Logistická regrese s vícekategoriální vysvětlovanou proměnnou

Iva Pecáková*

Obecný lineární model (GLM) zahrnující jednorozměrné i vícerozměrné varianty regresní analýzy, analýzy rozptylu či analýzy kovariance připouští použití kategoriálních proměnných jako vysvětlujících proměnných či faktorů. Na místě vysvětlovaných proměnných však nemohou vyhovovat podmínkám vysloveným v obecném lineárním modelu pro konkrétní výpočetní postupy. Cíle, pro které je konstruován, pak klasický model s kategoriální vysvětlovanou proměnnou nemůže splňovat. S rozvojem metodologie statistické analýzy kategoriálních dat proto byly navrženy modely regresního typu zohledňující specifika kategoriální vysvětlované proměnné podle jejího charakteru (binární, vícekategoriální nominální či ordinální). Cílem tohoto příspěvku je objasnění podstaty regresního modelu s kategoriální vysvětlovanou proměnnou. Užitečné bude proto rekapitulovat nejprve postup regresní analýzy používaný v situaci, kdy možné hodnoty vysvětlované proměnné jsou pouze dvě.

Binární vysvětlovaná proměnná

Uvažujme binární vysvětlovanou proměnnou Y, jež nabývá s pravděpodobností π hodnoty 1 a s pravděpodobností $(1 - \pi)$ hodnoty 0. Představuje-li vektor

$$\mathbf{x}_i' = \left[x_{i1},\ x_{i2}\ \ldots,\ x_{ik} \right],\ \mathrm{i} = 1, 2\ \ldots,\ \mathrm{n},$$

i-tou kombinaci hodnot k nenáhodných vysvětlujících proměnných $X_1, X_2\ \ldots, X_\mathrm{k}$, pak i-té podmíněné rozdělení veličiny Y je alternativní s parametrem (a střední hodnotou veličiny Y) π_i a pravděpodobnostní funkcí

$$P(y_i \mid \pi_i) = \pi_i^{\,y_i} (1 - \pi_i)^{1 - y_i}\ . \tag{1}$$

Zůstává-li pro různé vektory $\mathbf{x}_i$ hodnot veličin $X_1, X_2\ \ldots, X_k$ podmíněné rozdělení pravděpodobnosti veličiny Y (dané parametrem π_i) stejné, pak veličina Y na těchto proměnných nezávisí. Pokud však různé kombinace hodnot vysvětlujících proměnných vedou k různým pravděpodobnostem π_i, lze zřejmě uvažovat o nějakém typu závislosti Y na těchto vysvětlujících proměnných a pokusit se o její zobrazení regresním modelem.

* Doc. Ing. Iva Pecáková, CSc.; Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky, VŠE v Praze, pecakova@vse.cz.

Vektor **y** hodnot alternativní vysvětlované proměnné (o n prvcích) obsahuje pouze nuly a jedničky. Pokud jsou rovněž vysvětlující proměnné kategoriální a kombinace jejich hodnot se vyskytují opakovaně, což není nijak výjimečné, bývají údaje obvykle nejprve roztříděny do kontingenční tabulky. Jednotkami pro analýzu jsou v tomto případě pole v tabulce (jejichž počet označíme *C*), obsahující počty případů, kdy pro jednotlivé kombinace hodnot vysvětlujících proměnných veličina *Y* nabývá hodnoty 1. Tyto četnosti mají binomické rozdělení s pravděpodobnostní funkcí

$$P(y_i \mid n_i, \pi_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}, \tag{2}$$

označíme-li nyní y_i počet případů, kdy pro *i*-tou kombinaci hodnot vysvětlujících proměnných *Y* nabývá hodnoty 1, a n_i celkový počet případů pro *i*-tou kombinaci hodnot vysvětlujících proměnných. Vysvětlovanou proměnnou v regresním modelu je v takové situaci relativní četnost $p_i = y_i / n_i$ (s podmíněnou střední hodnotou π_i).

Úvahy o charakteru regresního vztahu mezi vysvětlujícími proměnnými a podmíněnou střední hodnotou binární vysvětlované proměnné jsou významně ovlivněny omezením jejích hodnot pouze na interval od 0 do 1. Použití lineární regresní funkce uvedený interval pro π nezajišťuje. V určitých omezených situacích (například vykazuje-li nepříliš velká či malá relativní četnost nastoupení sledovaného jevu v jednotlivých polích kontingenční tabulky nízkou variabilitu) to sice nemusí být na závadu, obecně však použití lineární regresní funkce působí pro některé možné kombinace hodnot vysvětlujících proměnných nesnáze.

Proti použití lineární regresní funkce v tomto případě lze vznést také věcnou námitku, neboť chápeme-li vztah mezi vysvětlovanou a vysvětlující proměnnou jako lineární, znamená to, že jednotkové absolutní změně vysvětlující proměnné odpovídá určitá vždy stejná změna střední hodnoty vysvětlované proměnné. Vliv vysvětlující proměnné na změnu pravděpodobnosti však v principu za lineární považovat nelze. Například stejný absolutní přírůstek příjmu v různých příjmových skupinách neznamená patrně stejnou změnu pravděpodobnosti realizace určitého většího vydání; ve vyšší příjmové skupině může být tato pravděpodobnost větší a nárůst příjmu se jí nemusí téměř dotknout, v nižší příjmové skupině může mít stejná změna vliv zásadnější.

Regresní funkci s tzv. logitovou transformací π,

$$g(\pi) = \ln \frac{\pi}{1-\pi} = \mathbf{x}'\boldsymbol{\beta}, \tag{3}$$

kde $\mathbf{x}' = [1, x_1, x_2 ..., x_k]$,

$\boldsymbol{\beta}' = [\beta_0, \beta_1 ..., \beta_k]$,

se říká logistická regresní funkce. Podmíněná střední hodnota binární vysvětlované proměnné je tak vyjádřena jako nelineární funkce vysvětlujících proměnných. Z (3) přitom vyplývá, že

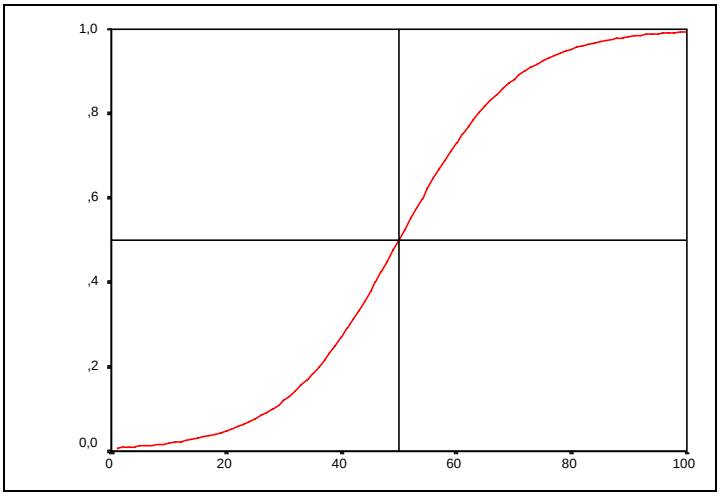
$$\frac{\pi}{1-\pi} = e^{\mathbf{x}'\boldsymbol{\beta}},$$

a dále

$$\pi = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}} = \left[1+e^{-\mathbf{x}'\boldsymbol{\beta}}\right]^{-1}. \tag{4}$$

Příklad takové funkce (pro $k = 1$) znázorňuje obrázek 1.

Obr. č. 1: $\pi = \left[1+\exp(5-0{,}1x)\right]^{-1}$



Jelikož funkce (4) je distribuční funkcí logistického rozdělení, lineární kombinace vysvětlujících proměnných $\mathbf{x}'\boldsymbol{\beta}$ (logit) je tedy jeho 100π-procentním kvantilem.

Použití distribuční funkce rozdělení pro modelování pravděpodobnosti π zajišťuje potřebné omezení jejích hodnot na interval $\langle 0,1\rangle$. Jiným podobně používaným rozdělením, jehož distribuční funkce je v grafu rovněž symetrickou s-křivkou, je normální rozdělení – v takovém případě je jako lineární kombinace vysvětlujících proměnných vyjádřen jeho 100π-procentní kvantil, tzv. probit (méně často normit). Odhady pravděpodobností pořízené s užitím logitového a probitového modelu jsou v mnoha situacích velmi podobné, logitové modely jsou však ve statistické literatuře preferovány, neboť jsou snáze interpretovatelné a v neposlední řadě mají velmi blízko k loglineárním modelům používaným často k analýze kontingenčních tabulek.

V logistické regresní funkci uvažujeme tedy obecně k vysvětlujících proměnných, od číselných spojitých až po kategoriální. Charakter vysvětlujících proměnných je podstatný pro konstrukci modelu, odhad a interpretaci jeho parametrů, hodnocení kvality modelu i jeho využití. Za účasti spojité proměnné (proměnných) v datové matici jsou jednotlivé kombinace hodnot vysvětlujících proměnných jedinečné a neopakují se. Jsou-li ovšem v datové matici pouze kategoriální proměnné, lze data uspořádat do

vícerozměrné kontingenční tabulky a využít pro logistické modelování četnosti získané tříděním.

Logistická regresní funkce o $k + 1$ parametrech je v těchto parametrech nelineární. K jejich odhadu se nejčastěji používá metoda maximální věrohodnosti. Postup hledání maxima věrohodnostní funkce výběrových údajů (resp. jejího logaritmu) vzhledem k neznámým parametrům vede k soustavě nelineárních věrohodnostních rovnic, samotné odhady parametrů jsou proto výsledkem použití vhodného iteračního algoritmu. Často se používá zejména Newtonova-Raphsonova metoda, kdy je logaritmus věrohodnostní funkce v okolí počátečního odhadu aproximován prvními třemi členy Taylorova rozvoje a určí se maximum pro tuto aproximaci; opravený odhad je pak vždy použit v dalším iteračním kroku. Počáteční odhad parametrů je získán například metodou nejmenších čtverců na základě vztahu mezi výběrovými logity a lineární kombinací vysvětlujících proměnných (spojité vysvětlující proměnné je pro určení výběrových logitů třeba kategorizovat). Algoritmus Newtonovy-Raphsonovy metody relativně rychle konverguje k maximálně věrohodnému odhadu parametrů. Jeho výhodou je rovněž to, že poskytuje informační matici, a tedy i kovarianční matici odhadů parametrů, na jejímž základě lze konstruovat odhady intervalové a rovněž testová kritéria pro ověřování hypotéz o parametrech.

Pro objasnění významu parametrů v lineární kombinaci vysvětlujících proměnných je podstatné, že vyjadřuje transformovanou střední hodnotu vysvětlované proměnné (alternativní či binomické) – logit. Logit je logaritmus podílu $\pi/(1 - \pi)$ vyjadřujícího *šanci* (*odds*), že veličina Y nabývá hodnoty 1. Parametr β_0 udává velikost logitu pro nulové hodnoty (resp. referenční kategorie) všech vysvětlujících proměnných. Pro $\beta_0 = 0$ je šance, že $Y = 1$, jedna ku jedné, neboli $\pi = 0{,}5$. Kladné hodnoty parametru β_0 znamenají, že tato šance je větší než jedna ($\pi > 0{,}5$), záporné hodnoty znamenají, že je menší než jedna ($\pi < 0{,}5$).

V závislosti na jedné nebo více vysvětlujících proměnných se logit může měnit. Míru této změny vyjadřují parametry β_j, $j = 1, 2 \ldots, k$. Při jednotkové změně j-té vysvětlující proměnné (a zůstanou-li ostatní veličiny beze změny), je potom šance, že $Y = 1$, $e^{\beta j}$-krát tak velká. Při použití indikátorových proměnných pro vícekategoriální vysvětlované proměnné závisí způsob interpretace parametrů na typu indikátorů – buď máme na mysli změnu logitu, a tedy také šance, oproti zvolené referenční kategorii (indikátory *dummy*), nebo oproti průměru všech použitých kategorií (indikátory *effect*).

Logistický regresní model lze hodnotit jednak podle toho, nakolik je model schopen na základě hodnot vysvětlujících proměnných rozlišovat jednotky podle hodnoty vysvětlované proměnné, jednak podle toho, nakolik se pro určité kombinace hodnot vysvětlujících proměnných shodují zjištěné a očekávané četnosti nastoupení sledovaného jevu (postupy vhodné pro tříděná data). S ohledem na řešení stejné úlohy jako u diskriminační analýzy nepřekvapí, že pro vyhodnocení klasifikační schopnosti regresní funkce se používají analogické nástroje. Patří k nim klasifikační tabulka a různé typy s ní souvisejících grafů, případně ROC křivka – cílem je vždy vyjádřit názorně podíl chybně zařazených jednotek.

Statistiky založené na konfrontaci zjištěných a očekávaných četností nastoupení sledovaného jevu ($Y = 1$) pro jednotlivé kombinace hodnot vysvětlujících proměnných lze použít za předpokladu, že takových kombinací není příliš mnoho, tedy že

vysvětlující proměnné mají malý počet hodnot či kategorií (data jsou tříděna v kontingenční tabulce). Nejvýhodnější vlastnosti má věrohodnostní poměr (*deviance*) G^2,

$$G^2 = 2\sum_{i}^{C}\left[y_i \ln \frac{p_i}{\hat{\pi}_i} + (n_i - y_i) \ln \frac{1-p_i}{1-\hat{\pi}_i} \right], \tag{5}$$

(stříškou je zde označena modelem odhadnutá pravděpodobnost π_i). Rozdělení této statistiky je asymptoticky chí-kvadrát s $(C - p)$ stupni volnosti (p je počet parametrů hodnocené funkce, $p = k + 1$). Spojité vysvětlující proměnné přímé použití této statistiky znemožňují, je však možné provést nějaké seskupení jednotek – například známý Hosmerův-Lemeshowův postup je založen na vytvoření obvykle deseti zhruba stejně obsazených skupin, v nichž jsou pro výpočet věrohodnostního poměru G^2 stanoveny průměrné odhadnuté pravděpodobnosti. Počet stupňů volnosti asymptotického chí-kvadrát rozdělení je v tomto případě počet skupin mínus dva.

Při rozhodování o vhodném modelu však statistika G^2 (podobně jako determinační koeficient v klasické regresní analýze) vede k upřednostňování složitějších modelů – čím více parametrů, tím lepší shody modelu s daty lze dosáhnout. Používají se proto různé modifikace této statistiky, jež počet parametrů zohledňují. Příkladem takové modifikace je Goodmannův index $GI = G^2 / df$, kde $df = C - p$ (počet stupňů volnosti), Akaikeho informační kritérium $AIC = G^2 + 2p = G^2 + 2(C - df)$ (případně bez konstanty $2C$) a jeho různě korigované varianty, případně bayesovské informační kritérium $BIC = G^2 - df(\ln n)$. Nižší hodnota kritérií znamená vždy vhodnější model (podle zvoleného způsobu penalizace mohou *AIC* a *BIC* nabývat i záporných hodnot).

Rozdíl věrohodnostních poměrů G^2 pro dva různé modely, model M_1 s p_1 parametry a model M_2 s p_2 parametry, $p_2 > p_1$, tedy

$$G^2_{M1/M2} = G^2_{M1} - G^2_{M2}, \tag{6}$$

má chí-kvadrát rozdělení s $p_2 - p_1$ stupni volnosti a přináší tak užitečnou informaci v situaci, kdy zvažujeme úlohu jednotlivých vysvětlujících proměnných. Lze jej totiž použít jako testové kritérium pro ověření hypotézy, že rozšíření regresního modelu o dalších $p_2 - p_1$ vysvětlujících proměnných je zbytečné, neboť nepřináší významné snížení deviance G^2 (nebo naopak odstranění proměnných je užitečné, neboť G^2 významně nezvýší). V tomto smyslu se vlastně jedná o analogii sekvenčních F-testů u klasického lineárního regresního modelu.

Při ověřování užitečnosti jednotlivých proměnných v logistické regresní funkci lze testovat hypotézu vždy o nulové hodnotě jednoho parametru β_j, $j = 0, 1 \ldots, k$, na základě Waldovy statistiky

$$\frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)} \tag{7}$$

(stříškami je označen odhad parametru, resp. jeho směrodatné chyby), s asymptoticky normovaným normálním rozdělením (analogie dílčího t-testu u klasického regresního

modelu). Poměrně časté případy selhávání Waldova testu však vedou k doporučování spíše předchozího postupu.

Na základě rozdílů věrohodnostních poměrů G^2 jsou konstruovány rovněž statistiky, jež lze chápat jako míry snížení neurčitosti v datech, kterého se podařilo dosáhnout hodnoceným regresním modelem. Jedná se tak o určité analogie determinačního indexu používaného pro lineární regresní funkce. Například Mc Faddenova statistika je definována jako

$$D_{MF} = \frac{G_0^2 - G_M^2}{G_0^2 - G_S^2}, \tag{8}$$

kde index 0 je použit k označení modelu pouze s parametrem β_0 a index S k označení modelu saturovaného (kdy odhadnuté hodnoty odpovídají zjištěným). Nevýhodné důsledky logaritmování věrohodností na hodnotu této statistiky, totiž tendence k jejímu nadhodnocování se zvětšováním rozsahu souboru při dané kontingenční tabulce, se snaží odstranit statistika Coxova-Snellova,

$$D_{CS} = 1 - \left[\frac{L_0}{L_M}\right]^{2/n}, \tag{9}$$

jejíž maximum však není jedna, ale $1 - L_0^{2/n}$, a její pro interpretaci tedy vhodnější modifikace (Nagelkerkeova statistika)

$$D_N = \frac{D_{CS}}{max(D_{CS})} = 1 - \left[\frac{L_0}{L_M}\right]^{2/n} /(1 - L_0^{2/n}). \tag{10}$$

L_0, L_M ve vzorcích (9) a (10) značí odpovídající věrohodnosti.

Multinomická vysvětlovaná proměnná – neuspořádané kategorie

Přirozeným zobecněním binomického logistického regresního modelu (nebo také logitového modelu s binární vysvětlovanou proměnnou) je multinomický logitový model. Předpokládejme nejprve, že vysvětlovaná proměnná Y je nominální a má $s \geq 2$ kategorií. V analogii na předchozí text pro ně použijeme kódy 0, 1 …, $s - 1$. Pro i-tou kombinaci hodnot vysvětlujících proměnných (tedy vždy z celkem n_i případů) nabývá Y jednotlivých hodnot s pravděpodobnostmi π_{ij}, $j = 0, 1 \ldots, s - 1$ a počty takových případů mají tedy podmíněné multinomické rozdělení s parametrem n_i a dále s parametry $\pi_{ij}, j = 0, 1 \ldots, s - 1$.

V případě binární vysvětlované proměnné jsme založili logit (3) na šanci nastoupení nějakého jevu ku jeho nenastoupení. Označíme-li si $\pi = \pi_1$ a $1 - \pi = \pi_0$, potom

$$\pi_1 = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_1)}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)}, \text{ kde } \boldsymbol{\beta}_1' = \left[\beta_{10}, \beta_{11}..., \beta_{1k}\right],$$

$$\pi_0 = \frac{1}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)} = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_0)}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)}, \text{ kde } \boldsymbol{\beta}_0' = [0,0...,0] = \mathbf{0}'.$$

Kategorii označenou indexem nula budeme i v dalším textu považovat za srovnávací (referenční).

Pro $s = 3$ můžeme tedy analogicky psát

$$\pi_0 = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_0)}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)+\exp(\mathbf{x}'\boldsymbol{\beta}_2)},$$

$$= \frac{1}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)+\exp(\mathbf{x}'\boldsymbol{\beta}_2)}, \text{ kde } \boldsymbol{\beta}_0 = \mathbf{0}.$$

$$\pi_1 = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_1)}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)+\exp(\mathbf{x}'\boldsymbol{\beta}_2)},$$

$$\pi_2 = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_2)}{1+\exp(\mathbf{x}'\boldsymbol{\beta}_1)+\exp(\mathbf{x}'\boldsymbol{\beta}_2)}.$$

Použijeme-li nyní pro multinomickou proměnnou šanci, že nastane nějaký jev a ne jev referenční (jedna zvolená možnost, zde $Y = 0$), pak můžeme zapsat dva logity se společným srovnávacím základem (*bazické* logity) jako

$$\ln\frac{\pi_1}{\pi_0} = \mathbf{x}'\boldsymbol{\beta}_1 \text{ a } \ln\frac{\pi_2}{\pi_0} = \mathbf{x}'\boldsymbol{\beta}_2.$$

Obecně tedy pro $s \geq 2$ neuspořádaných kategorií

$$\pi_j = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_j)}{\sum_{j=0}^{s-1}\exp(\mathbf{x}'\boldsymbol{\beta}_j)}, \text{ kdy pro referenční kategorii (zde } j = 0) \ \boldsymbol{\beta}_0 = \mathbf{0}. \tag{11}$$

Pro bazické logity pak platí

$$\ln\frac{\pi_j}{\pi_0} = \mathbf{x}'\boldsymbol{\beta}_j, j = 1, 2 \ldots, s-1. \tag{12}$$

Vysvětlující proměnné mohou být stejně jako v jakémkoliv jiném regresním modelu číselné i kategoriální, v druhém případě jsou jednotlivé kategorie vyjádřeny prostřednictvím indikátorů. Pro model s celkem k proměnnými to tedy znamená odhadnout $(k + 1)(s - 1)$ parametrů. Odhad lze opět pořídit metodou maximální věrohodnosti doplněnou o iterační algoritmus.

Parametry $\beta_{1,0}, \beta_{2,0} \ldots, \beta_{s-1,0}$ představují velikost jednotlivých logitů pro nulové hodnoty (resp. referenční kategorie) všech vysvětlujících proměnných. Jsou to tedy logaritmy šancí, že veličina *Y* nabude hodnoty 1 a nikoliv hodnoty 0, nabude hodnoty 2 a nikoliv hodnoty 0, … nabude poslední hodnoty a nikoliv hodnoty 0.

Jednotlivé parametry β_{ij}, $i = 1, 2 \ldots, s-1$ a $j = 1, 2 \ldots, k$ lze interpretovat analogicky k parametrům modelu s binomickou vysvětlovanou proměnnou. Představují vliv změny hodnoty či kategorie *i*-té vysvětlující proměnné na změnu šance, že vysvětlovaná proměnná *Y* nabude *j*-té kategorie a nikoliv kategorie referenční (zůstanou-li ostatní vysvětlující proměnné konstantní).

Při sestavování a vyhodnocování kvality regresního modelu s vícekategoriální vysvětlovanou proměnnou se používají analogické nástroje jako v případě vysvětlované proměnné binární. Věnujme se proto pouze situaci, kdy kategoriální vysvětlující proměnnou zastupujeme několika indikátory.

Rozhodnutí o zařazení či nezařazení takové proměnné do regresní funkce lze totiž učinit pouze v tom případě, shodují-li se výsledky použitých postupů: tedy jsou-li testy pro *všechny* indikátory významné, pak proměnnou zařadíme, jsou-li všechny testy nevýznamné, pak nikoliv. Často se ovšem výsledky u jednotlivých indikátorů rozcházejí. Taková situace naznačuje, že je vhodné zvážit u příslušné veličiny počet a vymezení jednotlivých kategorií. Některé z nich jsou si totiž zřejmě velmi podobné a jejich rozlišení je zbytečné. Problém pak může vyřešit například spojení takových kategorií a překódování dotčené proměnné.

Multinomická vysvětlovaná proměnná – uspořádané kategorie

Budiž nyní vysvětlovaná proměnná ordinální, jejích $s \geq 2$ kategorií lze tedy objektivně uspořádat. Na této skutečnosti lze založit definování logitu a zvolit i jiné způsoby konfrontace logitů, než jaký byl použit v předchozích odstavcích.

Vyjdeme-li při konstrukci modelu z kategorií v řadě sousedících, lze logity (*řetězové*) definovat jako

$$\ln \frac{\pi_j}{\pi_{j-1}}, j = 1, 2 \ldots, s-1$$

a těchto $s-1$ logitů pak vyjádřit jako lineární kombinaci vysvětlujících proměnných, tedy

$$\ln \frac{\pi_j}{\pi_{j-1}} = \mathbf{x}'\boldsymbol{\beta}_j \tag{13}$$

Mezi bazickými a řetězovými logity je jednoduchý vztah,

$$\ln \frac{\pi_j}{\pi_{j-1}} = \ln \frac{\pi_j}{\pi_0} - \ln \frac{\pi_{j-1}}{\pi_0} = \mathbf{x}'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}) \,, j = 1, 2 \ldots, s-1.$$

Je-li pro interpretaci zajímavá změna šance definované pro sousední kategorie vysvětlované proměnné, odpovídající parametry lze získat uvedeným způsobem z parametrů modelu založeného na logitech bazických.

Ordinální logistický regresní model lze založit rovněž na *kumulativních* logitech. Zatímco doposud byla konstrukce logitu založena na srovnání dvou hodnot pravděpodobnostní funkce podmíněného rozdělení vysvětlované proměnné, v tomto případě je využita hodnota distribuční funkce tohoto rozdělení (F_j), resp. její doplněk do jedné ($1 - F_j$). Kumulativní logit zapíšeme jako

$$\ln \frac{F_j}{1-F_j} = \ln \frac{P(Y \le y_j)}{P(Y > y_j)} = \ln \frac{\pi_0 + \pi_1 + \ldots + \pi_j}{\pi_{j+1} + \pi_{j+2} + \ldots + \pi_{s-1}} \,, j = 0, 1 \ldots, s-2, \qquad (14)$$

a regresní funkci s užitím kumulativního logitu jako

$$\ln \frac{F_j}{1-F_j} = \mathbf{x}'\boldsymbol{\beta}_j \,, \; j = 0, 1 \ldots, s-2. \qquad (15)$$

Parametry β_{0j} jsou prahové parametry pro jednotlivé kategorie veličiny Y, představují logaritmus šance, že Y nabývá nejvýše j-té kategorie, a nikoliv vyšší. Vzhledem ke způsobu definování kumulativního logitu přitom v tomto případě platí $\beta_{00} \le \beta_{01} \le \ldots \beta_{0,s-2}$. Kladné koeficienty ve vektoru $\boldsymbol{\beta}_j$ pak znamenají, že s *růstem* hodnot vysvětlujících proměnných roste převaha nižších, neboli *klesá* převaha vyšších kategorií veličiny Y nad kategoriemi nižšími, a naopak. V logitu použitý způsob konfrontace je vzhledem ke konvenčně vzestupně uspořádanému číslování kategorií vysvětlované proměnné oproti bazickým či řetězovým logitům vlastně opačný. V zájmu dosažení obvyklé interpretace parametrů se proto často model (15), kde

$$\mathbf{x}'\boldsymbol{\beta}_j = \beta_{0j} + \sum_{i=1}^{k} \beta_{ij} x_i \,,$$

zapisuje spíš jako

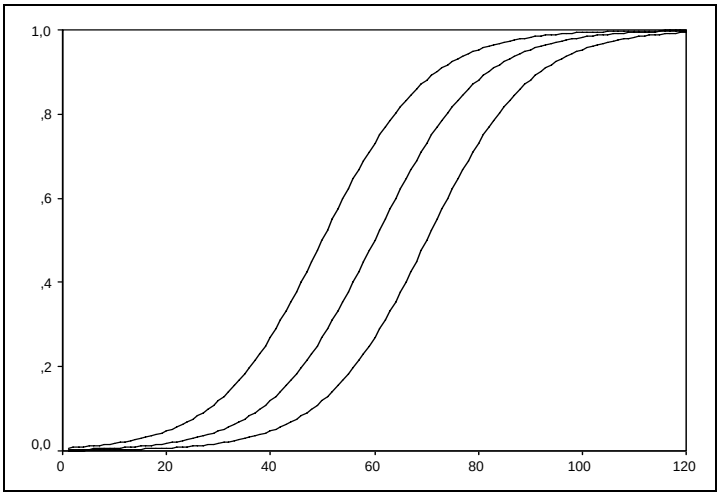
$$\ln \frac{F_j}{1-F_j} = \beta_{0j} - \sum_{i=1}^{k} \beta_{ij} x_i \,, j = 0, 1 \ldots, s-2.$$

Konečně *kombinovaným* logitem nazvěme logaritmus zlomku, v němž je použita hodnota pravděpodobnostní funkce i hodnota funkce distribuční, tedy například šance, že vysvětlovaná proměnná nabude hodnoty z j-té, a nikoliv z některé předchozí kategorie. Regresní model pak lze zapsat jako

$$\ln \frac{\pi_j}{\pi_0 + \pi_1 + \ldots + \pi_{j-1}} = \mathbf{x}'\boldsymbol{\beta}_j \,, j = 1, 2 \ldots, s-1.$$

Tento model lze vlastně odhadovat postupně pro jednotlivé kategorie na základě soustavy binárních logistických regresních funkcí.

Obr. č. 2: Soustava logistických křivek – příklad



Ve všech výše uvedených regresních funkcích odhadujeme (metodou maximální věrohodnosti s využitím iteračního algoritmu) celkem $(k + 1)(s - 1)$ parametrů. Výhodné proto je, můžeme-li vliv jednotlivých vysvětlujících proměnných na změnu různých šancí považovat za zhruba stejný. Pokud se parametry β pro všechny kategorie vysvětlované proměnné shodují, jsou změny logaritmů v modelu používaných šancí úměrné jen změnám hodnot vysvětlujících proměnných (model proporcionální šance). V dvourozměrném grafu jej představuje soustava $s - 1$ s-křivek (příklad viz obrázek 2). O smyslu použití modelu proporcionální šance lze rozhodnout na základě testu souběžnosti (*parallelism test*), kterým je zjišťována významnost snížení deviance modelu, pokud jsou namísto jednoho shodného odhadnuty různé vektory parametrů.

Většina statistických výpočetních systémů dnes procedury regresní analýzy tohoto typu běžně obsahuje. Doporučit lze například systém SPSS, v jehož nabídce lze nalézt binární i multinomickou logistickou regresi, ale také různé typy logitových (a také probitových) modelů pro binární, nominální i ordinální vysvětlovanou proměnnou.

Literatura

[1] AGRESTI, A., 1995: *Categorical Data Analysis*. New York, John Wiley & Sons, 1995.

[2] HOSMER, D. W. – LEMESHOW, S., 2000: *Applied Logistic Regression.* New York, John Wiley & Sons, 2000.

[3] JOBSON J. D., 1992: *Applied Multivariate Data Analysis.* 1992, Volume II, Categorical and Multivariate Methods, New York, Springer-Verlag.

[4] PECÁKOVÁ, I., 2006: Analýza a modelování souvislostí kategoriálních proměnných. Habilitační práce, 2006.

[5] SIMONOFF, J. S., 2000: *Analyzing Categorical Data*. New York, Springer-Verlag, 2000.

[6] POWERS, D. A. – XIE, Yu, 2000: *Statistical Methods for Categorical Data Analysis*. San Diego, Academic Press, 2000.

Logistická regrese s vícekategoriální vysvětlovanou proměnnou

Iva Pecáková

Abstrakt

Regresní model s vícekategoriální vysvětlovanou proměnnou je přirozeným zobecněním modelu s binární vysvětlovanou proměnnou. Založeno je na použití bazických logitů. Při jeho sestavování a vyhodnocování jeho kvality se používají analogické nástroje jako v případě vysvětlované proměnné binární. Jsou-li kategorie vysvětlované proměnné uspořádány (ordinální proměnná), může konstrukce modelu vycházet z řetězových, kumulativních či kombinovaných logitů. Způsob konstrukce modelu ovlivňuje význam a interpretaci parametrů.

Klíčová slova: kategoriální vysvětlovaná proměnná; logistická regrese; logitové modely.

Logistic regression with categorical dependent variable

Abstract

The regression model with categorical dependent variable is a natural generalization of the model with binary dependent variable. It is based on the use of baseline logits. For its building and for the evaluation of its quality, analogous procedures to the case of binary dependent variable are applied. When the categories of dependent variable are ordered (ordinal variable) the construction of model can be based on adjacent or cumulative logits or on proportional odds. The way of building of the model influences the meaning and the interpretation of its parameters.

Key words: categorical dependent variable; logistic regression; logit models.

JEL classification: G30