Aplikovaná logistická regrese
druhé vydání

David W. Hosmer, Stanley Lemeshow

24. dubna 2020

Obsah

Kapitola 1

Úvod do logistického regresního modelu

1.1 Úvod

V případě jednorozměrného lineárního regresního modelu předpokládáme, že střední hodnotu závislé veličiny Y podmíněnou hodnotou x lze vyjádřit jako

$$E(Y|x) = \beta_0 + \beta_1 x, \tag{1.1}$$

kde $E(Y|x)$ může nabývat libovolné hodnoty z intervalu $(-\infty, \infty)$.

V případě logistického regresního modelu má závislá veličina binární charakter, tj. může nabývat pouze dvou hodnot. Její střední hodnota tak omezena na interval $[0, 1]$. To je také patrné z obrázku 1.1, který ilustruje závislost mezi věkem pacienta a ischemickou chorobou srdeční. Z grafu, který má tvar písmene S, je patrné, že se střední hodnota vypočtena pro jednotlivé věkové kohorty postupně blíží jedné. Graf tak můžeme chápat ve smyslu kumulativní pravděpodobnostní funkce náhodné veličiny.

V následujícím textu budeme výraz $\pi(x) = E(Y|x)$ používat pro označení střední hodnoty závislé veličiny Y podmíněné hodnotu x pro jednorozměrný logistický regresní model, kde

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{1.2}$$

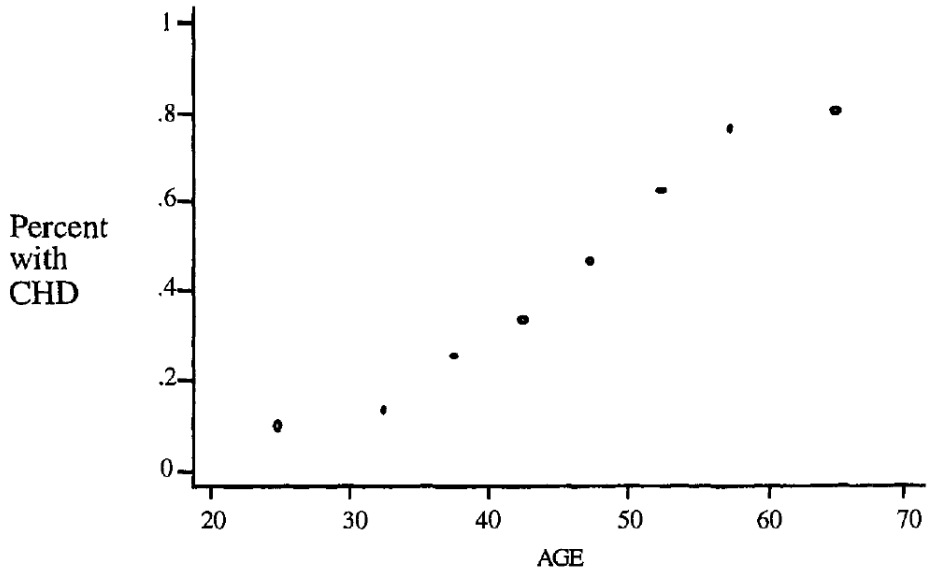
Tuto transformaci, která je klíčová pro studium logistického regresního modelu, nazýváme logit transformací a lze ji snadno upravit do tvaru

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x. \tag{1.3}$$

Funkce $g(x)$, kterou nazýváme logit funkcí, může nabývat hodnot z intervalu $(-\infty, \infty)$.

Hodnotu závislé veličiny y pro danou hodnotu x můžeme vyjádřit jako $y = \pi(x) + \varepsilon$, kde ε nabývá dvou stavů. Pokud $y = 1$, pak $\varepsilon = 1 - \pi(x)$ s pravděpodobností $\pi(x)$; pokud $y = 0$, pak $\varepsilon = -\pi(x)$ s pravděpodobností $1 - \pi(x)$.[1] Chyba ε tak sleduje binomické rozdělení s nulovou střední hodnotou a rozptylem $\pi(x)[1 - \pi(x)]$. To je další ze zásadních rozdílů oproti lineárnímu regresnímu modelu, ve kterém chyba ε sleduje normální rozdělení.

[1] Pokud Y nabývá hodnot 0 popř. 1, lze $\pi(x) = E(Y|x)$ interpretovat ve smyslu pravděpodobnosti, s jakou Y nabývá pro dané x hodnoty 1. Pravděpodobnost, s jakou Y nabývá pro dané x hodnoty 0, je pak $1 - \pi(x)$.



Obrázek 1.1: Vztah mezi výskytem ischemické poruchy srdeční a věkem pacienta

1.2 Kalibrace logistického regresního modelu

Základní metodou odhadu parametrů lineárního regresního modelu je metoda nejmenších čtverců. Podstatou této metody je volba takových hodnot parametrů β_0 a β_1, které minimalizují součet kvadrátu odchylek pozorovaných a predikovaných hodnot závislé veličiny Y. Tuto metodu však není v případě logistického regresního modelu možné použít.

Obecnější metodou odhadu parametrů je tzv. metoda maximální věrohodnosti, která odhadne hodnoty parametrů tak, aby výsledný model s maximální možnou pravděpodobností replikoval pozorovaná data. Za tímto účelem je třeba definovat tzv. funkci maximální věrohodnosti, která vyjadřuje pravděpodobnost výskytu pozorovaných dat v kontextu uvažovaného modelu jako funkci jeho parametrů.

Uvažujme jednorozměrný logistický regresní model a obecný pár (x_i, y_i). Pokud $y_i = 1$, je kontribuce páru (x_i, y_i) do funkce maximální věrohodnosti rovna $\pi(x)$. Pokud $y_i = 0$, je kontribuce páru (x_i, y_i) do funkce maximální věrohodnosti rovna $1 - \pi(x)$. Tyto dva stavy lze zkombinovat do podoby

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{1.4}$$

Pokud předpokládáme, že jednotlivá pozorování představovaná páry (x_i, y_i) jsou vzájemně nezávislá, lze funkci maximální věrohodnosti vyjádřit jako

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)^{1-y_i}]. \tag{1.5}$$

Z matematického a numerického hlediska je však snadnější pracovat s logaritmem funkce maximální věrohodnosti

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \Big(y_i \ln[\pi(x_i) + (1 - y_i) \ln[1 - \pi(x_i)] \Big), \tag{1.6}$$

kde $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. Abychom nalezli hodnotu $\boldsymbol{\beta}$, která maximalizuje $L(\boldsymbol{\beta})$, je třeba nejprve derivovat $L(\boldsymbol{\beta})$ dle β_0 a β_1 a tyto derivace položit rovny nule. Výsledné rovnice, které nazýváme rovnicemi maximální věrohodnosti, mají podobu

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \tag{1.7}$$

a

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \tag{1.8}$$

Rovnice maximální věrohodnosti jsou nelineární v parametrech β_0 a β_1, a proto je třeba je řešit iteračně pomocí optimalizační metody. Hodnotu $\boldsymbol{\beta}$, která je řešením výše uvedených rovnic, značíme $\hat{\boldsymbol{\beta}}$ a nazýváme ji odhadem maximální věrohodnosti. Podobně je $\hat{\pi}(x_i)$ odhadem maximální věrohodnosti pro $\pi(x_i)$, a platí

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}. \tag{1.9}$$

Odhadnutá logit funkce pak má tvar

$$\hat{g}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{1.10}$$

Rovnici (1.7) lze také vyjádřit ve tvaru

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i), \tag{1.11}$$

což lze interpretovat tak, že součet pozorovaných hodnot veličiny y je roven součtu predikovaných hodnot.

1.3 Testy významnosti odhadnutých parametrů

1.3.1 Věrohodnostní poměrový test

Při posuzování statistické významnosti odhadnutého parametru porovnáváme pozorované hodnoty s predikovanými hodnotami získaných na základě dvou logistických regresních modelů, z nichž jeden zahrnuje zkoumanou veličinu a druhý nikoliv. Samotné porovnání modelů je pak založené na porovnání logaritmů jejich funkce maximální věrohodnosti. Pro lepší pochopení principu je užitečné o pozorovaných hodnotách přemýšlet jako o hodnotách predikovaných tzv. saturovaným model. Saturovaný logistický regresní model je takový model, který obsahuje tolik parametrů kolik je pozorovaných párů (x_i, y_i).[2] Porovnání pozorovaných a predikovaných hodnot je pak založeno na statistice

$$D = -2\ln\Big[\frac{\textit{hodnota funkce maximální věrohodnosti kalibrovaného modelu}}{\textit{hodnota funkce maximální věrohodnosti saturovaného modelu}}\Big], \tag{1.12}$$

kterou nazýváme věrohodnostním poměrem (likelihood ratio) a test na ní založený pak věrohodnostním poměrovým testem (likelihood ratio test). S využitím (1.6), (1.12) a skutečnosti, že pro

[2] Jednoduchým příkladem saturovaného modelu je kalibrace jednorozměrného lineárního regresního modelu na dvou pozorováních.

hodnotu věrohodnostní funkce saturovaného modelu platí

$$l(\textit{saturovaný model}) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{1-y_i}, \tag{1.13}$$

lze statistiku D vyjádřit jako

$$D = -2\sum_{i=1}^{n}\left[u_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right], \tag{1.14}$$

kde $\hat{\pi}_i = \hat{\pi}(x_i)$. Pokud si dále uvědomíme, že hodnota věrohodnostní funkce saturovaného modelu je vždy rovna jedné[3], lze (1.12) zjednodušit do podoby

$$D = -2\ln(\textit{hodnota funkce maximální věrohodnosti kalibrovaného modelu}). \tag{1.15}$$

Pro účely zhodnocení významnosti nezávislé veličiny pak porovnáváme věrohodnostní poměr D pro model s a bez uvažované veličiny, tj.

$$G = D(\textit{hodnota funkce maximální věrohodnosti bez uvažované veličiny}) - D(\textit{hodnota funkce maximální věrohodnosti s uvažovanou veličinou}), \tag{1.16}$$

což lze dále upravit na

$$G = -2\ln\left[\frac{\textit{hodnota funkce maximální věrohodnosti bez uvažované veličiny}}{\textit{hodnota funkce maximální věrohodnosti s uvažovanou veličinou}}\right]. \tag{1.17}$$

Lze snadno dokázat, že v případě modelu, který zahrnuje pouze parametr β_0, je odhad tohoto parametru roven $\ln(n_1/n_0)$, kde $n_1 = \sum_{i=1}^{n} y_i$ a $n_0 = \sum_{i=1}^{n}(1-y_i)$ a predikovaná hodnota má charakter konstanty n_1/n. V tomto případě pak pro námi uvažovaný jednorozměrný logistický regresní model platí

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)^{1-y_i}}\right]. \tag{1.18}$$

Při platné nulové hypotéze $H_0 : \beta_1 = 0$ sleduje G chi-kvadrát rozdělení s jedním stupněm volnosti.

1.3.2 Waldův test

Alternativou k věrohodnostnímu poměrovému testu je Waldův test, který poměřuje hodnotu odhadnutého parametru $\hat{\beta}_1$ k jeho směrodatné odchylce. Výsledný poměr

$$W = \frac{\hat{\beta}_1}{\widehat{se}(\hat{\beta}_1)} \tag{1.19}$$

při platnosti nulové hypotézy $H_0 : \beta_1 = 0$ sleduje standardní normální rozdělení. Nevýhodou Waldova testu bohužel je, že často nezamítne nulovou hypotézu ani v případě, kdy je odhadnutý parametr významný. Proto je vhodnější používat věrohodnostní poměrový test.

[3]Toto tvrzení lze snadno ověřit dosazením $y_i = 1$ a $y_i = 0$ do (1.13).

1.3.3 Skóre test

Dalším možným testem významnosti parametru je tzv. skóre test, jehož hlavní výhodou je nižší výpočetní náročnost. Test je založen na teorii pravděpodobnostního rozdělení derivace logaritmu věrohodnostní funkce. Konkrétně v případě jednorozměrného logistického regresního modelu je test založen na znalosti pravděpodobnostního rozdělení derivace (1.8) podmíněné derivací (1.7). Test používá hodnotu rovnice (1.8) vypočtenou s pomocí $\beta_0 = \ln(n_1/n_0)$ a $\beta_1 = 0$. Jak již bylo zmíněno dříve, platí pro tento případ $\hat{\pi} = n_1/n = \overline{y}$. Dále lze dokázat, že odhadovaný rozptyl je $\overline{y}(1-\overline{y})\sum(x_i - \overline{x})^2$. To vede ke statistice

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum_{i=1}^{n}(x_i - \overline{x})^2}}, \tag{1.20}$$

která opět sleduje standardizované normální rozdělení. Opět nicméně platí, že věrohodnostní poměrový test je preferovaný před skóre testem.

1.4 Intervaly spolehlivosti

Intervaly spolehlivosti odhadnutých parametrů jsou založeny na Waldově testu a mají tvar

$$\hat{\beta}_i \pm z_{1-\alpha/2}\widehat{se}(\hat{\beta}_i). \tag{1.21}$$

Takto definované intervaly spolehlivosti lze použít nejen pro β_1 ale také pro konstantní člen modelu β_0.

Podobným způsobem lze také odhadnout intervaly spolehlivosti pro logit funkci. Střední hodnota logit funkce je definována jako

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1.22}$$

a její rozptyl v případě jednorozměrného logistického regresního modelu jako

$$\widehat{var}[g(x)] = \widehat{var}(\hat{\beta}_0) + x^2\widehat{var}(\hat{\beta}_1) + 2x\widehat{cov}(\hat{\beta}_0, \hat{\beta}_1). \tag{1.23}$$

Interval spolehlivosti logit funkce pak lze vypočíst pomocí

$$\hat{g}(x) \pm z_{1-\alpha/2}\widehat{se}[\hat{g}(x)], \tag{1.24}$$

kde $\widehat{se}[\hat{g}(x)] = \sqrt{\widehat{var}[g(x)]}$. Protože střední hodnota $\hat{\pi}(x)$ je definována jako $\frac{e^{\hat{g}(x)}}{1+e^{\hat{g}(x)}}$, lze interval spolehlivosti pro $\hat{\pi}(x)$ určit pomocí

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{se}[\hat{g}(x)]}}{1 + e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{se}[\hat{g}(x)]}}. \tag{1.25}$$

Kapitola 2

Vícerozměrný logistický regresní model

2.1 Vícerozměrný logistický regresní model

Logit funkce vícerozměrného logistického regresního modelu má podobu

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{2.1}$$

a samotný logistický regresní model pak podobu

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}}. \tag{2.2}$$

Pokud je některá z nezávislých veličin diskrétní numerická veličina, jejíž hodnoty označují jednotlivé kategorie (např. rasu nebo pohlaví), nelze s nimi nakládat jako s klasickými numerickými veličinami. V následujícím textu budeme tyto veličiny označovat jako kategorické. V případě numerické veličiny představují její hodnoty rozdílné úrovně; v případě kategorických veličin představují jejich hodnoty jednotlivé kategorie, které nemají ordinální význam. Pokud nabývá kategorická veličina k různých hodnot, je třeba ji nahradit $k-1$ pomocnými binárními proměnnými. Pokud např. nezávislá veličina X_i představující rasu nabývá hodnot 0 pro bělocha, 1 pro černocha a 2 pro asiata, je třeba ji nahradit pomocnými nezávislými veličinami D_{i1}, která nabývá hodnoty 1 pro černocha a 0 pro ostatní rasy, a D_{i2}, která nabývá hodnoty 1 pro asiata a hodnoty nula pro ostatní rasy. Je třeba zdůraznit, že není možné do modelu zahrnout také pomocnou proměnnou D_{i3}, která by nabývala hodnoty 1 pro bělocha, protože bychom tímto vytvořili perfektní multikolinearitu. Vícerozměrná logit funkce s kategorickou veličinou X_j, která nabývá k_j hodnot tak má podobu

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \beta_p x_p. \tag{2.3}$$

2.2 Kalibrace vícerozměrného logistického regresního modelu

Stejně jako v případě jednorozměrného logistického regresního modelu také v případě vícerozměrného logistického regresního modelu se pro jeho kalibraci používá metody maximální věro-

hodnosti.

Uvažujme model definovaný logit funkcí (2.1). Derivací logaritmu odpovídající věrohodnostní funkce podle $p+1$ parametrů získáme jednu rovnici maximální věrohodnosti ve tvaru

$$\sum_{i=1}^{n}[y_i - \pi(\boldsymbol{x}_i)] = 0 \tag{2.4}$$

a p rovnic maximální věrohodnosti ve tvaru

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\boldsymbol{x}_i)] = 0. \tag{2.5}$$

Tyto rovnice lze použít pro odhad hodnot parametrů jednotlivých veličin pomocí optimalizační metody.

Kromě bodového odhadu parametrů je třeba také získat odhad jejich směrodatných odchylek. Příslušné odhady lze vypočíst ze soustavy rovnic druhých parciálních derivací logaritmu věrohodnostní funkce, které mají podobu

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1-\pi_i) \tag{2.6}$$

a

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i(1-\pi_i), \tag{2.7}$$

pro $j, l = 0, 1, ..., p$, kde π_i označuje $\pi(\boldsymbol{x}_i)$. Matici $(p+1) \times (p+1)$ sestávající se ze záporných členů daných rovnicemi (2.6) a (2.7) označme jako $\boldsymbol{I}(\boldsymbol{\beta})$. Tato matice se nazývá pozorovaná informační matice (observed information matrix). Rozptyl a kovariance jednotlivých parametrů lze získat z inverze této matice, kterou budeme označovat jako $var(\boldsymbol{\beta}) = \boldsymbol{I}^{-1}(\boldsymbol{\beta})$. Bohužel až na výjimečné případy není možné explicitně vyjádřit členy matice $var(\boldsymbol{\beta})$. V následujícím textu budeme používat $var(\beta_j)$ k označení j-tého členu diagonály této matice, který představuje rozptyl odhadu $\hat{\beta}_j$, a $cov(\beta_j, \beta_l)$ k označení členů mimo diagonálu, které představují kovarianci mezi odhady $\hat{\beta}_j$ a $\hat{\beta}_l$. Jejich odhady jsou pak získány vyhodnocením matice $var(\boldsymbol{\beta})$ pro $\hat{\boldsymbol{\beta}}$.

Odhad informační matice lze vyjádřit ve tvaru $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}) = X'VX$, kde

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{2.8}$$

a

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}. \tag{2.9}$$

2.3 Testy významnosti odhadnutých parametrů

2.3.1 Věrohodnostní poměrový test

Stejně jako v případě jednorozměrného logistického regresního modelu lze také v případě vícerozměrného logistického regresního modelu použít pro posouzení významnosti odhadnutých

parametrů věrohodnostní poměrový test.

Uvažujme vícerozměrný logistický model s p nezávislými veličinami. Pro nulovou hypotézu předpokládáme, že hodnoty všech p odhadnutých parametrů jsou rovny nule a že distribuce odpovídající G statistiky sleduje chi-kvadrát rozdělení p stupni volnosti. Pro úplnost připomeňme, že G je podílem logaritmu věrohodnostní funkce modelu obsahujícího všech p veličin a modelu obsahujícího pouze konstantní člen.

Analogickou formu věrohodnostního poměrového testu lze použít také k testování menšího počet odhadnutých parametrů a to včetně testování významnosti jednoho parametru. Jediným rozdílem je, že se při výpočtu statistiky G namísto modelu obsahujícího pouze konstantní člen použije příslušný redukovaný model a odpovídajícím způsobem se upraví počet stupňů volnosti.

2.3.2 Waldův test

Jednorozměrný Waldův test

Při testování statistické významnosti jednoho odhadnutého parametru lze použít také Waldův test, jehož statistika má podobu

$$W_j = \frac{\hat{\beta}_j}{\widehat{se}(\hat{\beta}_j)}, \tag{2.10}$$

a která sleduje standardní normální rozdělení.

Jak již bylo zmíněno výše, pokud nezávislá proměnná nabývá $k > 2$ kategorií, je třeba ji "přeformulovat"pomocných $k - 1$ binárních veličin. Jestliže na tyto pomocné binární veličiny aplikujeme Waldův test, může jedna veličina vyjít jako statisticky významná, zatímco ostatní veličiny jako statisticky nevýznamné. V tomto případě je žádoucí v modelu buď ponechat všechny nebo naopak žádnou z těchto pomocných veličin. Toto pravidlo se vztahuje také na ostatní typy testů statistické významnosti.

Vícerozměrný Waldův test

Kromě výše popsaného Waldova testu významnosti existuje také jeho vícerozměrná varianta. Jeho statistika je pak vypočtena jako

$$W = \hat{\boldsymbol{\beta}}' \left[\widehat{var}(\hat{\boldsymbol{\beta}})\right]^{-1} \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'(\boldsymbol{X}'V\boldsymbol{X})\hat{\boldsymbol{\beta}} \tag{2.11}$$

a sleduje chi-kvadrát rozdělení s p stupni volnosti při hypotéze, že p odhadnutých parametrů je rovno nule. Testy pro méně než p parametrů jsou definovány analogicky.

Vyhodnocení vícerozměrného Waldova testu vyžaduje poměrně velkou výpočetní kapacitu, nenabízí tato metoda žádnou zásadnější výhodu oproti věrohodnostnímu poměrovému testu.

2.3.3 Skóre test

Vícerozměrná varianta skóre testu je založena na rozdělení p derivací $L(\boldsymbol{\beta})$ vzhledem k $\boldsymbol{\beta}$. Nicméně podobně jako v případě Waldova testu i vícerozměrná varianta skóre testu je poměrně výpočetně náročná, a proto se upřednostňuje věrohodnostní poměrový test.

2.4 Intervaly spolehlivosti

Intervaly spolehlivosti jednotlivých parametrů jsou konstruovány stejným způsobem jako v případě jednorozměrného logistického regresního modelu, tj.

$$\hat{\beta}_i \pm z_{1-\alpha/2}\widehat{se}(\hat{\beta}_i). \tag{2.12}$$

Podobně lze určit též interval spolehlivosti logit funkce jako

$$\hat{g}(\boldsymbol{x}) \pm z_{1-\alpha/2}\widehat{se}(\hat{g}(\boldsymbol{x})), \tag{2.13}$$

kde

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p \tag{2.14}$$

a

$$\widehat{se}[\hat{g}(\boldsymbol{x})] = \sqrt{\widehat{var}[\hat{g}(\boldsymbol{x})]} = \sqrt{\sum_{j=0}^{p} x_j^2 \widehat{var}(\hat{\beta}_j) + \sum_{j=0}^{p}\sum_{k=j+1}^{p} 2x_j x_k \widehat{cov}(\hat{\beta}_j, \hat{\beta}_k)}. \tag{2.15}$$

Maticově lze $\widehat{var}[\hat{g}(\boldsymbol{x})]$ vyjádřit také jako

$$\widehat{var}[\hat{g}(\boldsymbol{x})] = \boldsymbol{x}'\widehat{var}(\hat{\boldsymbol{\beta}})\boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{X'VX})^{-1}\boldsymbol{x}. \tag{2.16}$$

Kapitola 3

Interpretace logistického regresního modelu

3.1 Úvod

Předpokladem interpretace libovolného nakalibrovaného modelu je, že jsme schopni smysluplně interpretovat jeho odhadnuté parametry. Tato interpretace zahrnuje dva kroky - (a) odhad funkcionálního vztahu mezi závislou a nezávislou veličinou a (b) definování vhodné jednotky změny nezávislé veličiny.

První krok předpokládá, že vztah mezi závislou a nezávislou veličinou lze popsat pomocí lineární funkce. V případě jednorozměrného lineárního regresního modelu má tento vztah podobu $y(x) = \beta_0 + \beta_1 x$. V případě logistického regresního modelu je tento vztah definován na úrovni logit funkce, tj. jako $g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$.

Co se druhého kroku týče, je v případě jednorozměrného lineárního regresního modelu koeficient β_1 definován jako změna hodnoty závislé veličiny pro jednotkovou změnu závislé veličiny, tj. $\beta_1 = y(x+1) - y(x)$. Naproti tomu v případě logistického regresního modelu představuje koeficient β_1 změnu v logit funkci, tj. $\beta_1 = g(x+1) - g(x)$.

3.2 Nezávislá binární veličina

Nezávislou veličinu nazýváme binární, pokud nabývá pouze dvou hodnot. V následujícím textu budeme předpokládat, že taková veličina nabývá hodnot 0 a 1. Změna v logit funkci tak má podobu

$$g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1. \tag{3.1}$$

Možné kombinace logistické pravděpodobnosti jsou ilustrovány tabulkou na obrázku (3.1).

Důležitou statistikou používanou při analýze logistického regresního modelu je tzv. podíl rizik (odd ratio), které je definovaný jako

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}. \tag{3.2}$$

V případě jednorozměrné logistické regrese s binární nezávislou veličinou, která nabývá hodnot

Table 3.1 Values of the Logistic Regression Model When the Independent Variable Is Dichotomous

Outcome Variable (Y)	Independent Variable (X) $x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$y = 0$	$1-\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1-\pi(0) = \frac{1}{1+e^{\beta_0}}$
Total	1.0	1.0

Obrázek 3.1: Hodnoty jednorozměrného logistického regresního modelu pro binární nezávislou veličinu

0 nebo 1, je vztah mezi podílem rizik a hodnotou regresního parametru definován jako

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}. \tag{3.3}$$

Podíl rizik přibližně vyjadřuje kolikrát je pravděpodobnější výskyt pozitivní hodnoty závislé veličiny pro pozorování $x = 1$ než pro pozorování $x = 0$. Pokud např. y označuje výskyt / absenci rakoviny plic x obsahuje informaci o tom, zda-li je daná osoba kuřák, pak $\widehat{OR} = 2$ znamená, že pravděpodobnost výskytu rakoviny plic u kuřáka je dvakrát vyšší než u nekuřáka. Vzhledem k definici podílu rizik je zřejmé, že tato statistika přibližně vyjadřuje tzv. relativní riziko, které je definováno jako $\pi(1)/\pi(0)$. Z tabulky (3.1) je pak zřejmé, že tato aproximace platí, pokud $[1-\pi(0)]/[1-\pi(1)] \approx 1$. To je splněno, pokud $\pi(x)$ je malé jak pro $x = 1$ tak pro $x = 0$.

V řadě aplikací logistické regrese, se kterými se můžeme setkat v odborné literatuře, jsou spojité numerické veličiny převedeny na binární pomocí vhodně zvoleného hraničního bodu. Pro ilustraci uvádíme klasifikační tabulku (3.2), kdy veličina $AGED$ nabývá hodnoty 0 pokud je dané osobě méně než 55 let a hodnoty 1 v ostatních případech. Zkoumaná populace se dle tabulky (3.2) skládala z 21 osob s hodnotami $(x = 1, y = 1)$, 22 osob s hodnotami $(x = 0, y = 1)$, 6 osob s hodnotami $(x = 1, y = 0)$ a 51 osob s hodnotami $x = 0, y = 0$. Pokud bychom na základě těchto dat odhadli metodou maximální věrohodnosti parametr pro veličinu $AGED$, získali bychom hodnotu 2.094. Podíl rizik bychom pak odhadli na $\widehat{OR} = e^{2.094} = 8.1$. Podíl rizik však lze odhadnout také přímo s pomocí klasifikační tabulky, tj. jako

$$\widehat{OR} = \frac{\hat{\pi}(1)/[1-\hat{\pi}(1)]}{\hat{\pi}(0)/[1-\hat{\pi}(0)]} = \frac{21/6}{22/51} = 8.11. \tag{3.4}$$

Interval spolehlivosti podílu rizik se pak určí tak, že se nejprve vypočtou krajní hodnoty odpovídající intervalu spolehlivosti odhadu parametru β_1 a na ně se aplikuje exponenciála, tj.

$$e^{\beta_1 \pm z_{1-\alpha/2}\widehat{se}(\hat{\beta}_1)}. \tag{3.5}$$

Table 3.2 Cross-Classification of AGE Dichotomized at 55 Years and CHD for 100 Subjects

	AGED(x)		
CHD(y)	≥ 55 (1)	< 55 (0)	Total
Present (1)	21	22	43
Absent (0)	6	51	57
Total	27	73	100

Obrázek 3.2: Klasifikační tabulka pro binární nezávislou veličinu $AGED$ a závislou veličinu CHD představující výskyt ischemické choroby srdeční.

Vzhledem ke způsobu konstrukce je interval spolehlivosti podílu rizik nesymetrický.

V předchozím textu jsme předpokládali, že nezávislá binární veličina x nabývá veličin 0 a 1. Je třeba zdůraznit, že výše uvedené závěry jsou platné pouze pro tento případě. Uvažujme situaci, kdy x nabývá dvou různých obecných hodnot a a b. Pak platí

$$\ln[\widehat{OR}(a,b)] = \hat{g}(x=a) - \hat{g}(x=b) = (\hat{\beta}_0 + \hat{\beta}_1 a) - (\hat{\beta}_0 + \hat{\beta}_1 b) = \hat{\beta}_1(a-b) \tag{3.6}$$

neboli

$$\widehat{a,b} = e^{\hat{\beta}_1(a-b)}. \tag{3.7}$$

Analogie (3.2) pak má podobu

$$\widehat{OR}(a,b) = \frac{\hat{\pi}(x=a)/[1-\hat{\pi}(x=a)]}{\hat{\pi}(x=b)/[1-\hat{\pi}(x=b)]}. \tag{3.8}$$

Interval spolehlivosti podílu rizik lze vyjádřit jako

$$e^{\hat{\beta}_1(a-b) \pm z_{1-\alpha/2}|a-b|\widehat{se}(\hat{\beta}_1)}. \tag{3.9}$$

Je třeba zdůraznit, že kódování binární proměnné do 0 a 1 je zdaleka nejčastější, a proto ho budeme používat i v následujícím textu. Dalším používaným kódováním je pak -1 a 1, které je však méně frekventované.

3.3 Nezávislá kategorická veličina

Kategorickou veličinou rozumíme veličinu, která může nabývat $k > 2$ různých nominální hodnot. Jak jsme již zmínili výše, takovouto veličinu musíme převést na $k-1$ pomocných binárních veličin. Jako příklad uveďme veličinu X_i, která nabývá hodnot 0 pro bělocha, 1 pro černocha a 2 pro asiata. Tuto veličinu musíme nahradit dvojicí pomocných binárních veličin D_{i1} a D_{i2}. Při nejčastěji používaném kódování nabývá veličina D_{i1} hodnoty 1, pokud je daná osoba černoch, a 0 v ostatních případech. Podobně veličina D_{i2} nabývá hodnoty 1, pokud je daná osoba asiat, a 0 v ostatních případech. Je třeba zdůraznit, že pokud bych do modelu přidali ještě třetí veličinu, D_{i3}, která by nabývala hodnoty 1, pokud je daná osoba běloch, a 0 v ostatních případech, vytvořili bychom perfektní multikolinearitu.[1]

[1]Informaci o tom, že je daná osoba běloch totiž, lze získat také na základě veličin D_{i1} a D_{i2}. Konkrétně, pokud obě veličiny nabývají hodnoty 0, víme, že se nejedná ani o černocha a ani o asiata, takže se musí jednat o bělocha. Informace představovaná veličinou D_{i3} je tak duplicitní.

Pro ilustraci uvažujme model $y(\boldsymbol{x}) = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2}$, kde $y = 1$ představuje výskyt ischemické choroby srdeční. Pokud použijeme standardní kódování pomocných binárních veličin na hodnoty 0 a 1, platí

$$\ln[\widehat{OR}(black, white)] = \hat{g}(black) - \hat{g}(white) = (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 0) - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 0) = \hat{\beta}_1. \quad (3.10)$$

Jak vyplývá z označení $\widehat{OR}(black, white)$, vyjadřuje $\hat{\beta}_1$ odhad, kolikrát je výskyt ischemické choroby srdeční pravděpodobnější u černocha než u bělocha. Pokud bychom chtěli porovnat pravděpodobnosti výskytu ischemické choroby srdeční pro černocha vs. asiata, museli bychom výše uvedenou rovnici upravit do tvaru

$$\ln[\widehat{OR}(black, asian)] = \hat{g}(black) - \hat{g}(asian) = (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 0) - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1) = \hat{\beta}_1 - \hat{\beta}_2. \quad (3.11)$$

Interval spolehlivosti $\widehat{OR}(black, white)$ pak má podobu

$$e^{\hat{\beta}_1 \pm z_{1-\alpha/2}\sqrt{\widehat{var}\left(\ln[\widehat{OR}(black,white)]\right)}} = e^{\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{se}(\hat{\beta}_1)} \quad (3.12)$$

a interval spolehlivosti $\widehat{OR}(black, asian)$ podobu

$$e^{(\hat{\beta}_1 - \hat{\beta}_2) \pm z_{1-\alpha/2}\sqrt{\widehat{var}\left(\ln[\widehat{OR}(black,asian)]\right)}}, \quad (3.13)$$

kde

$$\widehat{var}\left(\ln[\widehat{OR}(black, asian)]\right) = \widehat{var}(\hat{\beta}_1) - 2\hat{\beta}_1\hat{\beta}_2\widehat{cov}(\hat{\beta}_1, \hat{\beta}_2) + \widehat{var}(\hat{\beta}_2). \quad (3.14)$$

Dalším možným způsobem kódování je použití hodnot -1, 0 a 1. Běloch je tak reprezentován hodnotami -1 pro D_{i1} a -1 pro D_i2, černoch hodnotami 1 pro D_{i1} a 0 pro D_{i2} a konečně asiat hodnotami 0 pro D_{i1} a 1 pro D_{i2}. Platí tedy

$$\ln[\widehat{OR}(black, white)] = \hat{g}(black) - \hat{g}(white) = (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 0) - (\hat{\beta}_0 + \hat{\beta}_1 \cdot (-1) + \hat{\beta}_2 \cdot (-1)) = 2\hat{\beta}_1 + \hat{\beta}_2 \quad (3.15)$$

a podobně

$$\ln[\widehat{OR}(black, asian)] = \hat{g}(black) - \hat{g}(asian) = (\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 0) - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1) = \hat{\beta}_1 - \hat{\beta}_2 \quad (3.16)$$

. Intervaly spolehlivosti podílů rizik pak mají podobu

$$e^{(2\hat{\beta}_1 + \hat{\beta}_2) \pm z_{1-\alpha/2}\sqrt{\widehat{var}\left(\ln[\widehat{OR}(black,white)]\right)}} \quad (3.17)$$

pro $\widehat{OR}(black, white)$ a

$$e^{(\hat{\beta}_1 - \hat{\beta}_2) \pm z_{1-\alpha/2}\sqrt{\widehat{var}\left(\ln[\widehat{OR}(black,asian)]\right)}} \quad (3.18)$$

pro $\widehat{OR}(black, asian)$ kde

$$\widehat{var}\left(\ln[\widehat{OR}(black, white)]\right) = 4\widehat{var}(\hat{\beta}_1) + 4\hat{\beta}_1\hat{\beta}_2\widehat{cov}(\hat{\beta}_1, \hat{\beta}_2) + \widehat{var}(\hat{\beta}_2). \quad (3.19)$$

a

$$\widehat{var}\left(\ln[\widehat{OR}(black, asian)]\right) = \widehat{var}(\hat{\beta}_1) - 2\hat{\beta}_1\hat{\beta}_2\widehat{cov}(\hat{\beta}_1, \hat{\beta}_2) + \widehat{var}(\hat{\beta}_2). \quad (3.20)$$

Tento způsob kódování je však používán spíše výjimečně, a proto budeme v následujícím textu používat kódování pomocí hodnot 0 a 1.