SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Andreja Vlahek

ANALIZA BINARNIH PODATAKA

Diplomski rad

Voditelj rada:
Izv. prof. dr. sc. Miljenko Huzak

Zagreb, rujan, 2017.

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. _____

Zahvaljujem se izv. prof. dr. sc. Miljenku Huzaku na vrhunskom mentorstvu te izdvojenom vremenu i korisnim savjetima tijekom izrade ovog diplomskog rada.
Neizmjerno hvala mojoj obitelji na podršci i razumijevanju.

Sadržaj

Uvod

U mnogim područjima ljudske djelatnosti, u kojima je statistika našla svoju primjenu, od interesa je prisutstvo ili odsutstvo nekog svojstva ili pojave. Na primjer, uspješnost zrna da proklija u određenim uvjetima, sposobnost biljnih nametnika da prežive tretiranje insekticidom ili uspješnost oporavka pacijenta nakon liječničkog tretmana. Podatke s kojima se susrećemo u takvim primjerima 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 opservaciju zasebno te nam tada uspjeh i neuspjeh odgovaraju realizacijama slučajne varijable s Bernoullijevom razdiobom. Drugi način je da grupiramo opservacije koje su jednake po ostalim karakteristikama mjerenim u eksperimentu te po grupama promatramo broj opservacija koje imaju promatrano svojstvo (ukupan broj jedinica) što odgovara realizaciji binomne slučajne varijable.

U povijesti prvi model za takve podatke uključivao je transformiranje podataka funkcijom $g(x) = \log(-\log(1-x))$, a razvio ga je Ronald Fisher 1922. godine. U svojim eksperimentima promatrao je otopine i smjese te ispitivao prisutnost kontaminanta. 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 je pokazao da je logistička funkcija također pogodna za modeliranje takvih podataka. Po uzoru na Blissov *probit* Berkson je model skraćeno prozvao *logit* modelom. Svi navedeni modeli bili su prilagođeni podacima minimizacijom χ^2 statistike težinskom metodom najmanjih kvadrata. Tijekom 60-ih i 70-ih godina razvijali su se prvi algoritmi za procjenu parametara takvih modela metodom maksimalne vjerodostojnosti. Prijelomna godina bila je 1972. kada su John Nelder i Robert Wedderburn razvili metodologiju statističkog modeliranja, a pripadne modele jednim imenom nazvali generaliziranim linearnim modelima. Tako 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 u odnosu na *probit* funkciju i lakše interpretacije. Koristi se u mnogim društvenim znanostima, zatim u biomedicini, radiologiji, aktuarstvu pa čak i u

strojnom učenju.

Ovaj rad ćemo započeti kratkom motivacijom i definiranjem osnovnih pojmova. U glavnom dijelu ćemo postaviti model za grupirani pristup podacima i objasniti interpretaciju parametara. Pretežno ćemo se baviti logističkim modelom. Nadalje, opisat ćemo dvije metode dobivanja procjena parametara modela i analizirati njihova svojstva, zatim predstaviti statistike koje se tiču prilagodbe modela i testiranja hipoteza te ukratko opisati dijagnostiku. Na samom kraju ćemo opisane metode ilustrirati trima primjerima.

Poglavlje 1

Uvod u model

1.1 Motivacija

Generalizirani linearni modeli su svojevrsno proširenje klasičnog linearnog modela primjenjivi za varijablu odziva iz neke eksponencijalne familije. Prisjetimo se, u linearnom se modelu promatra linearna veza varijable odziva (zavisna) Y i (neslučajnih) varijabli poticaja ili prediktora $X_1, X_2, \ldots, X_p$. Pretpostavimo da imamo n međusobno nezavisnih opservacija od Y koje označimo s $Y_1, \ldots, Y_n$. Tada linearnu vezu opisujemo s:

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

odnosno matrično

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

gdje su $\mathbf{Y}^T := (Y_1, \ldots, Y_n)$, $\boldsymbol{\beta}^T = (\beta_0, \ldots, \beta_p)$ vektor parametara modela, $X = [\mathbf{1}, \mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p}]$ matrica poticaja pri čemu je $\mathbf{x}_{\cdot j}$ vektor stupac realizacija 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:

$$\mathbb{E}[\mathbf{Y}] = X\boldsymbol{\beta} =: \boldsymbol{\mu}.$$

Dodatna pretpostavka je da greške dolaze iz normalne razdiobe što implicira normalnost promatrane varijable odziva Y. Time se značajno ograničavamo te nam je potreban novi model kojim bismo opisali diskretne podatke, štoviše binarne kod kojih su $y_i \in \{0, 1\}$. Skup takvih generaliziranih modela nazivamo još i modeli diskretnog izbora.

1.2 Osnovni pojmovi

Definicija 1.2.1. *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:

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

Definicija 1.2.2. *Neka je na statističkoj strukturi* $(\Omega, \mathcal{F}, \mathcal{P})$ *dan slučajni vektor* $\mathbf{Y} : \Omega \to \mathbb{R}^n$*. Za fiksni* $\theta \in \Theta$ *označimo s* $F(\,\cdot\,;\theta)$ *funkciju distribucije od* $\mathbf{Y}$ *u odnosu na vjerojatnost* $\mathbb{P}_\theta$*. Familiju* $\mathcal{P}' = \{F(\,\cdot\,;\theta) : \theta \in \Theta\}$ *nazivamo statističkim modelom, a za vektor* $\mathbf{Y}$ *kažemo da pripada tom statističkom modelu.*

Promatramo slučajni vektor $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ s gustoćom $f(\,\cdot\,;\theta)$ u odnosu na $\mathbb{P}_\theta$ i varijable poticaja $X_1, \ldots, X_p$. Slučajni vektor $\mathbf{Y}$ čine slučajne varijable Y_i koje modeliraju realizacije y_i varijable odziva Y. Zbog 1-1 korespondencije zakona razdiobe i funkcije distribucije [9] možemo poistovjetiti statistički model $\mathcal{P} = \{f(\,\cdot\,;\theta); \theta \in \Theta\}$.

Definicija 1.2.3. *Model* $\mathcal{P}$ *je k-parametarska eksponencijalna familija ako je gustoća* $f(\,\cdot\,;\theta) \in \mathcal{P}$ *dana kao*

$$f(x;\theta) = C(\theta)\, h(x)\, e^{\sum_{i=1}^{k} Q_i(\theta) t_i(x)}$$

gdje su $t_1, \ldots, t_k$ *linearno nezavisne nekonstantne funkcije* $t_i : \mathbb{R}^n \to \mathbb{R}$*, a* $C : \Theta \to [0, +\infty)$*,* $h : \mathbb{R}^n \to [0, +\infty)$*,* $\mathbb{Q}_i : \Theta \to \mathbb{R}, i = 1, \ldots, k$ *izmjerive funkcije.*

Nadalje, pretpostavimo da je vektor $\mathbf{Y}$ iz k-parametarske eksponencijalne familije $\mathcal{P}$ takav da je $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$.

Definicija 1.2.4. *Za parametre* $\beta_0, \beta_1, \ldots, \beta_p$ *i vrijednosti varijabli poticaja, vektore stupce* $\mathbf{x}_{\cdot 1}, \ldots, \mathbf{x}_{\cdot p}$ $(\mathbf{x}_{\cdot 0} = \mathbf{1})$*, izraz*

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

nazivamo linearnim prediktorom.

Definicija 1.2.5. *Monotonu diferencijabilnu funkciju* $g : \mathbb{R} \to \mathbb{R}$ *takvu da*

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

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

Zadaća funkcije povezivanja je uspostaviti vezu komponenata linearnog prediktora $\boldsymbol{\eta}$ i odgovarajućih komponenata očekivanja $\boldsymbol{\mu}$. Kod klasičnog linearnog modela zadovoljavajuća veza je funkcija identiteta. S druge strane, kod Bernoullijevog ili općenitijeg binomnog modela očekivanje zadovoljava $0 < \mu_i < 1$ te je zadaća funkcije povezivanja preslikati interval $(0, 1)$ na cijeli skup realnih brojeva. U tu svrhu naredne funkcije povezivanja dolaze u obzir:

1. logit $g(x) = \log \frac{x}{1-x}$,
2. probit $g(x) = \Phi^{-1}(x)$ gdje je $\Phi(\cdot)$ funkcija distribucije standardne normalne razdiobe,
3. komplementarna log-log $g(x) = \log(-\log(1-x))$,
4. log-log $g(x) = -\log(-\log(x))$,

pri čemu se posljednja nešto rjeđe koristi zbog nepovoljnih svojstava koje ima za vrijednosti $x < \frac{1}{2}$ koje su često od interesa te se ni u ovom radu zbog toga neće detaljnije izučavati.

U suštini, bilo koja funkcija koja je diferencijabilna (onda i neprekidna) i strogo monotona te takva da interval $(0, 1)$ preslikava na cijeli skup $\mathbb{R}$ može se koristiti za definiranje generaliziranog linearnog modela.
Pretpostavimo da je s $F(\cdot)$ dana strogo monotona vjerojatnosna funkcija distribucije neke slučajne varijable definirane na cijelom $\mathbb{R}$. Nadalje, pretpostavimo da za linearni prediktor i očekivanje komponenata početno zadanog slučajnog vektora $\mathbf{Y}$ vrijedi:

$$\mu_i = F(\eta_i), \quad -\infty < \eta_i < +\infty \quad \forall i \in \{1, \cdots, n\}.$$

Tada kao funkciju poveznicu možemo koristiti inverz funkcije distribucije.

$$\eta_i = F^{-1}(\mu_i), \quad 0 < \mu_i < 1 \quad \forall i \in \{1, \cdots, n\}.$$

Pokazat će se da gore navedene funkcije poveznice imaju ulogu inverza nekih dobro poznatih vjerojatnosnih funkcija distribucije.

Poglavlje 2

Logistički model

2.1 Bernoullijeva i binomna razdioba

Pretpostavimo da promatramo manifestaciju nekog svojstva, kojeg možemo kodirati s nulama i jedinicima, u skupu od n jedinki s obzirom na prediktore. Neka je Y_{ij} slučajna varijabla koja može poprimiti vrijednosti iz skupa $\{0, 1\}$ te označimo s y_{ij} njezine realizacije. Označimo s π_{ij} vjerojatnost s kojom Y_{ij} poprima vrijednost jedan.

Kažemo da slučajna varijabla X ima Bernoullijevu razdiobu s parametrom π ako je njezina razdioba dana s

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

Pišemo $X \sim B(1, \pi)$.

Primjer 2.1.1. *Matematičko očekivanje od $X \sim B(1, \pi)$ je $\mathbb{E}(X) = \pi$, a varijanca $Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \pi(1 - \pi)$.*

Iz prethodnog primjera možemo primijetiti da ukoliko neki prediktor mijenja promatranu vjerojatnost π, osim što mijenja očekivanje, mijenja i varijancu. Prema tome linearni model koji izvodimo pod pretpostavkom da je varijanca varijable odziva konstantna nije prikladan za modeliranje binarnih podataka.

Primjer 2.1.2. *Bernoullijev model je 1-parametarska eksponencijalna familija sukladno definiciji 1.2.3. Za $\theta = \pi$ imamo:*

$$\begin{aligned} f(x; \theta) &= \theta^x(1 - \theta)^{1-x}\mathbb{1}_{\{0,1\}} \\ &= (1 - \theta)\Big(\frac{\theta}{1 - \theta}\Big)^x \mathbb{1}_{\{0,1\}} \\ &= (1 - \theta)\mathbb{1}_{\{0,1\}} e^{x \log \frac{\theta}{1-\theta}} \end{aligned}$$

$\{x\}$ je linearno nezavisan skup, $C(\theta) = 1 - \theta$, $t_1(x) = x$, $h(x) = \mathbb{1}_{\{0,1\}}(x)$ i $Q_1(\theta) = \log \frac{\theta}{1-\theta}$ su izmjerive funkcije.

Nadalje, pretpostavimo da jedinke koje promatramo možemo klasificirati u k grupa na način da su u pojedinoj grupi one 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, tj. za koje je $Y_i = Y_{i1} + \cdots + Y_{in_i}$.

Kažemo da slučajna varijabla X ima binomnu razdiobu s parametrima n i π ako joj je razdioba dana s

$$\mathbb{P}(X = x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} \quad x \in \{0, 1, \ldots, n\}.$$

Pišemo $X \sim B(n, \pi)$.

Propozicija 2.1.3. *Neka su $X_1, X_2, \ldots, X_n$ nezavisne slučajne varijable takve da $X_i \sim B(1, \pi), \forall i = 1, \ldots, n$. Tada $\sum_{i=1}^{n} X_i \sim B(n, \pi)$.*

Primjer 2.1.4. *Po teoremu 10.1 i 11.6 u [9] te propoziciji 2.1.3 za slučajnu varijablu $X \sim B(n, \pi)$ matematičko očekivanje je $\mathbb{E}(X) = n\pi$, a varijanca $Var(X) = n\pi(1 - \pi)$.*

Primjer 2.1.5. *Binomni model je 1-parametarska eksponencijalna familija sukladno definiciji 1.2.3. Za $\theta = \pi$ i $D_n = \{0, 1, \ldots, n\}$ imamo:*

$$\begin{aligned} f(x; \theta) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \mathbb{1}_{D_n} \\ &= \binom{n}{x} (1 - \theta)^n \Big(\frac{\theta}{1 - \theta} \Big)^x \mathbb{1}_{D_n} \\ &= \binom{n}{x} \mathbb{1}_{D_n} (1 - \theta)^n e^{x \log \frac{\theta}{1-\theta}} \end{aligned}$$

$\{x\}$ je linearno nezavisan skup, $C(\theta) = (1-\theta)^n$, $t_1(x) = x$, $h(x) = \binom{n}{x} \mathbb{1}_{D_n}(x)$ i $Q_1(\theta) = \log \frac{\theta}{1-\theta}$ su izmjerive funkcije.

Pretpostavimo nezavisnost jedinki i to da sve jedinke unutar neke grupe s istom vjerojatnošću π_i imaju promatrano svojstvo. Tada je za svaki i, po propoziciji 2.1.3, $Y_i \sim B(n_i, \pi_i)$.

2.2 Izglednost i logit funkcija

U konstrukciji modela u centru promatranja je vjerojatnost $\pi_i \in (0, 1)$. Međutim, zbog ograničenosti na interval kojem ona mora pripadati potrebno je prikladnim transformacijama prijeći na cijelu realnu os. Kako bismo maknuli gornje ograničenje, promatramo izglednost.

Definicija 2.2.1. *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.*

Nadalje, logaritmiramo izglednost kako bismo maknuli donje ograničenje.

Funkciju $\text{logit} : (0, 1) \to \mathbb{R}$

$$\text{logit}\, x := \log \frac{x}{1-x}$$

nazivamo *logit* funkcija.

Uklanjanjem ograničenja navedenim transformacija dobivamo logaritmiranu izglednost (engl. *log-odds*):

$$\text{logit}\, \pi_i = \log \frac{\pi_i}{1-\pi_i} \tag{2.1}$$

Primijetimo jedno zanimljivo svojstvo transformacije. Naime, kad vjerojatnost ima vrijednost $1/2$ tada izraz u (2.1) ima vrijednost 0. Negativna logaritmirana izglednost reprezentira vjerojatnosti manje od $1/2$, a pozitivna veće od $1/2$.

Napomena 2.2.2. *Logaritmirana izglednost može se definirati i u smislu očekivanja varijable iz binomne razdiobe kao logaritmirani omjer očekivanih "uspjeha" $\mu_i = n_i\pi_i$ i očekivanih "neuspjeha" $n_i - \mu_i$. Rezultat je jednak gore dobivenom.*

Ako promatramo dva događaja A i B, tada definiramo omjer njihovih izlednosti (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.

2.3 Postavljanje modela i interpretacija parametara

Pretpostavimo da su za $i = 1, \ldots, k$, $Y_i \sim B(n_i, \pi_i)$ međusobno nezavisne slučajne varijable te $y_1, \ldots, y_k$ njihove realizacije. Neka je $n = n_1 + \cdots + n_k$.
Definirajmo slučajni vektor $\mathbf{Y}^T = (Y_1, \ldots, Y_k)$. Po primjeru 2.1.5 on pripada 1-parametarskoj eksponencijalnoj familiji. Dodatno pretpostavimo da je *logit* vjerojatnosti jednak linearnom prediktoru:

$$\begin{aligned} \operatorname{logit} \pi_i &= \eta_i \\ &= \mathbf{x}_{i\cdot}^T \boldsymbol{\beta} \quad \forall i \in \{1, \ldots, k\} \end{aligned} \tag{2.2}$$

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

Parametri β_i interpretiraju se na isti način kao kod linearnog modela, samo ne u smislu varijable odziva, nego u smislu logaritmirane izglednosti. Pretpostavimo da promijenimo j-ti prediktor na način $\mathbf{x}_{\cdot j} \to \mathbf{x}_{\cdot j} + \mathbf{1}$, a sve ostale držimo fiksnima. Kod linearnog modela ta se promjena očitovala kao promjena očekivanja varijable odziva za β_j. 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 ako 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 možemo dobiti ako promatramo izraz koji transformacijama dobivamo iz (2.2):

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

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. Donekle dobar odgovor možemo dobiti promatranjem vjerojatnosti π_i kao funkcije s argumentom x_{ij} te računanjem derivacije:

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

Također, primijetimo još da efekt j-te varijable poticaja na vjerojatnost π_i ovisi o parametru β_j i vrijednosti te vjerojatnosti. Taj se produkt 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 $\boldsymbol{\beta}$, uvršavajući konkretne vrijednosti varijabli poticaja mjerenih kod neke nove jedinke, ovaj model nam daje vjerojatnost s kojom ta jedinka ima promatrano svojstvo.

2.4 Latentna formulacija modela

Neka je Y_i slučajna varijabla koja reprezentira binarni ishod koji kodiramo s 0 i 1. Možemo je nazvati *opaženim (manifestiranim)* ishodom. Pretpostavimo da postoji neprekidna slučajna varijabla Y_i^* koja poprima vrijednosti na realnoj osi. Povezanost tih dviju varijabli je dana s:

$$Y_i = \begin{cases} 1, & Y_i^* > \theta \\ 0, & Y_i^* \leq \theta \end{cases} \tag{2.3}$$

gdje je $\theta \in \mathbb{R}$ unaprijed zadana granična vrijednost. Varijablu Y_i^* možemo zvati *latentnim* ishodom. Ona se direktno ne opaža, no smatra se da na neki način utječe na opaženu varijablu. U primjeni tumačenje varijabli Y_i i Y_i^* ovisi o konkretnom problemu koji se izučava. Na primjer, u ekonomiji varijablom Y_i može se modelirati izbor kao što je najam ili kupnja stana, a s Y_i^* razlika u iskoristivosti najma i kupnje.
Sukladno (2.3) slijedi:

$$\pi_i = \mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i^* > \theta). \tag{2.4}$$

Budući da se latentna varijabla ne opaža, bez smanjenja općenitosti za graničnu se vrijednost θ može uzeti 0. Također, povoljnim se transformacijama u (2.4) Y_i^* može standardizirati tako da joj je standardna devijacija 1 ili neka željena vrijednost c bez da se mijenja vjerojatnost događaja u (2.4).
Pretpostavimo sada da varijabla odziva ovisi o vektoru varijabli poticaja $\mathbf{x}_i$. Modeliramo tu ovisnost klasičnim linearnim modelom za latentnu varijablu:

$$\begin{aligned} Y_i^* &= \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \\ &= \eta_i + U_i \end{aligned} \tag{2.5}$$

te pretpostavimo da je U_i greška čija je vjerojatnosna funkcija distribucije $F(\cdot)$, ne nužno normalna. Iz (2.4) i (2.5) slijedi:

$$\begin{aligned} \pi_i &= \mathbb{P}(Y_i^* > 0) \\ &= \mathbb{P}(U_i > -\eta_i) \\ &= 1 - F(-\eta_i). \end{aligned} \tag{2.6}$$

Iz prošlog raspisa lako definiramo generalizirani model s funkcijom povezivanja $g = -F^{-1}$:

$$\eta_i = -F^{-1}(1 - \pi_i) \tag{2.7}$$

Često je distribucija grešaka simetrična oko 0 pa, koristeći $F(u) = 1 - F(-u)$, iz (2.6) slijedi

$$\pi_i = F(\eta_i)$$

te je generalizirani model dan nešto jednostavnijim izrazom:

$$\eta_i = F^{-1}(\pi_i).$$

2.5 Funkcije povezivanja

2.5.1 Logistička funkcija

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 logističku funkciju kod koje je $k = 1$, $x_0 = 0$ i $M = 1$, odnosno:

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

Za neprekidnu slučajnu varijablu X kažemo da ima logističku distribuciju s parametrima $m \in \mathbb{R}$ i $s > 0$ ako joj je funkcija gustoće dana s

$$f(x; m, s) = \frac{e^{-\frac{x-m}{s}}}{s\left(1 + e^{-\frac{x-m}{s}}\right)^2}.$$

Pišemo $X \sim Logist(m, s)$.
Standardna logistička funkcija je inverz *logit* funkcije. Odabiremo li standardnu logističku distribuciju $Logist(0, 1)$ za distribuciju grešaka U_i, tada je pripadna funkcija distribucije grešaka standardna logistička funkcija, a latentna formulacija modela odgovara modelu postavljenom u poglavlju 2.3.
Dakle, parametri logističkog regresijskog modela osim u terminima logaritmiranih izgleda mogu se interpretirati preko efekata koje varijable poticaja imaju na latentnu varijablu Y_i^* koja prati linearni model s grešakama iz logističke distribucije.

2.5.2 Probit

Za neprekidnu slučajnu varijabli X kažemo da ima normalnu distribuciju s parametrima $\mu \in \mathbb{R}$ i $\sigma^2 > 0$ ako joj je funkcija gustoće dana s

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

Pišemo $X \sim N(\mu, \sigma)$.
Postupamo li vođeni klasičnim linearnim modelom, očiti izbor distribucije grešaka je normalna razdioba $N(0, 1)$. Inverz funkcije distribucije jedinične normalne koji linearni prediktor prikazuje kao funkciju vjerojatnosti $\eta_i = \Phi^{-1}(\pi_i)$ nazivamo *probit*. On predstavlja jednu od alternativnih funkcija povezivanja za binarne podatke.
Promotrimo li nešto općenitiji slučaj gdje su greške $U_i \sim N(0, \sigma^2)$, dobivamo:

$$\begin{aligned}
\pi_i &= \mathbb{P}(Y_i^* > 0) \\
&= \mathbb{P}(U_i > -\mathbf{x}_i^T\boldsymbol{\beta}) \\
&= \mathbb{P}\Big(\frac{U_i}{\sigma} > \frac{-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\Big) \\
&= 1 - \Phi\Big(\frac{-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\Big) = \Phi\Big(\frac{\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}\Big)
\end{aligned}$$

Ovdje vidimo da $\boldsymbol{\beta}$ i σ ne možemo promatrati zasebno jer vjerojatnost ovisi o njihovom omjeru preko Φ. To je još jedan način da se vidi da vrijednosti koje latentna varijabla poprima nisu same po sebi određene. Zato uzimamo $\sigma = 1$ ili, ekvivalentno, interpretiramo efekt prediktora na način koliki je to dio standardne devijacije Y_i^*. Bez uvođenja latentne formulacije interpretacija tog modela nije moguća.
Mala mana korištenja poveznice *probit* je ta što ne postoji njezina zatvorena forma.

2.5.3 Komplementarna log-log

Kažemo da neprekidna slučajna varijabla ima Gumbelovu (log-Weibullovu) distribuciju s parametrima $\alpha \in \mathbb{R}$ i $\beta > 0$ ako joj je vjerojatnosna funkcija gustoće dana s

$$f(x;\alpha,\beta) = \frac{1}{\beta} e^{-\frac{x-\alpha}{\beta} - e^{-\frac{x-\alpha}{\beta}}}.$$

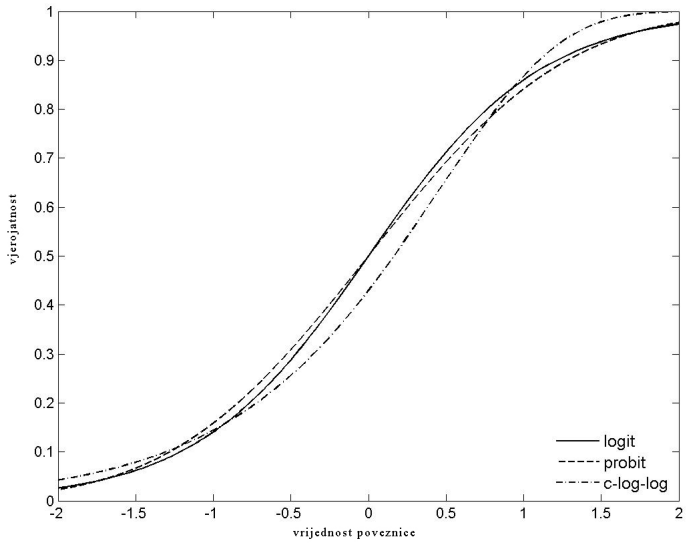
Pišemo $X \sim Gumbel(\alpha,\beta)$.
Ako pretpostavimo da greške $-U_i$ dolaze iz standardne Gumbelove distribucije $-U_i \sim Gumbel(0,1)$, onda za poveznicu uzimamo inverz njezine funkcije distribucije $F(x) = e^{-e^{-x}}$, odnosno iz (2.7) dobivamo:

$$g(x) = \log(-\log(1-x)).$$

Ni u ovom slučaju interpretacija nije moguća bez da promatramo latentnu formulaciju modela. Ova funkcija poveznica direktnu interpretaciju ima u hazardnim modelima kod hazardnih omjera.

2.5.4 Usporedba funkcija povezivanja

Sve tri poveznice su sigmoidalne funkcije pri čemu su *probit* i *logit* simetrične oko 0, a komplementarna log-log je asimetrična. Iz slike 2.1 je vidljivo da su *probit* i *logit* funkcije vrlo bliske te se zbog toga procijenjeni parametri u ta dva slučaja pretjerano ne razlikuju.



Slika 2.1: Funkcije distribucije koje odgovaraju poveznicama.

Prilikom usporedbe je zbog nejednakih varijanci standardne normalne i logističke distribucije potrebno promatrati procjene parametara koje smo prethodno standardizirali. Slično i kod usporedbe s procijenjenim parametrima koje su dobiju korištenjem komplementarne log-log funkcije kao poveznice (tablica 2.1).

poveznica	distribucija	μ	σ^2
logit	standardna logistička	0	$\frac{\pi^2}{3}$
probit	standardna normalna	0	1
c-log-log	standardna Gumbelova	-γ[1]	$\frac{\pi^2}{6}$

[1] Euler-Mascheronijeva konstanta, $\gamma \approx 0.577$

Tablica 2.1: Funkcije povezivanja i pripadne distribucije grešaka latentnog modela.

Komplementarna log-log funkcija je bliska *logit* funkciji za vjerojatnosti manje od 0.2. U praksi se često koristi u toksikologiji kod analize doživljenja gdje je vjerojatnost uspjeha vrlo velika ili vrlo mala.
Prilagođene modele koje dobijemo koristeći različite funkcije povezivanja uspoređujemo uz pomoć informacijskih kriterija. Najčešće su to Akaikeov i bayesovski informacijski kriterij pri čemu se najboljim modelom smatra onaj koji ima najmanje vrijednosti navedenih kriterija.

2.6 Procjena parametara

Sljedeće što nas zanima je procjena parametara logističkog modela. U ovom radu bit će predstavljena dva pristupa: metoda maksimalne vjerodostojnosti i minimalne χ^2 statistike.

2.6.1 Procjenitelj maksimalne vjerodostojnosti

Neka je $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$ 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 $\mathbf{y}^T = (y_1, \cdots, y_n)$ jedna njegova realizacija, tada je vjerodostojnost funkcija $L : \Theta \to \mathbb{R}$ definirana s:

$$L(\theta) \equiv L(\theta; \mathbf{y}) \coloneqq f_{\mathbf{Y}}(\mathbf{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta). \tag{2.8}$$

Definicija 2.6.1. *Statistika $\hat{\theta} = \hat{\theta}(\mathbf{Y})$ je procjenitelj maksimalne vjerodostojnosti (MLE) za θ ako vrijedi*

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

Maksimizacija funkcije L ekvivalentna je maksimizaciji log-vjerodostojnosti $l := \log L$ jer je $\log(\cdot)$ strogo rastuća injekcija te je to u praksi vrlo često puno lakše izvesti. Ponekad se koriste oznake L_n, odnosno l_n kako bi se označilo da se vjerodostojnost odnosi na uzorak duljine n.

Definicija 2.6.2. *Funkciju čije su komponente prve parcijalne derivacije log-vjerodostojnosti*

$$\mathbf{u}(\theta) = \frac{\partial l(\theta; \mathbf{y})}{\partial \theta}$$

nazivamo Fisherova funkcija pogotka.

Fisherovu funkciju pogotka možemo promatrati i kao vektor stupac parcijalnih derivacija od l po θ_i.

Definicija 2.6.3. *Fisherova informacijska matrica definirana je s*

$$I(\theta) = \mathbb{E}_\theta\Big[\mathbf{u}(\theta)\mathbf{u}(\theta)^T\Big].$$

Ovako definirana matrica naziva se i očekivana Fisherova informacijska matrica.

Definicija 2.6.4. *Za statistički model $\mathcal{P} = \{f(\,\cdot\,;\theta) \,:\, \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$, kod kojeg su gustoće $f(\cdot;\theta)$ izmjerive u odnosu na mjeru μ, kažemo da je regularan ako su zadovoljeni sljedeći uvjeti:*

1. $\operatorname{supp} f(\,\cdot\,;\theta) = \{y \in \mathbb{R} : f(\,\cdot\,;\theta) > 0\}$ *ne ovisi o θ;*
2. *Θ je otvoren skup u $\mathbb{R}^d$;*
3. *za svaki y preslikavanje $\theta \longmapsto f(y;\theta)$ je neprekidno diferencijabilno;*
4. *Fisherova informacijska matrica je pozitivno definitna;*
5. $\frac{d}{d\theta_i} \int\limits_{\operatorname{supp} f} f(y;\theta)d\mu(y) = \int\limits_{\operatorname{supp} f} \frac{d}{d\theta_i} f(y;\theta)d\mu(y)$, *za svaki $i = 1, \ldots, d$.*

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 $\theta = \boldsymbol{\pi}$:

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

$$\begin{aligned} l(\boldsymbol{\pi}; \mathbf{y}) &= \log L(\boldsymbol{\pi}; \mathbf{y}) \\ &= \sum_{i=1}^{k} \left(\log \binom{n_i}{y_i} + y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i) \right) \\ &= C + \sum_{i=1}^{k} \left(y_i \log g^{-1}(\eta_i) + (n_i - y_i) \log(1 - g^{-1}(\eta_i)) \right) \end{aligned} \tag{2.9}$$

gdje smo s C označili konstantni član $\sum_{i=1}^{k} \log \binom{n_i}{y_i}$ koji nam ne igra nikakvu ulogu u daljnjem izvodu. Koristeći definicije 1.2.4 i 1.2.5, log-vjerodostojnost je funkcija nepoznatih parametara modela $\beta_0, \beta_1, \ldots, \beta_p$.
Konkretno, promatramo li model iz potpoglavlja 2.3 i funkciju povezivanja:

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

dobivamo [6]:

$$l(\boldsymbol{\pi}; \mathbf{y}) = C + \sum_{i=1}^{k} \left(y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + n_i \log(1 - \pi_i) \right) \tag{2.10}$$

$$\begin{aligned} l(\boldsymbol{\pi}(\boldsymbol{\beta}); \mathbf{y}) &= C + \sum_{i=1}^{k} \left(y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + n_i \log(1 - \pi_i) \right) \\ &= C + \sum_{i=1}^{k} \left(y_i \eta_i + n_i \log \left(1 - \frac{e^{\eta_i}}{1 + e^{\eta_i}} \right) \right) \\ &= C + \sum_{i=1}^{k} \left(y_i \eta_i - n_i \log(1 + e^{\eta_i}) \right) \\ &= C + \sum_{i=1}^{k} \sum_{j=0}^{p} y_i x_{ij} \beta_j - \sum_{i=1}^{k} n_i \log \left(1 + e^{\sum_{j=0}^{p} x_{ij} \beta_j} \right) \end{aligned} \tag{2.11}$$

Deriviranjem izraza u (2.10) po pravilu kvocijenta dobivamo:

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - n_i \pi_i}{\pi_i (1 - \pi_i)}, \quad i \in \{1, \ldots, k\}.$$

Deriviranjem (2.11) i korištenjem lančanog pravila slijedi:

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{k} \frac{y_i - n_i \pi_i}{\pi_i (1 - \pi_i)} \cdot \frac{\partial \pi_i}{\partial \beta_r}, \quad r \in \{0, 1, \ldots, p\}. \tag{2.12}$$

U ovom trenutku prikladno je primijetiti da je moguće napraviti sljedeću supstituciju:

$$\frac{\partial \pi_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i} \cdot x_{ir} \tag{2.13}$$

pa uvrštavajući u (2.12) dobivamo:

$$\begin{aligned}
\frac{\partial l}{\partial \beta_r} &= \sum_{i=1}^{k} \frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_{ir} \\
&\overset{(*)}{=} \sum_{i=1}^{k} \frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)} \frac{d}{d\eta_i}\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right) x_{ir} \\
&= \sum_{i=1}^{k} \frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)} \frac{e^{\eta_i}}{(1+e^{\eta_i})^2} x_{ir} \\
&= \sum_{i=1}^{k} \frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)} \frac{e^{\eta_i}}{1+e^{\eta_i}} \frac{1}{1+e^{\eta_i}} x_{ir} \\
&\overset{(*)}{=} \sum_{i=1}^{k} (y_i - n_i\pi_i)\, x_{ir},
\end{aligned} \tag{2.14}$$

vektorski:

$$\mathbf{u}(\boldsymbol{\beta}) = X^T(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})). \tag{2.15}$$

Pritom se izrazom $\boldsymbol{\mu}(\boldsymbol{\beta})$ želi naglasiti da komponente vektora očekivanja ovise o vektoru parametara preko relacija $\mu_i = \frac{n_i}{1+e^{-\eta_i}}$, $\eta_i = \sum_{j=0}^{p} x_{ij}\beta_j$.
Promatramo ponovno parcijalne derivacije izraza u (2.14):

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_s \partial \beta_r} &= \sum_{i=1}^{k} -n_i x_{ir} \frac{\partial}{\partial \beta_s} \pi_i(\boldsymbol{\beta}) \\
&\overset{(*)}{=} -\sum_{i=1}^{k} n_i x_{ir} \frac{e^{-\eta_i}}{1+e^{-\eta_i}} \frac{1}{1+e^{-\eta_i}} x_{is} \\
&\overset{(*)}{=} -\sum_{i=1}^{k} x_{ir} n_i \pi_i (1-\pi_i) x_{is} \\
&= -[X^T W X]_{rs}.
\end{aligned} \tag{2.16}$$

Dakle, matrica parcijalnih derivacija drugog reda dana je s:

$$H(\boldsymbol{\beta}) = -X^T W X \tag{2.17}$$

gdje se $W \in \mathbb{R}^{k\times k}$ definira kao matrica težina $W = diag\{\, n_i\pi_i(1-\pi_i)\,\}$. Matrica $-H(\boldsymbol{\beta})$ često se naziva opažena Fisherova informacijska matrica.
S druge strane, po definiciji 2.6.3 je element Fisherove informacijske matrice:

$$\mathbb{E}\left(\frac{\partial l}{\partial \beta_s}\cdot\frac{\partial l}{\beta_r}\right) = \mathbb{E}\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{Y_i-n_i\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial \beta_s}\cdot\frac{Y_j-n_j\pi_j}{\pi_j(1-\pi_j)}\frac{\partial \pi_j}{\partial \beta_r}\right).$$

Korištenjem $Cov(Y_i, Y_j) = 0$ za $i \neq j$ slijedi:

$$\begin{aligned}
E\left(\frac{\partial l}{\partial \beta_s}\cdot\frac{\partial l}{\beta_r}\right) &= \sum_{i=1}^{k}\frac{Var(Y_i)}{\pi_i^2(1-\pi_i)^2}\frac{\partial \pi_i}{\partial \beta_r}\frac{\partial \pi_i}{\partial \beta_s}\\
&= \sum_{i=1}^{k}\frac{n_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial \beta_r}\frac{\partial \pi_i}{\partial \beta_s}\\
&\overset{(2.13)}{=} \sum_{i=1}^{k}\frac{n_i}{\pi_i(1-\pi_i)}\left(\frac{d\pi_i}{d\eta_i}\right)^2 x_{ir}x_{is}\\
&\overset{(*)}{=} \sum_{i=1}^{k}\frac{n_i}{\pi_i(1-\pi_i)}\left(\pi_i(1-\pi_i)\right)^2 x_{ir}x_{is}\\
&= \sum_{i=1}^{k} n_i\pi_i(1-\pi_i)\, x_{ir}x_{is}\\
&= [X^TWX]_{rs} \qquad (2.18)
\end{aligned}$$

Da bismo mogli tražiti procjenitelje maksimalne vjerodostojnosti kao stacionarne točke funkcije log-vjerodostojnosti, statistički model mora zadovoljavati uvjete regularnost. Sljedeći primjer to i pokazuje.

Primjer 2.6.5. *Binomni model zadovoljava uvjete regularnosti.*

1. $\operatorname{supp} f = \{0, 1, \ldots, n_i\}$ *ne ovisi o* $\boldsymbol{\pi}$, $\forall i \in \{1, \ldots, k\}$.
2. $\boldsymbol{\pi} \in (0,1)^k$ *što je otvoren skup u* $\mathbb{R}^k$.
3. $f(y;\boldsymbol{\pi}) = \begin{cases} (1-\pi_i)^{n_i} & y = 0\\ \pi_i^{n_i} & y = n_i\\ \binom{n_i}{y}\pi_i^y(1-\pi_i)^{n_i-y} & y \in \{1,\ldots,n_i-1\}\\ 0 & inače \end{cases}$

   Za fiksni y, $f(y;\boldsymbol{\pi})$ *je neprekidno diferencijabilna funkcija.*

4. *Pretpostavimo li da model ima manje parametara nego kovarijatnih razreda ($p+1 \leq k$) i da $0 < \pi_i < 1$, $\forall i$, tada je matrica $I(\boldsymbol{\pi}(\boldsymbol{\beta})) = X^T WX$ pozitivno definitna.*

5. *Model je diskretan pa se integrira u odnosu na konačnu brojeću mjeru. Tako promatramo parcijalnu derivaciju konačne sume te je zbog toga opravdano napraviti zamjenu derivacije i znaka sumacije.*

Po raspisu u (2.16) i (2.18) zaključujemo da vrijedi:

$$I(\boldsymbol{\beta}) = -H(\boldsymbol{\beta}), \tag{2.19}$$

odnosno da su očekivana i opažena informacijska matrica jednake. Također, Fisherovu informacijsku matricu iz definicije 2.6.3 možemo dobiti promatranjem očekivanja opažene infomacijske matrice.
Općenito, tvrdnju možemo izraziti pomoću sljedeće propozicije.

Propozicija 2.6.6. *Ako je model regularan i gustoća $f(\cdot;\theta)$ zadovoljava dodatna dva uvjeta:*

3'. za svaki $y \in \mathbb{R}$ preslikavanje $\theta \longmapsto f(y;\theta)$ je dvaput neprekidno diferencijabilno;

5'. $\frac{\partial^2}{\partial\theta_s\partial\theta_r} \int\limits_{\operatorname{supp} f} f(y;\theta)d\mu(y) = \int\limits_{\operatorname{supp} f} \frac{\partial^2}{\partial\theta_s\partial\theta_r} f(y;\theta)d\mu(y)$, za svaki $r, s = 1, \ldots, d$,

tada vrijedi

$$I(\theta) = -\mathbb{E}_\theta\Big[\frac{\partial \mathbf{u}(\theta)}{\partial\theta}\Big].$$

Dokaz. Za $r, s \in \{1, \ldots, d\}$ raspisujemo:

$$\begin{aligned}
\mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_s\partial\theta_r}\log f(\mathbf{Y};\theta)\Big] &= \sum_{i=1}^n \mathbb{E}_\theta\Big[\frac{\partial^2}{\partial\theta_s\partial\theta_r}\log f(Y_i;\theta)\Big] \\
&= \sum_{i=1}^n \int\limits_{\operatorname{supp} f} \frac{\partial}{\partial\theta_s}\Big(\frac{\frac{\partial}{\partial\theta_r}f(y_i;\theta)}{f(y_i;\theta)}\Big) f(y_i;\theta)d\mu(y_i) \\
&= \sum_{i=1}^n \int\limits_{\operatorname{supp} f} \Big[\frac{\partial^2}{\partial\theta_s\partial\theta_r}f(y_i;\theta) - \frac{\frac{\partial}{\theta_s}f(y_i;\theta)\frac{\partial}{\partial\theta_r}f(y_i;\theta)}{f^2(y_i;\theta)} f(y_i;\theta)\Big]d\mu(y_i) \\
&\overset{(5.')}{=} -\sum_{i=1}^n \int\limits_{\operatorname{supp} f} \frac{\partial}{\partial\theta_s}\log f(y_i;\theta)\frac{\partial}{\partial\theta_r}\log f(y_i;\theta) f(y_i;\theta)d\mu(y_i) \\
&= -\sum_{i=1}^n \mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_s}\log f(Y_i;\theta)\cdot\frac{\partial}{\partial\theta_r}\log f(Y_i;\theta)\Big]
\end{aligned}$$

$$\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_s}\log f(\mathbf{Y};\theta)\cdot\frac{\partial}{\partial\theta_r}\log f(\mathbf{Y};\theta)\Big]=\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_s}\log f(Y_i;\theta)\cdot\frac{\partial}{\partial\theta_r}\log f(Y_j;\theta)\Big]$$
$$\overset{\text{(nezav.,i}\neq\text{j)}}{=}\sum_{i=1}^{n}\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_s}\log f(Y_i;\theta)\cdot\frac{\partial}{\partial\theta_r}\log f(Y_i;\theta)\Big].$$

□

Newton-Raphsonov algoritam

Budući da ne možemo egzaktno odrediti nultočke funkcije pogotka, potrebno je koristiti iterativne metode. Zato promatramo Taylorov polinom prvog stupnja funkcije pogotka oko neke početne vrijednosti θ_0:

$$\mathbf{u}(\theta)\approx\mathbf{u}(\theta_0)+\frac{\partial\mathbf{u}(\theta)}{\partial\theta}(\theta-\theta_0).$$

Budući da za MLE procjenitelj $\hat{\theta}$ vrijedi $\mathbf{u}(\hat{\theta})=0$, dobivamo:

$$\hat{\theta}=\theta_0-H^{-1}(\theta_0)\mathbf{u}(\theta_0). \tag{2.20}$$

Ovim je izrazom dan općeniti Newton-Raphsonov algoritam. Za neku početnu procjenu θ_0 koristimo (2.20) kako bismo dobili nove procjenitelje te taj postupak ponavljamo dok proces ne počne konvergirati. Procedura brzo konvergira ako se funkcija lijepo ponaša u okolini maksimuma i ako je početno zadana procjena dovoljno blizu vrijednosti MLE-a.

Iterativna težinska metoda najmanjih kvadrata

Jedna inačice algoritma, koja se i najčešće koristi, je iterativna težinska metoda najmanjih kvadrata (engl. *IRLS-Iterative reweighted least squares*).
Kod nje ne promatramo realizacije y_i, nego novu varijablu Z koju nazivamo prilagođena zavisna varijabla i čije realizacije definiramo sa $z_i=g(y_i)$, a težine su funkcije prilagođenih vrijednosti $\hat{\mu}_i$. Proces je iterativan jer Z i težine kodirane u W ovise o prilagođenim vrijednostima koje dobivamo iz procjena poznatih u j-toj iteraciji algoritma.
U pogledu ovog algoritma korisno je promatrati funkciju povezivanja kao funkciju očekivanja μ_i.

$$\eta_i=g(\pi_i)=\log\frac{\pi_i}{1-\pi_i}=\log\frac{n_i\pi_i}{n_i-n_i\pi_i}$$
$$g(\mu_i)=\log\frac{\mu_i}{n_i-\mu_i}$$
$$\frac{d\eta_i}{d\mu_i}=\frac{n_i}{\mu_i(n_i-\mu_i)}=\frac{1}{n_i\pi_i(1-\pi_i)}.$$

Formuliramo realizacije varijable Z na način [6]:

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$

$$z_i = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i}.$$

Prikladne i "dobre" inicijalne vrijednosti možemo dobiti primjenom poveznice na originalne podatke. Čest slučaj u primjeni je $y_i = 0$ ili $y_i = n_i$ što nam stvara probleme po pitanju računanja log(0) ili nule u nazivniku. Njih rješavamo neznatnom promjenom brojnika i nazivnika, npr. dodavanjem $1/2$ pa i-tu komponentu vektora $\mathbf{z}$ računamo:

$$z_i = \log \frac{y_i + 1/2}{n_i - y_i + 1/2}.$$

Ako promotrimo liniju (9) Algoritma 1 za IRLS, možemo vidjeti gdje je skrivena srž Newton-Raphsonovog algoritma:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{(j+1)} &= (X^T W^{(j)} X)^{-1} X^T W^{(j)} \mathbf{z}^{(j)} \\
&= (X^T W^{(j)} X)^{-1} X^T W^{(j)} (\hat{\boldsymbol{\eta}}^{(j)} + (W^{(j)})^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(j)})) \\
&= \hat{\boldsymbol{\beta}}^{(j)} + (X^T W^{(j)} X)^{-1} X^T (\mathbf{y} - \hat{\boldsymbol{\mu}}^{(j)}) \\
&\overset{(2.15)}{=} \hat{\boldsymbol{\beta}}^{(j)} - (-(X^T W^{(j)} X)^{-1})\, \mathbf{u}(\hat{\boldsymbol{\beta}}^{(j)}).
\end{aligned}$$

Algoritam 1 IRLS

1: Neka je $\boldsymbol{\beta}^{T(0)} = (\beta_0^{(0)}, \ldots, \beta_p^{(0)})$ vektor inicijalnih vrijednosti parametara.
2: **for** $j = 0$ to n_{iter} **do**
3: $\quad \hat{\boldsymbol{\eta}}^{(j)} = X\boldsymbol{\beta}^{(j)}$
4: $\quad \hat{\boldsymbol{\mu}}^{(j)} = g^{-1}(\hat{\boldsymbol{\eta}}^{(j)})$ ▷ funkcija g se primjenjuje po komponentama vektora $\hat{\boldsymbol{\eta}}^{(j)}$
5: $\quad W^{(j)} = diag\{\frac{\hat{\mu}_i^{(j)}(n_i - \hat{\mu}_i^{(j)})}{n_i}\}$
6: $\quad$ **for** $i = 1$ to n **do**
7: $\quad\quad z_i^{(j)} = \hat{\eta}_i^{(j)} + (y_i - \hat{\mu}_i^{(j)}) \frac{n_i}{\hat{\mu}_i^{(j)}(n_i - \hat{\mu}_i^{(j)})}$
8: $\quad$ Regresija sa $\mathbf{z}^{(j)}$ kao zavisnom varijablom u odnosu na kovarijate kodirane u X.
9: $\quad \hat{\boldsymbol{\beta}}^{(j+1)} = (X^T W^{(j)} X)^{-1} X^T W^{(j)} \mathbf{z}^{(j)}$
10: $\quad$ **if** $\|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}\|_2 < \epsilon$ **then**
11: $\quad\quad$ **break**

Fisherova metoda pogađanja

Alternativnu proceduru prvi je predložio Fisher. Radi se o varijanti Newton-Raphsonovog algoritma gdje se napravi zamjena matrice $-H$ s njezinim očekivanjem što je po propoziciji 2.6.6 jednako Fisherovoj informacijskoj matrici. Dana je izrazom:

$$\hat{\theta} = \theta_0 + I^{-1}(\theta_0)\mathbf{u}(\theta_0)$$

i naziva se Fisherova metoda pogađanja (engl. *Fisher scoring method*). U ovom slučaju ona je očigledno ekvivaletna prethodno opisanoj metodi.

Svojstva procjenitelja maksimalne vjerodostojnosti

Prisjetimo se, za niz procjenitelja $(\theta_n)_n$ vektora parametara $\theta \in \Theta \subset \mathbb{R}^d$ kažemo da je (slabo) konzistentan ako

$$\forall \epsilon > 0,\ \forall \theta \in \Theta,\ \lim_{n\to\infty} \mathbb{P}_\theta\left(\|\theta_n - \theta\| \geq \epsilon\right) = 0, \text{u oznaci } \theta_n \xrightarrow{\mathbb{P}_\theta} \theta.$$

Općenito, niz statistika $(\theta_n)_n$ je asimptotski normalan ako postoje neslučajne funkcije parametra $A_n(\theta)$, $B_n(\theta) > 0$, $n \in \mathbb{N}$, takve da vrijedi

$$\forall \theta \in \Theta,\ \forall x \in \mathbb{R}^d,\ \lim_{n\to\infty} \mathbb{P}_\theta\left(B_n(\theta)^{-1}(\theta_n - A_n(\theta)) \leq x\right) = \Phi_d(x)$$

gdje s Φ_d označavamo funkciju distribucije multivarijatnog normalno distribuiranog slučajnog vektora $\mathbf{Z}$ s očekivanjem $\mathbf{0}$ i kovarijacijskom matricom $I \in \mathbb{R}^{d\times d}$.
Kraće pišemo:

$$B_n(\theta)^{-1}(\theta_n - A_n(\theta)) \xrightarrow{D-\mathbb{P}_\theta} N(0, I), \quad n \to \infty, \quad \forall \theta \in \Theta.$$

Napomena 2.6.7. *Ukoliko je $\Theta \subset \mathbb{R}^d$, $d > 1$, izraz $B_n(\theta) > 0$ označava pozitivno definitnu matricu.*

Uz proširene uvjete regularnosti iz propozicije 2.6.6 i dodatne pretpostavke na elemente matrice X, vektor procjenitelja dobiven metodom maksimalne vjerodostojnosti je nepristran s kovarijacijskom matricom:

$$cov(\hat{\boldsymbol{\beta}}) = (X^T W X)^{-1},$$

pri čemu je W matrica dobivena u posljednoj iteraciji Algoritma 1 za IRLS, konzistentan i asimptotski normalan kada $k \to \infty$ [6]. Pretpostavke teorema o asimptotskoj normalnost MLE-a u slučaju nejednako distribuiranih nezavisnih slučajnih varijabli mogu se pronaći u [8].

Isti rezultat vrijedi ako fiksiramo broj razreda k te u svakom od njih broj jedinki n_i puštamo u beskonačnost. Radi jednostavnosti pretpostavimo da je $n = n_i$ za svaki i. Tada po jakom zakonu velikih brojeva [9, teorem 12.14] vrijedi $\hat{\pi}_i := \frac{1}{n} Y_i \xrightarrow{g.s.} \pi_i$ kada $n \to \infty$. Po Moivre-Laplaceovom teoremu [9, teorem 5.6] za svaki $i = 1, \ldots, k$ slijedi:

$$\sqrt{n} \frac{\hat{\pi}_i - \pi_i}{\sqrt{\pi_i(1-\pi_i)}} \xrightarrow{D-\mathbb{P}_{\pi_i}} N(0,1),$$

odnosno

$$\sqrt{n}(\hat{\pi}_i - \pi_i) \xrightarrow{D-\mathbb{P}_{\pi_i}} N(0, \pi_i(1-\pi_i)).$$

Zbog nezavisnosti varijabli Y_i imamo:

$$\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{D-\mathbb{P}_{\pi}} N(\mathbf{0}, S), \quad S = diag\{\pi_i(1-\pi_i)\}.$$

Definiramo li funkciju $h : \mathbb{R}^k \to \mathbb{R}^k$ s $h(\boldsymbol{\pi}) = [\log \frac{\pi_1}{1-\pi_1}, \cdots, \log \frac{\pi_k}{1-\pi_k}]^T$, po Cramerovom teoremu [3, teorem 7] vrijedi:

$$\sqrt{n}(h(\hat{\boldsymbol{\pi}}) - h(\boldsymbol{\pi})) \xrightarrow{D-\mathbb{P}_{\pi}} N(\mathbf{0}, \dot{h}(\boldsymbol{\pi}) S \dot{h}^T(\boldsymbol{\pi})).$$

Pritom je $\dot{h}(\boldsymbol{\pi}) = [\frac{h_i}{\pi_j}]_{ij} = S^{-1}$. Nadalje, slijedi:

$$\sqrt{n}(X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}) \xrightarrow{D-\mathbb{P}_{\beta}} N(\mathbf{0}, S^{-1}),$$

odnosno

$$\hat{\boldsymbol{\beta}} \sim AN(\boldsymbol{\beta}, (X^T W X)^{-1}).$$

2.6.2 Procjenitelj minimalne χ^2 statistike

Drugi pristup dobivanja procjenitelja postavljenog modela dio je teorije kvadratnih formi asimptotski normalnih slučajnih varijabli. Problem dobivanja procjenitelja postavlja se kao problem minimizacije udaljenosti.

Neka su Z_n k-dimenzionalni slučajni vektori, $\theta \in \Theta$ vektor parametara te pretpostavimo da je Θ neprazan i otvoren podskup od $\mathbb{R}^d$, $d \leq k$. Pretpostavimo da su Z_n asimptotski normalno distribuirani, tj. da vrijedi:

$$\sqrt{n}(Z_n - A(\theta)) \xrightarrow{D-\mathbb{P}_{\theta}} N(0, C(\theta)), \tag{2.21}$$

gdje je $A(\theta) \in \mathbb{R}^k$ i $C(\theta) \in \mathbb{R}^{k \times k}$ kovarijacijska matrica, $\forall \theta \in \Theta$. Kvadratnom formom $Q_n(\theta)$ izražavamo udaljenost Z_n od $A(\theta)$:

$$Q_n(\theta) = n(Z_n - A(\theta))^T M(\theta)(Z_n - A(\theta)) \tag{2.22}$$

gdje je $M(\theta) \in \mathbb{R}^{k \times k}$ pozitivno definitna simetrična matrica.

Definicija 2.6.8. *Za niz procjenitelja $(\theta_n^*)_n$ kažemo da je niz procjenitelja minimalne χ^2 statistike ako:*

$$Q_n(\theta_n^*) - \inf_{\theta\in\Theta} Q_n(\theta) \xrightarrow{\mathbb{P}_\theta} 0. \tag{2.23}$$

Pretpostavimo da su komponentne funkcije od $A(\theta)$ diferencijabilne te matrica $\dot{A}(\theta) = \left[\frac{\partial A_i}{\partial \theta_j}\right]_{ij}$ punog ranga. Označimo: $\dot{A} \equiv \dot{A}(\theta_0)$, $M \equiv M(\theta_0)$, $C \equiv C(\theta_0)$.
U svrhu dokazivanja teorema potrebno je uvesti dodatne pretpostavke na komponentne funkcije od $A(\theta)$ i $M(\theta)$ koje se uglavnom tiču njihove diferencijabilnost, neprekidnosti i ograničenosti [3].

Teorem 2.6.9. *Za svaki niz procjenitelja minimalne χ^2 statistike $(\theta_n^*)_n$ vrijedi:*

$$\sqrt{n}(\theta_n^* - \theta_0) \xrightarrow{D-\mathbb{P}_{\theta_0}} N(0, \Sigma) \tag{2.24}$$

gdje je

$$\Sigma = (\dot{A}^T M \dot{A})^{-1} \dot{A}^T M C M \dot{A} (\dot{A}^T M \dot{A})^{-1}. \tag{2.25}$$

Dokaz. Dokaz u [3, teorem 23]. □

Oznakom $\Sigma(M)$ naglašavamo ovisnost matrice u (2.25) o matrici $M \equiv M(\theta_0)$. Sljedeći korolar pokazuje nam kako moramo odabrati matricu M da bi Σ bila minimalna kovarijacijska matrica u asimptotskom smislu.

Korolar 2.6.10. *Ako postoji pozitivno definitna simetrična $M_0 \in \mathbb{R}^{k\times k}$ takva da vrijedi $CM_0\dot{A} = \dot{A}$, tada je $\Sigma(M_0) = (\dot{A}^T M_0 \dot{A})^{-1}$. Štoviše, $\Sigma(M_0) \leq \Sigma(M)$, za svaki M.*

Dokaz. Iz pretpostavke zamjenom $CM_0\dot{A}$ s $\dot{A}$ odmah dobijemo:

$$\Sigma(M_0) = (\dot{A}^T M_0 \dot{A})^{-1} \dot{A}^T M_0 C M_0 \dot{A} (\dot{A}^T M_0 \dot{A})^{-1} = (\dot{A}^T M_0 \dot{A})^{-1}.$$

Nadalje, zbog pozitivne semidefinitnosti kovarijacijske matrice C imamo:

$$\begin{aligned}
0 &\leq (M\dot{A}(\dot{A}^T M\dot{A})^{-1} - M_0\dot{A}(\dot{A}^T M_0\dot{A})^{-1})^T\, C\, (M\dot{A}(\dot{A}^T M\dot{A})^{-1} - M_0\dot{A}(\dot{A}^T M_0\dot{A})^{-1}) \\
&= (\dot{A}^T M\dot{A})^{-1}\dot{A}^T MCM\dot{A}(\dot{A}^T M\dot{A})^{-1} - (\dot{A}^T M\dot{A})^{-1}\dot{A}^T MCM_0\dot{A}(\dot{A}^T M_0\dot{A})^{-1} - \\
&\quad (\dot{A}^T M_0\dot{A})^{-1}\dot{A}^T M_0CM\dot{A}(\dot{A}^T M\dot{A})^{-1} + (\dot{A}^T M_0\dot{A})^{-1}\dot{A}^T M_0CM_0\dot{A}(\dot{A}^T M_0\dot{A})^{-1} \\
&= (\dot{A}^T M\dot{A})^{-1}\dot{A}^T MCM\dot{A}(\dot{A}^T M\dot{A})^{-1} - (\dot{A}^T M_0\dot{A})^{-1} \\
&= \Sigma(M) - \Sigma(M_0).
\end{aligned}$$

Pritom prva jednakost vrijedi zbog $M = M^T$ i $M_0 = M_0^T$, a druga korištenjem pretpostavke na M_0. □

Iz danih rezultata možemo zaključiti da su procjenitelji minimalne χ^2 statistike također asimptotski normalno distribuirani te je moguće pronaći matricu M tako da njihova kovarijacijska matrica bude što je moguće manja.

Generalizacija χ^2 statistike

Moguće je promatrati matricu M i kao funkciju $M(Z_n, \theta)$ te minimizirati tzv. modificiranu χ^2 statistiku:

$$Q_n(\theta) = n(Z_n - A(\theta))^T M(Z_n, \theta)(Z_n - A(\theta)). \tag{2.26}$$

Tada zamjenom $M(Z_n, \theta)$ s njezinim limesom $M(A(\theta), \theta)$ i minimizacijom dobivamo asimptotski ekvivalentne procjenitelje onima koji minimiziraju kvadratnu formu (2.26).

Nadalje, neka je $g : \mathbb{R}^k \to \mathbb{R}^k$ takva da $g(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_k(\mathbf{x})]^T$. Pretpostavimo da su g_i klase C^1, $\forall i \in \{1, \ldots, k\}$. Označimo s $\dot{g} = \left[\frac{\partial g_i}{\partial x_j}\right]_{ij}$ te pretpostavimo da je matrica punog ranga. Tada po Cramerovom teoremu [3, teorem 7] slijedi:

$$\sqrt{n}(g(Z_n) - g(A(\theta))) \xrightarrow{D-\mathbb{P}_\theta} N(0, \dot{g}(A(\theta))C(\theta)\dot{g}(A(\theta))^T). \tag{2.27}$$

Kvadratna forma koju pritom promatramo je:

$$Q_n(\theta) = n(g(Z_n) - g(A(\theta)))^T[(\dot{g}(A(\theta))^T)^{-1}M(\theta)(\dot{g}(A(\theta)))^{-1}](g(Z_n) - g(A(\theta))) \tag{2.28}$$

i naziva se transformirana χ^2 statistika.

U praksi se često kombinira modificirani i transformirani oblik promatrane kvadratne forme u što ćemo se uvjeriti i primjenom na modelu koji smo razvili. Također, odabir funkcije g je takav da su komponentne funkcije $g(A(\theta))$ linearne funkcije komponenata vektora θ.
Prisjetimo se modela definiranog u potpoglavlju 2.3 uz pretpostavku da u svakom od k kovarijatnih razreda imamo jednaki broj jedinki n, dakle ukupno promatramo kn jedinki. Neka su $y_i = n_i$ realizacije slučajnih varijabli $Y_i \sim B(n, \pi_i)$ za $i \in \{1, \ldots, k\}$, odnosno u svakom od kovarijatnih razreda imamo n_i jedinki koje imaju promatrano svojstvo (n_i jedinica). Tada formiramo χ^2 statistiku na sljedeći način:

$$\chi^2 = \sum_{i=1}^{k}\left[\frac{(n_i - n\pi_i(\boldsymbol{\beta}))^2}{n\pi_i(\boldsymbol{\beta})} + \frac{((n - n_i) - n(1 - \pi_i(\boldsymbol{\beta})))^2}{n(1 - \pi_i(\boldsymbol{\beta}))}\right].$$

Stavljanjem na zajednički nazivnik i sređivanjem brojnika dobivamo:

$$\chi^2 = \sum_{i=1}^{k}\frac{(n_i - n\pi_i(\boldsymbol{\beta}))^2}{n\pi_i(\boldsymbol{\beta})(1 - \pi_i(\boldsymbol{\beta}))} = n\sum_{i=1}^{k}\frac{\left(\frac{n_i}{n} - \pi_i(\boldsymbol{\beta})\right)^2}{\pi_i(\boldsymbol{\beta})(1 - \pi_i(\boldsymbol{\beta}))} \tag{2.29}$$

što odgovara kvadratnoj formi (2.22) za $Z_n^T = [\frac{n_1}{n}, \ldots, \frac{n_k}{n}]$, $A^T(\boldsymbol{\beta}) = [\pi_1(\boldsymbol{\beta}), \ldots, \pi_k(\boldsymbol{\beta})]$ i $M(\boldsymbol{\beta}) = diag\{(\,\pi_i(\boldsymbol{\beta})(1 - \pi_i(\boldsymbol{\beta}))\,)^{-1}\}$.
Minimiziranje dobivenog izraza uključuje tehnički vrlo kompleksan postupak pa ga zato linearizirамо. Kako je $\pi_i(\boldsymbol{\beta}) = (1 + \exp\{\sum_{j=0}^{p} x_{ij}\beta_j\})^{-1}$, uzmemo li za funkcije g_i funkciju $\operatorname{logit}\pi_i = \log(\frac{\pi_i}{1-\pi_i})$, dobivamo $g(\boldsymbol{\pi}) = X\boldsymbol{\beta}$. Pritom je jasno da je $\frac{d}{d\pi_i}\operatorname{logit}\pi_i = \frac{1}{\pi_i(1-\pi_i)}$ neprekidno za $\pi_i \in (0, 1)$. Po izrazu (2.28) transformirana statistika izgleda:

$$\chi^2 = n\sum_{i=1}^{k} \pi_i(\boldsymbol{\beta})\,(1 - \pi_i(\boldsymbol{\beta}))\left(\operatorname{logit}\left(\frac{n_i}{n}\right) - \operatorname{logit}\pi_i(\boldsymbol{\beta})\right)^2.$$

Po jakom zakonu velikih brojeva [9, teorem 12.14] niz slučajnih vektora $(Z_n^T)_n$ je jako konzistentan niz procjenitelja za $\boldsymbol{\pi}(\boldsymbol{\beta})$ pa zamjenom prvih dvaju izraza u sumi s njihovim procjeniteljima, dobivamo modificiranu statistiku:

$$\operatorname{logit}\chi^2 = n\sum_{i=1}^{k} \frac{n_i}{n}\left(1 - \frac{n_i}{n}\right)\left(\operatorname{logit}\left(\frac{n_i}{n}\right) - \sum_{j=0}^{p} x_{ij}\beta_j\right)^2 \tag{2.30}$$

koju nazivamo Berksonov *logit* χ^2.
Promatramo li izraz u (2.30) kao funkciju $\Psi(\beta_0, \ldots, \beta_p)$, minimizirati ju možemo traženjem stacionarnih točaka što se u konačnici svodi na rješavanje linearnog sustava. Sustav linearnih jednadžbi dan je s:

$$\frac{\partial\Psi}{\partial\beta_0} = 0 \quad \Rightarrow \quad \sum_{j=0}^{p}\sum_{i=1}^{k} \frac{n_i}{n}\left(1 - \frac{n_i}{n}\right)x_{ij}\beta_j = \sum_{i=1}^{k} \frac{n_i}{n}\left(1 - \frac{n_i}{n}\right)\operatorname{logit}\frac{n_i}{n}$$

$$\frac{\partial\Psi}{\partial\beta_l} = 0 \quad \Rightarrow \quad \sum_{j=0}^{p}\sum_{i=1}^{k} \frac{n_i}{n}\left(1 - \frac{n_i}{n}\right)x_{il}x_{ij}\beta_j = \sum_{i=1}^{k} \frac{n_i}{n}\left(1 - \frac{n_i}{n}\right)x_{il}\operatorname{logit}\frac{n_i}{n}, \quad l \in \{1, \ldots, p\}.$$

Kod binarnih podataka možemo koristiti i neku od alternativnih funkcija povezivanja, npr. *probit*, za dobivanje generaliziranog modela. Promotrimo model $\pi_i = \Phi(\eta_i)$. Izraz u (2.29) je tada:

$$\chi^2 = n\sum_{i=1}^{k} \frac{\left(\frac{n_i}{n} - \Phi(\eta_i)\right)^2}{\Phi(\eta_i)(1 - \Phi(\eta_i))}.$$

S ciljem linearizacije g_i je potrebno definirati na način $g_i(\mathbf{x}) = \text{probit}\, x_i$. Pritom se derivacija dobije po pravilu deriviranja inverzne funkcije $\frac{d}{dx_i}\,\text{probit}\, x_i = \frac{1}{\Phi'(\Phi^{-1}(x_i))}$. Transformirana χ^2 statistika je onda jednaka:

$$\chi^2 = n\sum_{i=1}^{k}\frac{\left(\text{probit}\,\frac{n_i}{n}-\eta_i\right)^2(\Phi'(\eta_i))^2}{\Phi(\eta_i)(1-\Phi(\eta_i))} = n\sum_{i=1}^{k}\frac{\left(\text{probit}\,\frac{n_i}{n}-\eta_i\right)^2\varphi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))}$$

gdje smo s $\varphi(\cdot)$ označili funkciju gustoće standardne normalne razdiobe. Nadalje, zamjenom π_i s pripadnim procjeniteljima imamo:

$$\chi^2 = n\sum_{i=1}^{k}\frac{\left(\text{probit}\,\frac{n_i}{n}-\eta_i\right)^2\varphi(\Phi^{-1}(\pi_i))^2}{\pi_i(\boldsymbol{\beta})(1-\pi_i(\boldsymbol{\beta}))} = n\sum_{i=1}^{k}\frac{\left(\text{probit}\,\frac{n_i}{n}-\sum_{j=0}^{p}x_{ij}\beta_j\right)^2\varphi\left(\Phi^{-1}(\frac{n_i}{n})\right)^2}{\frac{n_i}{n}\left(1-\frac{n_i}{n}\right)}.$$

Dobiveni izraz se minimizira na isti način kao i u logističkom modelu.

2.7 Prilagodba modela podacima

Nakon procjene parametara modela, prirodno se nameće pitanje o tome 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 $i = 1, \ldots, k$. Ono što zapravo radimo jest da mjerimo "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 koje se temelje na funkciji vjerodostojnosti.

2.7.1 Statistika odstupanja

Prisjetimo se, za logistički model log-vjerodostojnost je:

$$l(\boldsymbol{\pi};\mathbf{y}) = C + \sum_{i=1}^{k}\left(y_i\log\left(\frac{\pi_i}{1-\pi_i}\right)+n_i\log(1-\pi_i)\right). \tag{2.31}$$

Za konkretnu realizaciju $\mathbf{y}$ funkcija vjerodostojnosti objedinjuje informaciju o nepoznatim parametrima promatranog modela. Vrijednost log-vjerodostojnosti koju dobivamo uvrštavanjem procjenitelja maksimalne vjerodostojnosti u (2.31) govori nam do koje se mjere promatrani model prilagodio podacima. Budući da ona ovisi o broju opservacija u uzorku, ne možemo ju 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 ugniježdenih modela s pripadnim pretpostavkama:

$$\begin{aligned} \mathcal{H}_0 &: \text{ Model } M \text{ je točan.} \\ \mathcal{H}_1 &: \text{ Model } M \text{ nije točan.} \end{aligned} \tag{2.32}$$

Pri tome je u $\mathcal{H}_0$ model koji se promatra ($p+1 < k$), a alternativna hipoteza reprezentira tzv. puni ili saturirani model M_f. On je egzaktno prilagođen podacima jer svakoj opservaciji odgovara jedan parametar ($p + 1 = k$). Imamo:

$$D = -2 \log \frac{\text{vjerodostojnost modela } M}{\text{vjerodostojnost modela } M_f}. \tag{2.33}$$

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

$$\begin{aligned} D &= -2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) + 2l(\tilde{\boldsymbol{\pi}}; \mathbf{y}) \\ &\overset{(2.31)}{=} (2C_2 - 2C_1) + 2 \sum_{i=1}^{k} \left(y_i \log \frac{\tilde{\pi}_i}{\hat{\pi}_i} + (n_i - y_i) \log \frac{1 - \tilde{\pi}_i}{1 - \hat{\pi}_i} \right) \\ &= (2C_2 - 2C_1) + 2 \sum_{i=1}^{k} \left(y_i \log \frac{y_i}{\hat{y}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \hat{y}_i} \right). \end{aligned}$$

Zanemarivanjem konstantnog člana dobivamo:

$$D(\mathbf{y}; \hat{\boldsymbol{\pi}}) = 2 \sum_{i=1}^{k} \left(y_i \log \frac{y_i}{\hat{y}_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i - \hat{y}_i} \right). \tag{2.34}$$

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 učiti što manjom. Izraz D poprima velike vrijednosti kada je brojnik u (2.33) relativno mali u odnosu na nazivnik što ukazuje na nedovoljno dobar model, odnosno model je presiromašan. S druge strane, vrijednosti D su manje kad su vrijednosti brojnika i nazivnika bliske što znači da je promatrani model dovoljno dobar.
Nadalje, zanima nas kojoj distribuciji pripada statistika $D(\mathbf{Y}; \hat{\boldsymbol{\pi}})$. Ako vrijedi $\mathcal{H}_0$, $D(\mathbf{Y}; \hat{\boldsymbol{\pi}})$ ima asimptotsku χ^2 distribuciju s $k-(p+1)$ stupnjeva slobode pri čemu je k broj kovarijatnih razreda, a $p+1$ broj nepoznatih parametara u modelu M [3, teorem 22]. Ovaj zaključak moguće je donijeti promatranjem uzoraka konačne duljine uz određene uvjete. Jedan od

uvjeta je nezavisnost opservacija iz binomne razdiobe koji je po pretpostavci našeg modela odmah zadovoljen. Drugi je taj da, neovisno o broju kovarijatnih razreda k, vrijedi da $n_i \to \infty$, odnosno $n_i\pi_i(1-\pi_i) \to \infty$, $\forall i = 1, \ldots, k$. Dakle, potrebno je prilagoditi π_i pa kad $n_i \to \infty$, broj takvih parametara je fiksan i iznosi k te možemo donijeti zaključak o asimpototskoj distribuciji statistike $D(\mathbf{Y}; \hat{\boldsymbol{\pi}})$.
Do odstupanja od asimptotske χ^2 razdiobe dolazi kad je $n_i = 1$, za svaki i. Tada je $k = n$ pa puštanjem $n \to \infty$ i broj prilagođenih vjerojatnosti $\hat{\pi}_i$ teži u beskonačno. Također, do odstupanja može doći ako podaci nisu jednoliko grupirani, odnosno ako su za neke i $n_i = 1$ ili vrlo mali u odnosu na ostale.
Budući da je očekivanje χ^2 distribucije s n stupnjeva slobode jednako n, zadovoljavajući model je onaj za kojeg je $D \approx n$. Zbog toga se srednje odstupanje definira kao omjer odstupanja i stupnjeva slobode te je prihvatljiv model onaj za kojeg je taj omjer približno 1.

Općenito, statistika D koristi se za usporedbu dvaju modela, nazovimo ih M i M_l, takvih da je $M \subset M_l$ te podatke modeliraju sa s, odnosno l varijabli poticaja. Želimo testirati poboljšavamo li značajno model M dodavanjem varijabli poticaja pa se alternativna pretpostavka u (2.32) mijenja na način:

$$\mathcal{H}_1' : \text{ Model } M_l \text{ je točan.} \tag{2.35}$$

Označimo s $\hat{\boldsymbol{\pi}}_0$ i $\hat{\boldsymbol{\pi}}_1$ prilagođene vrijednosti za modele M i M_l, respektivno. Razlika odstupanja je:

$$D(\mathbf{y}; \hat{\boldsymbol{\pi}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\pi}}_1) = 2l(\hat{\boldsymbol{\pi}}_1; \mathbf{y}) - 2l(\hat{\boldsymbol{\pi}}_0; \mathbf{y}) \tag{2.36}$$

što odgovara testu omjera vjerodostojnosti za testiranje $\mathcal{H}_0$ naprema $\mathcal{H}_1'$. Ako vrijedi $\mathcal{H}_0$, ova statistika ima asimptotsku χ^2 razdiobu s $df = (n-s)-(n-l) = l-s$ stupnjeva slobode, neovisno o veličini k i n_i.

2.7.2 Pearsonova χ^2 statistika

Najpoznatija alternativna mjera prilagođenosti modela je Pearsonova χ^2 statistika definirana s:

$$X_P^2 = \sum_{i=1}^{k} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{2.37}$$

koja ima asimptotsku χ^2 distribuciju s $k-(p+1)$ stupnjeva slobode. U praksi se numeričke vrijednosti statistike D i X_P^2 pretjerano ne razlikuju. Budući da za MLE log-vjerodostojnost poprima svoj maksimum, statistika D se minimizira tim procjeniteljem.

Vođeni tom činjenicom, ako smo vektor procjenitelja modela dobili metodom maksimizacije vjerodostojnosti, za mjeru odstupanja bolje je koristiti statistiku D. Također, još jedna prednost statistike D je mogućnost usporedbe dvaju modela, za što nije uvijek najbolje koristiti Pearsonovu χ^2 statistiku. U praksi se može dogoditi da se vrijednost X_P^2 poveća dodavanjem varijabli poticaja u model, dok razlika odstupanja pokazuje da je tako dobiven model značajno bolji.

2.7.3 Rijetkost podataka

Kažemo da su podaci rijetki (engl. *sparse*) ako su mnogi n_i manji od 5. To svojstvo jako utječe na aproksimativnu distribuciju dosad opisanih statistika. Ekstremni slučaj rijetkosti podataka je kada je u svakom razredu po jedna opservacija. U praksi je to čest slučaj kada u modelu postoji barem jedna neprekidna varijabla poticaja.
Promotrimo najprije statistiku odstupanja u slučaju $k = n$. Po (2.31) logaritmirana funkcija vjerodostojnosti je:

$$\begin{aligned} l(\boldsymbol{\pi}; \mathbf{y}) &= C + \sum_{i=1}^{k} \left(y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right) \\ &= C + \sum_{i=1}^{k} (y_i \log \pi_i + (1-y_i)\log(1-\pi_i)). \end{aligned} \tag{2.38}$$

Budući da kod punog modela y_i može poprimiti vrijednosti 0 ili 1, izraz $l(\tilde{\boldsymbol{\pi}}) = 0$ pa u izrazu za odstupanje preostane samo prvi član:

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\pi}}) &= -2l(\hat{\boldsymbol{\pi}}; \mathbf{y}) \\ &= -2 \sum_{i=1}^{k} (y_i \log \hat{\pi}_i + (1-y_i)\log(1-\hat{\pi}_i)). \end{aligned} \tag{2.39}$$

Deriviranjem (2.38) i ponavljanjem istog izvoda kao u potpoglavlju 2.6 (2.14) dobivamo:

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{k} (y_i - \pi_i)\, x_{ir},$$

$$\begin{aligned}\sum_{j=0}^{p}\beta_j\frac{\partial l}{\partial\beta_j} &= \sum_{j=0}^{p}\beta_j\sum_{i=1}^{k}(y_i-\pi_i)\,x_{ji}\\ &= \sum_{i=1}^{k}(y_i-\pi_i)\sum_{j=0}^{p}\beta_j x_{ji}\\ &= \sum_{i=1}^{k}(y_i-\pi_i)\log\frac{\pi_i}{1-\pi_i}.\end{aligned}$$

Ako na lijevoj strani uvrstimo komponente MLE-a $\hat{\boldsymbol{\beta}}$, prilagođene vjerojatnosti $\hat{\pi}_i$ moraju zadovoljavati sljedeće:

$$\sum_{i=1}^{k}(y_i-\hat{\pi}_i)\,\log\frac{\hat{\pi}_i}{1-\hat{\pi}_i}=0,$$

odnosno:

$$\sum_{i=1}^{k}y_i\,\operatorname{logit}\hat{\pi}_i=\sum_{i=1}^{k}\hat{\pi}_i\,\operatorname{logit}\hat{\pi}_i. \tag{2.40}$$

Uvršavanjem (2.40) u (2.39) dobivamo:

$$D(\mathbf{y};\hat{\boldsymbol{\pi}})=-2\sum_{i=1}^{k}(\hat{\pi}_i\operatorname{logit}\hat{\pi}_i+\log(1-\hat{\pi}_i)).$$

Vidimo da u ovom slučaju odstupanje nije dobra mjera prilagodbe modela jer o binarnom ishodu y_i ovisi samo preko $\hat{\pi}_i$ pa nikako ne opisuje razliku između opaženih i prilagođenih vrijednosti. Zbog toga je kod ovakih podataka potrebno koristiti nešto drugačiji pristup. Nije odmah vidljivo kako rijetkosti podataka utječe na Pearsonovu χ^2 statistiku. Dodatno pretpostavimo da su $\pi_i=\pi$, za svaki i. Tada je $\hat{\pi}=\bar{y}$ te se (2.37) svodi na:

$$X_P^2=\sum_{i=1}^{k}\frac{(y_i-\bar{y})^2}{\bar{y}(1-\bar{y})}=k,$$

što također nije dobra mjera prilagodbe modela podacima.

2.7.4 Hosmer-Lemeshowova statistika

Hosmer-Lemeshowova statistika je mjera prilagodbe modela koja se uglavnom koristi kod rijetkih podataka. Kada je $k<n$, potrebno je svaki razred razložiti i promatrati pojedinačne

jedinke.
U slučaju $k = n$ jedan razred čini jedna jedinka pa tada već imamo individualne podatke. Nakon procjene parametara modela dobijemo prilagođene vjerojatnosti $\hat{\pi}_i$ za svaku jedinku. Zatim ih poredamo u rastućem poretku. Tako uređene vjerojatnosti dijelimo u g grupa ($g < k$) približno istih veličina n'_j, za $j = 1, \ldots, g$. Za određivanje grupa mogu se koristiti percentili prilagođenih vjerojatnosti (u svakoj grupi približno n/g jedinki) ili jednostavnija podjela na $g = 10$ poluotvorenih intervala $(a, a + 0.1]$, za $a = 0.0, 0.1, \ldots, 0.9$. Ovakvim grupiranjem u prvoj grupi imamo jedinke s najmanjim prilagođenim vjerojatnostima, odnosno u zadnjoj s najvećim. Primijetimo da ako grupiramo obzirom na procjenjene vjerojatnosti i vrijedi da je $k \approx n$, ali $k < n$, može se dogoditi da dvije jedinke koje početno pripadaju istom razredu i ne budu više u istoj grupi. Za j-tu grupu izračunamo koliko jedinki ima promatrano svojstvo (o_j), a koliko nema ($n'_j - o_j$). Nadalje, računamo očekivani broj jedinki sa svojstvom kao sumu svih prilagođenih vjerojatnosti u toj grupi (e_j) i očekivani broj jedinki bez svojstva ($n'_j - e_j$). Tako dobivene vrijednosti uspoređujemo Pearsonovom χ^2 statistikom:

$$X^2_{HL} = \sum_{j=1}^{g} \frac{(o_j - n'_j \bar{\pi}_j)^2}{n'_j \bar{\pi}_j (1 - \bar{\pi}_j)} \tag{2.41}$$

te ju zovemo Hosmer-Lemeshowova statistika. Pritom je $\bar{\pi}_j = e_j / n'_j$, tj. prosječna prilagođena vjerojatnost za j-tu grupu.
Simulacijskim studijama pokazano je da X^2_{HL} ima približno χ^2 distribuciju s $g - 2$ stupnjeva slobode kada je prilagođen model prikladan. U mnogim istraživanjima pokazano je da je najbolje dijeliti podatke pomoću percentila i to u 10 grupa (engl. *deciles of risk*). Također, očekivane vrijednosti $n'_i \bar{\pi}_i$ moraju biti veće od 5.
Kod specifičnih skupova podataka za različiti broj grupa g može se dogoditi da p-vrijednost upućuje na različite odluke o valjanosti modela. Zbog toga dobivenu p-vrijednost ne smijemo tumačiti prekritično. U praksi se preporuča korištenje dodatnih testova: aproksimativna normalna distribucija Pearsonove χ^2 statistike (Osius i Rojek) i Stukelov test [5].

2.7.5 Generalizirani R^2

Kod klasičnog linearnog modela poznata veličina je koeficijent determinacije R^2 koji izražava koliko dobro prilagođeni model objašnjava varijabilnost varijable odziva Y. Dan je izrazom:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.42}$$

i poprima vrijednosti između 0 i 1. Njegova je vrijednost bliža 1 što je brojnik u (2.42) manji, odnosno model bolje prilagođen. Postoji više oblika generalizacije ove veličine od čega ćemo u ovom radu opisati tri najkorištenije [2].
Prirodna generalizacija je:

$$R_1^2 = 1 - \frac{\log L(\boldsymbol{\pi}(\hat{\boldsymbol{\beta}}))}{\log \hat{L}_0} = 1 - \frac{l(\hat{\boldsymbol{\pi}}; \mathbf{y})}{l(\hat{\boldsymbol{\pi}}_0; \mathbf{y})}, \tag{2.43}$$

gdje je $l(\hat{\boldsymbol{\pi}}; \mathbf{y})$ maksimum log-vjerodostojnosti dobiven uvršavanjem procjenitelja $\hat{\boldsymbol{\beta}}$ i korištenjem relacija $\hat{\pi}_i = (1 + \exp(-\sum_{j=0}^{p} x_{ij}\hat{\beta}_j))^{-1}$, a $l(\hat{\boldsymbol{\pi}}_0; \mathbf{y})$ uvrštavanjem procjenitelja dobivenog za model bez varijabli poticaja (samo konstantni član) preko $\hat{\pi}_i^0 = (1 + e^{-\hat{\beta}_0})^{-1}$, za svaki i. Vrijednosti koje R_1^2 može poprimiti su opet u intervalu $(0, 1)$.
Drugo poopćenje definira se kao:

$$R_2^2 = 1 - \left(\frac{\hat{L}_0}{L(\boldsymbol{\pi}(\hat{\boldsymbol{\beta}}))}\right)^{\frac{2}{n}}, \tag{2.44}$$

pri čemu je n ukupan broj binarnih ishoda. Maksimalna vrijednost koju R_2^2 može poprimiti je $R_{\max}^2 = 1 - (\hat{L}_0)^{\frac{2}{n}} < 1$ pa dodatno možemo promatrati njezinu standardiziranu verziju:

$$R_3^2 = \frac{R_2^2}{R_{\max}^2}. \tag{2.45}$$

2.8 Testiranje hipoteza

Jednom kada prilagodimo model podacima, zanima nas koja je varijabla poticaja značajna. Poznato je da kod običnog linearnog modela u tu svrhu provodimo t-test značajnosti parametara kojim za pojedini parametar testiramo:

$$\begin{aligned} &\mathcal{H}_0 : \beta_j = 0 \\ &\mathcal{H}_1 : \beta_j \neq 0, \end{aligned} \tag{2.46}$$

općenito:

$$\begin{aligned} &\mathcal{H}_0 : \beta_j = \beta_{j0} \\ &\mathcal{H}_1 : \beta_j \neq \beta_{j0}. \end{aligned} \tag{2.47}$$

Navedene hipoteze moguće je poopćiti za više parametara. Na taj način testiramo koji od dva modela je bolji, manji model M sa s prediktora i vektorom parametara $\boldsymbol{\beta}_1$ ili veći model M_l s ukupno l prediktora i vektorom parametara $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$. Testiramo:

$$\begin{aligned} &\mathcal{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0} \\ &\mathcal{H}_1 : \boldsymbol{\beta}_2 \neq \mathbf{0} \end{aligned} \tag{2.48}$$

te kod klasičnog linearnog modela koristimo F-test.
Kod logističkog modela koristimo tri testa za testiranje značajnosti parametara: test omjera vjerodostojnosti, Waldov test i test pogotka. Sva tri testa zasnovana su na vjerodostojnosti. Kada se koriste kod velikih uzoraka, daju vrlo slične rezultate. Nije poznato kako se ponašaju kod malih uzoraka, no neke simulacijske studije ipak sugeriraju korištenje testa omjera vjerodostojnosti naspram ostalih.

2.8.1 Test omjera vjerodostojnosti

Analizirajući statistiku odstupanja, spomenuli smo kako se uspoređuju dva ugniježdena modela M i M_l. Pritom je alternativna hipoteza u (2.36) ekvivalentna $\mathcal{H}_1$ u (2.48). Evaluiranjem log-vjerodostojnosti u prilagođenim vrijednostima $\hat{\boldsymbol{\pi}}_0$, odnosno $\hat{\boldsymbol{\pi}}_1$ i promatranjem njihove razlike dolazi do poništavanja log-vjerodostojnosti saturiranog modela te dobivamo statistiku [7]:

$$D(\mathbf{y}; \hat{\boldsymbol{\pi}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\pi}}_1) = 2l(\hat{\boldsymbol{\pi}}_1; \mathbf{y}) - 2l(\hat{\boldsymbol{\pi}}_0; \mathbf{y}) = -2 \log \frac{L(\hat{\boldsymbol{\pi}}_0; \mathbf{y})}{L(\hat{\boldsymbol{\pi}}_1; \mathbf{y})} \tag{2.49}$$

koja ima asimptotsku χ^2 razdiobu sa stupnjevima slobode jednakim razlici broja parametara uspoređivanih modela pod uvjetom da je $\mathcal{H}_0$ točna.

2.8.2 Waldov test

Prisjetimo se glavnog rezultata iz potpoglavlja 2.6. Uz proširene uvjete regularnosti procjenitelj maksimalne vjerodostojnosti $\hat{\boldsymbol{\beta}}$ ima asimptotsku multivarijatnu normalnu razdiobu s očekivanjem jednakim pravoj vrijednosti $\boldsymbol{\beta}$ i procijenjenom kovarijacijskom matricom $\hat{\Sigma}(\hat{\boldsymbol{\beta}}) = (X\hat{W}X)^{-1}$. Generaliziramo li (2.47) u više dimenzija, za testiranje:

$$\begin{aligned} &\mathcal{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ &\mathcal{H}_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \end{aligned} \tag{2.50}$$

definiramo testnu statistiku [7]:

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{2.51}$$

koju nazivamo Waldova statistika te koja, pod pretpostavkom da vrijedi $\mathcal{H}_0$, ima χ^2 razdiobu s $p + 1$ stupnjem slobode [3, lema 1].
Za usporedbu dvaju ugniježdeni modela opisanih ranije te testiranje hipoteza u (2.48) pripadnu statistiku prilagodimo na način:

$$W = \hat{\boldsymbol{\beta}}_2^T \hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}}_2)\hat{\boldsymbol{\beta}}_2.$$

Dobivena statistika ponovno ima χ^2 razdiobu, ali s $l - s$ stupnjeva slobode.
Ono za što se Waldova statistika ipak najčešće koristi je testiranje hipoteza u (2.46). Tada se za $j = 0, 1, \ldots, p$ promatra njezin korijen:

$$Z = \frac{\hat{\beta}_j}{\sqrt{(\hat{\Sigma}(\hat{\boldsymbol{\beta}}))_{jj}}}$$

koji ima asimptotsku standardnu normalnu razdiobu pod $\mathcal{H}_0$.

2.8.3 Test pogotka

Iz konstrukcije MLE-a $\hat{\boldsymbol{\beta}}$ jasno je da za funkciju pogotka vrijedi $\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. Ako je $\mathcal{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ istinita, za očekivati je da vrijedi $\mathbf{u}(\boldsymbol{\beta}_0) \approx \mathbf{0}$ te je lako dobiti $\mathbb{E}_{\boldsymbol{\beta}_0}[\mathbf{u}(\boldsymbol{\beta}_0)] = \mathbf{0}$.
Nadalje, prisjetimo se da kod našeg modela imamo:

$$\mathbf{u}(\boldsymbol{\beta}) = X^T(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})).$$

Fiksiramo li k te za svaki i pretpostavimo da je $n = n_i$, vrijedi:

$$\mathbf{u}(\boldsymbol{\beta}) = X^T(\mathbf{Y} - n\boldsymbol{\pi}(\boldsymbol{\beta})).$$

Tada sličnim zaključivanjem kao u potpoglavlju 2.6.1 dobivamo:

$$\mathbf{u}(\boldsymbol{\beta}_0) \sim AN(\mathbf{0}, I(\boldsymbol{\beta}_0)).$$

Za testiranje hipoteza u (2.50) koristimo statistiku:

$$Q = \mathbf{u}(\boldsymbol{\beta}_0)^T I^{-1}(\boldsymbol{\beta}_0)\, \mathbf{u}(\boldsymbol{\beta}_0)$$

koja, pod pretpostavkom $\mathcal{H}_0$, ima asimptotsku χ^2 razdiobu s $p+1$ stupnjem slobode. Ovako definirana statistika zapravo mjeri koliko je vrijednost funkcije pogotka daleko od 0 za vrijednost vektora parametara koja je zadana nultom hipotezom.

2.8.4 Pouzdani intervali

U višedimenzionalnom slučaju određivanju pouzdanih intervala možemo pristupiti na dva načina: poopćenjem pojma pouzdanog intervala na pouzdano područje ili promatranjem pouzdanog intervala pojedinog parametra β_j. U ovom radu ćemo opisati drugi pristup i to iz dviju perspektiva: omjera vjerodostojnosti i asimptotske normalnosti procijenjenih parametara iz koje proizlazi i Waldova statistika. Oba pristupa utemeljena su na asimptotskim rezultatima te su zbog toga pogodni za velike uzorke. S druge strane, kod malih uzoraka intervali dobiveni iz omjera vjerodostojnosti pokazuju se točnijima.

Waldovi pouzdani intervali

Za pojedini $j = 0, 1, \ldots, p$ i za $\alpha \in (0, 1)$ vrijedi:

$$\mathbb{P}(-z_{\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}})_{jj}}} \leq z_{\alpha/2}) \approx 1 - \alpha,$$

gdje je $z_{\frac{\alpha}{2}}$ $(1 - \frac{\alpha}{2})$-kvantil standardne normalne razdiobe. Iz toga dobivamo da je Waldov $(1 - \alpha) \cdot 100\%$ pouzdani interval dan s:

$$\left(\hat{\beta}_j - z_{\frac{\alpha}{2}} \sqrt{(\hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}}))_{jj}}, \hat{\beta}_j + z_{\frac{\alpha}{2}} \sqrt{(\hat{\Sigma}^{-1}(\hat{\boldsymbol{\beta}}))_{jj}}\right).$$

Pouzdani intervali utemeljeni na omjeru vjerodostojnosti

Za fiksni $j = 0, 1, \ldots, p$ definiramo funkciju vjerodostojnosti profila za $\beta_j = \gamma$:

$$l_j^*(\gamma) = \max_{\boldsymbol{\beta} \in \mathcal{B}_j(\gamma)} l(\boldsymbol{\beta})$$

pri čemu je $\mathcal{B}_j(\gamma)$ skup vektora $\boldsymbol{\beta}$ takvih da je $\boldsymbol{\beta}^T = [\beta_0, \ldots, \beta_{j-1}, \gamma, \beta_{j+1}, \ldots, \beta_p]$. Za svaku γ vrijednost $l_j^*(\gamma)$ je maksimum funkcije log-vjerodostojnosti preostalih nefiksiranih parametara. Testiramo:

$$\mathcal{H}_0 : \gamma = \gamma_0$$
$$\mathcal{H}_1 : \gamma \neq \gamma_0$$

statistikom iz (2.49). Pritom pretpostavimo da je veći model M_l s ukupno $p+1$ parametrom, a manji s p (j-ti je određen pretpostavkom $\mathcal{H}_0$). Označimo s $l_{max} = l(\hat{\boldsymbol{\beta}})$ gdje je $\hat{\boldsymbol{\beta}}$ MLE većeg modela. Tada statistika u (2.49) postaje:

$$-2\,(\,l_j^*(\gamma_0) - l_{max})$$

i ima asimpototsku χ^2 razdiobu s jednim stupnjem slobode. Nadalje, tražimo vrijednosti γ_0 tako da hipoteza $\mathcal{H}_0$ ne će biti odbačena na razini značajnosti $\alpha \in (0, 1)$, odnosno da vrijedi:

$$\mathbb{P}\Big(-2\,(l_j^*(\gamma_0) - l_{max}) \leq \chi^2_{1-\alpha}(1)\Big) \approx 1 - \alpha.$$

Ako označimo $l_0 = l_{max} - \frac{1}{2}\chi^2_{1-\alpha}(1)$, tada je $(1 - \alpha) \cdot 100\%$ pouzdani interval skup

$$\{\gamma : l_j^*(\gamma) \geq l_0\}.$$

Pouzdani interval može se očitati s grafa funkcije $l_j^*(\gamma)$ tako da se odrede γ za koje je graf iznad vrijednosti l_0 ili iterativnom metodom za rješavanje $l_j^*(\gamma) = l_0$ kojom dobijemo rubove pouzdanog intervala [1].

Pouzdani intervali parametara u logističkom modelu mogu se iskorisiti za dobivanje pouzdanih intervala izglednosti za jediničnu promjenu, odnosno promjenu $c > 0$, j-te varijable poticaja. Označimo li s D_j donju granicu dobivenog intervala, a s G_j gornju granicu, pripadni pouzdani intervali za omjer izglednosti su:

$$\frac{\omega(\pi_i(x_{ij}+1))}{\omega(\pi_i(x_{ij}))} \in \left(e^{D_j}, e^{G_j}\right),$$

odnosno:

$$\frac{\omega(\pi_i(x_{ij}+c))}{\omega(\pi_i(x_{ij}))} \in \left(e^{cD_j}, e^{cG_j}\right)$$

te se smatraju značajnima ako ne sadrže 1.

2.9 Dijagnostika

Završna provjera modela temelji se na razlici opaženih i prilagođenih vrijednosti, rezidualima. Za razliku od klasičnog linearnog modela kod logističkog modela u obzir je potrebno uzeti činjenicu nejednakih varijanci varijabli $Y_i \sim B(n_i, \pi_i)$. Najčešće se analiziraju dva tipa reziduala koji proizlaze iz već poznatih statistika prilagodbe modela podacima te njihove standardizirane verzije.

2.9.1 Reziduali

Pearsonovi reziduali

Najjednostavniji pristup računanja reziduala je promatranje razlike opaženih i prilagođenih vrijednosti $y_i - \hat{y}_i$ koje nazivamo osnovnim rezidualima. Budući da svaka od varijabli Y_i ima binomnu razdiobu s različitim brojem ishoda n_i i vjerojatnošću π_i, takvi se "sirovi" reziduali teško interpretiraju. U suštini, velika razlika između y_i i $\hat{y}_i$ je manje važna u slučaju kad je standardna greška dotične opservacije velika. Zato međusobno usporedive reziduale dobivamo dijeljenjem s procijenjenom standardnom greškom opaženih vrijednosti:

$$p_i = \frac{y_i - \hat{y}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}. \tag{2.52}$$

Izraz u nazivniku slijedi iz činjenice da je $Var(Y_i) = n_i\pi_i(1-\pi_i)$. Takve reziduale nazivamo Pearsonovim rezidualima jer kvadrirani i sumirani daju Pearsonovu statistiku (2.37). Dakle, Pearsonov rezidual pojedine opservacije je veličina doprinosa mjeri prilagodbe modela.

Reziduali odstupanja

Reziduali odstupanja definiraju se:

$$d_i = \text{sgn}\,(y_i - \hat{y}_i)\sqrt{2\left(y_i \log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right)} \tag{2.53}$$

pri čemu je d_i pozitivan kada je $y_i \geq \hat{y}_i$, odnosno negativan kada je $y_i < \hat{y}_i$. Također, njihovim kvadriranjem i sumiranjem dobivamo statistiku odstupanja (2.34).

Studentizirani reziduali

Do sada definirani reziduali nisu standardizirani. Njihove definicije uzimaju u obzir da Y_i imaju međusobno različite varijance, međutim ne objašnjavaju varijabilnost koja nastaje zbog procjene parametara, kao što je to slučaj kod studentiziranih reziduala linearnog modela. Prisjetimo se, za linearni model definira se matrica težina (engl. *hat matrix*) s $H = X(X^TX)^{-1}X^T$ te vrijedi $\hat{\mathbf{y}} = H\mathbf{y}$. Za logistički model matricu H definiramo na način:

$$H = W^{\frac{1}{2}}X(X^TWX)^{-1}X^TW^{\frac{1}{2}} \tag{2.54}$$

gdje je $W = diag\{n_i\pi_i(1-\pi_i)\}$ matrica iz iterativnog postupka dobivanja MLE-a. Matrica H je simetrična i idempotentna te je zbog toga linearni projektor. Motivirani linearnim modelom možemo ju shvatiti kao projektor koji k binomnih realizacija projicira na prostor varijabli poticaja. Vrijednost h_{ii} mjeri koliko opažena vrijednost y_i utječe na dobivanje procijenjene $\hat{y}_i$. Pravilo heuristike nalaže da se one opservacije za koje je $h_{ii} > \frac{2(p+1)}{k}$ smatrajaju jako utjecajnim točkama. Zbog idempotentnosti i simetričnosti vrijedi:

$$h_{ii} = \sum_{i\neq j} h_{ij}^2 + h_{ii}^2 \Rightarrow h_{ii} > h_{ii}^2$$

iz čega zaključujemo da su svi dijagonalni elementi iz intervala $(0, 1)$. Nadalje, računamo sumu dijagonalnih elemenata korištenjem činjenice da je $tr(AB) = tr(BA)$:

$$\text{tr}\,(H) = \text{tr}\,(W^{\frac{1}{2}}X\cdot(X^TWX)^{-1}X^TW^{\frac{1}{2}}) = \text{tr}\,((X^TWX)^{-1}X^TW^{\frac{1}{2}}\cdot W^{\frac{1}{2}}X) = \text{tr}\,(I_{p+1}) = p+1.$$

Slijedi da su prosječne vrijednosti dijagonalnih elemenata $(p+1)/k$ pa su h_{ii} uglavnom mali za dobre modele kod kojih je $p + 1 << k$. Može se pokazati da je standardna greška "sirovih" reziduala [2]:

$$\text{se}(y_i - \hat{y}_i) = \sqrt{(1-h_{ii})\,n_i\,\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Sukladno tome, Pearsonovi standardizirani reziduali se definiraju kao:

$$r_{Pi} = \frac{p_i}{\sqrt{1-h_{ii}}} = \frac{y_i - \hat{y}_i}{\sqrt{(1-h_{ii})n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \tag{2.55}$$

a standardizirani reziduali odstupanja:

$$r_{Di} = \frac{d_i}{\sqrt{1-h_{ii}}}. \tag{2.56}$$

Reziduali vjerodostojnosti

Još jedan tip reziduala može se dobiti usporedbom statistike odstupanja dobivene prilagodbom modela na cijeli skup od k binomnih opservacija i odstupanja dobivenog prilagodbom istog modela na skup od $k-1$ opservacija dobivenog izostavljanjem i-te, za $i = 1, 2, \ldots, k$. Tim postupkom dobijemo egzaktne vrijednosti. Dakako, taj postupak je računarski intenzivan pa se za razliku statistike odstupanja koristi aproksimacija [2]:

$$h_{ii}r_{Pi}^2 + (1-h_{ii})r_{Di}^2.$$

Kako su vrijednosti r_{Pi}, r_{Di} i h_{ii} dobivene prigodbom modela na svih k opservacija, ovakvom aproksimacijom na jednostavan način izbjegavamo dodatnih k prilagodbi. Definiramo reziduale vjerodostojnosti kao korijen konveksne kombinacije poznatih standardiziranih reziduala:

$$r_{Li} = \operatorname{sgn}(y_i - \hat{y}_i)\sqrt{h_{ii}r_{Pi}^2 + (1-h_{ii})r_{Di}^2}. \tag{2.57}$$

Budući da su dijagonalni elementi matrice H mali, vrijednosti r_{Li} bit će slične r_{Di}.

Anscombeovi reziduali

Za razliku od linearnog modela gdje znamo da su reziduali normalno distribuirani, kod logističkog modela egzaktna distribucija dosad opisanih nije poznata. Zbog toga postoji još jedan pristup kojem je cilj naći funkciju A koja će binomne vrijednosti transformirati u vrijednosti s približno normalnom razdiobom. Tada se prikladni standardizirani reziduali definiraju kao [2]:

$$r_{Ai} = \frac{A(y_i) - A(\hat{y}_i)}{\operatorname{se}\{A(y_i) - A(\hat{y}_i)\}} \tag{2.58}$$

gdje je $\operatorname{se}\{A(y_i) - A(\hat{y}_i)\}$ standardna greška. Takvi reziduali nazivaju se Anscombeovi reziduali. Prikladna funkcija za binomne ishode je:

$$A(u) = \int_0^{u/n_i} t^{-1/3}(1-t)^{-1/3}dt, \quad 0 \le u \le n_i,$$

a standardna greška $(\hat{\pi}_i(1-\hat{\pi}_i))^{1/6}\sqrt{(1-h_{ii})/n_i}$. Vrijednost funkcije A računa se pomoću generalizirane (nepotpune) beta funkcije:

$$I_z(\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\int_0^z t^{\alpha-1}(1-t)^{\beta-1}dt$$

iz izraza $A(u) = B(\frac{2}{3},\frac{2}{3})I_{u/n_i}(\frac{2}{3},\frac{2}{3})$.

U praksi se pokazalo da su vrijednosti standardiziranih reziduala odstupanja, reziduala vjerodostojnosti i Anscombeovih reziduala vrlo slične. Mnoge studije pokazuju da se oni mogu vrlo dobro aproksimirati standardnom normalnom razdiobom u slučajevima kada vrijednosti n_i nisu premale. Zbog takve aproksimacije reziduali će se uglavnom nalaziti između -2 i 2 ako je model zadovoljavajući. S druge strane, standardizirani Pearsonovi reziduali ne prate tako dobro standardnu normalnu razdiobu pa za $|y_i - \hat{y}_i| > 2$ njihove vrijednost odstupaju od ostalih. Zaključak koji se na kraju nameće je taj da je preporučljivo koristiti standardizirane reziduale odstupanja r_{Di} i reziduale vjerodostojnosti r_{Li}.

2.9.2 Grafički prikazi reziduala

Popis reziduala po opservacijama je vrlo koristan u analizi modela, no njihovi grafički prikazi su reprezentativniji i sažetije prikazuju kvalitetu prilagođenog modela. Najjednostavniji grafički prikaz je prikaz reziduala po opservacijama, odnosno njihovom indeksu. Takav je prikaz namjenjen detektiranju odstupajućih opservacija (engl. *outliers*) koje imaju neobično velike reziduale. Druga vrsta grafičkog prikaza je prikaz reziduala naprema vrijednostima linearnog prediktora $\hat{\eta}_i = \sum_{j=0}^{p} x_{ij}\hat{\beta}_j$. Pojavljivanje pravilnosti na grafu (točke lako interpolirane polinomom) ukazuje na nedovoljno dobar model. Koristan je i prikaz reziduala naprema vrijednostima pojedine varijable poticaja. Prisutnost trenda u prikazu reziduala naprema varijabli poticaja koja nije u modelu upućuje na moguće poboljšavanje modela ukoliko se ona uključi.
Prisjetimo se, reziduali r_{Di} i r_{Li} imaju približnu standardnu normalnu razdiobu pa je očekivano da informaciju o adekvatnosti modela sadržava i normalni vjerotnostni graf. Iako se on u osnovi koristi za ispitivanje pripadnosti normalnoj distribuciji, i na njemu je moguće prepoznati odstupajuće vrijednosti ili neželjene anomalije prilagodbe. Ipak, te karakteristike bolje se izražavaju kod tzv. polunormalnog vjerojatnostnog grafa kod kojeg radimo prikaz reziduala u rastućem poretku naprema kvantila $\Phi^{-1}\{(i+n-\frac{3}{8})/(2n+\frac{1}{2})\}$ (kod normalnog vjerojatnostnog grafa računamo $\Phi^{-1}\{(i-\frac{3}{8})/(n+\frac{1}{4})\}$). Odstupajuće vrijednosti se pojavljuju desno na vrhu grafa. U nekim se primjerima može dogoditi da takvim prikazom ne dobivamo približno ravnu liniju iako je model dobar. Zato se na prikazu dodatno konstruira simulirana pruga grafa (engl. *envelope*) unutar koje u tom slučaju upadaju sve točke

prikaza.
Pretpostavimo da simuliramo prugu polunormalnog grafa standardiziranih reziduala r_{Di}. Za svaku od k opservacija simuliramo dodatnih 19 opservacija iz $B(n_i, \hat{\pi}_i)$. Tada model koji smo prilagodili originalnim podacima prilagodimo simuliranim vrijednostima i za svaku prilagodbu računamo apsolutne vrijednosti reziduala $|r_{Di}|$. Zatim poredamo te vrijednosti i dobivamo uređajne statistike $|r_D|_{(i)}$ takve da vrijedi $|r_D|_{(1)} < |r_D|_{(2)} < \cdots < |r_D|_{(k)}$. Potom računamo aritmetičku sredinu, minimum i maksimum vrijednosti $|r_D|_{(i)}$ po svih 19 simuliranih skupova, za $i = 1, 2, \ldots, k$, te ih dodamo početnom grafu. Minimumi i maksimumi određuju prugu. Tako dobijemo kriterij kojim neku opservaciju s veliki rezidualom proglasimo odstupajućom vrijednošću. Također, ako točke odstupaju od sredina simuliranih vrijednosti ili ih je mnogo izvan pruge, model nije prikladan. Naravno, podrobniju analizu možemo dobiti većim brojem simulacija.
Svi opisani grafički prikazi su poprilično jednostavni i daju nam samo osnovnu informaciju o valjanosti modela. Postoje sofisticiranije grafičke metode koje dublje zaziru u samu strukturu linearnog prediktora [2].

2.9.3 Rijetkost podataka

Budući da dva osnovna tipa reziduala potječu od mjera prilagodbe modela podacima, slučaj $n_i = 1$ za svaki i potrebno je ponovno zasebno komentirati. Jednostavnim sređivanjem izraza u (2.52) i (2.53) dobivamo Pearsonove reziduale:

$$p_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}},$$

odnosno reziduale odstupanja:

$$d_i = \operatorname{sgn}(y_i - \hat{\pi}_i)\sqrt{-2[y_i \log \hat{\pi}_i + (1 - y_i)\log(1 - \hat{\pi}_i)]}.$$

Oba tipa reziduala, kao i njihove standardizirane verzije, poprimaju pozitivne vrijednosti kad je $y_i = 1$, odnosno negativne kad je $y_i = 0$. To znači da distribucija takvih reziduala ne može imati približno normalnu razdiobu. Mnogi prethodno opisani grafički prikazi imaju svojstva koja su rezultat takve prirode podataka. Na primjer, u prikazu reziduala naprema vrijednostima linearnog prediktora podaci se odvajaju u dvije putanje što je nemoguće interpretirati u smislu ispravnosti modela. S druge strane, običan graf reziduala po opservacijama i polunormalan graf zajedno sa simuliranom prugom lijepo odvajaju odstupajuće vrijednosti od ostalih.
Kod ostalih naprednijih grafičkih prikaza dolazi do grupiranja podataka u dvije skupine što rezultira izostankom valjane interpretacije. Ono što se tada najčešće koristi kako bi se opisala struktura podataka jesu metode zaglađivanja kao što su lokalna težinska regresija [2] i zaglađivanje splajnom.

2.9.4 ROC analiza

ROC analiza (engl. *Receiver Operator Characteristic*) razvila se kao dio teorije detekcije signala te se prvotno koristila u proučavanju radarskih signala. Utemeljena je na ROC krivulji kojom se ispitivala moć prijamnika da ispravno odvoji radarski signal od šuma. Tijekom 70-ih godina prošlog stoljeća ROC analiza zauzima važnu ulogu u medicini gdje se i danas intenzivno koristi za ispitivanje valjanosti dijagnostičkih testova.

ROC krivulja

Promatramo realizacije varijable odziva Y te one jednake 1 nazovimo uspjehom, odnosno 0 neuspjehom. Kod logističke regresije model prilagođavamo podacima radi dobivanja prediktivnog modela koji će dati dobru procjenu vjerojatnosti uspjeha nove jedinke s obzirom na njezine vrijednosti kovarijata. Na temelju prilagodbe određujemo kolika je njegova moć predikcije tako da uspoređujemo procijenjene vjerojatnosti uspjeha i postojeće binarne ishode promatrane varijable odziva. Većina statističkih alata navedenu usporedbu izvodi na skupu podataka na kojem je model prilagođen (engl. *resubstitution*). Pri tome su mjere asocijacije procijenjenih vjerojatnosti i opaženih realizacija djelomično pristrane. Radi dobivanja nepristranih mjera koristimo naprednije metode kao što su metoda validacije (engl. *split-sample validation*) i unakrsne validacije te metoda *bootstrap* [4]. Tako tražene mjere asocijacije dobivamo na temelju podataka koji su nezavisni od podataka na kojima je model prilagođen. Nadalje, za procijenjene vjerojatnosti određujemo graničnu vrijednost (engl. *cut-off*). Sve vjerojatnosti veće od granične vrijednosti kodiramo jedinicom (uspjeh), a manje nulom (neuspjeh). Za graničnu vrijednost najčešće se uzima 0.5. Od velikog značaja su opservacije kod kojih model griješi, odnosno opservacije koje smo na temelju modela pogrešno proglasili uspjehom (engl. *false positive*, FP) i koje smo pogrešno proglasili neuspjehom (engl. *false negative*, FN). One određuju osjetljivost i specifičnost modela. Osjetljivost modela definira se kao:

$$p_{\text{os}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

što odgovara vjerojatnosti da model ispravno klasificira uspjehe (engl. *true positive*, TP). Specifičnost modela definira se kao:

$$p_{\text{sp}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

što odgovara vjerojatnosti da model ispravno klasificira neuspjehe (engl. *true negative*, TN). Što su brojevi FN i FP manji, to su osjetljivost i specifičnost veće te je model bolji.

	opaženi ishod		
ishod na temelju modela	1	0	
1	TP	FP	$\frac{\text{TP}}{\text{TP+FP}}$
0	FN	TN	$\frac{\text{TN}}{\text{FN+TN}}$
	$\frac{\text{TP}}{\text{TP+FN}}$	$\frac{\text{TN}}{\text{TN+FP}}$	$\frac{\text{TP+TN}}{\text{UKUPNO}}$

Tablica 2.2: Tablica mjera valjanosti modela.

Promotrimo tablicu 2.2. Prepoznajemo izraze za osjetljivost i specifičnost u zadnjem retku tablice. Primijetimo da nazivnici navedenih omjera ne ovise o iznosu granične vrijednosti te ih možemo lakše interpretirati nego ostale mjere valjanosti prikazane u zadnjem stupcu tablice. To su redom: pozitivna prediktivna vrijednost (preciznost), negativna prediktivna vrijednost i točnost modela. Uz točnost modela veže se i greška modela koja se definira kao 1−točnost.
Omjer koji odgovara osjetljivosti modela poznat je još pod nazivom omjer ispravno klasificiranih uspjeha (engl. *true positive rate*):

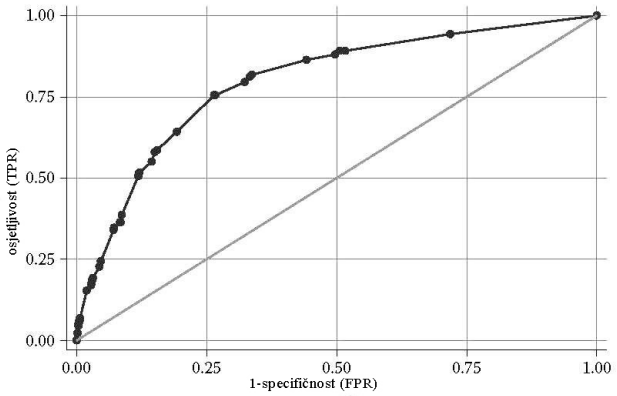
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

a 1−specifičnost naziva se omjer neispravno klasificiranih uspjeha (engl. *false positive rate*):

$$\text{FPR} = 1 - \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{FP}}{\text{TN} + \text{FP}}.$$

ROC krivulja je grafički prikaz omjera TPR naprema FPR za svaki iznos granične vrijednosti iz $(0, 1)$. Površina ispod ROC krivulje je mjera prediktivne moći modela i naziva se Harrellova C statistika. Njezinu realizaciju označavamo s c. Može poprimiti vrijednosti od 0 do 1 te je pokazatelj koliko dobro model diskriminira jedinke za koje je realizacija promatrane varijable odziva 1, odnosno one za koje je 0. Ukoliko je $c = 0.5$, nema diskriminacije. Ta površina odgovara pravcu nagiba 1 te se dobiva za slučajno generirane vjerojatnosti (slučajan model). Vrijednosti $c \in [0.7, 0.8)$ smatramo prihvatljivom, a $c \in [0.8, 0.9)$ izuzetno dobrom diskriminacijom. Vrijednosti iznad 0.9 rijetko dobivamo u praksi. Ponekad ne promatramo cijelu površinu ispod ROC krivulje, već analiziramo udaljenost ROC krivulje od već spomenute ROC krivulje slučajnog modela (slika 2.2). Što je

ta udaljenost veća, to je pripadni model bolji.



Slika 2.2: Primjer empirijske ROC krivulje [2].

Površinu ispod ROC krivulje možemo dobiti i na sljedeći način. Označimo s N_1 broj uspjeha, odnosno s N_0 broj neuspjeha. Sparimo po jednu opservaciju iz svake skupine. Tako dobivamo $N_1 \cdot N_0$ različitih parova. Par u kojem jedinka s ishodom 1 ima veću od dviju promatranih vjerojatnosti nazivamo suglasni par (engl. *concordant*). Parovi u kojima jedinka s ishodom 1 ima manju vjerojatnost nazivaju se nesuglasni (engl. *discordant*), a kada su dvije vjerojatnosti jednake vezani parovi (engl. *tied*). Tada je površina ispod ROC krivulje jednaka:

$$c = \frac{n_{\text{suglasni}} + 0.5\, n_{\text{vezani}}}{N_1 N_0}$$

što odgovara udjelu suglasnih parova u ukupnom broju parova. Pritom pola vezanih parova smatramo suglasnim, a pola nesuglasnim parovima. ROC krivulju koju dobijemo iz podataka na neki od opisanih načina nazivamo empirijska ROC krivulja.
Koristeći takve skupine parova, definiramo još neke poznate statistike kao što su Somersova D statistika, Goodman-Kruskalova γ i Kendallov koeficijent τ koje su mjere asocijacije procijenjenih vjerojatnosti i opaženih ishoda varijable odziva.

Zaglađena ROC krivulja

Kod ispitivanja valjanosti dijagnostičkih testova često se analizira utjecaj samo jednog neprekidnog prediktora X na ishod varijable odziva Y. U tom se slučaju empirijska ROC krivulja aproksimira glatkom funkcijom koju nazivamo zaglađena ROC krivulja. Pritom se najčešće koristi pristup koji uključuje pretpostavku binormalnosti. Promatramo zasebno

distribuciju neprekidnog prediktora kod skupa opservacija za koje je $Y = 0$ te distribuciju prediktora za opservacije kod kojih je $Y = 1$. Tada se za fiksnu graničnu vrijednost t omjeri dobivaju na način:

$$\text{FPR} = 1 - \mathbb{P}(X \leq t \,|\, Y = 0) = \int_t^\infty f(s \,|\, Y = 0)\, ds$$
$$\text{TPR} = \mathbb{P}(X > t \,|\, Y = 1) = \int_t^\infty f(s \,|\, Y = 1)\, ds$$

gdje su $f(s \,|\, Y = 0)$ i $f(s \,|\, Y = 1)$ uvjetne gustoće prediktora. Nadalje, pretpostavimo da prediktor (ili njegova monotona transformacija) na skupu opservacija gdje je $Y = 0$ ima distribuciju $N(\mu_0, \sigma_0^2)$, a na skupu gdje je $Y = 1$ distribuciju $N(\mu_1, \sigma_1^2)$. U praksi će se rijetko dogoditi da je prediktor normalno distribuiran. Zbog invarijantnosti ROC krivulje na monotone transformacije za zadovoljavanje pretpostavke normalnosti koristi se monotona transformacija prediktora [4]. Slijedi:

$$\text{FPR} = 1 - \Phi\left(\frac{t - \mu_0}{\sigma_0}\right) = \Phi\left(\frac{\mu_0 - t}{\sigma_0}\right), \tag{2.59}$$

$$\text{TPR} = 1 - \Phi\left(\frac{t - \mu_1}{\sigma_1}\right) = \Phi\left(\frac{\mu_1 - t}{\sigma_1}\right). \tag{2.60}$$

Označimo s x neku vrijednost omjera FPR iz intervala $(0, 1)$. Tada iz (2.59) dobivamo:

$$t = \mu_0 - \sigma_0 \Phi^{-1}(x).$$

Izraz u (2.60) postaje:

$$\Phi\left(\frac{\mu_1 - t}{\sigma_1}\right) = \Phi\left(\frac{\mu_1 - \mu_0}{\sigma_1} + \frac{\sigma_0}{\sigma_1}\, \Phi^{-1}(x)\right)$$

te se zaglađena ROC krivulja dobiva crtanjem točaka $(x, \Phi\,(a + b\Phi^{-1}(x))$ za $a = (\mu_1 - \mu_0)/\sigma_1$ i $b = \sigma_0/\sigma_1$. Vrijednost c može se dobiti u zatvorenoj formi [10]:

$$c = \int_0^1 \Phi[a + b\,\Phi^{-1}(x)]dx = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right).$$

Poglavlje 3

Primjeri

Na primjerima ćemo demonstrirati prilagodbu logističkog modela, usporediti procjene parametara dobivene maksimizacijom log-vjerodostojnosti i minimizacijom χ^2 statistike te analizirati empirijsku i zaglađenu ROC krivulju. Podaci u primjerima preuzeti su iz medicine i agronomije. Za sprovođenje primjera koristili smo softvere SAS i MATLAB.

3.1 Primjer 1

[2] Hoblyn i Palmer (1934) su istraživali vegetativnu reprodukciju podloga šljive na temelju reznica koje su uzete od korijenja starijih stabala. Reznice su uzimali s vrste *Common mussel* u razdoblju od rujna 1931. do veljače 1932. Pola reznica zasadili su odmah nakon rezidbe, a drugu polovicu čuvali u pijesku te zasadili u proljeće. Uzimali su reznice duljine 12 cm i 6 cm. Za svaku od 4 kovarijatna razreda uzeli su 240 reznica te u rujnu 1932. promatrali koja od reznica je uspješno izrasla u podlogu. Podaci su dani tablicom 3.1.

duljina	vrijeme	uspjelo	neuspjelo
kratke	odmah	107	133
	u proljeće	31	209
duge	odmah	156	84
	u proljeće	84	156

Tablica 3.1: Tablica podataka.

Za prilagodbu logističkog modela koristili smo program SAS i proceduru LOGISTIC. SAS koristi iterativnu Fisherovu metodu pogađanja za dobivanje procjenitelja maksimalne

vjerodostojnosti te u izlaznoj datoteci procedure vidimo da je metoda iskonvergirala. Za dobivanje procjenitelja minimalne χ^2 statistike koristili smo MATLAB u kojem smo metodom najmanjih kvadrata minimizirali χ^2 statistiku iz potpoglavlja 2.6.2. Prilagodili smo model:

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_{vrijeme} + \hat{\beta}_2 \cdot x_{duljina}$$

gdje je $\hat{\pi}$ procijenjena vjerojatnost da reznica uspješno izraste u podlogu. Dobiveni procjenitelji su dani u tablici 3.2.

procjenitelj \metoda	MLE	χ^2
$\hat{\beta}_0$	-0.3039	-0.3019
$\hat{\beta}_1$	-1.4275	-1.4213
$\hat{\beta}_2$	1.0177	1.0125

Tablica 3.2: Tablica procijenjenih parametara.

Zaključujemo da su procjenitelji približno jednaki. Naime, radi se o povećem uzorku i jednakom broju reznica u svakom od kovarijatnih razreda. Na temelju Waldovog testa značajnosti parametara svi parametri su značajni na razini značajnosti od 1% ($p < 0.01$). U tablici 3.3 možemo vidjeti Waldove 95%-tne pouzdane intervale i pouzdane intervala dobivene na temelju omjera vjerodostojnosti (engl. *profile-likelihood confidence intervals*). Zbog velikog uzorka oni su približno jednaki.

	procjenitelj	Waldov 95%-tni p. i.		PL 95%-tni p. i.	
$\hat{\beta}_0$	-0.3039	-0.5336	-0.0743	-0.5350	-0.0753
$\hat{\beta}_1$	-1.4275	-1.7146	-1.1405	-1.7178	-1.1433
$\hat{\beta}_2$	1.0177	0.7325	1.3028	0.7347	1.3053

Tablica 3.3: Tablica pouzdanih intervala procijenjenih parametara.

Iz prethodne tablice lagano dobivamo tablicu omjera izglednosti. Pritom je važno spomenuti način na koji smo kodirali kovarijate. Vrijeme sadnje reznice odmah nakon rezidbe definirali smo kao referentnu vrijednost i kodirali s 0, a sadnju u proljeće s 1. Analogno, ako je u pitanju kratka reznica, kodirali smo s 0, a ako je dulja, onda s 1. Pripadni omjeri izglednosti dani su tablicom 3.4.

kovarijata	procjena	Waldov 95%-tni p. i.	
vrijeme 1 vs 0	0.240	0.180	0.320
duljina 1 vs 0	2.767	2.080	3.680

Tablica 3.4: Tablica omjera izglednosti.

Iz tablice vidimo da su izgledi da reznica uspije 1/0.24=4.166 puta veći ako ju sadimo odmah kad je odrezana, nego da sa sadnjom čekamo do proljeća. Isto tako, ako sadimo dulju reznicu, izglednost da uspije je 2.767 puta veća nego izglednost uspjeha kraće reznice. Promotrimo li prikazane pouzadane intervale, u oba slučaja oni ne sadrže jedinicu pa su opisane razlike u izglednostima značajne na razini značajnosti od 5%.

3.2 Primjer 2

[10] U ovom primjeru podaci su dobiveni mjerenjem aktivnosti izoenzima CK-BB u cerebrospinalnoj tekućini kod pacijenata unutar 24 sata od teške ozlijede glave. Uzorak čini 60 pacijenata od kojih se nakon ozlijede 19 djelomično ili potpuno oporavilo, a 41 se slabo oporavilo ili uopće nije. Zanima nas je li CK-BB izoenzim dobar prediktor oporavka nakon teške traume na glavi. Osim CK-BB izoenzima poznate su nam i godine pacijenata.

neuspješan oporavak						uspješan oporavak			
dob	CKBB	dob	CKBB	dob	CKBB	dob	CKBB	dob	CKBB
4	140	19	303	29	156	6	136	24	253
7	1087	19	193	30	356	6	286	28	70
8	230	20	76	40	350	7	281	35	40
11	183	20	1370	41	323	8	23	38	6
15	1256	20	543	45	1560	8	200	46	46
16	700	20	913	45	120	10	146		
16	16	20	230	50	216	11	220		
16	800	21	463	51	443	12	96		
17	253	22	60	56	523	12	100		
18	740	23	509	59	76	16	60		
18	126	23	576	61	303	17	17		
18	153	24	671	61	353	18	27		
19	283	29	80	62	206	18	126		
19	90	29	490			19	100		

Tablica 3.5: Tablica podataka.

Varijable *dob* i *CKBB* su neprekidne kovarijate pa je prema tome za očekivati da imamo mali broj podataka u svakom od kovarijatnih razreda. Ovo je tipičan primjer rijetkosti podataka ($n_i = 1$, za svaki i) zbog čega je nemoguće dobiti procjenitelje minimizacijom χ^2 statistike. Ponovno, Fisherovom metodom pogađanja dobivamo procjenitelje modela:

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = 2.6489 - 0.0648 \cdot x_{dob} - 0.0101 \cdot x_{CKBB}. \tag{3.1}$$

Ispuštanjem varijable *dob* dobivamo model:

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = 1.1354 - 0.00935 \cdot x_{CKBB}. \tag{3.2}$$

U oba modela $\hat{\pi}$ je procijenjena vjerojatnost uspješnog oporavka. Nadalje, $-2\log L$ kod modela u (3.2) iznosi 54.138, a kod modela u (3.1) 47.938. Njihova razlika je $6.2 > \chi^2_{0.05}(1) = 3.8415$ pa zaključujemo da je dodavanje varijable *dob* značajno poboljšanje modela. U nastavku promatramo model u (3.1). Usporedbom pouzdanih intervala parametara modela vidimo da se oni neznatno razlikuju, a kako se radi o manjem uzorku, boljima se smatraju pouzdani intervali utemeljeni na omjeru vjerodostojnosti.

	procjenitelj	Waldov 95%-tni p. i.		PL 95%-tni p. i.	
$\hat{\beta}_0$	2.6489	0.7310	4.5667	0.9487	4.8771
$\hat{\beta}_1$	-0.0648	-0.1246	-0.0050	-0.1346	-0.0123
$\hat{\beta}_2$	-0.0101	-0.0174	-0.0028	-0.0188	-0.0042

Tablica 3.6: Tablica pouzdanih intervala procijenjenih parametara.

Program SAS kao dio izlazne datoteke ima tablicu statistika koje se tiču asocijacije procijenjenih vjerojatnosti i opaženih vrijednosti varijable odziva, a koje su dobivene na temelju skupa podataka na kojem je model prilagođen. Najvažnija među njima je realizacija Harrellove C statistike koju označavamo s c i koja odgovara površini ispod empirijske ROC krivulje. Osim toga, za model u (3.1) radili smo metodu krosvalidacije. Jedan od pristupa koji smo koristili je krosvalidacija izostavljanjem jedne opservacije (engl. *leave-one-out*) te smo izvodili proceduru LOGISTIC na ostalim podacima i tako po svim opservacijama. Pomoću makro funkcije %roc dostupne u [4] dobivamo c. Drugi pristup koji smo koristili je krosvalidacija na temelju particije skupa podataka na k podskupova (engl. *K-fold cross-validation*). Pritom smo podatke podijelili na k slučajno određenih podskupova te nam je jedan od njih služio kao skup za testiranje, a ostalih $k-1$ kao skup za prilagodbu modela i tako po svim k. Vrijednost c dobili smo pomoću makro funkcije %xval dostupne u [4].

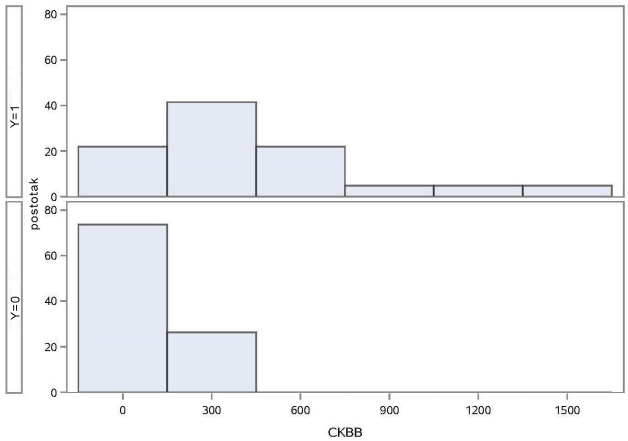
metoda	c
originalni podaci	0.8628
izostavljanje jednog	0.8241
krosvalidacija s k=5	0.8107

Tablica 3.7: Tablica vrijednosti c po metodama.

U tablici 3.7 vidimo da je prva vrijednost najveća i dosta se razlikuje od preostale dvije dobivene metodama krosvalidacije. Zbog njezine djelomične pristranosti bolje je koristiti neku od preostalih dviju metoda za dobivanje mjere moći predikcije logističkog modela.

3.3 Primjer 3

Koristimo podatke iz prethodnog primjera te neuspješan oporavak kodiramo jedinicom, a uspješan nulom. Zanima nas koliku prediktivnu moć ima model u kojem imamo samo jedan neprekidni prediktor, a to je aktivnost izoenzima CK-BB. Osim empirijske ROC krivulje želimo dobiti i zaglađenu ROC krivulju te vrijednosti statistike C.
Procedurom LOGISTIC dobili smo da vrijednost površine ispod empirijske ROC krivulje iznosi $c = 0.8286$. Na temelju histograma na slici 3.1 i Kolmogorov-Smirnovljevog testa normalnosti na razini značajnosti od 5% zaključili smo da aktivnost enzima CK-BB nije normalno distribuirana ni kod skupine pacijenata s uspješnim niti neuspješnim oporavkom.



Slika 3.1: Histogrami CK-BB-a.

Pomoću procedure TRANSREG našli smo pogodnu λ Box-Coxove transformacije. Box-Coxove transformacije [4] su monotone transformacije definirane s:

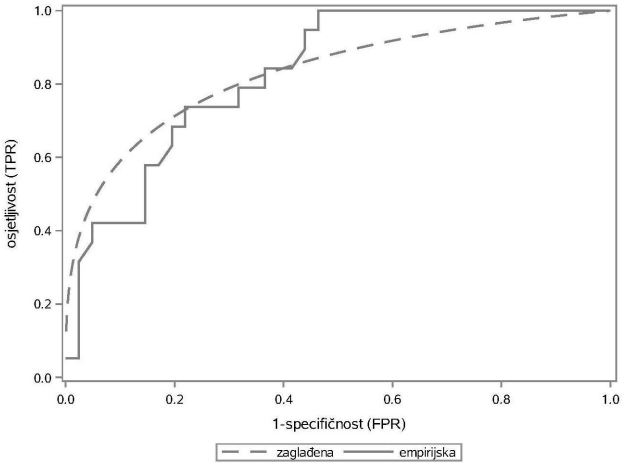
$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, & \lambda > 0 \\ \log(x), & \lambda = 0. \end{cases}$$

Rezultati prije i nakon transformacije ($\lambda = 0.25$) dani su tablicom 3.8.

	oporavak	$\bar{x}$ / $\overline{x^{(0.25)}}$	sd / $sd_{x^{(0.25)}}$	$\hat{a}$	$\hat{b}$	c
prije	neuspješan	427.2927	372.6351	0.8313	0.2445	0.7903
	uspješan	117.5263	91.1143			
nakon	neuspješan	12.9962	3.8671	1.1976	0.7469	0.8313
	uspješan	8.3648	2.8884			

Tablica 3.8: Tablica procjena.

Nakon što su podaci transformirani tako da budu normalno distribuirani, površina ispod zaglađene ROC krivulje bliska je površini ispod emprijske ROC krivulje.



Slika 3.2: Empirijska i zaglađena ROC krivulja.

Bibliografija

[1] *Confidence intervals for parameters*, https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect040.htm, posjećena 12.7.2017.

[2] D. Collet, *Modelling binary data*, Chapman & Hall, Boca Raton, 2003.

[3] T. S. Ferguson, *A course in large sample theory*, Chapman & Hall, London, 1996.

[4] M. Gönen, *Analyzing receiver operating characteristic curves with SAS*, SAS Institute, 2007.

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

[6] P. McCullagh i J. A. Nelder, *Generalized linear models*, Chapman & Hall/CRC, Boca Raton, 1989.

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

[8] A. N. Philippou i G. G. Roussas, *Asymptotic normality of the maximum likelihood estimate in the independent not identically distributed case*, Annals of the Institute of Statistical Mathematics **27** (1975), br. 1, 45–55.

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

[10] X. H. Zhou, D. K. McClish i N. A. Obuchowski, *Statistical methods in diagnostic medicine*, John Wiley & Sons, 2011.

Sažetak

U ovom radu bavili smo se modelima koje koristimo u analizi binarnih podataka. Takvi modeli motivirani su klasičnim linearnim modelom, a pripadaju široj klasi modela koje nazivamo generalizirani linearni modeli. Predstavili smo tri funkcije povezivanja koje se mogu koristiti u njihovom definiranju.
Pretežno smo se koncentrirali na logistički model, utemeljen na *logit* transformaciji podataka, koji u praksi prednjači pred ostalima zbog svoje jednostavnosti i lakše interpretacije parametara preko omjera izglednosti. Opisali smo dvije metode procjenjivanja parametara modela: metodu maksimalne vjerodostojnosti i minimalne χ^2 statistike. Također, dvama glavnim teoremima iskazali smo svojstva i asimptotsko ponašanje procjenitelja. Nakon toga predstavili smo mjere prilagodbe modela podacima te statistike koje služe za testiranje značajnosti parametara. Na kraju smo definirali nekoliko tipova reziduala i ukratko opisali ROC analizu. Kroz nekoliko primjera ilustrirali smo opisane statističke metode.

Summary

This thesis deals with the models which are used in the analysis of binary data. This kind of models are motivated by the classical linear model and belong to a broader class of models known as the generalized linear models. In this thesis, there were presented three link functions commonly used for defining this models.
Mainly, this paper focuses on the logistic model, based on the *logit* transformation of data, which has precendence among other transfomations in practice because of its simplicity and easier interpretation of parameters using odds ratios. There were described two estimation methods: method of maximum likelihood and minimum χ^2 method. Moreover, using two main theorems there were presented attributes of estimators as well as their asymptotic behavior. Furthermore, there were presented some goodness of fit statistics and the statistics for testing significance of the parameters. Besides, in the thesis there were defined several types of residuals and briefly described the ROC analysis. A few examples were used to illustrate the application of the statistical techniques presented in the thesis.

Životopis

Rođena sam 4. siječnja 1994. godine u Zagrebu. Osnovnu školu završila sam u Loboru, a opću gimnaziju u Zlataru. Tijekom osnovnoškolskog i srednjoškolskog obrazovanja uspješno sam se natjecala u poznavanju hrvatskog jezika i matematici. Na državnom natjecanju iz matematike 2010. i 2011. godine osvojila sam prvo mjesto, a 2012. drugo mjesto u B kategoriji natjecatelja. Preddiplomski studij matematike na Prirodoslovno-matematičkom fakultetu upisala sam 2012., a završila 2015. godine te sam školovanje nastavila na Matematičkom odsjeku upisavši diplomski studij Matematičke statistike. Tijekom studija držala sam demonstrature iz kolegija Matematička analiza 1 i 2, Obične diferencijalne jednadžbe, Matematičke metode u fizici i Statistika. Pri završetku diplomskog studija nagrađena sam za izniman uspjeh od Matematičkog odsjeka.