Regresijska analiza višekategorijskih varijabli odziva

Barišić Lučić, Ivana

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Ivana Barišić Lučić

REGRESIJSKA ANALIZA VIŠEKATEGORIJSKIH VARIJABLI ODZIVA

Diplomski rad

Voditelj rada:
prof. dr. sc. Miljenko Huzak

Zagreb, 2018

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. _____

Hvala mom mentoru, prof. dr. sc. Miljenku Huzaku, na razumijevanju i strpljenju tijekom izrade ovog rada. Hvala mom liječniku, dr. Marku Brinaru, na suradnji te na potpori i pomoći u liječenju.

Veliko hvala mojim roditeljima, bratu i sestri na strpljenju i ljubavi. Hvala svim prijateljima koji su uvijek bili uz mene i kolegama koji su sada prijatelji.

Posebno hvala mom suprugu Krševanu koji mi je bio najveća podrška i koji je najviše vjerovao u moj uspjeh.

Sadržaj

Uvod

Logistička regresija je statistička metoda za analizu skupa podataka u kojem postoji jedna ili više nezavisnih varijabli, odnosno prediktora, koje određuju ishod, odnosno zavisnu varijablu. Pogodna je za situacije kada relacija između zavisne i nezavisne varijable nije linearna.

Ishod može biti dihotomna varijabla, odnosno imati dva moguća ishoda. Ukoliko imamo samo jednu nezavisnu varijablu kažemo da je riječ o jednostavnoj logističkoj regresiji, te o tome pišemo u prvom poglavlju. Ukoliko imamo dvije ili više nezavisnih varijabli, kažemo da je riječ o složenoj logističkoj regresiji, i to je sadržaj drugog poglavlja.

Ukoliko ishod nije dihotomna varijabla nego ima tri ili više kategorija, odnosno kada je ishod multinomna varijabla, takvu logističku regresiju zovemo polinomna logistička regresija, te o njoj pišemo u trećem poglavlju.

Glavna zadaća logističke regresije je odrediti omjer izgleda promjene kategorije zavisne varijable (prelazak iz jedne kategorije u drugu) ukoliko se promjeni neka nezavisna varijabla. Upravo to ćemo probati odrediti u primjeru iz biomedicine koja se bazira na pacijentima s kroničnom upalom crijeva, a to je u ovom primjeru Crohnova bolest.

Poglavlje 1

Jednostavan logistički regresijski model

Jednostavan logistički regresijski model primijenjujemo kada zavisnu varijablu (*varijablu odziva*), koja ima dva moguća ishoda, odnosno koja je binomna ili dihotomna, želimo opisati pomoću jedne nezavisne varijable, tj. jednog *prediktora*.

1.1 Omjer izgleda

Važni pojmovi u logističkoj regresiji su upravo izgledi (*engl. odds*) te omjeri izgleda (*engl. odds ratio*). Pokušat ćemo ih objasniti pomoću sljedeće frekvencijske tablice.

		Bolestan	Zdrav	Ukupno
Faktor	Pušač	a	b	$a+b$
	Nepušač	c	d	$c+d$
	Ukupno	$a+c$	$b+d$	$a+b+c+d$

Tablica 1.1: Ocjena rizika

Neka je p vjerojatnost nekog događaja, odnosno u našem slučaju vjerojatnost bolesti. Izgled (*odds*) da razvijemo neku bolest je omjer između vjerojatnosti da ju imamo i vjerojatnosti da ju nemamo:

$$\omega = \frac{p}{1-p}. \tag{1.1}$$

Na primjer, ako je vjerojatnost da imamo neku bolest $p = 0.6$, onda je izgled za razvoj te bolesti 1.5, tj. $\omega = \dfrac{0.6}{0.4} = 1.5$.

Izgledi nam služe radi lakšeg razmijevanja. Dakle, ukoliko je vjerojatnost da netko (osoba A) ima neku bolest 0.3, dok je vjerojatnost osobe B 0.6 da ima tu istu bolest, razumno je zaključiti da osoba B ima duplo veću vjerojatnost oboljenja. No, kada bi se vjerojatnost osobe A povećala na 0.6, tada bismo dobili da je vjerojatnost da osoba B ima bolest jednaka 1.2, što je nemoguće jer je najveća vjerojatnost 1. Stoga je za uspoređivanje razumnije koristiti izglede. Ako je $p = 0.3$, $\omega = \dfrac{0.3}{0.7} = 0.43$, dok je za $p = 0.6$, $\omega = 1.5$. Dakle, izgled da osoba A razvije bolest je 0.43, dok je izgled da osoba B razvije istu bolest 1.5.

Omjer izgleda će biti upravo omjer između izgleda izloženih i izgleda neizloženih određenom faktoru, u našem slučaju pušenju. Dakle, to je

$$OR = \frac{\frac{p_{\text{pušači}}}{1-p_{\text{pušači}}}}{\frac{p_{\text{nepušači}}}{1-p_{\text{nepušači}}}}.$$

U prethodnom primjeru bi vrijedilo $p_{\text{pušači}} = \frac{a}{a+b}$, $1 - p_{\text{pušači}} = \frac{b}{a+b}$, $p_{\text{nepušač}} = \frac{c}{c+d}$, $1 - p_{\text{nepušač}} = \frac{d}{c+d}$, time dobivamo

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

.

		Bolestan	Zdrav	Ukupno
Faktor	Pušač	36	25	61
	Nepušač	30	38	68
	Ukupno	66	63	129

Tablica 1.2: Primjer računanja ocjena rizika

Izračunajmo izglede da je pušač bolestan, da je nepušač bolestan, te omjer izgleda oboljenja između pušača i nepušača.

ω(pušač) $= \frac{36}{25} = 1.44$. Odnosno, pušači imaju 1.44 puta veću vjerojatnost da dobiju bolest nego da ju ne dobiju.

ω(nepušač) $= \frac{30}{38} = 0.79$. Odnosno, nepušači imaju 0.79 puta veću vjerojatnost da dobiju bolest nego da ju ne dobiju.

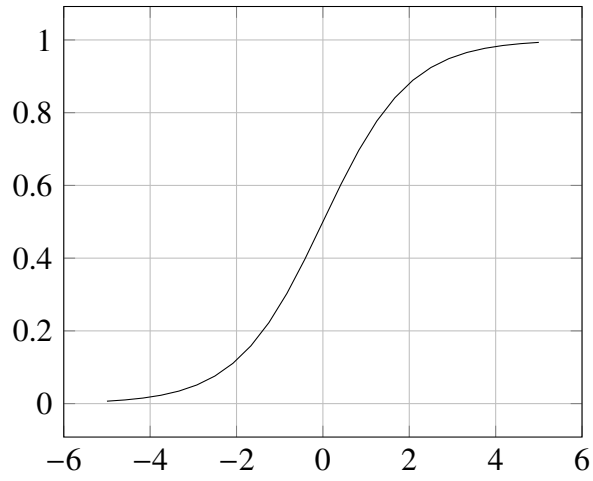
$OR = \frac{36 \cdot 38}{30 \cdot 25} = 1.82$. Dakle, pušači imaju 1.82 puta veći izgled oboljenja u odnosu na nepušače.

1.2 Model jednostavne logističke regresije

Definirajmo funkciju $p : \mathrm{R} \to \langle 0, 1\rangle$ s

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

Tu funkciju zovemo *logistička funkcija*.



Slika 1.1: Logistička funkcija

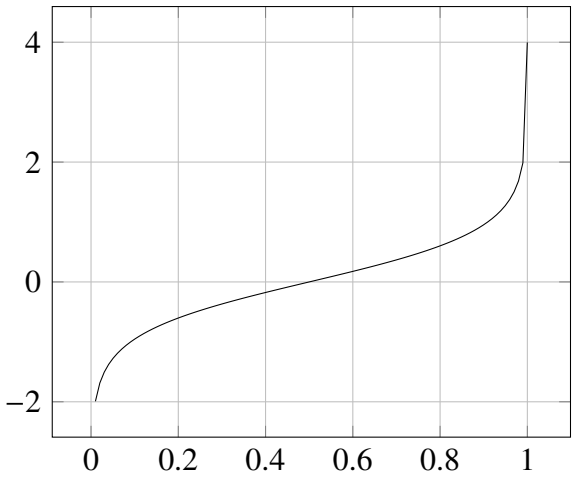
Logistička funkcija ima svoju inverznu funkciju koju zovemo *logit funkcija* i označava se s logit (slika 1.2). Dakle, vrijedi

$$\mathrm{logit} : \langle 0, 1\rangle \to \mathrm{R}, \quad \mathrm{logit}(y) = \log \frac{y}{1 - y} = \log(y) - \log(1 - y).$$

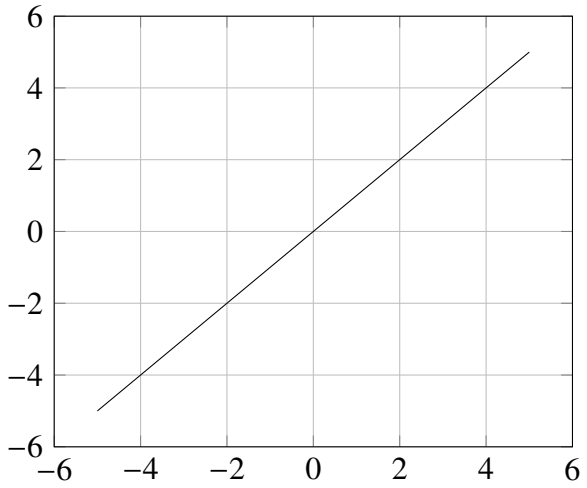
Transformacijom vjerojatnosti $p(x)$ u izgled $\left[\dfrac{p(x)}{1 - p(x)}\right]$, odnosno korištenjem definicije logističke funkcije, mičemo gornju granicu na transformiranu veličinu, a logaritmiranjem donju granicu, te tada dobivamo linearnu transformaciju (slika 1.3).

Općenito, to možemo zapisati kao:

$$P(x) = \mathrm{logit}p(x) = \beta_0 + \beta_1 x \Leftrightarrow \frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 x) \Leftrightarrow p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{1.2}$$



Slika 1.2: Logit funkcija



Slika 1.3: Transforamcija logit funkcije

Pretpostavimo da imamo zadanih I različitih vrijednosti kovarijate $x_1, x_2, \ldots, x_I$. Za svaku od tih vrijednosti x_i opažamo uzorak od n_i nezavisnih Bernoullijevih slučajnih varijabli $Y'_1, Y'_2, ..., Y'_{n_i}$. Dakle, $Y_i = Y'_1 + Y'_2 + \cdots + Y'_{n_i} \sim B(p(x_i), n_i)$. Prisjetimo se da je $E(Y_i) = m_i = n_i p(x_i)$ i da su izgledi $\frac{p(x_i)}{1-p(x_i)}$. Logistička regresija određuje linearnu strukturu za logaritmirane izglede s

$$P(x_i) := \text{logit}\,(p(x_i)) \equiv \log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \beta_1 x_i, \qquad i = 1, 2, \ldots, I. \tag{1.3}$$

Uočite da promjenom $x = x_i$, izraz na desnoj strani jednadžbe (1.3) može poprimiti bilo koju realnu vrijednost. Iako su vjerojatnosti $p(x_i)$ ograničene između 0 i 1, logaritmirani izgledi mogu poprimiti bilo koju realnu vrijednost.

Procjena parametara modela

Za procjenu parametara koristimo metodu maksimalne vjerodostojnosti. Jednadžba (1.3) se može transformirati u

$$p_i := p(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$
$$1 - p_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Neka je y_i opažena vrijednost od Y_i. Uz notaciju $p_{i1} \equiv p_i, p_{i2} \equiv (1 - p_i), n_{i1} \equiv y_i, n_{i2} \equiv n_i - y_i$, odnosno $n_i = n_{i1} + n_{i2}$, dobivamo funkciju vjerodostojnosti parametara β_0 i β_1:

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{I} \left[\frac{n_i!}{\prod_{j=1}^{2} n_{ij}!} \left\{ \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\}^{n_{i1}} \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\}^{n_{i2}} \right] \\ &= \prod_{i=1}^{I} \left[\frac{n_i!}{\prod_{j=1}^{2} n_{ij}!} \{p_{i1}\}^{n_{i1}} \{p_{i2}\}^{n_{i2}} \right]. \end{aligned} \tag{1.4}$$

Funkciju L zovemo *vjerodostojnost*. Nadalje, jednadžbu (1.4) možemo logaritmirati te dobivamo *log-vjerodostojnost*:

$$LL(\beta_0, \beta_1) = \sum_{i=1}^{I} [n_{i1} \log(p_{i1}) + n_{i2} \log(p_{i2})] + \sum_{i=1}^{I} \log \frac{n_i!}{\prod_{j=1}^{2} n_{ij}!}. \tag{1.5}$$

Sada ćemo (1.5) maksimizirati tako da parcijalno deriviramo p_{i1}, odnosno, p_{i2} po β_0 i β_1, te to izjednačimo s 0.

Za početak izračunajmo:

$$\begin{aligned} \frac{\partial p_{i1}}{\partial \beta_0} &= \frac{\exp(\beta_0 + \beta_1 x_i)}{(1 + \exp(\beta_0 + \beta_1 x_i))^2} = p_{i1} p_{i2}, \\ \frac{\partial p_{i2}}{\partial \beta_0} &= \frac{-\exp(\beta_0 + \beta_1 x_i)}{(1 + \exp(\beta_0 + \beta_1 x_i))^2} = -p_{i1} p_{i2}, \\ \frac{\partial p_{i1}}{\partial \beta_1} &= \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{(1 + \exp(\beta_0 + \beta_1 x_i))^2} = x_i p_{i1} p_{i2}, \\ \frac{\partial p_{i2}}{\partial \beta_1} &= \frac{-x_i \exp(\beta_0 + \beta_1 x_i)}{(1 + \exp(\beta_0 + \beta_1 x_i))^2} = -x_i p_{i1} p_{i2}. \end{aligned}$$

Sada vrijedi:

$$\frac{\partial LL}{\partial \beta_0} = \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_{i1}}p_{i1}p_{i2} - n_{i2}\frac{1}{p_{i2}}p_{i1}p_{i2}\right] = \sum_{i=1}^{I}\left[n_{i1}p_{i2} - n_{i2}p_{i1}\right] = \sum_{i=1}^{I}\left[n_{i1} - n_i p_{i1}\right]$$

$$\frac{\partial LL}{\partial \beta_1} = \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_{i1}}x_i p_{i1}p_{i2} - n_{i2}\frac{1}{x_i}p_{i2}p_{i1}p_{i2}\right] = \sum_{i=1}^{I} x_i\left[n_{i1}p_{i2} - n_{i2}p_{i1}\right] = \sum_{i=1}^{I} x_i\left[n_{i1} - n_i p_{i1}\right]$$

Dakle, dobivamo:

$$\sum_{i=1}^{I}\left[n_{i1} - n_i p_{i1}\right] = 0 \qquad \sum_{i=1}^{I} x_i\left[n_{i1} - n_i p_{i1}\right] = 0. \tag{1.6}$$

Rješavanjem nelinearnog sustava (1.6) po β_0, β_1 dobivamo MLE[1] procjenu parametara $\hat{\beta}_0$ i $\hat{\beta}_1$, iz čega slijedi procjena vjerojatnosti

$$\hat{p}_i = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}, \quad i = 1, \ldots, I.$$

Nije teško pokazati da je funkcija $(\beta_0, \beta_1) \to LL(\beta_0, \beta_1)$ konkavna. Budući da je $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ stacionarna točka od LL, ona je ujedno i točka maksimuma.

Dakle, ova formula nam daje $\hat{p}_i$ koje predviđamo pomoću x_i. Također, pomoću toga dobivamo i procjenu očekivanja, odnosno, $\hat{m}_{ij} = n_i \hat{p}_{ij}$.
Adekvatnost modela (*engl. Goodness of fit*) možemo testirati u usporedbi sa saturiranim modelom. Saturirani model je model koji sadrži onoliko parametara koliko ima podataka. Naime, u saturiranom modelu p_{ij} procjenjujemo pomoću relativnih frekvencija: $\tilde{p}_{ij} = \frac{n_{ij}}{n_i}$.

Koristimo statistiku G^2:

$$\begin{aligned} G^2 &= 2\sum_{i=1}^{I}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \\ &= 2\sum_{i=1}^{I}\left[n_{i1}\log\left(\frac{n_{i1}}{\hat{m}_{i1}}\right) + n_{i2}\log\left(\frac{n_{i2}}{\hat{m}_{i2}}\right)\right] \\ &= 2\sum_{i=1}^{I}\left[n_{i1}\log\frac{\tilde{p}_i}{\hat{p}_i} + (n_i - y_i)\log\frac{(1-\tilde{p}_i)}{(1-\hat{p}_i)}\right] \end{aligned} \tag{1.7}$$

Ukoliko je $y_i = 0$ uzima se da je $y_i \log(y_i) = 0$.
Primijetimo da vrijedi

$$G^2 = 2\log\left[\frac{L(model)}{L(saturirani\ model)}\right],$$

[1] engl. Maximum Likelihood Estimator

gdje je

$$L(\textit{saturirani model}) = \prod_{i=1}^{I} \left[\frac{n_i!}{\prod_{j=1}^{2} n_{ij}!} \left\{ \frac{y_i}{n_i} \right\}^{n_{i1}} \left\{ \frac{n_i - y_i}{n_i} \right\}^{n_{i2}} \right].$$

Također, možemo testirati razliku modela s ($\beta_1 \neq 0$) i bez varijabli prediktora ($\beta_1 = 0$), te tada koristimo statistiku D^2:

$$\begin{aligned} D^2 &= G^2 \left[\text{ model bez prediktora} \right] - G^2 \left[\text{model s prediktorom} \right] \\ &= 2 \log \left[\frac{L(\text{ model bez prediktora})}{L(\text{saturirani model})} \right] - \left\{ 2 \log \left[\frac{L(\text{model s prediktorom})}{L(\text{saturirani model})} \right] \right\} \\ &= 2 \log \left[\frac{L(\text{ model bez prediktora})}{L(\text{model s prediktorom})} \right] \end{aligned} \tag{1.8}$$

Alternativno, test istih hipoteza možemo sprovesti i pomoću Pearsonove statistike:

$$\begin{aligned} X^2 &= \sum_{i=1}^{I} \sum_{j=1}^{2} \frac{\left(n_{ij} - n_i \hat{p}_{ij} \right)^2}{n_i \hat{p}_{ij}} \\ &= \sum_{i=1}^{I} \left[\frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i} + \frac{\left[(n_i - y_i) - n_i (1 - \hat{p}_i) \right]^2}{n_i (1 - \hat{p}_i)} \right] \\ &= \sum_{i=1}^{I} \left[\frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i} + \frac{(y_i - n_i \hat{p}_i)^2}{n_i (1 - \hat{p}_i)} \right] \\ &= \sum_{i=1}^{I} \frac{(y_i - n_i \hat{p}_i)^2}{n_i \hat{p}_i (1 - \hat{p}_i)}. \end{aligned} \tag{1.9}$$

Sve tri statistike, G^2, D^2, X^2 su prikladne testne statistike, ali, nažalost, za njihove asimptotske nul-distribucije (χ^2), potrebno je da n_i bude što je moguće veći. Općenito, χ^2 *testovi hipoteza o modelu su prikladni* ***samo*** *kada su veličine* n_i *velike za sve I populacije*.

Testiranje hipoteza o parametarima modela

Bitan test u logističkoj regresiji je test značajnosti modela.

Općenito, hipoteze za testiranje značajnost nekog modela su:

$$H_0 : A^T \beta = c$$
$$H_1 : A^T \beta \neq c$$

gdje matrica $A \in M_{k+1,r}$ punog ranga $r(A) = r \leq k + 1$ i vektor $c \in \mathbf{R}^r$. Tada je Waldova statistika:

$$W = A^T(\hat{\beta} - c)^T(A^T I^{-1}(\hat{\beta})A)^{-1}(A^T(\hat{\beta} - c)) \tag{1.10}$$

gdje je $I(\beta) = X^T VX$ očekivana informacijska matrica, X je matrica dizajna, a V je dijagonalna matrica kojoj je na $i - tom$ mjestu vrijednost $n_i p_i(1 - p_i)$, $i = 1, 2, \ldots, I$.
Vrijedi da je $W \sim A\chi^2(r)$ ako vrijedi nul-hipoteza H_0. [1]

Za testiranje značajnosti jednostavnog logističkog modela, odnosno hipoteze:

$$\begin{aligned} H_0 &: \beta_1 = 0 \qquad \text{(reducirani model)} \\ H_1 &: \beta_1 \neq 0 \qquad \text{(puni model)} \end{aligned}$$

Waldova statistika će biti:

$$W = \frac{\hat{\beta}^2}{\sqrt{\hat{var}(\beta)}} \sim A\chi^2(1).$$

Interpretacija parametara β_0 i β_1

Iz (1.3) slijedi:

$$\begin{aligned} \text{logit}\,(p\,(x)) =& g\,(x) = \beta_0 + \beta_1 x \\ & g\,(x+1) = \beta_0 + \beta_1\,(x+1) \\ \Rightarrow & g\,(x+1) - g\,(x) = \beta_1 \\ \Rightarrow & \text{logit}\,(p\,(x+1)) - \text{logit}\,(p\,(x)) = \beta_1 \\ \Rightarrow & \log\,(\omega\,(p\,(x+1))) - \log\,(\omega\,(p\,(x))) = \beta_1 \\ \Rightarrow & \log\left(\frac{\omega\,(p\,(x+1))}{\omega\,(p\,(x))}\right) = \beta_1 \end{aligned}$$

Ukoliko su i zavisna i nezavisna varijabla dihotomne (s vrijednostima 0 i 1), tada imamo:

$$\begin{aligned} x = 1 \qquad & \omega_1 = \frac{p(1)}{1 - p(1)} \\ x = 0 \qquad & \omega_0 = \frac{p(0)}{1 - p(0)} \end{aligned}$$

Pa dobivamo:

$$\begin{aligned} g(0) &= \beta_0 \\ g(1) - g(0) &= \log\left(\frac{\omega_1}{\omega_0}\right) \\ &= \log\left(\frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}}\right) \\ &= \log(OR) = \beta_1 \Rightarrow OR = \exp(\beta_1) \end{aligned}$$

Ovo nam govori da prelaskom iz bazne kategorije (odnosno iz kategorije 0), u alternativnu kategoriju (odnosno 1), omjer izgleda iznosi $\exp(\beta_1)$. Ukoliko se sjetimo primjera iz tablice 1.2, te nam $x = 0$ predstavlja nepušača, dok je $x = 1$ pušač, tada omjer izgleda iznosi 1.82. Kažemo: *ukoliko nepušač postane pušač, omjer izgleda da će oboljeti od te bolesti poveća se 1.82 puta*.

Situacija je ponešto drugačija ukoliko je nezavisna varijabla *kontinuirana*. Interpretacija u tom slučaju varira ovisno o tome što nezavisna varijabla predstavlja. Na primjer, ako nezavisna varijabla predstavlja sistolički tlak izražen u mmHg, pomak za 1mmHg nije dovoljno velik da bi nam biološki nešto značio, no pomak za 10mmHg može značajno utjecati. U tom slučaju omjer izgleda možemo zapisati kao:

$$g(x + c) - g(x) = c\beta_1 \Rightarrow OR = \exp(c\beta_1)$$

Tada možemo reći: *ukoliko x promijeni vrijednost za c jedinica, omjer izgleda se mijenja* $\exp(c\beta_1)$ *puta*.

Poglavlje 2

Složeni logistički regresijski model

Složenu logističku regresiju koristimo kada zavisnu varijablu (*varijablu odziva*), koja ima dva moguća ishoda, odnosno koja je binomna ili dihotomna, želimo opisati pomoću više nezavisnih varijabli, tj. više *prediktora*.

2.1 Omjer izgleda

Kao i u slučaju jednostavne logističke regresije, izglede i omjere izgleda pokušat ćemo objasniti pomoću frekvencijske tablice. Radi jednostavnijeg prikaza koristit ćemo samo dva kategorijska prediktora.

Faktor 1	Faktor 2	Normalan	Povišen	Ukupno
A	Normalan	n_{111}	n_{211}	$n_{.11}$
	Povišen	n_{112}	n_{212}	$n_{.12}$
B	Normalan	n_{121}	n_{221}	$n_{.21}$
	Povišen	n_{122}	n_{222}	$n_{.22}$
	Ukupno	$n_{1..}$	$n_{2..}$	$n_{...}$

Tablica 2.1: Ocjena rizika

U tablici 2.1 promatramo ponašanje krvnog tlaka (normalan ili povišen), obzirom na tip osobe (Faktor 1), te razinu kolesterola (Faktor 2). U ovom slučaju možemo promatrati

za svaki faktor izglede povišenog krvnog tlaka ovisno o faktoru. Dakle imamo:

$$\begin{aligned}
\omega(\text{tip A}) &= \frac{n_{211}+n_{212}}{n_{111}+n_{112}}\\
\omega(\text{tip B}) &= \frac{n_{221}+n_{222}}{n_{121}+n_{122}}\\
\omega(\text{normalan kolesterol}) &= \frac{n_{211}+n_{221}}{n_{111}+n_{121}}\\
\omega(\text{povišen kolesterol}) &= \frac{n_{212}+n_{222}}{n_{112}+n_{122}}.
\end{aligned}$$

No, ono što je zanimljivije promatrati je izgled povišenog krvnog tlaka u kombinaciji oba faktora, odnosno:

$$\begin{aligned}
\omega(\text{tip A i normalan kolesterol}) &= \frac{n_{211}}{n_{111}}\\
\omega(\text{tip A i povišen kolesterol}) &= \frac{n_{212}}{n_{112}}\\
\omega(\text{tip B i normalan kolesterol}) &= \frac{n_{221}}{n_{121}}\\
\omega(\text{tip B i povišen kolesterol}) &= \frac{n_{222}}{n_{122}}.
\end{aligned}$$

Na isti način možemo promatrati i omjere izgleda. Dakle, možemo promatrati omjere izgleda povišenog tlaka po tipu osobe, po razini kolesterola, te u kombinaciji oba faktora. Tada imamo:

$$\begin{aligned}
OR(\text{tip osobe}) &= \frac{\omega(\text{tip A})}{\omega(\text{tip B})} = \frac{\frac{n_{211}+n_{212}}{n_{111}+n_{112}}}{\frac{n_{221}+n_{222}}{n_{121}+n_{122}}} = \frac{(n_{211}+n_{212})(n_{121}+n_{122})}{(n_{111}+n_{112})(n_{221}+n_{222})}\\
OR(\text{razina kolesterola}) &= \frac{\omega(\text{normalan kolesterol})}{\omega(\text{povišen kolesterol})} = \frac{\frac{n_{211}+n_{221}}{n_{111}+n_{121}}}{\frac{n_{212}+n_{222}}{n_{112}+n_{122}}} = \frac{(n_{211}+n_{221})(n_{112}+n_{122})}{(n_{111}+n_{121})(n_{212}+n_{222})}\\
OR(\text{tip A i razina kolesterola}) &= \frac{\omega(\text{tip A i normalan kolesterol})}{\omega(\text{tip A i povišen kolesterol})} = \frac{\frac{n_{211}}{n_{111}}}{\frac{n_{212}}{n_{112}}} = \frac{n_{211}n_{112}}{n_{111}n_{212}}\\
OR(\text{tip B i razina kolesterola}) &= \frac{\omega(\text{tip B i normalan kolesterol})}{\omega(\text{tip B i povišen kolesterol})} = \frac{\frac{n_{221}}{n_{121}}}{\frac{n_{222}}{n_{122}}} = \frac{n_{221}n_{122}}{n_{121}n_{222}}
\end{aligned}$$

$$OR(\text{normalan kolesterol i tip}) = \frac{\omega(\text{normalan kolesterol i tip A})}{\omega(\text{normalan kolesterol i tip B})} = \frac{\frac{n_{211}}{n_{111}}}{\frac{n_{221}}{n_{121}}} = \frac{n_{211}n_{121}}{n_{111}n_{221}}$$
$$OR(\text{povišen kolesterol i tip}) = \frac{\omega(\text{povišen kolesterol i tip A})}{\omega(\text{povišen kolesterol i tip B})} = \frac{\frac{n_{212}}{n_{112}}}{\frac{n_{222}}{n_{122}}} = \frac{n_{212}n_{122}}{n_{112}n_{222}}.$$

Pokažimo primjer računanja izgleda te omjera izgleda u sljedećoj tablici.

Faktor 1	Faktor 2	Normalan	Povišen	Ukupno
A	Normalan	716	79	795
	Povišen	207	25	232
B	Normalan	819	67	886
	Povišen	186	22	208
	Ukupno	1928	193	2121

Tablica 2.2: Primjer računanja ocjena rizika

Izračunajmo najprije izglede da tip A ima povišen tlak, da tip B ima povišen tlak, da osobe s normalnim kolesterolom imaju povišen tlak, te da osobe s povišenim kolesterolom imaju povišen tlak, te omjere tih izgleda.
Dakle, osobe tipa A imaju $\omega(\text{tip A}) = \frac{79+25}{716+207} = \frac{104}{923} = 0.11$ puta veću vjerojatnost da imaju povišen krvni tlak nego da nemaju. Dok, osobe tipa B imaju $\omega(\text{tip B}) = \frac{67+22}{819+186} = \frac{89}{1005} = 0.09$ puta veću vjerojatnost da imaju povišen krvni tlak nego da nemaju. Osobe s normalnim kolesterolom imaju $\omega(\text{normalan kolesterol}) = \frac{79+67}{716+819} = \frac{146}{1535} = 0.10$ puta veću vjerojatnost da imaju povišen krvni tlak nego da nemaju. Dok osobe s povišenim kolesterolom imaju $\omega(\text{povišen kolesterol}) = \frac{25+22}{207+186} = \frac{47}{393} = 0.12$ puta veću vjerojatnost da imaju povišen krvni tlak nego da nemaju. Dakle, osobe tipa A imaju $OR(\text{tip osobe}) = \frac{104 \cdot 1005}{923 \cdot 89} = 1.27$ puta veći izgled od povišenog krvnog tlaka od osobe tipa B. Dok osobe s normalnim kolesterolom imaju $OR(\text{razina kolesterola}) = \frac{146 \cdot 393}{1535 \cdot 47} = 0.80$ puta veći izgled od povišenog krvnog tlaka od osoba s povišenim kolesterolom.

Zatim izračunajmo izglede povišenog krvnog tlaka u kombinaciji faktora, te omjere izgleda.
Dakle, osobe tipa A s normalnim kolesterolom imaju $\omega(\text{tip A i normalan kolesterol}) = \frac{79}{716} = 0.11$ puta veću vjerojatnost da imaju povišen tlak nego da nemaju. Dok osobe tipa A s povišenim kolesterolom imaju $\omega(\text{tip A i povišen kolesterol}) = \frac{25}{207} = 0.12$ puta veću vjerojatnost da imaju povišen tlak nego da nemaju. Osobe tipa B s normalnim kolesterolom imaju $\omega(\text{tip B i normalan kolesterol}) = \frac{67}{819} = 0.08$ puta veću vjerojatnost da

imaju povišen tlak nego da nemaju. Dok osobe tipa B s povišenim kolesterolom imaju ω(tip B i povišen kolesterol) $= \frac{22}{186} = 0.12$ puta veću vjerojatnost da imaju povišen tlak nego da nemaju.
Dakle, osoba tipa A s normalnim kolesterolom ima OR(tip A i razina kolesterola) $= \frac{79 \cdot 207}{716 \cdot 25} = 0.91$ puta veći izgled od povišenog krvnog tlaka od osobe s povišenim kolesterolom. Dok osoba tipa B s normalnim kolesterolom ima OR(tip B i razina kolesterola) $= \frac{67 \cdot 186}{819 \cdot 22} = 0.69$ puta veći izgled od povišenog krvnog tlaka od osobe s povišenim kolesterolom. Također, osoba s normalnim kolesterolom koja je tip A ima OR(normalan kolesterol i tip) $= \frac{79 \cdot 819}{716 \cdot 67} = 1.35$ puta veći izgled od povišenog krvnog tlaka od osobe s normalnim kolesterolom koja je tip B. Dok osoba s povišenim kolesterolom koja je tip A ima OR(povišen kolesterol i tip) $= \frac{25 \cdot 186}{207 \cdot 22} = 1.02$ puta veći izgled od povišenog krvnog tlaka od osobe s povišenim kolesterolom koja je tip B.

2.2 Model složene logističke regresije

Pretpostavimo da imamo k kovarijata $X_1, X_2, \ldots, X_k$, $k \in \mathbf{N}$, te da svaka kovarijata ima zadanih I različitih vrijednosti $x_{1j}, x_{2j}, \ldots, x_{Ij}$, $j = 1, \ldots, k$. Za svaku od tih vrijednosti $x_{i\cdot} := (x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, \ldots, I$ imamo uzorak od n_i nezavisnih Bernoullijevih slučajnih varijabli $Y'_1, Y'_2, \ldots, Y'_{n_i}$. Dakle, $Y_i = Y'_1 + Y'_2 + \cdots + Y'_{n_i} \sim B(p(x_{i\cdot}), n_i)$. Prisjetimo se da je $E(Y_i) = m_i = n_i p(x_{i\cdot})$ i da su izgledi $\frac{p(x_{i\cdot})}{1-p(x_{i\cdot})}$. Sada logistička regresija određuje linearnu strukturu za logaritmirane izglede s

$$P(x_{i\cdot}) := \operatorname{logit}(p(x_{i\cdot})) \equiv \log\left(\frac{p(x_{i\cdot})}{1 - p(x_{i\cdot})}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \qquad i = 1, 2, \ldots, I. \quad (2.1)$$

Procjena parametara modela

U odnosu na jednostavnu logističku regresiju, gdje smo imali parametre β_0 i β_1, kod složene logističke regresije imamo parametre $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. Sada jednadžbu (2.1) možemo transformirati u

$$p_{i1} := p(x_{i\cdot}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}$$
$$p_{i2} := p(x_{i\cdot}) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}.$$

Primijetimo da je $p_{i1} + p_{i2} = 1$.

Neka je y_i opažena vrijednost od Y_i. Uz notaciju $p_{i1} \equiv p_i, p_{i2} \equiv (1 - p_i), n_{i1} \equiv y_i, n_{i2} \equiv n_i - y_i$, odnosno $n_i = n_{i1} + n_{i2}$, dobivamo funkciju vjerodostojnosti parametara β:

$$
\begin{aligned}
L(\beta) = \prod_{i=1}^{I} \Bigg[& \frac{n_i!}{\prod_{j=1}^{2} n_{ij}!} \left\{ \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \right\}^{n_{i1}} \\
& \cdot \left\{ \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \right\}^{n_{i2}} \Bigg] \\
= \prod_{i=1}^{I} & \left[\frac{n_i!}{\prod_{j=1}^{2} n_{ij}!} \{p_{i1}\}^{n_{i1}} \{p_{i2}\}^{n_{i2}} \right].
\end{aligned} \tag{2.2}
$$

Logaritmiranjem funkcije vjerodostojnosti dobivamo:

$$
LL(\beta_0, \beta_1) = \sum_{i=1}^{I} \left[n_{i1} \log(p_{i1}) + n_{i2} \log(p_{i2}) \right] + \sum_{i=1}^{I} \log \frac{n_i!}{\prod_{j=1}^{2} n_{ij}!}. \tag{2.3}
$$

Sada ćemo (2.2) maksimizirati tako da parcijalno deriviramo p_{i1}, odnosno, p_{i2} po $\beta_0, \beta_1, \ldots, \beta_k$, te to izjednačimo s 0.

Za početak izračunajmo:

$$
\begin{aligned}
\frac{\partial p_{i1}}{\partial \beta_0} &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2} = p_{i1} p_{i2} \\
\frac{\partial p_{i1}}{\partial \beta_1} &= \frac{x_{i1} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2} = x_{i1} p_{i1} p_{i2} \\
&\vdots \\
\frac{\partial p_{i1}}{\partial \beta_k} &= \frac{x_{ik} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2} = x_{ik} p_{i1} p_{i2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial p_{i2}}{\partial \beta_0} &= \frac{-\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2} = -p_{i1} p_{i2} \\
\frac{\partial p_{i2}}{\partial \beta_1} &= \frac{-x_{i1} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2} = -x_{i1} p_{i1} p_{i2} \\
&\vdots \\
\frac{\partial p_{i2}}{\partial \beta_k} &= \frac{-x_{ik} \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{(1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))^2} = -x_{ik} p_{i1} p_{i2}
\end{aligned}
$$

Sada vrijedi:

$$\frac{\partial LL}{\partial \beta_0} = \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_{i1}}p_{i1}p_{i2} - n_{i2}\frac{1}{p_{i2}}p_{i1}p_{i2}\right] = \sum_{i=1}^{I}\left[n_{i1}p_{i2} - n_{i2}p_{i1}\right] = \sum_{i=1}^{I}\left[n_{i1} - n_i p_{i1}\right]$$

$$\frac{\partial LL}{\partial \beta_1} = \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_{i1}}x_i p_{i1}p_{i2} - n_{i2}\frac{1}{x_i}p_{i2}p_{i1}p_{i2}\right] = \sum_{i=1}^{I} x_i\left[n_{i1}p_{i2} - n_{i2}p_{i1}\right] = \sum_{i=1}^{I} x_{i1}\left[n_{i1} - n_i p_{i1}\right]$$

$$\vdots$$

$$\frac{\partial LL}{\partial \beta_k} = \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_{i1}}x_i p_{i1}p_{i2} - n_{i2}\frac{1}{x_i}p_{i2}p_{i1}p_{i2}\right] = \sum_{i=1}^{I} x_i\left[n_{i1}p_{i2} - n_{i2}p_{i1}\right] = \sum_{i=1}^{I} x_{ik}\left[n_{i1} - n_i p_{i1}\right]$$

Dakle, dobivamo:

$$\begin{aligned}
\sum_{i=1}^{I}\left[n_{i1} - n_i p_{i1}\right] &= 0 \\
\sum_{i=1}^{I} x_{i1}\left[n_{i1} - n_i p_{i1}\right] &= 0 \\
&\vdots \\
\sum_{i=1}^{I} x_{ik}\left[n_{i1} - n_i p_{i1}\right] &= 0.
\end{aligned} \tag{2.4}$$

Rješavanjem nelinearnog sustava (2.4) po $\beta_0, \beta_1, \ldots, \beta_k$ dobivamo MLE procjenu parametara $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, iz čega slijedi procjena vjerojatnosti

$$\hat{p}_i = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_k x_{ik}\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_k x_{ik}\right)}.$$

I u ovom slučaju može se pokazati da je funkcija $(\beta_0, \beta_1, \ldots \beta_k) \to LL(\beta_0, \beta_1, \ldots \beta_k)$ konkavna. Budući da je $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k\right)$ stacionarna točka od LL, ona je ujedno i točka maksimuma.

Također, pomoću procjene vjerojatnosti dobivamo i procjenu očekivanja, odnosno $\hat{m}_{ij} = n_i \hat{p}_{ij}$, za $i = 1, 2, \ldots, I$, $j = 1, 2$.

Analogno jednostavnom modelu logističke regresije, i u slučaju polinomne logističke regresije možemo testirati adekvatnost modela u usporedbi sa saturiranim modelom pomoću statistike G^2:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) = 2\sum_{i=1}^{I}\left[n_{i1}\log\frac{\tilde{p}_i}{\hat{p}_i} + (n_i - y_i)\log\frac{(1-\tilde{p}_i)}{(1-\hat{p}_i)}\right], \tag{2.5}$$

gdje je $\tilde{p}_{ij} = \frac{n_{ij}}{n_i}$.

Također, možemo testirati razliku punog modela s ($\beta_i \neq 0, i = 1, 2, \ldots, k$) i bez varijabli prediktora ($\beta_i = 0$, za sve $i = 1, 2, \ldots, k$), te tada koristimo statistiku D^2:

$$\begin{aligned} D^2 &= G^2\left[\text{ model bez prediktora}\right] - G^2\left[\text{model s } k \text{ prediktora}\right] \\ &= 2\log\left[\frac{L(\text{ model bez prediktora})}{L(\text{saturirani model})}\right] - \left\{2\log\left[\frac{L(\text{model s } k \text{ prediktora})}{L(\text{saturirani model})}\right]\right\} \\ &= 2\log\left[\frac{L(\text{ model bez prediktora})}{L(\text{model s } k \text{ prediktora})}\right] \end{aligned} \tag{2.6}$$

Alternativno, test istih hipoteza možemo sprovesti i pomoću Pearsonove statistike:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{2}\frac{\left(n_{ij} - n_i\hat{p}_{ij}\right)^2}{n_i\hat{p}_{ij}} = \sum_{i=1}^{I}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i\,(1-\hat{p}_i)}. \tag{2.7}$$

Testiranje hipoteza o parametarima modela

Kao i u slučaju jednostavnog modela logističke regresije, i sada možemo pomoću Waldovog testa testirati značajnost modela. Analogno 1.10 za testiranje značajnosti kovarijate i, $i = 1, 2, \ldots, k$, tj. ako su hipoteze:

$$\begin{aligned} H_0 &: \beta_i = 0 \\ H_1 &: \beta_i \neq 0 \end{aligned}$$

testna je statistika:

$$W_i = \frac{\hat{\beta}_i}{\sqrt{\hat{\text{var}}(\beta_i)}}.$$

Također, možemo testirati set linearnih kombinacija više kovarijata. Tada je testna statistika:

$$K^T\hat{\beta} - K^T\beta_0 \overset{H_0}{\sim} AN(0, K^T I^{-1}((\hat{\beta}))K),$$

gdje I očekivana informacijska matrica, dok je K matrica koja daje linearnu kombinaciju više kovarijata (naprimjer, ukoliko testiramo linearnu kombinaciju $\beta_2 + \beta_3 + \beta_5 = 0$, tada će stupac matrice K biti oblika $[0, 1, 1, 0, 1, 0, \ldots, 0]^T$).

2.3 Metode odabira modela

Postoje tri metode za odabir modela: *unaprijed (engl. forward), unatrag (engl. backward) i koračna (engl. stepwise)*.
Zajedničko ovim metodama je to što u svakom koraku koriste kriterij za uključivanje ili isključivanje prediktora u model. Vjerojatno najpopularniji kriterij za logističke regresijske modele je *p-vrijednost* testa značajnosti za određeni prediktor. U bilo kojoj fazi razvoja modela, prediktori koji imaju *p-vrijednost* koja prelazi određenu razinu isključeni su iz modela. Isto tako, prediktori koji imaju *p-vrijednost* manju od neke razine mogu se zadržati u modelu.

Metoda odabira unaprijed

Metoda odabira unaprijed počinje modelom koji sadrži samo konstantu, a zatim dodaje prvi prediktor, odnosno najznačajniju varijablu. U svakom sljedećem koraku, svaki prediktor koji nije u modelu procjenjuje se, odnosno testira, zajedno s prethodno prihvaćenim prediktorom je li za uključivanje u model. Najznačajniji od preostalih prediktora dodaje se u model, sve dok je *p-vrijednost* ispod određene razine. Uobičajeno je postaviti vrijednost te razine na 0.05, 0.10 ili 0.15.

Najveći nedostatak metode unaprijed je što dodavanjem novog prediktora možemo jednu ili više već uključenih prediktora učiniti neznačajnim, no ne možemo ga izbaciti iz modela. [6]

Metoda odabira unatrag

Alternativni pristup forward metodi je *metoda odabira unatrag*. Metoda odabira unatrag počinje s modelom koji sadrži sve prediktore od interesa, odnosno sve prediktore čija je $p-vrijednost < 0.25$. Svi prediktori čija je $p-vrijednost > 0.25$ isključeni su inicijalno iz modela. Sada testiramo prediktore koji su u modelu, te iz modela isključujemo onaj koji je najmanje značajan. Taj postupak ponavljamo sve dok ne dobijemo *p-vrijednost* modela ispod određene razine. Uobičajeno je, kao i u metodi unaprijed, postaviti tu razinu na 0.05, 0.10 ili 0.15.

Također, i metoda unatrag ima nedostatke. Najveći nedostatak je što jednom izbačen prediktor ne možemo više vratiti u model. [5]

Koračna metoda odabira

Koračna metoda odabira je kombinacija metoda unaprijed i unatrag. U svakom koraku omogućuje uključivanje novih prediktora ili isključivanje već uključenih prediktora. Odnosno, koračna metoda naizmjenično isključuje najmanje značajan prediktor te uključuje

najznačaniji prediktor koji u tom koraku nije u modelu. U svakom koraku računaju se *p-vrijednosti* koje moraju biti ispod određene razine. Kao i kod metode unatrag, odnosno, unaprijed, i u ovoj metodi najčešće vrijednosti razine su 0.05, 0.10 ili 0.15.

Prediktori su obično povezani, dok mi unaprijed ne znamo kako će uključivanje ili isključivanje jednog od njih utjecati na značajnost drugog prediktora. No, upravo ova metoda eliminira mogućnost da smo mogli nekim isključenim prediktorom značajno pridonijeti modelu ukoliko isključimo neki drugi prediktor, te je to razlog zašto se ona najčešće koristi. [7]

Kriteriji za određivanje modela

U prethodnim metodama *p-vrijednost* može se odrediti obzirom na z, t ili *Waldovu* statistiku ili izračunavanjem omjera vjerodostojnosti. *P-vrijednosti* bazirane na tim statistikama, tj. metodama, su obično slične, ali prednost ima statistički test omjera vjerodostojnosti. Iako takvo testiranje zahtjeva više vremena, rezultati su vrijedni toga. Međutim, ako netko ima 100 varijabli, test omjera vjerodostojnosti može potrajati dulje nego što je prihvatljivo. U takvim situacijama korištenje z, t ili *Waldove* statistike kao osnovu za izračun *p-vrijednosti* može biti efikasnija taktika.

Osim korištenja *p-vrijednosti* kao kriterija za određivanje modela, postoje još dva kriterija: *Akaikeov (AIC) i Bayesov (BIC) kriterij*.

Formula za određivanje AIC indeksa je:

$$AIC = -2LL + 2(k + 1),$$

gdje je LL log-vjerodostojnost modela koji testiramo, a k broj prediktora u modelu.

Ukoliko imamo mali broj opservacija n ($\frac{n}{k} < 40$), tada koristimo formulu:

$$AIC = -2LL + 2(k + 1) + (2k(k - 1)/(n - k)),$$

gdje je LL log-vjerodostojnost modela koji testiramo, k broj prediktora u modelu, a n je broj opservacija.

Dakle, kada uspoređujemo dva modela pomoću Akaikeovog kriterija, reći ćemo da je bolji onaj model čiji je AIC indeks manji.

Formula za određivanje BIC indeksa je:

$$BIC = -2LL + k\log(n),$$

gdje je LL log-vjerodostojnost modela koji testiramo, k broj prediktora u modelu i n je broj opservacija.

Dakle, kada uspoređujemo dva modela pomoću Bayesovog kriterija, reći ćemo da je bolji onaj model čiji je BIC indeks manji.

Općenito, usporedimo li AIC i BIC indeks za isti model BIC indeks će biti veći.

Poglavlje 3

Polinomni logistički regresijski model

Polinomna logistička regresija razlikuje se od složene logističke regresije u tome što zavisna varijabla (*varijabla odziva*), može imati više od dva moguća ishoda. I u slučaju polinomne logističke regresije, varijablu odziva želimo opisati pomoću više nezavisnih varijabli, tj. više *prediktora*.

3.1 Omjer izgleda

Kao i u slučaju složene logističke regresije, izglede i omjere izgleda pokušat ćemo objasniti pomoću frekvencijske tablice. Radi jednostavnijeg prikaza koristit ćemo jedan kategorijski prediktor, te tri moguća ishoda varijable odziva.

Faktor	Nizak	Normalan	Povišen	Ukupno
Žena	n_{11}	n_{21}	n_{31}	$n_{\cdot 1}$
Muškarac	n_{12}	n_{22}	n_{32}	$n_{\cdot 2}$
Ukupno	$n_{1\cdot}$	$n_{2\cdot}$	$n_{3\cdot}$	$n_{\cdot\cdot}$

Tablica 3.1: Ocjena rizika

U tablici 3.1 promatramo ponašanje krvnog tlaka (nizak, normalan ili povišen), obzirom na spol.

Izglede možemo promatrati na tri načina.

Prvi način je sličan izgledima u prethodnim modelima. Odnosno, izgled za razvoj bolesti računamo kao omjer vjerojatnosti za razvoj bolesti i vjerojatnosti da ju ne razvijemo ($\omega = \frac{p}{1-p}$). Na ovaj način odabiremo kategoriju varijable odziva za koju želimo izračunati izgled. Na primjer, želimo li računati izgled povišenog krvnog tlaka po spolu to računamo

na način:

$$\omega(\text{žena}) = \frac{n_{31}}{n_{.1} - n_{31}}$$
$$\omega(\text{muškarac}) = \frac{n_{32}}{n_{.2} - n_{32}}.$$

Analogno dobivamo izgled za normalan, odnosno nizak tlak:

$$\omega_{norm}(\text{žena}) = \frac{n_{21}}{n_{.1} - n_{21}}$$
$$\omega_{norm}(\text{muškarac}) = \frac{n_{22}}{n_{.2} - n_{22}}$$
$$\omega_{low}(\text{žena}) = \frac{n_{11}}{n_{.1} - n_{11}}$$
$$\omega_{low}(\text{muškarac}) = \frac{n_{12}}{n_{.2} - n_{12}}.$$

Drugi način bi bio promatrati tzv. izglede u odnosu na susjedne kategorije. Dakle, možemo promatrati omjer da će žena, odnosno muškarac, imati nizak krvni tlak u odnosu da ima normalan krvni tlak:

$$\omega_S(\text{žena}) = \frac{n_{11}}{n_{21}}$$
$$\omega_S(\text{muškarac}) = \frac{n_{12}}{n_{22}}.$$

Ili možemo promatrati omjer da žena, odnosno muškarac, ima povišen krvni tlak u odnosu da ima normalan krvni tlak:

$$\omega_S(\text{žena}) = \frac{n_{21}}{n_{31}}$$
$$\omega_S(\text{muškarac}) = \frac{n_{22}}{n_{32}}.$$

Takve omjere zvat ćemo *izgledima u odnosu na susjedne kategorije*.

Za **treći način** potrebno je odabrati baznu kategoriju. Ako je p_0 vjerojatnost bazne kategorije, odgovarajući omjer za svaku od ostalih i kategorija računamo s $\omega_{B_i} = \frac{p_i}{p_0}$. Na primjer, neka je bazna kategorija normalan krvni tlak. Na ovaj način promatramo omjer da

će žena, odnosno muškarac, imati nizak krvni tlak u odnosu da ima normalan krvni tlak, te da žena, odnosno muškarac, ima povišen krvni tlak u odnosu da ima normalan krvni tlak. To računamo na način:

$$\begin{aligned}\omega_{B_{low}}(\text{žena}) &= \frac{n_{11}}{n_{21}}\\ \omega_{B_{high}}(\text{žena}) &= \frac{n_{31}}{n_{21}}\\ \omega_{B_{low}}(\text{muškarac}) &= \frac{n_{12}}{n_{22}}\\ \omega_{B_{high}}(\text{muškarac}) &= \frac{n_{32}}{n_{22}}.\end{aligned}$$

Takve omjere zvat ćemo *izgledima u odnosu na baznu kategoriju*.

Analogno tome, možemo na tri načina promatrati i omjere izgleda.

Prvi način bi bio promatrati omjere izgleda između žena i muškaraca po tipu krvnog tlaka (nizak, normalan, odnosno povišen). To činimo na način:

$$\begin{aligned}OR(\text{nizak}) &= \frac{\frac{n_{31}}{n_{\cdot 1}-n_{31}}}{\frac{n_{32}}{n_{\cdot 2}-n_{32}}} = \frac{n_{31}(n_{\cdot 2}-n_{32})}{n_{32}(n_{\cdot 1}-n_{31})}\\ OR(\text{normalan}) &= \frac{\frac{n_{21}}{n_{\cdot 1}-n_{21}}}{\frac{n_{22}}{n_{\cdot 2}-n_{22}}} = \frac{n_{21}(n_{\cdot 2}-n_{22})}{n_{22}(n_{\cdot 1}-n_{21})}\\ OR(\text{povišen}) &= \frac{\frac{n_{11}}{n_{\cdot 1}-n_{11}}}{\frac{n_{12}}{n_{\cdot 2}-n_{12}}} = \frac{n_{11}(n_{\cdot 2}-n_{12})}{n_{12}(n_{\cdot 1}-n_{11})}.\end{aligned}$$

Drugi način bi bio promatrati omjere izgleda između žena i muškaraca po susjednim tipovima tlaka, odnosno:

$$\begin{aligned}OR_S(\text{nizak i normalan}) &= \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}\\ OR_S(\text{normalan i povišen}) &= \frac{\frac{n_{21}}{n_{31}}}{\frac{n_{22}}{n_{32}}} = \frac{n_{21}n_{32}}{n_{31}n_{22}}.\end{aligned}$$

Na **treći način** možemo promatrati omjere izgleda između žena i muškaraca za nizak, odnosno povišen, krvni tlak ovisno o baznom (normalnom) krvnom tlaku. To računamo:

$$OR_B(\text{nizak i normalan}) = \frac{\frac{n_{11}}{n_{21}}}{\frac{n_{12}}{n_{22}}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$
$$OR_B(\text{povišen i normalan}) = \frac{\frac{n_{31}}{n_{21}}}{\frac{n_{32}}{n_{22}}} = \frac{n_{31}n_{22}}{n_{21}n_{32}}.$$

Promotrimo to na primjeru.

Faktor	Nizak	Normalan	Povišen	Ukupno
Žena	28	51	16	95
Muškarac	19	44	42	105
Ukupno	47	95	58	200

Tablica 3.2: Primjer računanja ocjene rizika

Prvi način nam daje izglede da žene, odnosno muškarci, imaju nizak, normalan, odnosno povišen krvni tlak. Također, dobivamo omjere izgleda niskog, normalnog, te povišenog krvnog tlaka između žena i muškaraca.
U ovom slučaju možemo reći da žene imaju $\omega(\text{žena}) = \frac{28}{95-28} = \frac{28}{67} = 0.4$ puta veću vjerojatnost da imaju nizak krvni tlak, nego da imaju normalan ili povišen. Dok muškarci imaju $\omega(\text{muškarac}) = \frac{19}{105-19} = \frac{19}{86} = 0.22$ puta veću vjerojatnost da imaju nizak krvni tlak, nego da imaju normalan ili povišen. Dakle, žene imaju $OR = \frac{0.42}{0.22} = 1.91$ puta veći izgled da imaju nizak krvni tlak u odnosu na muškarce. Analogno računamo izglede, te omjere izgleda, za normalan i povišen krvni tlak.

Drugi način nam daje izgled da žena ima nizak krvni tlak u odnosu da žena ima normalan krvni tlak, te izgled da žena ima normalani krvni tlak u odnosu da žena ima povišen krvni tlak. Analogno računamo i za muškarce. Također, možemo izračunati omjere izgleda između žena i muškaraca.
Dakle, žene imaju $\omega_S(\text{žena}) = \frac{51}{16} = 3.19$ puta veću vjerojatnost da imaju normalan krvni tlak, nego da imaju povišen krvni tlak. Dok muškarci imaju $\omega_S(\text{muškarac}) = \frac{44}{42} = 1.05$ puta veću vjerojatnost da imaju normalan krvni tlak, nego da imaju povišen krvni tlak. Dakle, žene imaju $OR_S = \frac{3.19}{1.05} = 3.04$ puta veći izgled da imaju normalan krvni tlak, nego da imaju povišen krvni tlak, u odnosu na muškarce. Analogno računamo izglede, te omjere izgleda, za nizak krvni tlak u odnosu na normalan krvni tlak.

Treći način nam daje izgled da žena ima nizak krvni tlak u odnosu da žena ima normalan krvni tlak, te izgled da žena ima povišen krvni tlak u odnosu da žena ima normalan krvni tlak. Analogno računamo i za muškarce. Također, možemo izračunati omjere izgleda između žena i muškaraca.
Dakle, žene imaju $\omega_B(\text{žena}) = \frac{16}{51} = 0.31$ puta veću vjerojatnost da imaju povišen krvni tlak, nego da imaju normalan krvni tlak. Dok muškarci imaju $\omega_B(\text{muškarac}) = \frac{42}{44} = 0.95$ puta veću vjerojatnost da imaju povišen krvni tlak, nego da imaju normalan krvni tlak. Dakle, žene imaju $OR_B = \frac{0.31}{0.95} = 0.33$ puta veći izgled da imaju povišen krvni tlak, nego da imaju normalan krvni tlak, u odnosu na muškarce. Analogno računamo izglede, te omjere izgleda, za nizak krvni tlak u odnosu na normalan krvni tlak.

3.2 Model polinomne logističke regresije

Pretpostavimo da imamo k kovarijata $X_1, X_2, \ldots, X_k$, $k \in \mathbf{N}$, te da svaka kovarijata ima zadanih I različitih vrijednosti $x_{1j}, x_{2j}, \ldots, x_{Ij}$, $j = 1, \ldots, k$. Za svaku od tih vrijednosti $x_{i\cdot} := (x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, \ldots, I$ imamo uzorak od n_i nezavisnih slučajnih kategorijalnih varijabli $Y'_1, Y'_2, \ldots, Y'_{n_i}$ gdje su

$$Y'_r \sim \begin{pmatrix} 1 & 2 & 3 & \ldots & m \\ p_1 & p_2 & p_3 & \ldots & p_m \end{pmatrix},$$

takvi da vrijedi $p_1 + p_2 + p_3 + \cdots + p_m = 1$, te je $m \in \mathbf{N}$.

Neka je $Y_{il} = \sum_{r=1}^{n_i} \mathbf{1}_{(Y'_r = l)} =$ broj opservacija (među njih n_i) koje su bile u kategoriji l, $l = 1, \ldots, m$. Iz toga dobivamo slučajan vektor s polinomijalnom distribucijom $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im}) \sim M(n_i, p_1, p_2, \ldots, p_m)$, $i = 1, 2, \ldots, I$, za koji je $\mathbf{P}(Y_{i1} = k_1, Y_{i2} = k_2, \ldots, Y_{im} = k_m) = \frac{n_i!}{k_1!k_2!\ldots k_m!} p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m}$.

Prisjetimo se da izglede možemo računati na tri načina:

$$\begin{aligned}
\omega_l &= \frac{p_l(x_{i\cdot})}{1 - p_l(x_{i\cdot})}, \quad i = 1, 2, \ldots, m, \\
\omega_{S_l} &= \frac{p_l(x_{i\cdot})}{p_{l+1}(x_{i\cdot})}, \quad l = 1, 2, \ldots, m-1, \\
\omega_{B_l} &= \frac{p_l(x_{i\cdot})}{p_B(x_{i\cdot})}, \quad l = 1, 2, \ldots, m, \quad l \neq B
\end{aligned}$$

Za svaku od metoda računanja izgleda logistička regresija određuje linearnu strukturu za logaritmirane izglede s

$$P_l(x_{i\cdot}) = \log(\omega_l) = \beta_{l0} + \beta_{l1} x_{i1} + \cdots + \beta_{lk} x_{ik}, \quad i = 1, 2, \ldots, I, \quad l = 1, 2, \ldots, m. \tag{3.1}$$

U nastavku, za računanje izgleda koristimo model bazne kategorije, odnosno,

$$\omega_{B_l} = \frac{p_l(x_{i\cdot})}{p_B(x_{i\cdot})}, \quad l = 1, 2, \ldots, m, l \neq B, \quad i = 1, 2, \ldots, I.$$

Procjena parametara modela

Pretpostavimo da je $m = 3$, odnosno $Y' \sim \begin{pmatrix} 1 & 2 & 3 \\ p_1 & p_2 & p_3 \end{pmatrix}$, tako da vrijedi $p_1 + p_2 + p_3 = 1$.

Neka je 2 bazna kategorija. Tada su logaritmirani izgledi:

$$\begin{aligned} \log \frac{p_1(x_{i\cdot})}{p_2(x_{i\cdot})} &= \beta_{10} + \beta_{11} x_{i1} + \cdots + \beta_{1k}(x_{ik}) \\ \log \frac{p_3(x_{i\cdot})}{p_2(x_{i\cdot})} &= \beta_{30} + \beta_{31} x_{i1} + \cdots + \beta_{3k}(x_{ik}). \end{aligned}$$

Označimo s $\beta_1 := (\beta_{10}, \beta_{11}, \ldots, \beta_{1k})$ i $\beta_3 := (\beta_{30}, \beta_{31}, \ldots, \beta_{3k})$, te s $\beta = (\beta_1, \beta_3)$.

Iz toga slijedi:

$$\begin{aligned} p_1(x_{i\cdot}) &= p_2(x_{i\cdot}) \exp(\beta_1^T x_{i\cdot}) \\ p_3(x_{i\cdot}) &= p_2(x_{i\cdot}) \exp(\beta_3^T x_{i\cdot}) \end{aligned} \tag{3.2}$$

Pa dobivamo:

$$\begin{aligned} p_{i1} := p_1(x_{i\cdot}) &= \frac{\exp(\beta_1^T x_{i\cdot})}{\exp(\beta_1^T x_{i\cdot}) + \exp(\beta_3^T x_{i\cdot}) + 1} \\ p_{i2} := p_2(x_{i\cdot}) &= \frac{1}{\exp(\beta_1^T x_{i\cdot}) + \exp(\beta_3^T x_{i\cdot}) + 1} \\ p_{i3} := p_3(x_{i\cdot}) &= \frac{\exp(\beta_3^T x_{i\cdot})}{\exp(\beta_1^T x_{i\cdot}) + \exp(\beta_3^T x_{i\cdot}) + 1} \end{aligned} \tag{3.3}$$

Primijetimo da je $p_{i1} + p_{i2} + p_{i3} = 1$.

Neka je y_i opažena vrijednost od Y_i. Označimo $p_{i3} \equiv 1 - p_{i1} - p_{i2}$, $n_{i1} \equiv y_{i1}$, $n_{i2} \equiv y_{i2}$, $n_{i3} \equiv n_i - y_{i1} - y_{i2}$, odnosno $n_i = n_{i1} + n_{i2} + n_{i3}$. Tada funkciju vjerodostojnosti parametara β možemo zapisati na sljedeći način:

$$
\begin{aligned}
L(\beta) = \prod_{i=1}^{I} \Bigg[& \frac{n_i!}{\prod_{j=1}^{3} n_{ij}!} \left\{ \frac{\exp(\beta_1^T x_{i\cdot})}{\exp(\beta_1^T x_{i\cdot}) + \exp(\beta_3^T x_{i\cdot}) + 1} \right\}^{n_{i1}} \\
& \cdot \left\{ \frac{1}{\exp(\beta_1^T x_{i\cdot}) + \exp(\beta_3^T x_{i\cdot}) + 1} \right\}^{n_{i2}} \\
& \cdot \left\{ \frac{\exp(\beta_3^T x_{i\cdot})}{\exp(\beta_1^T x_{i\cdot}) + \exp(\beta_3^T x_{i\cdot}) + 1} \right\}^{n_{i3}} \Bigg] \\
= \prod_{i=1}^{I} & \left[\frac{n_i!}{\prod_{j=1}^{3} n_{ij}!} \{p_{i1}\}^{n_{i1}} \{p_{i2}\}^{n_{i2}} \{p_{i3}\}^{n_{i3}} \right].
\end{aligned}
\tag{3.4}
$$

Logaritmiranjem funkcije vjerodostojnosti dobivamo:

$$
LL(\beta) = \sum_{i=1}^{I} \left[n_{i1} \log(p_{i1}) + n_{i2} \log(p_{i2}) + n_{i3} \log(p_{i3}) \right] + \sum_{i=1}^{I} \log \frac{n_i!}{\prod_{j=1}^{3} n_{ij}!}. \tag{3.5}
$$

Sada ćemo (3.5) maksimizirati tako da parcijalno deriviramo p_{i1}, p_{i2}, te p_{i3} po β_{10}, $\beta_{11}, \ldots, \beta_{1k}$, te po $\beta_{30}, \beta_{31}, \ldots \beta_{3k}$, te to izjednačimo s 0.

Jednostavno je dobiti parcijalne derivacije, a iz njih će slijediti:

$$
\begin{aligned}
\frac{\partial LL(\beta)}{\partial \beta_{10}} &= \sum_{i=1}^{I} \left[n_{i1} \frac{1}{p_{i1}} p_{i1} (p_{i2} + p_{i3}) + n_{i2} \frac{1}{p_{i2}} (-p_{i1} p_{i2}) + n_{i3} \frac{1}{p_{i3}} (-p_{i1} p_{i3}) \right] = \sum_{i=1}^{I} \left[n_{i1} - n_i p_{i1} \right] \\
\frac{\partial LL(\beta)}{\partial \beta_{11}} &= \sum_{i=1}^{I} \left[n_{i1} \frac{1}{p_{i1}} x_{i1} p_{i1} (p_{i2} + p_{i3}) + n_{i2} \frac{1}{p_{i2}} x_{i1} (-p_{i1} p_{i2}) + n_{i3} x_{i1} \frac{1}{p_{i3}} (-p_{i1} p_{i3}) \right] \\
&= \sum_{i=1}^{I} x_{i1} \left[n_{i1} - n_i p_{i1} \right] \\
&\vdots \\
\frac{\partial LL(\beta)}{\partial \beta_{1k}} &= \sum_{i=1}^{I} \left[n_{i1} \frac{1}{p_{i1}} p_{i1} x_{ik} (p_{i2} + p_{i3}) + n_{i2} \frac{1}{p_{i2}} x_{ik} (-p_{i1} p_{i2}) + n_{i3} \frac{1}{p_{i3}} x_{ik} (-p_{i1} p_{i3}) \right] \\
&= \sum_{i=1}^{I} x_{ik} \left[n_{i1} - n_i p_{i1} \right]
\end{aligned}
$$

$$\begin{aligned}
\frac{\partial LL(\beta)}{\partial \beta_{30}} &= \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_{i1}}(-p_{i1}p_{i3}) + n_{i2}\frac{1}{p_{i2}}(-p_{i2}p_{i3}) + n_{i3}\frac{1}{p_{i3}}p_{i3}\,(p_{i1}+p_{i2})\right] = \sum_{i=1}^{I}\left[n_{i3} - n_i p_{i3}\right] \\
\frac{\partial LL(\beta)}{\partial \beta_{31}} &= \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_i}x_{i1}(-p_{i1}p_{i3}) + n_{i2}\frac{1}{p_{i2}}x_{i1}(-p_{i2}p_{i3}) + n_{i3}\frac{1}{p_{i3}}x_{i1}p_{i3}\,(p_{i1}+p_{i2})\right] \\
&= \sum_{i=1}^{I} x_{i1}\left[n_{i3} - n_i p_{i3}\right] \\
&\vdots \\
\frac{\partial LL(\beta)}{\partial \beta_{3k}} &= \sum_{i=1}^{I}\left[n_{i1}\frac{1}{p_i}x_{ik}(-p_{i1}p_{i3}) + n_{i2}\frac{1}{p_{i2}}x_{ik}(-p_{i2}p_{i3}) + n_{i3}\frac{1}{p_{i3}}x_{ik}p_{i3}\,(p_{i1}+p_{i2})\right] \\
&= \sum_{i=1}^{I} x_{ik}\left[n_{i3} - n_i p_{i3}\right]
\end{aligned}$$

Dakle, dobivamo:

$$\begin{aligned}
\sum_{i=1}^{I}\left[n_{i1} - n_i p_{i1}\right] &= 0 & \sum_{i=1}^{I}\left[n_{i3} - n_i p_{i3}\right] &= 0 \\
\sum_{i=1}^{I} x_{i1}\left[n_{i1} - n_i p_{i1}\right] &= 0 & \sum_{i=1}^{I} x_{i1}\left[n_{i3} - n_i p_{i3}\right] &= 0 \\
&\vdots & &\vdots \\
\sum_{i=1}^{I} x_{ik}\left[n_{i1} - n_i p_{i1}\right] &= 0 & \sum_{i=1}^{I} x_{ik}\left[n_{i3} - n_i p_{i3}\right] &= 0.
\end{aligned} \tag{3.6}$$

Rješavanjem nelinearnog sustava (3.6) po $\beta_{10}, \beta_{11}, \ldots, \beta_{1k}$, te po $\beta_{30}, \beta_{31}, \ldots \beta_{3k}$ dobivamo MLE procjenu parametara $\hat{\beta}_{10}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{1k}$, $\hat{\beta}_{30}, \hat{\beta}_{31}, \ldots \hat{\beta}_{3k}$, odnosno procjenu parametara $\hat{\beta}_1 = (\hat{\beta}_{10}, \hat{\beta}_{11}, \ldots, \hat{\beta}_{1k})$ i $\hat{\beta}_3 = (\hat{\beta}_{30}, \hat{\beta}_{31}, \ldots \hat{\beta}_{3k})$ iz čega slijedi procjena vjerojatnosti:

$$\hat{p}_{i1} = \frac{\exp(\hat{\beta}_1^T x_{i\cdot})}{\exp(\hat{\beta}_1^T x_{i\cdot}) + \exp(\hat{\beta}_3^T x_{i\cdot}) + 1}$$
$$\hat{p}_{i2} = \frac{1}{\exp(\hat{\beta}_1^T x_{i\cdot}) + \exp(\hat{\beta}_3^T x_{i\cdot}) + 1}$$
$$\hat{p}_{i3} = \frac{\exp(\hat{\beta}_3^T x_{i\cdot})}{\exp(\hat{\beta}_1^T x_{i\cdot}) + \exp(\hat{\beta}_3^T x_{i\cdot}) + 1}$$

Može se pokazati da je funkcija $(\beta_1, \beta_3) \to LL(\beta_1, \beta_3)$ konkavna. Budući da je $\left(\hat{\beta}_1, \hat{\beta}_3\right)$ stacionarna točka od LL, ona je ujedno i točka maksimuma.

U slučaju kada je $m \geq 4$ postupak za dobivanje procjenjenih vjerojatnosti $\hat{p}_{i1}$, $\hat{p}_{i2}$, …, $\hat{p}_{i,m-1}$ je analogan.

Neka je $m_{ij} = \mathbf{E}Y_{ij} = n_i p_{ij}$. Pomoću procjene vjerojatnosti dobivamo i procjenu očekivanja, odnosno $\hat{m}_{ij} = n_i \hat{p}_{ij}$, za $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, m$.

Analogno jednostavnom modelu logističke regresije, i u slučaju polinomne logističke regresije možemo testirati adekvatnost modela u usporedbi sa saturiranim modelom pomoću statistike G^2:

$$\begin{aligned} G^2 &= 2 \sum_{i=1}^{I} \sum_{j=1}^{m} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right) \\ &= 2 \sum_{i=1}^{I} \left[n_{i1} \log \frac{\tilde{p}_{i1}}{\hat{p}_{i1}} + n_{i2} \log \frac{\tilde{p}_{i2}}{\hat{p}_{i2}} + \cdots + n_{im} \log \frac{\tilde{p}_{im}}{\hat{p}_{im}} \right], \end{aligned} \tag{3.7}$$

gdje je $\tilde{p}_{ij} = \frac{n_{ij}}{n_i}$.

Također, možemo testirati značajnost modela s prediktorima u odnosu na model bez prediktora, te tada koristimo statistiku D^2:

$$\begin{aligned} D^2 &= G^2 \left[\text{ model bez prediktora} \right] - G^2 \left[\text{model s prediktorima} \right] \\ &= 2 \log \left[\frac{L(\text{ model bez prediktora})}{L(\text{saturirani model})} \right] - \left\{ 2 \log \left[\frac{L(\text{model s prediktorima})}{L(\text{saturirani model})} \right] \right\} \\ &= 2 \log \left[\frac{L(\text{ model bez prediktora})}{L(\text{model s prediktorima})} \right] \end{aligned} \tag{3.8}$$

Alternativno, test istih hipoteza možemo sprovesti i pomoću Pearsonove statistike:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{m} \frac{\left(n_{ij} - n_i \hat{p}_{ij}\right)^2}{n_i \hat{p}_{ij}}. \tag{3.9}$$

Testiranje hipoteza o parametarima modela

Kao i u slučaju jednostavnog modela logističke regresije, i sada možemo pomoću Waldovog testa testirati značajnost modela. Ukoliko želimo testirati značajnost kovarijate j, $j = 1, 2, \ldots, k$, tj. hipoteze su:

$$H_0 : \forall i = 1, \ldots, m-1 \qquad \beta_{ij} = 0,$$
$$H_1 : \exists i, i \in \{1, \ldots, m-1\} \quad \beta_{ij} \neq 0,$$

tada će testna statistika 1.10 biti oblika:

$$W = (A^T(\hat{\beta}))^T (A^T I^{-1}(\hat{\beta}) A)^{-1} (A^T(\hat{\beta})) \tag{3.10}$$

gdje je (za $m = 3$) očekivana informacijska matrica dana s $I(\beta) = \mathbf{X}^T \mathbf{V} \mathbf{X}$, $\mathbf{X} = \begin{bmatrix} X & 0 \\ 0 & X \end{bmatrix}$, $V = \begin{bmatrix} \mathbf{V_{11}} & \mathbf{V_{12}} \\ \mathbf{V_{12}} & \mathbf{V_{22}} \end{bmatrix}$, $\mathbf{V_{11}} = diag(n_i p_{i1}(1 - p_{i1}))$, $\mathbf{V_{12}} = diag(-n_i p_{i1} p_{i3})$, $\mathbf{V_{22}} = diag(n_i p_{i3}(1 - p_{i3}))$, $A^T = \begin{bmatrix} e_j^T & 0 \\ 0 & e_j^T \end{bmatrix}$, gdje je e_j^T element kanonske baze za $\mathbf{R}^{k+1}$.

3.3 Metode odabira modela

Metode odabira modela jednake su metodama odabira u slučaju složenog logističkog regresijskog modela.

Poglavlje 4

Primjena polinomne logističke regresije

4.1 Uvod

Primjena polinomne logističke regresije bazirat će se na pacijentima s Crohnovom bolešću. Crohnova bolest je upalna bolest probavnog sustava koja pogađa odrasle i djecu. Način na koji se dijanosticira je kolonoskopski pregled te različiti testovi krvi ili stolice. Obzirom da kolonoskopija dodatno iscrpljuje pacijente, kako bi se pacijentima olakšalo liječenje koriste se testovi u praćenju aktivnosti bolesti.

Najčešći parametri koji se promatraju u liječenju su razine sedimentacije krvi, C-reaktivnog proteina (CRP), te kalprotektina. CRP se dobiva iz uzorka krvi, dok se kalprotektin dobiva iz uzorka stolice.

Kako bi se pacijentima olakšalo liječenje i kako bi se smanjila potreba za kolonoskopskim pregledima potrebno je ustanoviti koje kovarijate će najbolje opisivati aktivnost bolesti, odnosno stupanj bolesti. Dosadašnja istraživanja su pokazala kako su CRP i kalprotektin dobri prediktori aktivnosti bolesti [4], stoga očekujemo da će neki od tih prediktora biti značajan i u našoj analizi.

Analizu radimo u programu R.

4.2 Podaci

Podaci su o 61 pacijenatu koji su trenutno uključeni u kliničko istraživanje o kroničnim upalama crijeva (dobiveni uz dozvolu liječnika dr. Marka Brinara). Svakom pacijentu izmjerena je razina kalprotektina, CRP-a, sedimentacije, indeks tjelesne mase (engl. *body mass index - BMI*) te znamo kojeg je spola i koliko ima godina. Također, postoje tri razine aktivnosti bolesti: 1 (remisija, tj. trenutna neaktivnost bolesti), 2 (blaga upala), 3 (jaka upala). Aktivnost bolesti određena je prema kliničkim indeksima, odnosno na osnovi kliničke slike koja uključuje biopsiju crijeva, kolonoskopsku pretragu, krvnu sliku,

učestalost stolice, bolova, itd. U ovom poglavlju pokušat ćemo opisati aktivnost bolesti pomoću prethodno navedenih varijabli.

Opisna statistika

Koeficijenti korelacije

Na početku provjerimo postoji li korelacija među prediktorima. U R-u ćemo to provjeriti koristeći funkciju *cor()*.

Koeficijenti korelacije među prediktorima su:

	Dob	**Spol**	**Sedimentacija**	**CRP**	**Kalprotektin**	**BMI**
Dob	1	0.180	-0.117	-0.208	-0.209	0.236
Spol	0.180	1	-0.219	-0.161	-0.101	-0.398
Sedimentacija	-0.117	-0.219	1	0.580	0.554	-0.239
CRP	-0.208	-0.161	0.580	1	0.644	-0.097
Kalprotektin	-0.209	-0.101	0.553	0.644	1	-0.199
BMI	0.236	-0.398	-0.239	-0.097	-0.199	1

Tablica 4.1: Koeficijenti korelacije

S obzirom na to da niti jedan koeficijent nije značajno velik/malen, odnosno $> 0.98/< -0.98$, ne možemo zaključiti da postoje korelirani prediktori. No, možda postoji prediktor koji bismo mogli opisati linearnom kombinacijom preostalih prediktora. Sada ćemo provjeriti postoji li korelacija između jednog prediktora te linearne kombinacije preostalih prediktora. Sada je koeficijent korelacije [3]:

$$\hat{R}_i = \sqrt{\frac{\mathbf{v}_{2i}^T V_{ii} \mathbf{v}_{2i}}{v_{ii}}}, \qquad i = 1, \ldots, 6,$$

gdje je $\mathbf{v}_{2i}$ kovarijacijski vektor i-tog i preostalih prediktora, V_{ii} kovarijacijska matrica preostalih prediktora, te v_{ii} varijanca i-tog prediktora.

Kod u R-u:

```
X<-matrix(0,6,6)
for(i in 1:6){
X[i,i]=var(pod2[,i])
for(j in i:6){
X[i,j]=cov(pod2[,i], pod2[,j])
```

```
X[j,i]=cov(pod2[,i], pod2[,j])}}
R<-numeric(6)
for(k in 1:6){
v<-X[k,k]
vv<-numeric(5)
br=1
j=1
while(br<6 & j<7){
if(j==k){j=j+1}
else{
vv[br]<-X[k,j]
j=j+1
br=br+1}}
Vpom<-X[,-k]
V<-Vpom[-k,]
Vinv<-solve(V)
R[k]<-sqrt(t(vv)%*%Vinv%*%vv/v)}
```

U sljedećoj tablici prikazane su vrijednosti koeficijenata korelacije između pojedinog prediktora te linearne kombinacije preostalih.

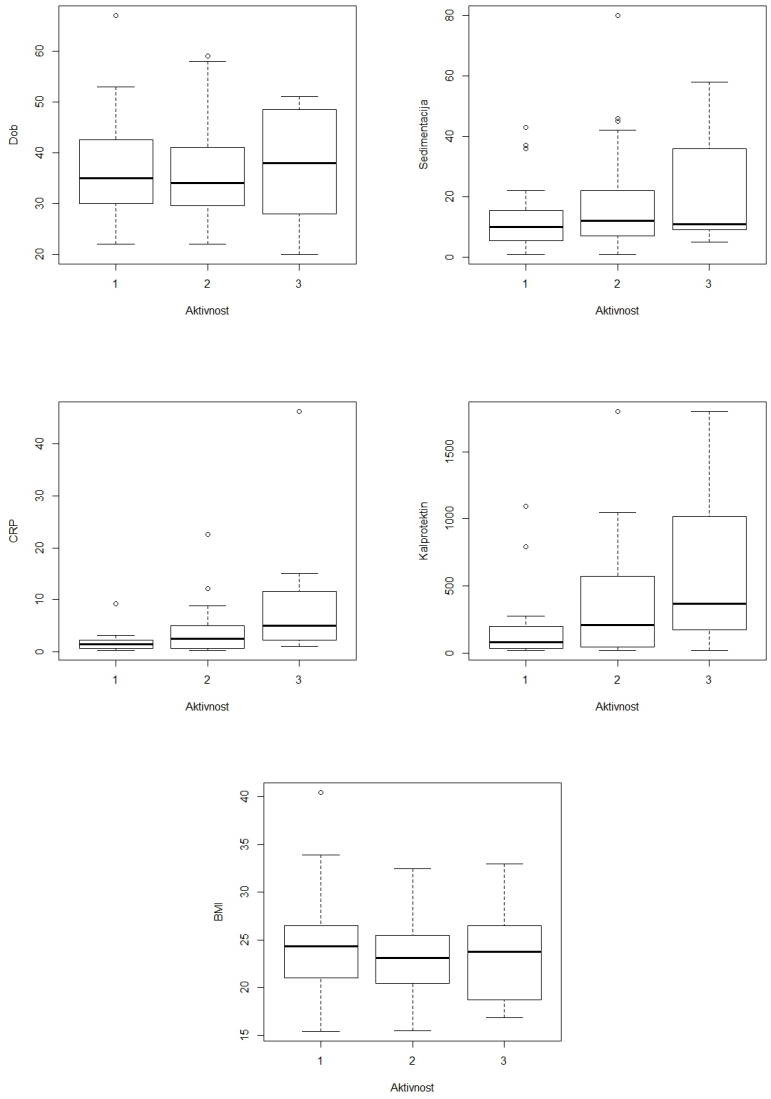
Dob	Spol	Sedimentacija	CRP	Kalprotektin	BMI
0.352	0.567	0.686	0.765	0.696	0.459

Tablica 4.2: Koeficijenti korelacije navedenog s linearnom kombinacijom ostalih prediktora

S obzirom na to da ni sada niti jedan koeficijent nije značajno velik/malen, odnosno $> 0.98/< -0.98$ ne možemo reći kako neki prediktor može biti opisan linearnom kombinacijom preostali. Dakle, sve ćemo prediktore uzeti u obzir u daljnjoj analizi.

Grafički prikaz

Prikažimo grafički vrijednosti kontinuiranih prediktora po razinama aktivnosti bolesti. Kontinuirani prediktori su razina sedimentacije, kalprotektina, CRP-a, BMI, te dob. Ovime želimo vidjeti postoji li pravilno ponašanje, primjerice povećanje vrijednosti, povećanjem kategorije aktivnosti bolesti.



Slika 4.1: Boxplot prikaz kontinuiranih prediktora

Uočimo kako raspon vrijednosti sedimentacije, CRP-a, te kalprotektina raste s obzirom na stupanj aktivnosti bolesti. Također, uočimo kako se i prosječna vrijednost CRP-a, te kalprotektina povećava s obzirom na stupanj aktivnosti bolesti.

4.3 Izgledi i omjeri izgleda

Za početak ćemo prikazati podatke o aktivnosti bolesti po spolu.

Spol	Remisija	Blaga upala	Jaka upala	Ukupno
Muškarac	11	12	2	25
Žena	12	19	5	36
Ukupno	23	31	7	61

Tablica 4.3: Ocjena rizika

Promotrimo sada izglede u odnosu na baznu kategoriju. Za baznu kategoriju odaberimo kategoriju blage upale, odnosno kategoriju stupnja 2, te dobivamo:

- Izgledi da je bolest u fazi remisije u odnosu na blagu upalu:

$$\omega_B(\text{muškarac}) = \tfrac{11}{12} = 0.92$$
$$\omega_B(\text{žena}) = \tfrac{12}{19} = 0.63$$

- Izgledi da je bolest u fazi jake upale u odnosu na blagu upalu:

$$\omega_B(\text{muškarac}) = \tfrac{2}{12} = 0.17$$
$$\omega_B(\text{žena}) = \tfrac{5}{19} = 0.26$$

Iz toga slijedi da su omjeri izgleda između muškaraca i žena da iz blage upale pređe u remisiju, odnosno iz blage upale u fazu jaku upale:

$$OR_B = \frac{0.92}{0.63} = 1.46$$
$$OR_B = \frac{0.17}{0.26} = 0.65$$

Iz prikazanog možemo zaključiti kako muškarci imaju veće izglede prelaska iz blage upale u remisiju, dok žene imaju veće izglede prelaska iz blage upale u jaku upalu.

4.4 Procjena parametara punog modela

Prisjetimo se, koristeći izglede u odnosu na baznu kategoriju, tj. kategoriju blage upale, dobivamo jednadžbe:

$$\log \frac{p_1(x_{i\cdot})}{p_2(x_{i\cdot})} = \beta_{10} + \beta_{11}x_{i1} + \cdots + \beta_{16}(x_{i6})$$
$$\log \frac{p_3(x_{i\cdot})}{p_2(x_{i\cdot})} = \beta_{30} + \beta_{31}x_{i1} + \cdots + \beta_{36}(x_{i6}),$$

gdje su $x_{\cdot 1}$, $x_{\cdot 2}$, $x_{\cdot 3}$, $x_{\cdot 4}$, $x_{\cdot 5}$, $x_{\cdot 6}$ redom dob, spol, sedimentacija, CRP, kalprotektin, te BMI.

Kako bi dobili procijenjene parametre $\beta_{10}, \beta_{11}, \cdots, \beta_{16}, \beta_{30}, \beta_{31}, \cdots, \beta_{36}$, u R-u koristimo funkciju *multinom()*, tj. za procjenu parametara punog modela koristimo:

```
modelp<-multinom(Aktivnost ~ ., data=podaci)
```

Vrijednosti procjenjenih parametara $\beta = (\beta_1, \beta_3)$ su:

Slobodan član	**Dob**	**Spol**	**Sedimentacija**	**CRP**	**Kalprotektin**	**BMI**
β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
-1.06	0.01	-0.89	-0.01	-0.09	-0.001	0.07
β_{30}	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
-3.78	0.05	-0.48	-0.04	0.21	0.0003	0.01

Tablica 4.4: Procjena parametara β

Pomoću procijenjenih koeficijenata $\hat{\beta}_1$ i $\hat{\beta}_3$ možemo dobiti procijenjene vjerojatnosti $\hat{p}_{i1}$, $\hat{p}_{i2}$, te $\hat{p}_{i3}$, $\forall i$, $i = 1, 2, \ldots, 61$.

Testovi značajnosti procjenjenih parametara β

Sada kada smo dobili procijenjene koeficijente β_1 i β_3 testirajmo njihove značajnosti, odnosno hipoteze:

$$H_0: \quad \beta_{ij} = 0$$
$$H_1: \quad \beta_{ij} \neq 0$$

Za svaki ij, $i \in 1, 3$, $j \in 1, 2, \ldots, 6$, hipoteze testiramo pomoću Waldove statistike. Rezultati su prikazani u sljedećoj tablici.

	Slobodan član	**Dob**	**Spol**	**Sedim.**	**CRP**	**Kalpr.**	**BMI**
	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
Wald	-0.560	0.282	-1.160	-0.355	-0.569	-1.111	1.003
pv	0.575	0.778	0.246	0.723	0.569	0.267	0.316
	β_{30}	β_{31}	β_{32}	β_{33}	β_{34}	β_{35}	β_{36}
Wald	-1.232	1.027	-0.384	-0.863	1.184	0.249	0.057
pv	0.218	0.305	0.701	0.388	0.236	0.803	0.954

Tablica 4.5: Waldove statistike i p-vrijednosti

Iz ovih vrijednosti Waldove statistike s $df = 1$ stupnjom slobode te pripadnih p-vrijednosti, na bilo kojoj razumnoj razini značajnosti, primjerice 5%, niti za jedan ij, $i \in 1, 3$, $j \in 1, 2, \ldots, 6$, ne bismo mogli odbaciti nultu hipotezu u korist alternativne hipoteze.

S obzirom na to da test značajnost pojedinog koeficijenta nije pokazao kako postoji značajan koeficijent, no, možda postoji značajna kovarijata, odnosno prediktor. Dakle, sada ćemo provjeriti postoji li kovarijata koja ima značajan utjecaj u našem modelu, odnosno, testiramo hipoteze:

$$\begin{aligned} H_0 &: \quad \beta_{1j} = 0 \ \ \& \ \ \beta_{3j} = 0 \\ H_1 &: \quad \exists\, \beta_{ij} \neq 0, \quad i \in \{1, 3\} \end{aligned}$$

za $j = 1, 2, \ldots, 6$. Za testiranje ovih hipoteza koristimo Waldovu statistiku (za kod u R-u vidi chapter 4.4) te su rezultati prikazani u sljedećoj tablici:

	Dob	**Spol**	**Sedimentacija**	**CRP**	**Kalprotektin**	**BMI**
Wald	0.961	1.142	0.725	1.642	0.873	0.916
pv	0.619	0.565	0.696	0.440	0.646	0.633

Tablica 4.6: Waldove statistike i p-vrijednosti

Na osnovi dobivenih vrijednosti Waldove statistike s $df = 2$ stupnja slobode te pripadne p-vrijednosti, niti za jednu kovarijatu, na razumnoj razini značajnosti, primjerice 5%, ne možemo odbaciti nultu hipotezu u korist alternativne hipoteze. Dakle, ne možemo reći kako neka kovarijata ima značajan utjecaj u našem modelu.

Test adekvatnosti punog modela

Kako prethodni testovi nisu pokazali da postoji značajna kovarijata, sada ćemo testirati adekvatnost punog modela u usporedbi s modelom bez kovarijata. Dakle, pomoću G^2 statistike testirat ćemo hipoteze:

$$H_0: \quad \forall i \in \{1,3\}, j \in \{1,2,\ldots,6\} \quad \beta_{ij} = 0$$
$$H_1: \quad \exists i \in \{1,3\}, j \in \{1,2,\ldots,6\} \quad \beta_{ij} \neq 0$$

Za početak procijenimo koeficijente β modela bez kovarijata.

β_{10}	β_{30}
-0.298	-1.488

Tablica 4.7: Procjena parametara β modela bez kovarijata

Usporedimo adekvatnost modela bez kovarijata i punog modela. Dakle, hipoteze su:

H_0 : model bez kovarijata je dovoljan
H_1 : puni model je dovoljan

U R-u test ovih hipoteza provodimo pomoću funkcije *lrtest*:

```
lrtest(modelpr, modelp)
```

Dobivamo $G^2 = 16.071$ s $df = 12$ stupnjeva slobode za koju je $p - vrijednost = 0.188$. No, ni u ovom slučaju ne možemo odbaciti nultu hipotezu u korist alternativne na razumnoj razini značajnosti, primjerice 5%.

Usporediti model sa svim kovarijatama i model bez kovarijata možemo i koristeći Akaikeov kriterij. Vrijednosti Akaikeovog kriterija za pripadne modele prikazani su u sljedećoj tablici.

Model bez kovarijata	Model sa svim kovarijatama
121.144	129.073

Tablica 4.8: Vrijednosti AIC kriterija

Koristeći i Akaikeov kriterij, reći ćemo kako je model bez kovarijata dovoljan, odnosno nemamo kovarijatu za koju bismo mogli reći kako značajno opisuje zavisnu varijablu, tj. aktivnost bolesti.

Iako smo na ovim podacima kao rezultat polinomne regresijske analize dobili da niti jedan prediktor značajno ne opisuje stupanj aktivnosti bolesti, želimo naglasiti kako općenito ne možemo isto zaključiti. Dakle, istu analizu trebalo bi ponoviti na većem skupu podataka.

Prilog 1: Kod u R-u za računanje Waldove statistike

```
x<-read.csv(file="Crohn_zadnji.csv", header=T, sep=",")
x<-data.matrix(x, rownames.force = NA)
for(i in 1:61){
x[i,1]<-1}
N<-matrix(0,61,7)
X1<-rbind(x,N)
X2<-rbind(N,x)
X<-cbind(X1,X2)
beta <- as.vector(t(coef(modelp)))
beta1<-beta[1:7]
beta2<-beta[8:14]
p_i1<-numeric(61)
p_i2<-numeric(61)
p_i3<-numeric(61)
for(i in 1:61){
y<-c(1,0,0,0,0,0,0)
y[2]=podaci$Dob[i]
y[3]=podaci$Spol[i]
y[4]=podaci$SE[i]
y[5]=podaci$CRP[i]
y[6]=podaci$KALPRO[i]
y[7]=podaci$BMI[i]
p_i1[i]=exp(y%*%beta1)/(exp(y%*%beta1)+exp(y%*%beta2)+1)
p_i2[i]=1/(exp(y%*%beta1)+exp(y%*%beta2)+1)
p_i3[i]=exp(y%*%beta2)/(exp(y%*%beta1)+exp(y%*%beta2)+1)
}
V_11<-diag(p_i1*(1-p_i1))
V_12<-diag(-p_i1*p_i3)
```

```
V_22<-diag(p_i3*(1-p_i3))
V1<-cbind(V_11,V_12)
V2<-cbind(V_12,V_22)
V<-rbind(V1,V2)
I<-t(X)%*%V%*%X
Is<-solve(I)
#dob
A<-rbind(c(0,1,0,0,0,0,0,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,0,0,1,0,0,0,0,0))
wald<-t(A %*% beta) %*% solve(A %*% Is %*% t(A))%*%(A %*% beta)
wald
pchisq(wald, 2, lower=FALSE)
#spol
A<-rbind(c(0,0,1,0,0,0,0,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,0,0,0,1,0,0,0,0))
wald<-t(A %*% beta) %*% solve(A %*% Is %*% t(A))%*%(A %*% beta)
wald
pchisq(wald, 2, lower=FALSE)
#SE
A<-rbind(c(0,0,0,1,0,0,0,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,0,0,0,0,1,0,0,0))
wald<-t(A %*% beta) %*% solve(A %*% Is %*% t(A))%*%(A %*% beta)
wald
pchisq(wald, 2, lower=FALSE)
#CRP
A<-rbind(c(0,0,0,0,1,0,0,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,0,0,0,0,0,1,0,0))
wald<-t(A %*% beta) %*% solve(A %*% Is %*% t(A))%*%(A %*% beta)
wald
pchisq(wald, 2, lower=FALSE)
#Kalp
A<-rbind(c(0,0,0,0,0,1,0,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,0,0,0,0,0,0,1,0))
wald<-t(A %*% beta) %*% solve(A %*% Is %*% t(A))%*%(A %*% beta)
wald
pchisq(wald, 2, lower=FALSE)
#BMI
A<-rbind(c(0,0,0,0,0,0,1,0,0,0,0,0,0,0),
c(0,0,0,0,0,0,0,0,0,0,0,0,0,1))
```

```
wald<-t(A %*% beta) %*% solve(A %*% Is %*% t(A))%*%(A %*% beta)
wald
pchisq(wald, 2, lower=FALSE)
```

Bibliografija

[1] Charles E. McCulloch, Shayle R.Searle. *Generalized, Linear, and Mixed Models*. A Wiley-Interscience Publication, Kanada, 2001, 148-149.

[2] Ronald Christensen. *Log-Linear Models and Logistic Regression*. Springer, New York, 1997.

[3] Martin Bilodean, David Brenner. *Theory of Multivariate Statistics*. Springer, New York, 1999, 109.

[4] Maria Cappello, Gaetano Cristian Morreale. *The Role of Laboratory Test in Crohn's Disease*. Clin Med Insights Gastroenterol, SAD, 2016.

[5] Backward selection, `https://www.stat.ubc.ca/ rollin/teach/643w04/lec/node42.html` (kolovoz 2018.).

[6] Forward selection, `https://www.stat.ubc.ca/ rollin/teach/643w04/lec/node41.html` (kolovoz 2018.).

[7] Stepwise selection, `https://www.stat.ubc.ca/ rollin/teach/643w04/lec/node43.html` (kolovoz 2018.).

Sažetak

Logistička regresija korisna je metoda u istraživačkim problemima u kojima želimo opisati zavisnu varijablu s dvije ili više kategorija pomoću nezavisnih varijabli. Područja u kojima možemo koristiti logističku regresiju su raznolika (biomedicina, osiguranje, itd.), te upravo to logističku regresiju čini praktičnom.

U ovom radu logističku regresiju objašnjavamo od najjednostavnijeg modela, odnosno kada pomoću jedne nezavisne varijable opisujemo dihotomnu zavisnu varijablu. Zatim preko složenog logističkog modela u kojem imamo k nezavisnih varijabli pomoću kojih želimo opisati dihotomnu zavisnu varijblu dolazimo do polinomnog logističkog modela u kojem imamo k nezavisnih varijabli pomoću kojih želimo opisati kategorijsku (m kategorija) zavisnu varijablu.

Glavna zadaća logističke regresije, složene i polinomne, je pronaći koje nezavisne varijable najbolje opisuju zavisnu. Naravno, moramo imati na umu da će uvijek puni model logističke regresije bolje opisivati zavisnu varijablu, no od interesa je pronaći što je moguće manje varijabli koji će biti dobri prediktori zavisne varijable.

Polinomnu logističku regresiju provodimo na podacima o 61-om bolesniku s Crohnovom bolesti. Iako je analiza u ovom radu pokazala kako niti jedan od prediktora (dob, spol, sedimentacija, CRP, kalprotektin, te BMI) značajno ne opisuju stupanj aktivnosti Crohnove bolesti (remisija, blaga upala, te jaka upala) preporučit ćemo ponovnu analizu na većem uzorku.

Summary

Logistic regression is useful method in researches which analyze categorical (two or more categories) dependent variable using independent variables. The areas where we can use logistic regression are diverse (biomedicine, insurance, etc.), and this makes logistic regression practical.

This thesis explain the logistic regression from the simplest model, ie when using an independent variable we describe a dichotomous dependent variable. Then, through a multiple logistic model in which we have k independent variables by which we want to describe a dichotomous dependent variable, we come to a polynomial logistic model in which we have k independent variables by which we want to describe the categorical (m category) dependent variables.

The main task of logistic regression, multiple and polynomial, is to find which independent variables best describe the dependent variable. Of course, we must keep in mind that a full logistic regression model will always better describe the dependent variable, but it is of interest to find as less as possible variables which will be good predictors.

Polynomial logistic regression is performed on data of 61 patients with Crohn's disease. Although the analysis in this thesis did not get any predictor (age, gender, sedimentation, CRP, calprotectin, and BMI) which significantly describe the degree of Crohn's disease activity (remission, mild and severe disease), we would recommend re-analysis on a larger sample.

Životopis

Ivana Barišić Lučić, rođena je 25. kolovoza 1992. u Zagrebu, gdje je pohađala Osnovnu školu dr. Vinka Žganca, a potom II gimnaziju. Svoje obrazovanje 2011. godine nastavlja na Matematičkom odsjeku Prirodslovno-matematičkog fakulteta upisom preddiplomskog studija Matematika. Po završetku preddiplomskog studija, 2016. godine, upisuje diplomski studij Matematička statistika, na istom odsjeku.

Tijekom završne godine studija zapošljava se u osiguravajućoj kući Allianz Zagreb d.d. kao asistent glavnom pricing aktuaru u sektoru neživotnih osiguranja. Nakon završetka studija, prelazi na poziciju poslovnog analitičara u sektoru za motorna vozila.