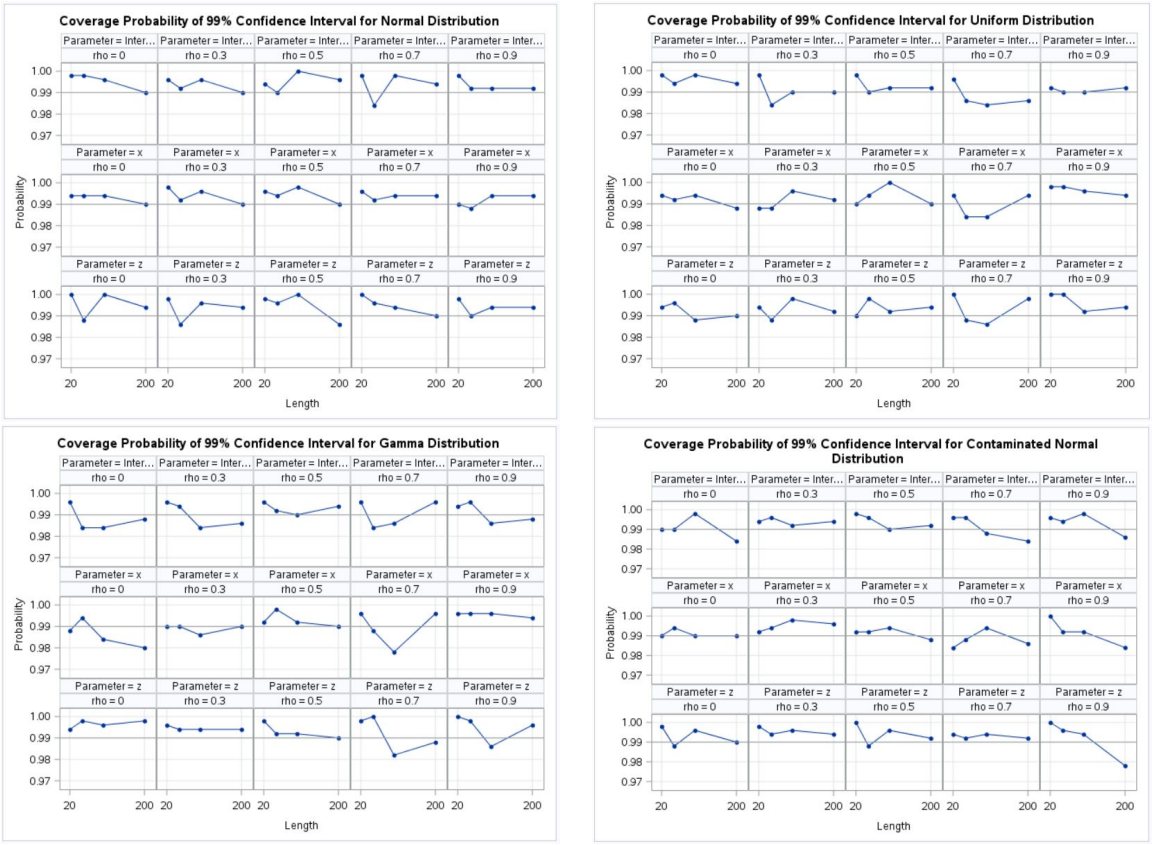
Rezultati analize vjerojatnosti pokrivanja 99% pouzdanih intervala su potpuno drugačiji za razliku od prethodnih. Promotrimo sljedeći grafički prikaz svih distribucija.



Slika 4.14: Grafički prikaz vjerojatnosti pokrivanja 99% pouzdanih intervala

Vrijednosti u svim kombinacijama se prilično razlikuju, stoga linije koje predstavljaju različite distribucije izgledaju ”raštrkano”. U svakom slučaju vrijednosti su u puno slučajeva manje od nominalnih i ne možemo reći da padaju kako se povećava broj n, već se ponašaju nepravilno i u nekim situacijama rastu, a nekada naglo padaju. Ni s povećavanjem koeficijenta korelacije među varijablama ne dešava se promjena, od kud zaključujemo da koreliranost ne utječe na vjerojatnosi pokrivanja pouzdanih intervala.

Do istih zaključaka ćemo doći promatramo li svaku od distribucija posebno (Slika 4.15)



Slika 4.15: Grafički prikaz vjerojatnosti pokrivanja 99% pouzdanih intervala

4.4 Zaključak

U većini kombinacija koje smo proveli analizirajući vjerojatosti pokrivanja 95% i 99% Waldovih pouzdanih intervala uočili smo da su vjerojatnosti veće za uzorke malih duljina od onih za uzorke duljina $n = 100$ ili 200. Često su vrijednosti bile veće od nominalnih, osobito za uzorke malih duljina. Razlog tomu je problem potpune separacije (*engl. complete separation* je česta poruka koja se pojavljuje izvodeći programe u SAS-u).

Kako bismo shvatili o čem se radi, koristit ćemo jedan jednostavan primjer u kojem je Y varijabla odaziva, a X_1 i X_2 varijable poticaja s vrijednostima:

Y	X_1	X_2
0	1	3
0	2	2
0	3	-1
0	3	-1
1	5	2
1	6	4
1	10	1
1	11	0

Vidimo da sva opažanja s $Y = 0$ imaju vrijednosti $X_1 \leq 3$, a opažanja s $Y = 1$ imaju vrijednosti $X_1 > 3$. Drugim riječima, Y savršeno razdvaja X_1. Također, možemo reći da X_1 savršeno predviđa Y jer $X_1 \leq 3$ odgovara $Y = 0$, a $X_1 > 3$ odgovara $Y = 1$. U situaciji smo da smo slučajno pronašli savršen prediktor X_1 za varijablu ishoda Y. U kontekstu predviđenih vjerojatnosti imamo $\mathbb{P}(Y = 1|X_1 \leq 3) = 0$ i $\mathbb{P}(Y = 1|X_1 > 3) = 1$, bez potrebe za procjenom modela.
Potpuna separacija je problem koji može nastati iz nekoliko razloga. Jedan od uobičajenih primjera je upotreba nekoliko kategoričnih varijabli čije su kategorije kodirane pomoću indikatora. Na primjer, ako se proučava bolest povezana sa starenjem (prisutna / odsutna) i starost je jedan od prediktora, mogu postojati podskupine (npr. žene starije od 55 godina) od kojih svaka ima bolest. Drugi mogući scenarij je kada je veličina uzorka vrlo mala. U našem gornjem primjeru nema razloga zašto Y mora biti 0 kada je $X_1 \leq 3$. Ako bi uzorak bio dovoljno velik, vjerojatno bismo imali neka opažanja s $Y = 1$ i $X_1 \leq 3$, razbivši na taj način potpunu separaciju, tj. odvajanje varijable X_1.
Pokušamo li primijeniti logistički model na danim podacima, SAS će nas obavijestiti o problemu potpune separacije uz daljnje poruke upozorenja koje pokazuju da procjena maksimalne vjerodostojnosti ne postoji, no usprkos tome nastavit će s izvršavanjem programa. Neće nas obavijestiti koja je varijabla ili koje su varijable u potpunosti odvojene varijablom ishoda te da je procjena parametra za neku varijablu netočna.

Upravo to se dešavalo pri svakom izvođenju programa pomoću kojih smo vršili analizu. Iz tog razloga vjerojatnosti pokrivanja pouzdanih intervala su ispadale jako visoke. Naime, često su se pojavljivale spomenute poruke uz činjenicu da procjena maksimalne vjerodostojnosti ne postoji pa se dobivaju jako široki intervali pouzdanosti i koeficijenti β_{10}, β_{11}, β_{20}, β_{21}, β_{22} najčešće budu unutar intervala, posebno za male uzorke (kada je duljina uzoraka n do 200). Zato je primjena logističke regresije upitna za male uzorke i jednostavne modele s jednom do dvije prediktorske varijable.

Navest ćemo nekoliko tehnika za rješavanje problema potpune separacije. Recimo da je prediktorska varijabla uključena u potpunu separaciju nazvana X.
Jedan način je da se pobrinemo da varijabla ishoda nije dihotomna varijabla u modelu. Druga jednostavna strategija je ne uključiti X u model, što dovodi do pristranih procjena za druge prediktorske varijable u modelu pa to nije preporučena strategija. Još jedan alternativan pristup pruža Firthova logistička regresija koja je bazirana na "penalized likelihood estimation" metodi i smatra se idealnim rješenjem za ovakve probleme, međutim nju nećemo detaljnije opisivati.

Bibliografija

[1] P. McCullagh i J. A. Nelder, *Generalized Linear Models*, Chapman & Hall/CRC, Boca Raton, 1989.

[2] C. E. McCulloch i S. R. Searle, *Generalized, linear and mixed models*, John Wiley & Sons, New York, 2001.

[3] B. Basrak, *Generalizirani linearni modeli*, https://web.math.pmf.unizg.hr/ bbasrak/pdf_files/FinPrak/FPchap7.pdf, PMF-MO slideovi, 2016.

[4] N. Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 1987.

[5] A. Agresti, *Categorical Data Analysis*, University of Florida, Gainesville, Florida, 2002.

[6] D. W. Hosmer i S. Lemeshow, *Applied logistic regression*, John Wiley & Sons, 2000.

[7] R. Wicklin, *Simulating Data with SAS®*, SAS Institute Inc., Cary, NC, 2013.

[8] *7.1.1 - Example - The Donner Party*, https://onlinecourses.science.psu.edu/stat504/node/159/

[9] *Gamma distribucija*, http://mathworld.wolfram.com/PearsonTypeIIIDistribution.html?fbclid=IwAR151JF070WfcR25jGO6FqG24LLxV51YKso-X_obs-Nyrb3VgvYdQEDmbWo

[10] *Complete separation*, https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/

Sažetak

U ovom radu bavili smo se modelom logističke regresije, koja pripada široj klasi modela pod imenom generalizirani linearni modeli. Koristi se u analizi binarnih podataka, a može biti multinomna - u slučaju da varijabla odaziva poprima vrijednosti koje se mogu rasporediti u više od dvije kategorije te dihotomna, ako varijabla odaziva poprima samo dvije vrijednosti. Fokusirali smo se na dihotomnu logističku regresiju pri čemu smo prisutstvo i odsutstvo promatranog svojstva interpretirali jedinicom i nulom. Opisali smo način postavljanja modela te objasnili interpretacjiu parametara, kao i metodu dobivanja procjene parametara. Koliko dobro model opisuje podatke te kako tumačimo određena svojstva pokazali smo na primjeru. Također, proveli smo analizu vjerojatnosti pokrivanja Waldovih pouzdanih intervala za regresijske koeficijente i iz rezultata zaključili da je primjena logističke regresije upitna za male uzorke (do duljine 200) i jednostavne modele s jednom do dvije prediktorske varijable zbog mogućeg problema s potpunom separacijom koji se često javalja pri izvođenju programa u SAS-u.

Summary

In this paper, we are dealing with a model of logistic regression, which belongs to a wider model class called generalized linear models. It is used in binary data analysis, and can be multinational - in case the dependent variable is taking values that can be deployed in more than two categories and dichotomous, if the responding variable getting only two values. We focused on the dichotomous logistic regression in which the presence and absence of the observed property interpreted by one and zero. We described the fitting model and explained the interpretation of parameters as well as the method of parameter estimation. How good a model describes the data and how we interpret some of the properties we showed in the example. Also, conducted an analysis of the coverage probability of Wald's confidence intervals for regression coefficients and concluded that the application of logistic regression is questionable for small samples (to length 200) and simple models with one to two prediction variables due to a potential problem with complete separation often occurring when doing the program in SAS.

Životopis

Rođena sam 7. rujna 1993. godine u Zagrebu. 2008. godine završila sam osnovnoškolsko obrazovanje u Osnovnoj školi Retkovec, nakon čega sam upisala Gornjogradsku gimnaziju u Zagrebu. Po završetku srednje škole, 2012. u istom gradu upisala sam Prirodoslovno-matematički fakultet, smjer Matematika, gdje sam 2016. godine završila preddiplomski studij. Diplomski studij nastavila sam na istom fakultetu, upisavši smjer Matematička statistika.