Generalizirani linearni modeli I
(teorija GLM-a, logistička regresija, modeliranje proporcija)

H. Planinić

Prosinac 2023.

Problem

- Problem kod raznih "black box" modela u strojnom učenju je što ne dopuštaju jednostavnu interpretaciju; npr. koje kovarijate (i kako) najviše utječu na odziv? Ako i dobijemo da neka kovarijate utječe na odziv, kako provjeriti je li to moglo biti samo plod slučajnosti u prikupljanju uzorka?
- U raznim disciplinama, interpretabilnost je još uvijek nešto što je jako poželjno i važno (npr. u medicini), ali u zadnje vrijeme se i u strojnom učenju sve veći naglasak stavlja na interpretabilnost ("Interpretable ML").

Klasični statistički pristup

- Imamo uzorak

$$\tau = \{(x^{(i)}, y_i),\ i = 1, \ldots, n\}$$

  gdje je y_i odziv, a $x^{(i)} = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ kovarijate.

- Pretp. da je odziv Y slučajna varijabla, kovarijate $X = (X_1, \ldots, X_p)$ su slučajni vektor te tražimo dobar model za (X, Y), a (tipično) pretpostavljamo da je $(x^{(i)}, y_i), i = 1, \ldots, n$ njd uzorak iz (X, Y)

- ipak, u nastavku ćemo pretposaviti tzv. fixed design, tj. pretpostavit ćemo da su vrijednosti kovarijata $X^{(i)} = x^{(i)}$ zadane, tj. neslučajne, a odziv y_i je realizacija slučajne varijable Y_i koja ima istu distribuciju kao i Y, ali uvjetno na $X = x^{(i)}$.

- tipično pretpostavljamo da su $Y_1, \ldots, Y_n$ nezavisne

Primjer 1 – normalni linearni model (NLM)

- NLM pretpostavlja da je

$$Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i = (x^{(i)})^\tau \beta + \epsilon_i \,,\ i = 1, \ldots, n \,,$$

  gdje su $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ nepoznati koeficijenti, a $\epsilon_1, \ldots, \epsilon_n$ njd $\mathrm{N}(0, \sigma^2)$ slučajne varijable za neki $\sigma > 0$.

- Problem: Što ako je $Y_i \in \{0, 1\}$ ili $Y_i \in \{0, 1, 2, \ldots\}$?

Reformulacija NLM-a i GLM

- Pretpostavke NLM-a možemo zapisati i na sljedeći način: $Y_1, \ldots, Y_n$ su nezavisni t.d.
  1. $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, uz
  2. $\mu_i := \mathbb{E}[Y_i \mid X^{(i)} = x^{(i)}] = (x^{(i)})^\tau \beta =: \eta_i$, pri čemu η_i zovemo linearnim prediktorom.
- Generalizirani linearni model (GLM) generalizira (očito!) gornji model u dva smjera
  1. $Y_i \sim$ distribucija iz neke eksponencijalne familije razdioba (binomna, Poissonova, itd.), te
  2. $\mu_i = h(\eta_i)$, za moguće nelinearnu funkciju h.

Eksponencijalne familije distribucija

– Familija distribucija

$$\{P_{\theta,\phi} : \theta \in \Theta \subseteq \mathbb{R}, \phi \in \Phi \subseteq (0,\infty)\},$$

na $\mathbb{R}$ je eksp. familija distribucija ako je gustoća (diskretna ili neprekidna) $f_{\theta,\phi}$ od $P_{\theta,\phi}$ oblika[1]

$$f_{\theta,\phi}(y) = h(y,\phi)\exp\left\{\frac{1}{\phi}(\theta y - b(\theta))\right\}, \; y \in \mathbb{R},$$

za neke funkcije h i b.

– θ zovemo prirodni parametar, a ϕ parametar disperzije.

[1] Ako je $Y \sim P_{\theta,\phi}$ diskretna varijabla, $f_{\theta,\phi}(y) := \mathbb{P}(Y = y)$, za sve $y \in \mathbb{R}$.

Neka svojstva

– Ako je $Y \sim P_{\theta,\phi}$ vrijedi

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

– Uz pretpostavku da je $\mathrm{Var}(Y) > 0$ za sve θ, ϕ, imamo da b' strogo rastuća pa dakle i invertibilna, pa možemo umjesto parametra θ koristiti parametar očekivanja $\mu \in \mathcal{M}$, uz $\theta = \theta(\mu) = (b')^{-1}(\mu)$.

– Tada imamo vezu između varijance i očekivanja

$$\mathrm{Var}(Y) = \phi b''(\theta(\mu)) =: \phi V(\mu)\,,$$

pri čemu funkciju $V(\mu) := b''(\theta(\mu))$ zovemo funkcijom varijance.

Primjer 2 – $\mathrm{N}(\alpha, \sigma^2)$

– za sve $y \in \mathbb{R}$ imamo

$$f_{\mathrm{N}(\alpha,\sigma^2)}(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(y-\alpha)^2}{2\sigma^2}\right\}$$
$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{y^2}{2\sigma^2}\right\}}_{=:h(y,\sigma^2)} \exp\left\{\frac{1}{\sigma^2}(\alpha y - \alpha^2/2)\right\}$$

– dakle, $\{\mathrm{N}(\alpha, \sigma^2) : \alpha \in \mathbb{R}, \sigma > 0\}$ je eksponencijalna familija razdioba uz $\theta = \alpha$, $\phi = \sigma^2$ i $b(\theta) = \theta^2/2$.

– iz $b'(\theta) = \theta$, imamo $\mu = \mathbb{E}[Y] = \theta \ (= \alpha)$

— iz $b''(\theta) = 1$, imamo $V(\mu) = 1$, pa je $\mathrm{Var}(Y) = \phi V(\mu) = \phi \ (= \sigma^2)$

Primjer 3 – Poiss(λ)

- za sve $y \in \mathbb{R}$ imamo

$$f_{\text{Poiss}(\lambda)}(y) = 1_{\{y \in \mathbb{N}_0\}} \frac{\lambda^y}{y!} \exp\{-\lambda\} = \underbrace{\frac{1}{y!} 1_{\{y \in \mathbb{N}_0\}}}_{=:h(y)} \exp\{y \log(\lambda) - \lambda\}$$

- dakle, $\{\text{Poiss}(\lambda) : \lambda > 0\}$ je eksponencijalna familija razdioba uz $\theta = \log(\lambda)$ $(\in \mathbb{R})$, $\phi = 1$ i $b(\theta) = \exp(\theta)$ $(= \lambda)$.

- iz $b'(\theta) = \exp(\theta)$, imamo $\mu = \mathbb{E}[Y] = \exp(\theta)$ $(= \lambda)$, te $\theta = \log(\mu)$

- iz $b''(\theta) = \exp(\theta)$, imamo $V(\mu) = b''(\log(\mu)) = \mu$, pa je $\text{Var}(Y) = \phi V(\mu) = \mu$ $(= \lambda)$

Primjer 4 – $\mathrm{B}(m, p)$

- ispostavlja se da je umjesto $Y \sim \mathrm{B}(m, p)$, potrebno gledati proporciju uspjeha $\tilde{Y} := Y/m$.
- za sve $y \in \mathbb{R}$ imamo

$$\tilde{f}_{m,p}(y) = f_{m,p}(my) = 1_{\{y \in \{0,1/m,\ldots,1\}\}} \binom{m}{my} p^{my}(1-p)^{m-my}$$

$$= \underbrace{1_{\{y \in \{0,1/m,\ldots,1\}\}} \binom{m}{my}}_{=:h(y,1/m)} \exp\left\{ m\left(y \log\left(\tfrac{p}{1-p}\right) - \log\left(\tfrac{1}{1-p}\right)\right)\right\}$$

Primjer 4 – $\mathrm{B}(m, p)$

- dakle, $\{\frac{1}{m}\mathrm{B}(m,p) : m \in \mathbb{N}, p \in (0,1)\}$ je eksponencijalna familija razdioba uz

$$\theta = \log(\tfrac{p}{1-p}) =: \operatorname{logit}(p) \in \mathbb{R}\,,$$

  te $\phi = 1/m$ i $b(\theta) = \log(1 + e^{\theta})$ $(= \log(\frac{1}{1-p}))$.

- vrijedi $\mu = \mathbb{E}[\tilde{Y}] = \frac{e^{\theta}}{1+e^{\theta}} =: \operatorname{expit}(\theta)$ $(= p)$, te
- $V(\mu) = \mu(1-\mu)$ (DZ).

GLM

– Za danu eksponencijalnu familiju $\{P_{\mu,\phi}\}$, GLM pretpostavlja da su $Y_1, \ldots, Y_n$ nezavisne i takve da

$$Y_i \sim P_{\mu_i,\phi_i}\ ,\ i = 1, \ldots, n$$

pri čemu

1. za neki $\beta \in \mathbb{R}^p$, očekivanje μ_i i linearni prediktor $\eta_i = (x^{(i)})^\tau \beta$ su povezani preko strogo monotone funkcije veze g t.d. $g(\mu_i) = \eta_i$, tj.

$$\mu_i = \mathbb{E}[Y_i \mid X^{(i)} = x^{(i)}] = g^{-1}(\eta_i)$$

2. parametar disperzije je oblika $\phi_i = \frac{\phi}{w_i}$ pri čemu su "težine" $w_1, \ldots, w_n$ poznate, a parametar $\phi > 0$ potencijalno nepoznat.[2]

[2] ϕ također zovemo parametar disperzije (modela). Npr. kod $\mathrm{B}(m, p)$ odziva, $w_i = m_i$.

Funkcije veze

– bitno je da je funkcija veze takva da je $g^{-1}(\eta_i)$ dopustiva vrijednost za μ_i, za sve vrijednosti koeficijanata β, npr.

1. kod normalnih odziva, $\mu_i \in \mathbb{R}$ pa možemo uzeti $\mu_i = \eta_i$ (tj. $g(\mu) = \mu$) i tako dobiti NLM kao specijalan slučaj GLM-a.
2. kod binomnih odziva, $\mu = p \in (0, 1)$ pa se može uzeti

$$\mu_i = \text{expit}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} ,$$

tj. $\eta_i = g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \text{logit}(\mu) \rightsquigarrow$ logistički model

3. kod Poissonovih odziva, $\mu = \lambda \in (0, \infty)$ pa se može uzeti

$$\mu_i = e^{\eta_i} ,$$

tj. $\eta_i = g(\mu_i) = \log(\mu_i)$.

Kanonska funkcija veze

– često za g uzmimamo tzv. kanonsku funkciju veze g_c – to je funkcija $g_c(\mu) := \theta(\mu)$, tj. u tom slučaju je

$$\theta_i = \theta(\mu_i) = \theta(g_c^{-1}(\eta_i)) = \eta_i \,.$$

– lako se provjeri da su funkcije veze u gornja tri primjera upravo kanonske funkcije veze (DZ)

– korištenje kanonskih funkcija veze ima neka dobra statistička svojstva, ali nisu nužno uvijek najbolji izbor

Procjena parametara

- u nastavku pretpostavljamo da je ϕ poznat (tako je npr. u binomnom i Poissonovom slučaju), pa preostaje procijeniti koeficijente β.
- za procjenu β koristimo procjenitelj maksimalne vjerodostojnosti: ako je

$$\ell(\beta) := \sum_{i=1}^{n} \log f_{\mu_i(\beta),\phi_i}(y_i)\,,\ \beta \in \mathbb{R}^p$$

  log-vjerodostojnost, $\widehat{\beta} := \arg\max_\beta \ell(\beta)$.
- Osim za NLM, $\widehat{\beta}$ aproksimiramo numeričkim (iterativnim) algoritmima (IRLS).
- Općenita teorija ML procjenitelja daje da, pod nekim dodatnim uvjetima, kada $n \to \infty$, $\widehat{\beta}$ asimptotski ima (višedimenzionalnu) normalnu distribuciju što omogućuje izradu intervala pouzdanosti npr. za svaki β_j.

Usporedba modela – test omjera vjerodostojnosti

– Pretpostavimo da imamo model ω_0 koji koristi samo kovarijate $X_1, \ldots, X_{p_0}$, te model ω_1 koji koristi $X_1, \ldots, X_{p_1}$ (uz $1 \leq p_0 < p_1 \leq p$)[3], i želimo testirati

$$H_0 : \text{model } \omega_0 \text{ je točan, tj. } \beta_{p_0+1} = \cdots = \beta_{p_1} = 0$$
$$H_1 : \text{potreban je model } \omega_1$$

– Može se pokazati da, ako je H_0 točan, za velike n vrijedi

$$T := 2(\ell(\widehat{\beta}_{\omega_1}) - \ell(\widehat{\beta}_{\omega_0})) \sim \chi^2_{p_1 - p_0} .$$

– Intitivno, ako nam je potreban ω_1, tj. ako nam je potrebna bar jedna od kovarijata $X_{p_0+1}, \ldots, X_{p_1}$, očekujemo da će log-vjerodostojnost modela ω_1 biti puno veća nego za ω_0.

[3] Kažemo da su modeli *ugnježđeni*.

Devijanca

- Kada smo procjenili koeficijente $\widehat{\beta}$, zanima nas koliko se dobro naš model prilagođava podacima (goodness-of-fit).
- Mjera koju možemo koristiti je devijanca modela, a definiramo je s

$$D = 2(\widehat{\ell}_{\text{sat}} - \widehat{\ell})\phi\,,$$

gdje je

1. $\widehat{\ell} = \ell(\widehat{\beta})$ maksimizirana log-vjerodostojnost, te
2. $\widehat{\ell}_{\text{sat}}$ je vrijednost koja se dobije kada u izrazu za log-vjerodostojnost $\sum_{i=1}^{n} \log f_{\mu_i,\phi_i}(y_i)$ uvrstimo $\mu_i := y_i$, $i = 1, \ldots, n$ (tzv. saturirani model koji ima po jedan parametar za svaki podatak $(x^{(i)}, y_i)$).

- Intuitivno, manja devijanca znači bolju prilagodbu, ali ne nužno i bolji model (jer je moguć overfitting).

Devijanca

– npr., za normalne odzive lako se pokaže (DZ) da je

$$D = \sum_{i=1}^{n} (y_i - (x^{(i)})^\tau \widehat{\beta})^2$$

tj. devijanca je točno suma kvadrata reziduala (RSS).

– Može se pokazati da u nekim slučajevima, za velike n i ako je naš model točan, tzv. skalirana devijanca D/ϕ ima približno χ^2 razdiobu s brojem stupnjeva slobode jednakim $n - p$, što omogućava testiranje hipoteze da je naš model (približno) točan.[4]

– Uočimo, ako je $\phi = 1$ (kao u binomnom i Poissonovom modelu), skalirana devijanca i devijanca su jednake, te se statistika $T = 2(\ell(\widehat{\beta}_{\omega_1}) - \ell(\widehat{\beta}_{\omega_0}))$ za usporedbu dva ugnježđena modela može zapisati kao razlika devijanci

$$T = D_{\omega_0} - D_{\omega_1} .$$

[4] Nažalost, ova tvrdnja ne vrijedi u slučaju binarnih odziva, tj. kada je $Y \in \{0, 1\}$.

Binarni/Bernoullijevi odzivi

- ako imamo $y_i \in \{0, 1\}$, pretpostavljamo da $Y_i \sim B(p_i) = B(1, p_i)$ uz

$$p_i = \mathbb{P}(Y_i = 1 \mid X^{(i)} = x^{(i)}) = \mu_i \, .$$

- ako koristimo kanonsku funkciju veze $g(p) = \log(\frac{p}{1-p}) =: \operatorname{logit}(p)$ dobivamo tzv. logističku regresiju – dakle, pretpostavka je

$$p = \mathbb{P}(Y = 1 \mid X = X) = \operatorname{expit}(\eta) = \frac{e^\eta}{1 + e^\eta}$$

  gdje je $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, za $X = (x_1, \ldots, x_p)$.

- ako pretpostavimo da je $p = \Phi(\eta)$ gdje je Φ funkcija distribucije $\mathrm{N}(0, 1)$ razdiobe (tj. $g(p) = \Phi^{-1}$), govorimo o probit regresiji.

Primjer u R-u

Interpretacija parametara

- interpretacija parametara β sada vise nije jednostavna kao u NLM-u jer vjerojatnost p na nelinearan način ovisi o linearnom prediktoru $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- u logističkoj regresiji moguća je interpretacija koristeći koncept izgleda (*engl.* odds).
- Za Bernoulijevu slučajnu varijablu $Y \sim \mathrm{B}(p)$, izgledi se definiraju kao

$$\mathrm{odds}(Y) = \frac{p}{1-p} = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)}\,.$$

- npr. ako je $p = 1/2$ imamo $\mathrm{odds}(Y) = 1$, ako je $p = 2/3$ imamo $\mathrm{odds}(Y) = 2/1 = 2$.

Interpretacija parametara

- neka je $p(x) := \mathbb{P}(Y = 1 \mid X = x)$
- u logističkoj regresiji imamo $\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, tj.

$$\text{odds}(Y \mid X = x) := \frac{p(x)}{1 - p(x)} = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\,.$$

- npr., ako imamo samo jednu kvantitativnu kovarijatu x (npr. visina, temperatura ili slično), linearni prediktor je oblika $\eta = \beta_0 + \beta_1 x$ te je

$$\frac{\text{odds}(Y \mid X = x + 1)}{\text{odds}(Y \mid X = x)} = \exp\{\beta_1\}$$

$\rightsquigarrow$ dakle, faktor $\exp\{\beta_1\}$ je promjena izgleda kada se X poveća za 1.
$\rightsquigarrow$ $\beta_1 > 0$ povlači $\exp\{\beta_1\} > 1$, a $\beta_1 < 0$ povlači $\exp\{\beta_1\} < 1$.

Interpretacija parametara – kategorijalne kovarijate

– ako imamo samo jednu kategorijalnu kovarijatu $x \in \{0, 1\}$ (npr. spol uz 0 = muško, 1 = žensko)

$$\eta = \beta_0 + \beta_1 1_{\{x=\text{žensko}\}} = \begin{cases} \beta_0\,, & x = \text{muško}\,, \\ \beta_0 + \beta_1\,, & x = \text{žensko}\,. \end{cases}$$

– u ovom slučaju imamo

$$\frac{\text{odds}(Y \mid X = \text{žensko})}{\text{odds}(Y \mid X = \text{muško})} = \exp\{\beta_1\}$$

$\rightsquigarrow$ dakle, faktor $\exp\{\beta_1\}$ omjer izgleda za žene u odnosu na muškarce

$\rightsquigarrow$ koristeći npr. test omjera vjerodostojnosti možemo testirati ima li statistički značajne razlike u izgledima između muškaraca i žena, tj. testirati $H_0 : \beta_1 = 0$.

Interpretacija parametara – kategorijalne kovarijate

– ako imamo samo jednu kategorijalnu kovarijatu s npr. 3 različite vrijednosti, tj. $x \in \{0, 1, 2\}$,

$$\eta = \beta_0 + \beta_1 1_{\{x=1\}} \beta_2 1_{\{x=2\}} = \begin{cases} \beta_0\,, & x = 0\,, \\ \beta_0 + \beta_1\,, & x = 1\,, \\ \beta_0 + \beta_2\,, & x = 2\,. \end{cases}$$

– u ovom slučaju imamo

$$\frac{\text{odds}(Y \mid X = 1)}{\text{odds}(Y \mid X = 0)} = \exp\{\beta_1\}\,, \quad \frac{\text{odds}(Y \mid X = 2)}{\text{odds}(Y \mid X = 0)} = \exp\{\beta_2\}$$

$\rightsquigarrow$ dakle, β_1 (β_2) mjeri promjenu za slučaj $X = 1$ ($X = 2$) u odnosu na $X = 0$.

$\rightsquigarrow$ u ovom slučaju 0 je tzv. bazna kategorija, ali taj izbor je proizvoljan – ako izaberemo neku drugu baznu kategoriju, dobijemo isti model, samo se mijenja interpretacija parametara.

Interpretacija parametara – općeniti slučaj

– u općenitom logističkom modelu imamo p kovarijata i

$$\text{odds}(Y \mid X = x) := \frac{p(x)}{1 - p(x)} = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\,.$$

– ako je X_j npr. kvantitativna kovarijata, te X' dobijemo iz X tako da samo x_j povećamo za 1, a sve ostale kovarijate ostavimo nepromijenjene, imamo da je

$$\frac{\text{odds}(Y \mid X = x')}{\text{odds}(Y \mid X = x)} = \exp\{\beta_j\}$$

$\rightsquigarrow$ dakle, kada se X_j poveća za 1 i sve ostale kovarijate ostanu iste, izgledi se mijenjaju za faktor $\exp\{\beta_j\}$.
$\rightsquigarrow$ gornja interpretacija ima manje (ili nimalo) smisla ukoliko postoji barem jedna kovarijata X_i koja je (jako) zavisna s X_j (ili čak funkcija od X_j, npr. $X_i = X_j^2$)

Interpretacija parametara – slobodni član

– za slobodni član imamo da je

$$\mathbb{P}(Y = 1 \mid X = (0, \ldots, 0)) = \operatorname{expit}(\beta_0)$$

– ipak, slučaj $X = (0, \ldots, 0)$ često nema smisla pa β_0 u tom slučaju nema neku posebnu interpetaciju.

Zadatak 1

U datoteci `upisi.csv` nalaze se podaci o uspješnosti upisa 400 studenata na poslijediplomske studije. Za svakog su aplikanta dani rezultati GRE testa, prosjek ocjena (GPA) i rang fakulteta na koji se aplicirao; rang je kategorijalna kovarijata (1,2,3 ili 4).

(i) Izračunajte relativne frekvencije upisa (koriste funkcije `table` i `prop.table`) u odnosu na rang apliciranog fakulteta te grafički prikažite ovisnost upisa s obzirom na GPA, odnosno GRE test (koristite `jitter`). Ima li indikacija da neke od ovih kovarijata utječu na vjerojatnost upisa?

(ii) Prilagodite logistički model za dane podatke koristeći samo kovarijatu GPA. Napišite model te interpretirajte dobivene koeficijente. Na temelju smanjenja devijance u odnosu na nul-model, zaključite je li ova kovarijata statistički značajna u modeliranju vjerojatnosti upisa (u odnosu na nul-model). Prikažite grafički vjerojatnost upisa u ovisnosti o GPA u ovom modelu te na istom grafu prikažite i odgovarajuće stvarne podatke.

Zadatak 1

(iii) Procijenite parametre logističkog modela za dane podatke koristeći sve kovarijate; bitno je koristiti naredbu `factor(rank)` kako bi R znao da se radi o kategorijalnoj kovarijati. Napišite koji je model te interpretirajte dobivene koeficijente. Usporedite koeficijent uz kovarijatu GPA u odnosu na model iz (ii) – možete li objasniti zašto je došlo do razlike? Testirajte statističku značajnost svake od kovarijata tako što ćete koristeći test omjera vjerodostojnosti usporediti puni model s modelom u kojem ste ispustili jednu kovarijatu. Koje su kovarijate statistički značajne na razini značajnosti 5%, a koje na razini 1%. Izračunajte vjerojatnosti upisa u ovom modelu za studenta s GRE rezultatom 750 i prosjekom 3.88 koji želi upisati poslijediplomski program na sveučilištu ranga 1,2,3, odnosno 4.

Zadatak 2

Promatramo podatke `Whickam` iz paketa `mosaicData`. Za 1314 žena u Velikoj Britaniji, u periodu od 1972-1974 uzeta je informacija o tome da li je pušač ili ne (kovarijata `smoker`) te koliko ima godina (`age`). Dvadeset godina nakon provjereno je jesu li još uvijek žive (odziv `outcome`). Želimo ispitati utjecaj pušenja na smrtnost koristeći logistički model.

(i) Prilagodite logistički model za dane podatke koristeći samo kovarijatu `smoker`. Napišite model te interpretirajte dobivene koeficijente; koristite naredbu `contrasts(outcome)` kako biste vidjeli kako je `R` kodirao odziv. Jesu li rezultati očekivani?

(ii) Podijelite kovarijatu `age` u 5 grupa koristeći naredbu `cut`. Izračunajte relativne frekvencije odziva, te kovarijate `smoker`, u odnosu na tako grupirane godine. Kako godine utječu na smrtnost dvadeset godina nakon? Kakva je proprcija pušača ovisno o godinama? Možete li sada objasniti što se dogodilo u (i)?

Zadatak 2

(iii) Prilagodite logistički model za dane podatke koristeći obje kovarijate. Napišite model te interpretirajte dobivene koeficijente. Jesu li sada rezultati intuitivniji? Testirajte statističku značajnost kovarijate `smoker` na razini značajnosti 5% tako što ćete usporediti puni model i model u kojem je ta kovarijata ispuštena, koristeći test omjera vjerodostojnosti.

Napomena

– Fenomen iz prethodnog zadatka naziva se Simpsonov paradoks – u model nismo bili uključili kovarijatu koja značajno utječe na odziv (confounder).

Modeliranje proporcija

- Pretpostavimo da imamo binarne odzive $y_i \in \{0, 1\}$, ali da se neke vrijednosti kovarijata $x^{(i)}$ za različite i-eve ponavljaju.
- Tada takve parove možemo grupirati te kao odziv uzeti proporciju y-ona koji su bili jednaki 1.
- Sve skupa, modeliramo odzive kao $\frac{1}{m_i}\mathrm{B}(m_i, p_i)$ slučajne varijable, gdje je m_i ukupan broj parova koje smo grupirali za danu vrijednost kovarijata. Uočimo da je ukupan broj podataka m nužno manji ili jednak n,
- interpretacija parametara je ista kao i u običnoj logističkoj regresiji
- razlika je što u ovom slučaju možemo koristiti devijancu za provjeru prilagodbe modela, tj. devijanca za velike m i ako je naš model točan ima približno χ^2 razdiobu s brojem stupnjeva slobode jednakim $m - p$, što omogućava testiranje hipoteze da je naš model (približno) točan; m je broj podataka nakon grupiranja.[5]

[5] Ipak, ovo ne vrijedi uvijek, ali diskusija o tome izlazi van okvira našeg kolegija.

Primjer 5

– Istražujemo vezu određenog krvnog enzima i pojave srčanog udara. Dani su podaci za 360 pacijenata – za svakog je dana razina tog enzima te je li ili nije imao srčani udar. Razine enzima su grupirane u $m = 12$ grupa tako da možemo koristiti binomni model.

Primjer u R-u

Zadatak 3

(i) Nastavno na Primjer 5, ubacite u linearni prediktor kubični član. Testirajte statističku značajnost dodane kovarijate koristeći test omjera vjerodostojnosti. Provjerite i prilagodbu modela uspoređujući devijancu modela s odgovarajućom χ^2 razdiobom. Što vam se čini, treba li nam model s kubičnim članom?

(ii) Dodajte u model iz (i) i potenciju reda 4, te ponovite analizu iz dijela (i).

(iii) Prikažte prilagodbu modela iz (i) i (ii) grafički, po uzoru na Primjer 5.