10 Zobecněný lineární model (GLM)

Průvodce studiem

Zobecněný lineární model (GLM) projděte spíše pro celkový přehled, snažte se však důkladně pochopit část o logistické regresi včetně řešeného příkladu. Na tuto kapitolu počítejte nejméně se čtyřmi hodinami studia s tím, že se k probírané látce budete ještě vracet po pochopení dalších souvislostí.

Zobecněný lineární model (Generalized Linear Model) označovaný jako GLM nebo GLIM, je podrobněji popsán v knize McCullagh a Nelder [19] a do základních pojmů tohoto modelu nyní nahlédneme.

Zobecněný lineární modle (GLM) zahrnuje:

- lineární regresi
- různé modely analýzy rozptylu (ANOVA)
- logistickou regresi
- probitový model
- log-lineární model (multinomický model pro četnosti v analýze mnohorozměrných kontingenčních tabulek)

Označení datových struktur a význam symbolů v GLM:

- Pozorování závisle proměnné (response) je sloupcový vektor náhodných veličin a je typu $(n \times 1)$, tedy $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T$.
- Pokud z kontextu je zřejmé, že se jedná o libovolný prvek vektoru $\mathbf{y}$, bude označován y (netučná kursiva bez indexu)
- Matice $\mathbf{X}$ nezávislých proměnných (regresorů, covariates) je typu $(n \times p)$. Její j-tý sloupec označujeme $\mathbf{x}_j$.
- Vektor parametrů je $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_p]^T$.
- Náhodná složka modelu má vektor středních hodnot $\mathrm{E}(\mathbf{Y}) = \mu$ typu $(n \times 1)$ a kovarianční matici $\mathrm{cov}(\mathbf{Y})$

Lineární prediktor $\boldsymbol{\eta}$ je systematická složka v lineárním modelu, tedy

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \mathbf{x}_j \beta_j$$

kde $\mathbf{x}_j$ je j – tý sloupec matice $\mathbf{X}$, tj. vektor $(n \times 1)$.

Každá složka vektoru $\mathbf{Y}$ má rozdělení z exponenciální rodiny rozdělení s hustotou

$$f_Y(y, \theta, \phi) = \exp\{(y\theta - b(\theta))/a(\theta) + c(y, \theta)\}, \tag{32}$$

kde θ a ϕ jsou parametry rozdělení, $a(.), b(.), c(.)$ jsou funkce, jejichž tvar je dán konkrétním rozdělením z exponenciální rodiny. Pokud ϕ je známé, je rov.(32) hustota rozdělení z exponenciální rodiny a má *kanonický* parametr θ. Pokud ϕ je neznámé, pak to může, ale nemusí být dvouparametrické rozdělení z exponenciální rodiny.

Např. pro normální rozdělení, $Y \sim N(\mu, \sigma^2)$

$$f_Y(y, \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(y-\mu)^2/2\sigma^2\} =$$
$$= \exp\{(y\mu - \mu^2/2)/\sigma^2 - \tfrac{1}{2}(y^2/\sigma^2 + \ln(2\pi\sigma^2))\},$$

takže v tomto případě

$$\theta = \mu, \quad \phi = \sigma^2,$$

$$a(\phi) = \phi \quad b(\theta) = \theta^2/2, \quad c(y, \phi) = -\frac{1}{2}(y^2/\sigma^2 + \ln(2\pi\sigma^2))$$

Logaritmus věrohodnostní funkce (při známém y funkce parametrů θ, ϕ) je

$$l(\theta, \phi, y) = \ln f_Y(y, \theta, \phi)$$

Střední hodnota a rozptyl mohou pak být určeny ze vztahů známých pro věrohodnostní funkci:

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0 \qquad \mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right) + \mathrm{E}\left(\frac{\partial l}{\partial \theta}\right)^2 = 0$$

Věrohodnostní funkci pro jedno pozorování z jakéhokoli rozdělení z exponenciální rodiny lze zapsat

$$l(\theta, \phi, y) = (y\theta - b(\theta))/a(\phi) + c(y, \phi)$$

Derivace podle kanonického parametru jsou $\partial l/\partial\theta = (y - b'(\theta))/a(\phi)$ a $\partial^2 l/\partial\theta^2 = b''(\theta)/a(\phi)$.

Položíme-li je rovny nule, můžeme vyjádřit střední hodnotu a rozptyl vysvětlované náhodné veličiny:

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = (\mu - b'(\theta))/a(\phi) = 0 \Rightarrow \mathrm{E}(Y) = \mu = b'(\theta)$$

a

$$\mathrm{var}(Y) - \frac{b''(\theta)}{a(\phi)} = 0 \Rightarrow \mathrm{var}(Y) = b''(\theta)a(\phi).$$

Střední hodnota je funkcí pouze kanonického parametru θ. Rozptyl náhodné veličiny Y je součinem dvou funkcí. Jedna, $b''(\theta)$ závisí pouze na kanonickém parametru rozdělení (a tedy na střední hodnotě μ náhodné veličiny Y). Nazývá se varianční funkce (variance function) a můžeme ji zapsat jako funkci střední hodnoty, $V(\mu)$. Druhá funkce v součinu je nezávislá na kanonickém parametru θ a závisí jen na ϕ.

Funkce $a(\phi)$ má obvykle tvar $a(\phi) = \phi/w$. Parametr ϕ se nazývá disperzní parametr a je konstantní pro všechna pozorování, w je apriorně známá váha pozorování, může být různá pro různá pozorování.

GLM tedy dovoluje i jiná rozdělení z exponenciální rodiny než jen normální užité v klasickém modelu. Další zobecnění je v tom, že lineární prediktor nemusí vysvětlovat jen (podmíněnou) střední hodnotu náhodné veličiny, ale i nějakou její funkci. Vztah mezi lineárním prediktorem η a střední hodnotou μ vysvětlované náhodné veličiny Y vyjadřuje spojovací funkce (link):

$$\eta_i = g(\mu_i)$$

Spojovací funkce $g(.)$ může být jakákoli monotónní diferencovatelná funkce.

V klasickém lineárním modelu je spojovací funkcí identita, tj. $\eta = \mu$. U jiných modelů se užívají zejména tyto spojovací funkce:

logit	$\eta = \ln\{\mu/(1-\mu)\}$	$0 < \mu < 1$
probit	$\eta = \Phi^{-1}(\mu)$	$0 < \mu < 1$
	$\Phi(.)$ je distribuční funkce	
	rozdělení N(0,1)	
komplementární		
log-log	$\eta = \ln\{-\ln(1-\mu)\}$	$0 < \mu < 1$
mocninové funkce	$\eta = \begin{cases} \mu^\lambda & \text{pro } \lambda \neq 0 \\ \ln\mu & \text{pro } \lambda = 0 \end{cases}$	$\mu > 0$

Jelikož střední hodnota $\mu = b'(\theta)$, je tedy jen funkcí kanonického parametru θ a spojovací funkce je monotónní, existuje inverzní funkce, kterou můžeme vyjádřit jako funkci střední hodnoty, $\theta(\mu)$. Některá rozdělení mají zvláštní spojovací funkce, kdy kanonický parametr rozdělení je roven lineárnímu prediktoru, $\theta = \eta$. Tyto spojovací funkce se nazývají kanonické (canonical link). Pro běžná rozdělení jsou kanonickými následující spojovací funkce:

normální rozdělení,$N(\mu, \sigma^2)$	$\eta = \mu$
Poissonovo, $P(\mu)$	$\eta = \ln\mu$
alternativní, $A(\pi)$	$\eta = \ln\{\pi/(1-\pi)\}$
gamma, $G(\mu, v)$	$\eta = \mu^{-1}$
inversní Gaussovo, $IG(\mu, \sigma^2)$	$\eta = \mu^{-2}$

V klasickém modelu se výstižnost modelu (těsnost proložení) vyjadřuje obvykle pomocí koeficientu determinace R^2, tj. jako podíl variability závislé veličiny vysvětlené modelem na celkové variabilitě.

$$R^2 = 1 - \frac{RSS}{\sum(y_i - \bar{y})^2}$$

Celková variabilita $\sum(y_i - \bar{y})^2$ odpovídá RSS lineárního modelu s jedním parametrem, jehož odhad je $b_0 = \bar{y}$. Pro takový model je

$$R^2 = 1 - \frac{\sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 0.$$

Pro model vysvětlující variabilitu veličiny y úplně je $RSS = 0$ a tedy je $R^2 = 1$.

Ve zobecněném lineárním modelu lze model (těsnost proložení) posuzovat analogicky. Uvažujme tak zvaný úplný model s n parametry, který by vysvětloval pozorované hodnoty y přesně, tzn. $y_i = \mu_i$. Jelikož můžeme kanonický parametr vyjádřit jako funkci střední hodnoty, $\theta(\mu)$, můžeme věrohodnostní funkci zapsat jako $l(\mathbf{y}, \phi, \mathbf{y})$. To je maximálně dosažitelná hodnota věrohodnostní funkce.

Pro model jen s jedním parametrem, kdy $\mu_i = konst$ (nulový model, obsahuje jen intercept), bychom dostali věrohodnostní funkci minimální hodnoty pro daná data. Dvojnásobek rozdílu mezi těmito věrohodnostními funkcemi je analogií k celkové variabilitě v klasickém modelu. Označíme-li odhad středních hodnot v modelu s p parametry jako $\hat{\boldsymbol{\mu}}$ a odhad kanonického parametru pro tento model jako $\hat{\theta} = \theta(\mu)$ a $\tilde{\theta} = \theta(\mathbf{y})$ a předpokládáme-li $a_i(\phi) = \phi/w_i$, pak dvojnásobek rozdílu věrohodnostních funkcí $l(\mathbf{y}, \phi, \mathbf{y})$ a $l(\hat{\mu}, \phi, \mathbf{y})$ je

$$\sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}/\phi = D(\mathbf{y}, \hat{\mu})/\phi.$$

$D(\mathbf{y}, \hat{\mu})/\phi$, která je funkcí pozorovaných dat, se nazývá deviance a je to analogie residuální sumy čtverců, RSS. Klasický lineární model je zvláštním případem zobecněného modelu, kdy spojovací funkce je identita a pak pro normálně rozdělenou náhodnou složku modelu je deviance rovna residuálnímu součtu čtverců, $D(\mathbf{y}, \hat{\mu})/\phi = \sum(y_i - \hat{\mu_i})^2 = RSS$

$D^*(\mathbf{y}, \hat{\mu}) = D(\mathbf{y}, \hat{\mu})/\phi$ je tzv. scaled deviance, je to deviance vyjádřená jako násobek disperzního parametru.

Ve zobecněném lineárním modelu je tedy cílem nalézt model, který zmenšuje celkovou devianci (úměrnou rozdílu logaritmů věrohodnostních funkcí mezi úplným modelem a nulovým modelem s jedním parametrem). Takový model může být vytvářen i postupně, mohou být do modelu zařazovány ty regresory, které nejvíce snižují devianci vzhledem k aktuálnímu modelu se zařazenými k parametry, tedy může být použit krokový (stepwise) postup pro vyhledávání regresního modelu. Regresory ve

zobecněném lineárním modelu mohou být i kvalitativní (faktory) a regresory mohou být i interakce (součiny) původních regresorů, takže pomocí zobecněného modelu je možné odhadovat parametry i složitých modelů analýzy rozptylu.

10.1 Logistická regrese

K logistickému regresnímu modelu dojdeme ze zobecněného lineárního modelu (GLM)

$$g[E(Y|\mathbf{x})] = \mathbf{x}^T\boldsymbol{\beta}, \tag{33}$$

ve kterém nějaká funkce g podmíněné střední hodnoty náhodné veličiny Y je vyjádřena jako lineární funkce vektoru regresorů $\mathbf{x}^T = (1, x_1, x_2, \ldots, x_s)$ s regresními koeficienty $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_s)$. Pokud má náhodná veličina Y alternativní rozdělení, tedy $Y \sim A(p)$, které má, jak známo, střední hodnotu $E(Y) = p$, a jako spojovací (t. zv. link) funkci ve zobecněném lineárním modelu zvolíme *logit*,

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{34}$$

dojdeme k logistickému regresnímu modelu

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{x}^T\boldsymbol{\beta}, \tag{35}$$

ve kterém logit podmíněné střední hodnoty je vyjádřen jako lineární funkce regresorů.

Parametry $\beta_0, \beta_1, \ldots, \beta_s$ regresního modelu (35) lze odhadovat metodou maximální věrohodnosti. Algoritmy pro nalezení těchto odhadů $b_0, b_1, \ldots, b_s$ jsou již řadu let implementovány v dostupných statistických programech. Logistický regresní model má poměrně snadnou a přímočarou interpretaci. Poměr $p/(1-p)$, tedy poměr pravděpodobnosti „úspěchu“ ku pravděpodobnosti „neúspěchu“, je v anglosaském světě označován jako *odds* a je zcela samozřejmě používán i mimo statistiku, např. při sázkách. Česká terminologie není ustálená, užívá se poměr šancí nebo sázkové riziko.

Nechť tedy

$$odds_0 = \frac{p_0}{1-p_0} \text{ při hodnotách regresorů } \mathbf{x} = \mathbf{x_0}$$

$$odds_1 = \frac{p_1}{1-p_1} \text{ při hodnotách regresorů } \mathbf{x} = \mathbf{x_1}$$

Poměr dvou *odds* je označován jako *odds ratio*, zkratkou *OR*.

$$OR = \frac{odds_1}{odds_0} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \tag{36}$$

Odhad regresního koeficientu b_i, $i \in [1, s]$, znamená odhad změny logitu při změně regresoru x_i o jedničku a při konstantních hodnotách regresorů ostatních, tedy

$$b_i = ln(\widehat{OR}), \text{ jestliže } x_{1,i} - x_{0,i} = 1 \text{ a } x_{1,j} = x_{0,j}, \quad j \neq i, \quad j = 1, 2, \ldots, s$$

Odhad OR při změně regresoru x_i o jedničku lze spočítat jednoduše jako

$$\widehat{OR} = e^{b_i}$$

Interpretaci výsledků logistické regrese ilustruje následující příklad nejjednoduššího logistického modelu s jedním dichotomickým regresorem. Pro větší názornost si představme, že regresor X znamená expozici (vystavení riziku), vysvětlovaná proměnná Y znamená přítomnost příznaku nemoci. Četnosti pozorovaných případů pak můžeme zapsat do čtyřpolní tabulky

Nemoc	Expozice	
	$X = 1$	$X = 0$
$Y = 1$	a	b
$Y = 0$	c	d

Pak, je-li $a, b, c, d > 0$

$$odds_1 = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}, \quad odds_0 = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

$$\widehat{OR} = \frac{ad}{bc}, \quad b_1 = \ln\left(\frac{ad}{bc}\right)$$

Pro rozptyl tohoto odhadu asymptoticky platí - viz na př. [5]

$$\text{var}(b_1) = \text{var}\left(\ln\left(\frac{ad}{bc}\right)\right) = 1/a + 1/b + 1/c + 1/d$$

Je tedy zřejmé, že logistickou regresi je možno aplikovat i v případech, kdy regresor je diskrétní dichotomická veličina. Pokud je regresor nominální, lze takovou proměnnou transformovat na dichotomické veličiny s hodnotami $\{0, 1\}$, tzv. indikátory (dummy variables). Uvažujme regresor $\mathbf{x}_i$, $i \in [1, s]$, který je nominální s k_i kategoriemi a má pozorované hodnoty $x_{li}, \quad l = 1, 2, \ldots, n$, n je počet pozorování. Hodnoty kategorií můžeme označit číselnými kódy $\{0, 1, \ldots, k_i - 1\}$. Kategorii s kódem 0 zvolíme jako referenční (t.zv. baseline category) a vytvoříme $k_i - 1$ indikátorů s ohodnocením podle následujícího pravidla

$$(d_{ij})_l = \begin{cases} 1 & \text{když} \quad x_{li} = j \\ & \\ 0 & \text{jinak} \end{cases} \qquad j = 1, 2, \ldots, k_i - 1, \quad l = 1, 2, \ldots, s \qquad (37)$$

Regresní koeficienty korespondující s těmito indikátory můžeme označit β_{ij}, $j = 1, 2, \ldots, k_i - 1, \quad i = 1, 2, \ldots, s$, jejich odhady pak označíme b_{ij}. Odhady regresních koeficientů u jednotlivých indikátorů jsou vlastně logaritmem odhadovaného poměru *odds* příslušné kategorie k *odds* kategorie referenční, tedy logaritmem příslušného *odds ratio*. Pro velké výběry můžeme $100(1-\alpha)$-procentní oboustranný interval spolehlivosti pro regresní koeficient β_{ij} vyjádřit jako

$$\langle b_{ij} - u(1-\alpha/2)SE(b_{ij}), \quad b_{ij} + u(1-\alpha/2)SE(b_{ij})\rangle$$

a interval spolehlivosti pro *OR*

$$\langle \exp\left[b_{ij} - u(1-\alpha/2)SE(b_{ij})\right], \quad \exp\left[b_{ij} + u(1-\alpha/2)SE(b_{ij})\right]\rangle, \tag{38}$$

kde $u(1-\alpha/2)$ je kvantil normovaného normálního rozdělení $N(0,1)$ a $SE(b_{ij})$ je směrodatná odchylka odhadu regresního koeficientu. Neobsahuje-li interval spolehlivosti pro *OR* jedničku, lze *odds* v této kategorii považovat za odlišný od *odds* kategorie referenční, takže interpretace výsledků regresního modelu je velice přímočará. Pokud máme regresní model s více regresory, odhad regresního parametru vyjadřuje lineární závislost predikované veličiny na daném regresoru po adjustování vlivu ostatních regresorů. Tedy v logistické regresi je odhad regresního koeficientu roven logaritmu odhadovaného *odds ratio* po adjustaci vlivu ostatních regresorů.

Příklad 10.1 Data pro tuto úlohu jsou v souboru LOGREG2.XLS. Vysvětlovaná veličina Y je dichotomická s hodnotami $\{0,\ 1\}$. Hodnota 1 znamená, že pozorovaná osoba je nemocná, hodnotu 0 má osoba zdravá. Regresory jsou veličiny *expozice* (dichotomická, hodnota 1 znamená, že osoba pracuje v rizikovém provozu, hodnota 0 znamená opak), *vek* (roky) a *koureni* (počet cigaret za den) jsou spojité, resp. i počet cigaret za spojitý můžeme považovat.

Zkrácený výstup z modulu Logistic Regression [14] následuje:

```
Logistic Regression Report

Response       Y

Parameter Estimation Section
           Regression        Standard  Chi-Square  Prob
Variable   Coefficient       Error       Beta=0    Level
Intercept  -25.35205         5.291554     22.95    0.000002
expozice     2.285141        0.4990213    20.97    0.000005
vek          0.6799906       0.1548485    19.28    0.000011
koureni      5.641818E-02    1.658742E-02 11.57    0.000671
```

```
Model in Transformation Form
-25.35205 + 2.285141*expozice + .6799906*vek +
+ 5.641818E-02*koureni
Note that this is XB. Prob(Y=1) is 1/(1+Exp(-XB)).

Odds Ratio Estimation Section
            Regression      Odds     Lower 95%   Upper 95%
Variable    Coefficient     Ratio    Conf.Limit  Conf.Limit
Intercept   -25.352054
expozice      2.285141    9.827076   3.695359    26.133163
vek           0.679991
koureni       0.056418

Model Summary Section
Model       Model      Model          Model
R-Squared    D.F.      Chi-Square     Prob
0.345596      3        77.63        0.000000
```

V první části jsou odhady parametrů logistického modelu a statistiky pro test hypotéz o nulovosti parametrů. Vidíme, že u všech čtyř parametrů zamítáme nulovou hypotézu $\beta_j = 0$, odhadované parametry u všech tří regresorů jsou kladné, tzn. logit roste s hodnotou regresoru, je tedy vyšší u exponovaných, roste s věkem a počtem vykouřených cigaret. V části Odds Ratio Estimation jsou znovu uvedeny odhady parametrů a pro dichotomický regresor je uveden i $\widehat{OR}$ a 95%-ní interval spolehlivosti. Jelikož tento interval neobsahuje hodnotu 1 (dolní hranice intervalu je 3,7), znamená to, že $\widehat{OR}$ je významně větší než 1 i po odečtení (adjustaci) vlivu věku a kouření a že expozice významně zvyšuje riziko onemocnění. Model Summary Section je analogií sekce ANOVA v lineární regresi a slouží k testu hypotézy, že všechny parametry jsou nulové.

Shrnutí

- zobecněný lineární model (GLM)
- spojovací funkce, kanonické spojovací funkce pro běžná rozdělení
- logistická regrese, odds ratio

Kontrolní otázky

1. Vysvětlete hlavní myšlenky zobecněného modelu.
2. Co je lineární prediktor?
3. Jaké rozdělení má vysvětlovaná veličina v logistické regresi?
4. Co je to logit? Je funkcí střední hodnoty vysvětlované veličiny?

Korespondenční úloha

Korespondenční úlohy budou zadávány ke každému kursu samostatně.