ROBUST'98, 187 – 194

LOGISTICKÁ REGRESE A VYHLEDÁVÁNÍ MODELŮ

Josef TVRDÍK

KI PřF OU, Ostrava

Abstract. Basic concepts of logistic regression with respect to the analysis of categorial data are explained. An example of searching risk factors of respiratory symptoms from a large questionnaire data by logistic models is described and a new attempt to the presentation of modelling results in a very brief form by one-character mnemonic code is shown.

Резюме: В работе описаны основные концепты логистической регресии с точки зрения применений в анализе категорияьных данных. Применение поиска логистических моделей показано на примере анализа данных из крупной анкетной студии респираторных симптомов. Описан новый способ сжатой презентации резуьтатов из моделирования с применением однобуквового мнемонического кода.

1. Logistický regresní model

K logistickému regresnímu modelu můžeme dojít na př. ze zobecněného lineárního modelu (GLM)

$$g[E(Y|\mathbf{x})] = \mathbf{x}'\beta, \tag{1}$$

ve kterém nějaká funkce g podmíněné střední hodnoty náhodné veličiny Y je vyjádřena jako lineární funkce vektoru regresorů $\mathbf{x}' = (1, x_1, x_2, \ldots, x_s)$ s regresními koeficienty $\beta' = (\beta_0, \beta_1, \ldots, \beta_s)$. 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*,

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right), \tag{2}$$

dojdeme k logistickému regresnímu modelu

$$\ln\left(\frac{p}{1-p}\right) = \mathbf{x}'\beta, \tag{3}$$

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

Parametry $\beta_0, \beta_1, \ldots, \beta_s$ regresního modelu (3) lze odhadovat metodou maximální věrohodnosti. Algoritmy pro nalezení těchto odhadů $b_0, b_1, \ldots, b_s$ jsou již řadu let implementovány v běžně dostupných statistických programech. Některé softwarové prostředky, na př. S-PLUS nebo SAS, dovolují

dokonce výběr mezi různě formulovanými věrohodnostními funkcemi (podmíněná nebo nepodmíněná věrohodnostní funkce), viz [2]. Dostupnost programové podpory je patrně jedním z hlavních faktorů, které přispívají k častému využívání logistické regrese v analýze dat. Druhým důležitým faktorem je poměrně snadná a přímočará interpretace odhadů parametrů logistického regresního modelu. 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, na př. při sázkách. Česká terminologie není ustálená, užívá se křížový poměr, interakce, také 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*.

$$OR = \frac{odds_1}{odds_0} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \tag{4}$$

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. [1]

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

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 $\{0, 1\}$, t.zv. 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

$$(5) \quad (d_{ij})_l = \begin{cases} 1 & \text{když} \quad x_{li} = j \\ & \qquad\qquad j = 1, 2, \ldots, k_i - 1, \quad l = 1, 2, \ldots, s \\ 0 & \text{jinak} \end{cases}$$

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

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

a interval spolehlivosti pro *OR*

$$(6) \qquad \left(e^{b_{ij} - u(1-\alpha/2)SE(b_{ij})}, \quad e^{b_{ij} + u(1-\alpha/2)SE(b_{ij})}\right),$$

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.

Studie CESAR

Logistická regrese byla využita jako metoda modelování v analýze dotazníků o zdravotním stavu v rozsáhlé mezinárodní studii *CESAR* (*C*entral *E*uropean *S*tudy on *A*ir pollution and *R*espiratory health) v rámci projektu

PHARE [4]. Studie probíhala v létech 1995-97 za spolupráce institutů veřejného zdraví v šesti zemích střední a východní Evropy (Bulharsko, Česko, Maďarsko, Polsko, Rumunsko, Slovensko) a partnerských institucí v zemích Evropské unie, mezi kterými byly Národní ústav veřejného zdraví a životního prostředí v Bilthovenu a Londýnská fakulta hygieny a tropických nemocí (LSHTM). Českým partnerem byla Krajská hygienická stanice v Ostravě.

Z vybraných oblastí lišících se úrovní znečištění prostředí bylo v každé participující zemi do šetření vybráno téměř 4000 dětí ve věku 7 až 11 let. Rodiče každého z těchto dětí vyplňovali obsáhlý dotazník. Otázky se týkaly zdravotního stavu dítěte se zaměřením na choroby dýchacího systému a alergická onemocnění, zdravotního stavu rodičů a sourozenců, životních podmínek rodiny, stravovacích návyků, kouření, pohybové aktivity a socioekonomického statutu.

Jedním z dílčích cílů studie CESAR bylo získat co nejspolehlivější data s co nejméně chybějícími údaji. Mimo jiné data z dotazníku byla vkládána dvakrát dvěma různými osobami a neshody řešeny v některých případech i dokonce telefonickými dotazy na rodiče, který dotazník vyplňoval. Rovněž značná pozornost byla věnována předběžné analýze, zejména možnému slučování kategorií a nalezení vhodných odvozených veličin, kdy bylo nutno uvažovat protichůdné požadavky: ztratit co nejméně informací z původních dat a dosáhnout přijatelně vysoké četnosti ve všech kategoriích, i při vícenásobném třídění. Navíc bylo nutno nalézt taková pravidla pro slučování a transformace, která vyhovují pro data ze všech zúčastněných zemí.

Po tomto předzpracování dat bylo v souboru 3670 dětí stanoveno celkem 44 kategoriálních veličin (některé dichotomické, některé s více kategoriemi), které byly považovány za veličiny vysvětlující, tedy regresory, a 10 dichotomických veličin znamenajících přítomnost některého respiračního symptomu.

2. Statistické modelování

Logistické regresní modely měly nalézt v datech rizikové a protektivní faktory výskytu respiračních symptomů. Cílem tohoto vyhledávání bylo nalézt faktory, které pak budou využity jako t.zv. matoucí proměnné (confounders) v modelech vlivu znečištění ovzduší na zdravotní stav dětí. Vzhledem k vysokému počtu vysvětlujících proměnných bylo nutno rezignovat na hledání úplných modelů s interakcemi. Po různých kompromisních úvahách bylo rozhodnuto, že rizikové a protektivní faktory budou vyhledávány ve třech různých modelech:

Model	fixně zařazené regresory	postup vyhledávání	počet modelů
1	age, sex, area	přidán vždy jen 1 z 41	41
2	11 vybraných	přidán vždy jen 1 z 33	33
3	11 vybraných	stepwise	1

Účelem této práce není publikovat dílčí výsledky studie [4], ale ilustrovat situaci vedoucí k rozsáhlým výstupům, které je nutno strukturovat a stručně prezentovat. Proto je vynechán podrobnější popis modelů i analýzovaných veličin.

Výpočty byly provedeny s pomocí programového paketu STATA [3], kterým byla v rámci studie CESAR vybavena všechna zúčastněná pracoviště. Tento paket byl vybrán zejména z těch důvodů, že v něm lze zpracovávat rozsáhlá data, zadávat dlouhé výpočty dávkově z příkazových souborů a velmi dobře podporuje právě logistickou regresi [5]. Dovoluje velmi pohodlné zadávání transformací kategoriálních regresorů na indikátory při uživatelově volbě referenční kategorie a volbu výpočetního modulu, který prezentuje výsledky přímo jako odhady odds ratio s intervalem spolehlivosti – viz vztah (6). Přes tyto pro podobnou aplikaci logistické regrese přátelské rysy statistického paketu bylo však nutno se vypořádat s rozsahem výstupů a jejich přehlednou prezentací. Z údajů v tabulce je zřejmé, že výstupem bylo $10(41 + 33 + 1) = 750$ logistických modelů, tedy několik set stran textu. K tomu, aby bylo možno snadněji nahlédnout do výsledků, vhodně je strukturovat a porovnávat výsledky pro všech 10 vysvětlovaných proměnných i pro všech šest zúčastněných zemí, bylo nutno uvažovat o co největší kondenzaci a zkratkové prezentaci výsledků.

3. Jak prezentovat výsledky stručně

Počítačový výstup z logistické regrese (pro jeden model, jednu vysvětlovanou proměnnou) má pro interpretaci přinejmenším tolik zajímavých řádků, kolik je v modelu indikátorů – viz rov.(5). Každý z těchto řádků obsahuje mimo jiné odhad příslušného *odds ratio* a jeho interval spolehlivosti. V úvahách o stručnější prezentaci je nutné se zabývat možnostmi snížení počtu řádků a snížením počtu sloupců výstupu, obojí s cílem, aby na okem přehlédnutelném počtu řádků mohly být v jedné tabulce vedle sebe stručně zaznamenány výsledky z různých modelů pro různé vysvětlované veličiny.

Počet řádků můžeme redukovat na počet původních vysvětlujících veličin (kategoriálních regresorů před transformací na indikátory). U dichotomických původních regresorů k žádnému snížení počtu řádků oproti počítačovému výstupu nedojde, ale vzhledem k tomu, že v uvažované úloze měla řada původních regresorů více než dvě kategorie, některé dokonce i pět kategorií, je výsledek zkrácení počtu řádků podstatný, v našem případě ze 72 na 41. Ve zkrácené prezentaci má tedy každý původní regresor pouze jeden řádek bez ohledu na to, kolik má kategorií.

Na každém řádku je pak nutno co nejúsporněji zobrazit výsledek z regresního modelu. Pro každý původní regresor lze testovat pomocí poměru věrohodnosti hypotézu, zda regresor významně ovlivňuje odhadovaný logit. Pro dichotomický regresor je pro velké výběry tento test ekvivalentní testu

hypotézy, zda $\beta_{i1} = 0$. Tato hypotéza se zamítá, nepokrývá-li interval spolehlivosti pro *OR* jedničku. U regresoru s více kategoriemi je to poněkud složitější: buď se můžeme spokojit s obvyklým výsledkem testu poměrem věrohodnosti, tedy s výsledkem ano–ne, s možností kódování znakem $*$ nebo mezera (tím ale nevyužijeme informaci o znaménku odhadů regresních koeficientů b_{ij}), nebo i ve zkratkovitém výstupu zachytit více informací, které je možné zakódovat jedním znakem. Nechť i je zvolený index označující původní regresor, $i \in [1, s],\ j = 1, 2, \ldots, k_i - 1$. Pak můžeme rozlišit tyto situace:

(1) Všechny β_{ij} jsou nulové (*OR* rovny jedné)
(2) Alespoň jeden $\beta_{ij} < 0$, t.j. $OR < 1$, žádný $\beta_{ij} > 0$
(3) Alespoň jeden $\beta_{ij} > 0$, t.j. $OR > 1$, žádný $\beta_{ij} < 0$
(4) Alespoň jeden $\beta_{ij} > 0$, alespoň jeden $\beta_{ij} < 0$ (může nastat jen u vícekategoriálních regresorů, kdy $k_i > 2$)

Pro tyto čtyři možné výsledky se celkem přirozeně nabízí kódování $\{mezera, -, +, \sim\}$, případně i nějaké jiné kódování, které bude uchovávat tuto přirozenou mnemotechniku a jehož znaky ve výsledných tabulkách budou výraznější a přehlednější. V projektu CESAR bylo zvoleno kódování $\{mezera, \mathrm{v}, \#, \sim\}$. Pro zachycení výsledků z jednoho modelu potřebujeme pro každý původní regresor jen jeden znak a přitom máme uchovanou informaci i o směru vlivu ve srovnání s referenční kategorií, tedy to, zda je vliv regresoru rizikový či protektivní.

Ukázka takové prezentace výsledků je v tabulce 1. Jde jen o výsek z výsledků [4]. Trojice znaků ve sloupci jednotlivých vysvětlovaných veličin kóduje výsledek získaný modelem 1, 2 a 3. Sledované respirační symptomy (vysvětlované veličiny) uvedené ve sloupcích tabulky byly postupně kašel v zimě (B3), noční kašel (B4), sípání nebo hvízdání na prsou někdy v minulosti (B5), sípání nebo hvízdání na prsou v posledním roce (B6) , buzení pro dechové potíže (B7r) a zahlenění (B12). Pro čtení tabulky je u vícekategoriálních regresorů důležité, která kategorie je zvolena jako referenční. U těch regresorů, kde referenční kategorie není zřejmá z názvu regresoru, je tato kategorie uvedena v závorkách v prvním sloupci tabulky.

Uvedený způsob komprimace výsledků je vhodný pro explorativní analýzu dat, kdy v datech vyhledáváme zajímavé či podezřelé vztahy, nikoli pro konfirmační analýzu, kdy zkratkový výsledek k interpretaci nepostačuje.

Z tabulky 1 vidíme, které regresory v různých modelech ovlivňují výskyt sledovaných respiračních symptomů. Na příklad vlhkost a plíseň je rizikový faktor vždy kromě modelu 3 a symptomu B3, zatímco konzumace ovoce neovlivňuje žádný z uvedených respiračních symptomů.

V tab. 1 se neobjevuje ani jednou symbol $\sim$. Důvod je celkem pochopitelný: Výskyt tohoto znaku indikuje nevhodnou volbu referenční kategorie, při nejmenším s hlediska interpretace, a tab. 1 je výsekem ze závěrečných výsledků, kdy referenční kategorie už byly voleny i s ohledem na interpretaci.

TABULKA 1. Ukázka části výsledků logistických modelů

Respir. symptomy Rizikové faktory	B3	B4	B5	B6	B7r	B12
Kouř_ mat(ne)	#	#				
Vzděl_mat(S+VŠ)	# # #	# # #	#	#	#	v v v
Kouř_ot (ne)	#			#		
Vzděl_ot (S+VŠ)	# # #	#		#	#	v v
Plyn (kuchyň)	# #	# #	# # #	#		
Vlhk./plíseň (ne)	# #	# # #	# # #	# # #	# # #	# # #
Poč.st.souroz.(0)	v v			# #	#	
Narozen dříve(ne)	#	# #	#	# #	# # #	# #
Por.váha<2,5kg	#	# # #	#		#	
Perinat_kompl(ne)	# #	#	# # #	# # #	#	# # #
Kojení						
Ovoc_ léto(denně)						
Ovoc_ zima(denně)						
Zel_léto (denně)	v				v v v	
Zel_zima (denně)		v v v		v v v	v v	
Ryby (často)	#	#				
Užív.synt vit.	# #	# # #	#	# #	# # #	# # #
Pobyt venku(max)	# #	# #	# # #	# # #	#	# #

Za povšimnutí stojí výsledek týkající se konzumace zeleniny v zimě. Mezi lékaři se považuje pravidelná konzumace čerstvé zeleniny v zimě za protektivní faktor respiračních onemocnění. Při zběžném pohledu na tabulku vidíme v tomto řádku u některých modelů a symptomů opravdu symboly znamenající snížení rizika oproti referenční kategorii. Ale pozor - jako referenční kategorie byla zvolena kategorie s největší konzumací čerstvé zeleniny, takže zjištěné výsledky jsou v rozporu s očekáváním. Zdá se, že se v tomto případě projevilo to, co je známo z mnoha aplikací regresních modelů, totiž záměna příčiny a důsledku. Nabízí se vysvětlení, že rodiče dětí majících potíže dopřejí potomkům dietu podle doporučení lékařů. S podobnou záměnou příčiny a důsledku je možné se setkat v analýze biomedicínských dat poměrně často. Na př. před léty byl vyšetřován zdravotní stav vybraných profesních skupin a jeho závislost na spoustě socioekonomických a dalších veličin. V analýze dat byla užita i metoda GUHA, u které výsledky mohou mít formu implikací, platných v analýzovaných datech. Vygenerované výsledky obsahovaly tvrzení typu „kouření & alkohol & káva implikuje dobré zdraví“. Lékaři však neinterpretovali tyto výsledky tak, jak nabízel výstup z počítače, ale k zármutku nás hříšníků jako záměnu příčiny a důsledku, t.j. dobrý zdravotní stav vede k méně úzkostlivé životosprávě.

4. Závěr

V příspěvku byly připomenuty základy logistické regrese s zřetelem na aplikace v analýze kategoriálních dat. Na příkladu analýzy rozsáhlých dat byly ilustrovány některé problémy prezentace výsledků a uveden postup umožňující kondenzovat výsledky regresního modelování a tím usnadnit porovnávání výsledků na hrubé rozlišovací úrovni a poskytnout orientaci pro další interpretaci.

Příspěvek může posloužit jako jedna z provokací k důslednějšímu řešení problémů prezentace rozsáhlých výsledků statistického modelování a případnému zařazení vhodných způsobů stručné prezentace výsledků mezi statistickým softwarem podporované postupy analýzy dat.

Za trpělivé čtení a užitečné připomínky děkuji původně anonymnímu recenzentovi, později známému recenzentovi M. Malému.

Literatura

[1] Armitage P., Berry G.: *Statistical Methods in Medical Research*, Blackwell Sci Publ., 1987 (reprinted 1994)

[2] Kleinbaum D.G.: *Logistic Regression: A Self-Learning Text*, Springer-Verlag, New York Berlin Heidelberg, 1994

[3] StataCorp.: *Stata Statistical Software: Release 5*, College Station, Texas, Stata Corporation, 1997

[4] Šplíchalová A., Volf J., Tomášková H., Tvrdík J.: *Studie vlivu znečištění ovzduší na dýchací systém dětí*, závěrečná zpráva dílčí části úkolu, KHS Ostrava, 1997

[5] Tvrdík J.: *STATA a NCSS z pohledu uživatele*, Inf.bul.ČStS **8**(3), 5-16, 1997