M5VM05 Statistické modelování

11. Konkrétní GLM modely – II.

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

Motivace

Na minulé přednášce jsme si uvedli zobecněné lineární modely pro alternativní, binomická a poissonovská data. Tato přednáška navazuje na přednášku minulou. Nejprve budeme zkoumat problémy příliš velkého nebo příliš malého rozptylu v datech. Dále pak nastíníme modelování multinomických dat a jeho využití v testování nezávislosti v kontingenčních tabulkách.

Overdispersion, underdispersion

Předpokládáme, že náhodný výběr $\mathbf{Y}_n = (Y_1, \dots, Y_n)^{\mathrm{T}}$ z rozdělení exponenciálního typu se řídí GLM modelem, tj.

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \theta_i) = \exp\left\{ \sum_{i=1}^{n} \frac{y_i\theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right\}.$$

Předpokládejme, že platí

$$\psi_i(\phi) = \frac{\phi}{\omega_i} > 0,$$

kde $\omega_i > 0$ jsou známé **apriorní váhy** a $\phi > 0$ je neznámý **rušivý parametr**.
Škálová deviace

$$\begin{aligned} D &= 2\left[l^*(\widehat{\boldsymbol{\beta}}_{max}; \mathbf{Y}) - l^*(\widehat{\boldsymbol{\beta}}; \mathbf{Y}) \right] \\ &= \frac{1}{\phi} 2 \sum_{i=1}^{n} \omega_i \left[Y_i(\hat{\theta}_{i,max} - \hat{\theta}_i) - \gamma(\hat{\theta}_{i,max}) + \gamma(\hat{\theta}_i) \right] \\ &= \frac{1}{\phi} D^* \end{aligned}$$

a D^* nazveme **neškálovou deviací** (unscaled deviance).

Overdispersion, underdispersion

Protože platí

$$D = \frac{1}{\phi} D^* \overset{A}{\sim} \chi^2(n-k) \quad \Rightarrow \quad ED = \frac{1}{\phi} ED^* \approx n-k,$$

pak

$$\hat{\phi}_{D^*} = \frac{D^*}{n-k}.$$

Další často používanou mírou vhodnosti modelu je tzv. zobecněná Pearsonova statistika

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \overset{A}{\sim} \chi^2(n-k)$$

a proto dalším momentovým odhadem založeným na této statistice je

$$\hat{\phi}_{X^2} = \frac{X^2}{n-k}.$$

Overdispersion, underdispersion

Přehled rušivých parametrů

Rozdělení	ϕ
Normální rozdělení	σ^2
Poissonovo rozdělení	1
Binomické rozdělení	1
Gamma rozdělení	$1/\alpha$

Overdispersion, underdispersion

V prostředí R je k řešení tohoto problému k dispozici modifikovaná volba pro třídu exponenciálního rozdělení. V případě binomického rozdělení máme možnost volby

`family=quasibinomial`

a pro Poissonovo rozdělení

`family=quasipoisson`.

Nejde o nový typ exponenciálního rozdělení, ale o změnu ve výpočtu druhého momentu, pro jehož odhad se použije jednoduchý momentový odhad disperzního parametru ϕ. Výsledná korekce rozptylu je pak důležitá při testování hypotéz, neboť zohledňuje vyšší/nižší variabilitu v datech a zabraňuje tak nadbytku/nedostatku falešně pozitivních výsledků testů hypotéz o parametrech modelu.

Příklad

Příklad 1

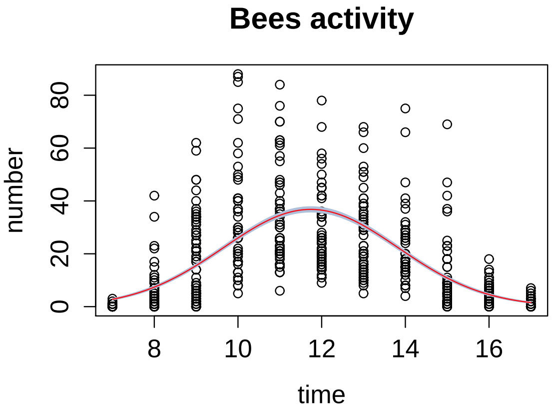
V souboru „`bees.RData`“ jsou uvedeny údaje o aktivitě včel v závislosti na čase. Jednou z důležitých charakteristik při zkoumání včelí aktivity je počet včel, které opustí úl kvůli práci ve vnějším prostředí. Studie se zabývala měřením této veličiny během několika slunečných dní v závislosti na čase během dne. Datový soubor obsahuje tyto proměnné

`number`	*počet včel, které opustily úl*
`time`	*čas, kdy byl tento údaj zaznamenán*

Modelujte závislost počtu včel, které opustí úl, na čase během dne.

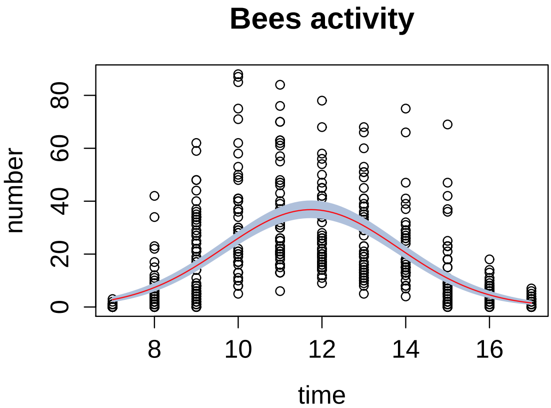
Řešení. Pro modelování závislosti použijeme poissonovský model s kanonickou linkovací funkcí. Do modelu vstupuje jediná vysvětlující proměnná `time` a přidáme také její druhou mocninu.
Hodnota reziduální deviace (4 879,3) je nepoměrně vyšší než počet stupňů volnosti (501). Je zřejmé, že došlo k „overdispersion“ a v jazyce R je třeba volit `family=quasipoisson`. Použití této volby neovlivňuje odhady koeficientů, ale mění jejich odhady variability, což se projeví např. v intervalu spolehlivosti.

Příklad



Obrázek : Odhad regresní funkce **bez** vyrovnání se s problematikou velkého rozptylu.

Příklad



Obrázek : Odhad regresní funkce **s** vyrovnáním se s problematikou velkého rozptylu.

Modely pro multinomická data

Náhodný výběr $\mathbf{Y} = \mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, pro který $n = J \cdot K$, tj.

$$\mathbf{Y} = \mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}} = (Y_{11}, \ldots, Y_{1K}, \ldots, Y_{J1}, \ldots, Y_{JK})^{\mathrm{T}}.$$

Předpokládejme, že náhodný výběr $\mathbf{Y}$ je z Poissonova rozdělení, tj.

$$Y_{jk} \sim Po(\lambda_{jk}) \qquad j = 1, \ldots, J;\ k = 1, \ldots, K$$

s tzv. **celkovou** dodatečnou podmínkou

$$N = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} \qquad N \in \mathbb{N}^+,$$

kde y_{jk} jsou realizace náhodných veličin Y_{jk}.

Modely pro multinomická data

Rozdělení náhodného vektoru $\mathbf{Y}$ za podmínky $Z_{..} = N$ je **multinomické**

$$p_{\mathbf{Y}|Z_{..}=N}(\mathbf{y}) = \begin{cases} N! \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{\pi_{jk}^{y_{jk}}}{y_{jk}!} & \textit{pro} \quad y_{jk} = 0, 1, \dots, N; \quad j = 1, \dots, J; \\ & \qquad\qquad\qquad\qquad\qquad k = 1, \dots, K, \\ & \quad \sum_{j=1}^{J} \sum_{k=1}^{K} y_{jk} = N \qquad \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{jk} = 1 \\ 0 & \text{jinak} \end{cases},$$

tj.

$$\mathbf{Y}|Z_{..} = N \quad \sim \quad Mn(N, \pi_{11}, \dots, \pi_{1K}, \dots, \pi_{J1}, \dots, \pi_{JK}) \,,$$

přičemž

$$\begin{aligned} EY_{jk} =& N\pi_{jk} \\ DY_{jk} =& N\pi_{jk}(1 - \pi_{jk}) \\ C(Y_{jk}, Y_{j'k'}) =& -N\pi_{jk}\pi_{j'k'} \end{aligned}$$

Kontingenční tabulky

Realizace náhodných veličin i teoretické pravděpodobnosti lze uspořádat do tzv. **kontingenční tabulky**:

Kontingenční tabulka četností

faktor A	faktor B				$\sum$
	B_1	B_2	$\cdots$	B_K	
A_1	y_{11}	y_{12}	$\cdots$	y_{1K}	$N_{1.}$
A_2	y_{21}	y_{22}	$\cdots$	y_{2K}	$N_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	y_{J1}	y_{J2}	$\cdots$	y_{JK}	$N_{J.}$
$\sum$	$N_{.1}$	$N_{.2}$	$\cdots$	$N_{.K}$	$N = N_{..}$

Kontingenční tabulky

Kontingenční tabulka pravděpodobností

faktor A	faktor B				$\sum$
	B_1	B_2	$\cdots$	B_K	
A_1	π_{11}	π_{12}	$\cdots$	π_{1K}	$\pi_{1.}$
A_2	π_{21}	π_{22}	$\cdots$	π_{2K}	$\pi_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	π_{J1}	π_{J2}	$\cdots$	π_{JK}	$\pi_{J.}$
$\sum$	$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.K}$	$\pi_{..} = 1$

Kontingenční tabulky

Nejčastěji se v kontingenčích tabulkách testuje hypotéza, že

$$\boxed{\textbf{faktory A a B jsou nezávislé}}$$

faktor A	faktor B $\cdots$	B_k	$\cdots$	$\sum$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_j	$\cdots$	$\pi_{j.}\pi_{.k}$	$\cdots$	$\pi_{j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum$	$\cdots$	$\pi_{.k}$	$\cdots$	1

tj.

$$\pi_{jk} = \pi_{j.}\pi_{.k}, \text{ takže potom } EY_{jk} = N\pi_{j.}\pi_{.k}\,, \text{ přičemž } \sum_{j=1}^{J} \pi_{j.} = \sum_{k=1}^{K} \pi_{.k} = 1.$$

Pro model s **celkovou dodatečnou podmínkou** lze hypotézu o **nezávislosti dvou faktorů** definovat takto

$$EY_{jk} = N\pi_{j.}\pi_{.k}\,, \qquad \text{přičemž} \qquad \sum_{j=1}^{J} \pi_{j.} = 1 \qquad \text{a} \qquad \sum_{k=1}^{K} \pi_{.k} = 1.$$

V GLM s **log-lineární linkovací funkcí** máme $\eta_{jk} = \log EY_{jk} = \mathbf{x}_{jk}^{\mathrm{T}}\boldsymbol{\beta}$, tedy

$$\eta_{jk} = \log EY_{jk} = \log(N\pi_{j.}\pi_{.k}) = \underbrace{\mu}_{=\log N} + \underbrace{\alpha_j}_{=\log \pi_{j.}} + \underbrace{\beta_k}_{=\log \pi_{.k}}\,.$$

Pokud bychom nepředpokládali nezávislost faktorů A a B, dostaneme **maximální model**

$$\eta_{jk} = \log EY_{jk} = \log(N\pi_{jk}) = \underbrace{\mu}_{=\log N} + \underbrace{\alpha_j + \beta_k + (\alpha\beta)_{jk}}_{=\log \pi_{jk}}$$

Hypotéza nezávislosti dvou faktorů v kontingenčních tabulkách je ekvivalentní s hypotézou neexistence interakcí v analýze rozptylu (deviace), tj.

$$H_0: \quad (\alpha\beta)_{jk} = 0 \qquad\qquad j = 1, \ldots, J; \;\; k = 1, \ldots, K.$$

Příklad

Příklad 2

V následující kontingenční tabulce jsou obsaženy údaje studie 400 pacientů o počtech různých typů onemocnění rakovinou kůže (Malignant Melanoma) v závislosti na části těla, kde se vyskytují.

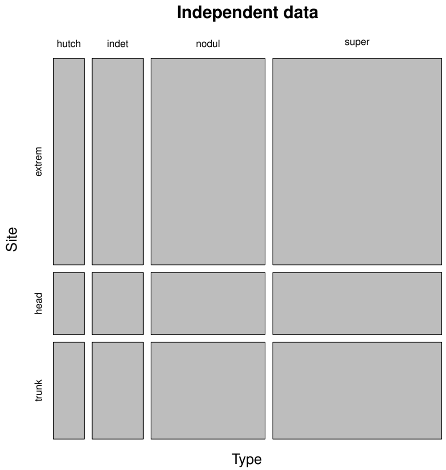
	Část těla		
Typ rakoviny	*končetiny*	*hlava a krk*	*trup*
Hutchinson's melanotic freckle	10	22	2
neurčitý	28	11	17
Nodular	73	19	33
Superficial spreading melanoma	115	16	54

Na hladině významnosti $\alpha = 0,05$ *testujte hypotézu, zda typ rakoviny kůže závisí na části těla, kde se vyskytuje.*

Příklad

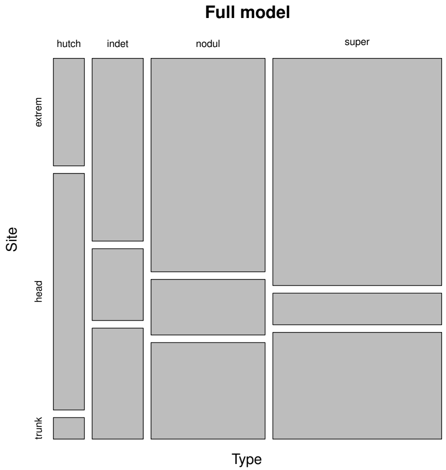
Řešení Nejprve definujeme oba log-lineární modely, tj. model `m1`, který předpokládá nezávislost obou faktorů a model `m2`, který počítá i s interakcemi. Model `m1` je tedy submodelem modelu `m2`. K testování využijeme analýzu deviace, Pearsonův test. Jeho p-hodnota vychází $2,05 \times 10^{-9}$ a proto zamítáme hypotézu o nezávislosti typu rakoviny kůže na části těla, kde se vyskytuje. Výsledky obou modelů lze také znázornit pomocí mozaikového grafu.

Příklad



Obrázek : Mozaikový graf pro model, který předpokládá nezávislost.

Příklad



Obrázek : Mozaikový graf pro model s interakcemi.

Úlohy k procvičení I

Příklad 2.1

V souboru „`sharks.RData`“ jsou k dispozici data, která popisují počty napadení žraloky na Floridě v letech 1946 až 1999. Známe také velikost populace. Datový soubor obsahuje tyto proměnné:

`Year`	*rok*
`Population`	*velikost populace*
`Attacks`	*počet napadení žraloky*
`Fatalities`	*počet úmrtí způsobených žraloky*

Nejprve vykreslete bodový graf počtu napadení na 1 milión obyvatel v závislosti na čase. Pro modelování použijte binomický i poissonovský model s kanonickou linkovací funkcí. Pro matici plánu uvažujte kubický polynom v proměnné `Year`.

Úlohy k procvičení II

Příklad 2.1

Predikce obou modelů i s intervalem spolehlivosti pro regresní funkci vykreslete do obrázku. Zkoumejte také, jestli nenastal problém příliš velkého nebo příliš malého rozptylu. Pokud ano, předefinujte model a výsledky znovu vykreslete do obrázku. Pomocí výsledného modelu odhadněte, kolik útoků (na 1 milión obyvatel) způsobí žraloci na Floridě v roce 2013 a také v jakém intervalu se tato hodnota s 95% pravděpodobností bude pohybovat.

[Nastal problém příliš velkého rozptylu. Odhad: 33,96 útoků na 1 milión obyvatel, interval spolehlivosti: $[3,207; 359,55]$.]

Úlohy k procvičení

Příklad 2.2

V následující kontingenční tabulce jsou obsaženy údaje o počtech různých typů onemocnění horních cest dýchacích (Respiratory Tract Infections) v závislosti na čase.

	Časové období				
Diagnóza	*1-3/96*	*4-6/96*	*7-9/96*	*10-12/96*	*1-3/97*
Acute bronchitis	113	58	40	108	100
Acute sinusitis	99	37	23	50	32
URI	410	228	125	366	304
Pneumonia	60	43	30	56	45

Na hladině významnosti $\alpha = 0{,}05$ *testujte hypotézu, zda onemocnění horních cest dýchacích závisí na čase.*

[závisí]