

M5VM05 Statistické modelování

10. Konkrétní GLM modely

Jan Koláček (kolacek@math.muni.cz)

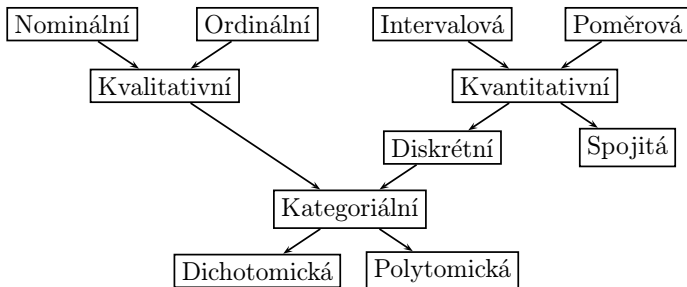
Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

podzim 2013



Na minulé přednášce jsme si uvedli obecnou definici zobecněného lineárního modelu a obecné konstrukce testů hypotéz o parametrech těchto modelů. Na této přednášce se již budeme zabývat zobecněnými lineárními modely pro konkrétní případy podle toho, jaké rozdělení má závisle proměnná Y .

Typy veličin:



Modely pro alternativní a binomická data

Předpokládejme, že $U_i \sim A(\pi_i)$ ($i = 1, \dots, N$) nabývá pouze dvou hodnot 0 a 1,

$$P(U_i = u) = \begin{cases} \pi_i & u = 1 \\ 1 - \pi_i & u = 0 \\ 0 & \text{jinak} \end{cases} = \begin{cases} \pi_i^u (1 - \pi_i)^{1-u} & u = 0, 1 \\ 0 & \text{jinak.} \end{cases}$$

Předpokládejme, že náhodná veličina U_i závisí na k veličinách x_{i1}, \dots, x_{ik} , tzv. **kovariáty**.

Data můžeme mít zadána různým způsobem:

- **jednotlivá pozorování U_i :**

hodnoty kovariát	pozorované binární veličiny
x_{i1}, \dots, x_{ik}	U_i

Modely pro alternativní a binomická data

- **skupinově**, kdy známe **absolutní** četnosti úspěchů Y_j a celkový počet pokusů n_j , tedy máme k dispozici binomická data $Y_j \sim Bi(n_j, \pi_j)$

$$P(Y_j = y) = \begin{cases} \binom{n_j}{y} \pi_j^y (1 - \pi_j)^{n_j - y} & y = 0, 1, \dots, n_j \\ 0 & \text{jinak} \end{cases}$$

kde

$$j = 1, \dots, n; \quad N = n_1 + \dots + n_n$$

a data můžeme zapsat formou tabulky

hodnota kovariát	počet úspěchů	počet pokusů
x_{j1}, \dots, x_{jk}	Y_j	n_j

Modely pro alternativní a binomická data

- **skupinově**, kdy známe **relativní** četnost úspěchů $Z_j = \frac{Y_j}{n_j}$ a celkový počet pokusů n_j

$$P(Z_j = y) = \begin{cases} \binom{n_j}{n_j y} \pi_j^{n_j y} (1 - \pi_j)^{n_j - n_j y} & y = 0, \frac{1}{n_j}, \dots, 1 \\ 0 & \text{jinak} \end{cases}$$

kde

$$j = 1, \dots, n; \quad N = n_1 + \dots + n_n$$

Data lze zapsat do tabulky

kovariáty	relativní úspěšnost	počet pokusů
x_{j1}, \dots, x_{jk}	$Z_j = \frac{Y_j}{n_j}$	n_j

Hlavním úkolem statistické analýzy je nalézt vztah mezi Z_i , (tj. i Y_i) a x_{i1}, \dots, x_{ik} ,
tj. funkci

$$\pi_i = \pi(\mathbf{x}_i) = \pi(x_{i1}, \dots, x_{ik}).$$

Protože chceme použít GLM modely, modelujeme pravděpodobnosti π_i pomocí linkovacích funkcí

$$g(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Nejjednodušším modelem je **lineární model**

$$\pi_i = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Avšak tento model má řadu nevýhod, především je třeba zajistit, aby $\mathbf{x}_i^T \boldsymbol{\beta}$ nabývala hodnot mezi 0 a 1, tedy je třeba přidat nějaké dodatečné podmínky. Proto, abychom tuto podmínku dodrželi, využijeme nějakou **distribuční funkci**

$$F(t) = \int_{-\infty}^t f(s) ds \quad f(s) \geq 0 \quad \int_{-\infty}^{\infty} f(s) ds = 1$$

s odpovídající **hustotou** $f(s)$, která se v tomto případě nazývá **toleranční funkce** (toleranční distribuce).

Modely dávka – odpověď

Typickým příkladem těchto modelů je vztah mezi dávkou toxické látky a odezvy (kladná-přežití, záporná-smrt) jedince na tuto dávku. Odezvy bývají obvykle udávány jako procenta kladné odezvy (quantal responses).

Symetrické modely

Jestliže uvažujeme toleranční distribuci jako **rovnoměrně spojitou** na nějakém intervalu (a, b) , tj

$$f_0(s) \sim Rs(a, b) \quad f(s) = \begin{cases} \frac{1}{b-a} & s \in (a, b) \\ 0 & \text{jinak} \end{cases}$$

pak

$$\pi_0(x) = F_0(x) = \int_a^x f(s) ds = \frac{x-a}{b-a} \quad \text{pro } x \in (a, b)$$

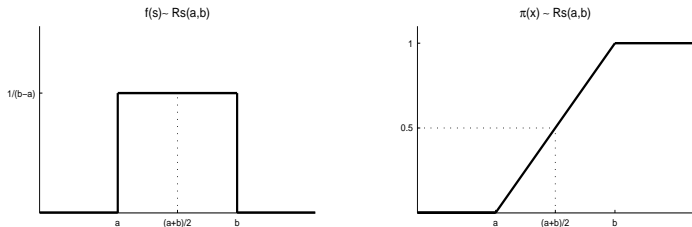
a tento model je **lineárním modelem**

$$\pi_0(x) = \frac{x-a}{b-a} = \beta_0 + \beta_1 x \quad \text{tj. } \beta_0 = -\frac{a}{b-a}, \beta_1 = \frac{1}{b-a} > 0$$

s identickou linkovací funkcí

$$g_0(\pi) = \pi.$$

Symetrické modely



Obrázek : Rovnoměrné rozdělení na (a, b) .

Další možností je vzít **normální hustotu** jako **toleranční funkci**. V tomto případě mluvíme o tzv. **probitovém modelu**:

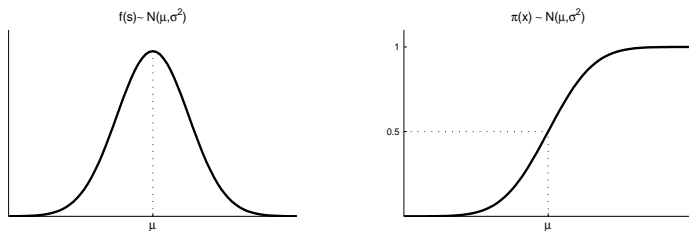
$$\pi_1(x) = F_1(x) = \int_{-\infty}^x f_1(s) ds = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2} ds = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

kde Φ je distribuční funkce standardizovaného normálního rozdělení. Pak tzv. **probitovou linkovací funkcí** je kvantilová funkce normálního rozdělení

$$g_1(\pi) = \Phi^{-1}(\pi) = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \quad \text{tj.} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

Hodnota mediánu $x = \mu$ se nazývá **mediánová smrtící dávka** (*median lethal dose* - LD50) a odpovídá dávce, při které polovina jedinců má kladnou a polovina zápornou odezvu.

Probitový model



Obrázek : Normální rozdělení $N(\mu, \sigma^2)$.

Jiným velmi podobným modelem je tzv. **logistický model**, kde toleranční funkce je hustota **logistického rozdělení**

$$f_2(s) = \frac{1}{\sigma} \frac{\exp\left(\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{s-\mu}{\sigma}\right)\right]^2} = \frac{1}{\sigma} \frac{\exp\left(-\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(-\frac{s-\mu}{\sigma}\right)\right]^2},$$

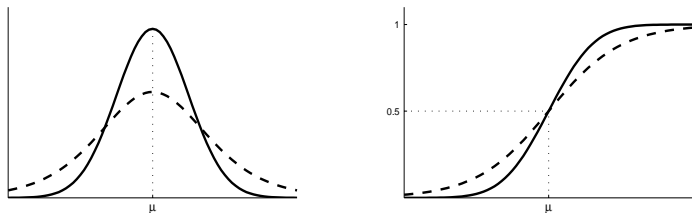
takže

$$\pi_2(x) = F_2(x) = \int_{-\infty}^x \frac{1}{\sigma} \frac{\exp\left(\frac{s-\mu}{\sigma}\right)}{\left[1+\exp\left(\frac{s-\mu}{\sigma}\right)\right]^2} ds = \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)} = \frac{1}{1+\exp\left(-\frac{x-\mu}{\sigma}\right)}$$

s tzv. **logit linkovací funkcí**

$$g_2(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \quad \text{tj.} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

Logistický model



Obrázek : Srovnání probitového a logistického (- -) modelu při stejných parametrech μ a σ .

Asymetrické (extremální) modely

Pokud za toleranční funkci zvolíme **Log-Weibullovo rozdělení** (*extreme-minimal-value distribution*) ve tvaru

$$f_3(s) = \frac{1}{\sigma} \exp\left(\frac{s-\mu}{\sigma}\right) \exp\left[-\exp\left(\frac{s-\mu}{\sigma}\right)\right],$$

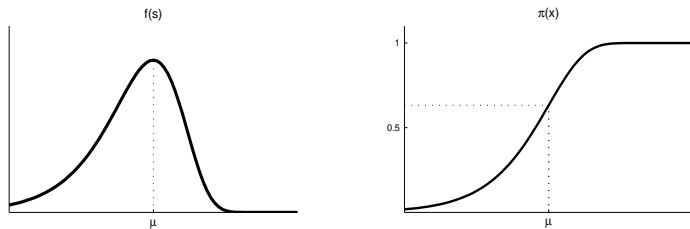
pak

$$\pi_3(x) = \int_{-\infty}^x \frac{1}{\sigma} \exp\left(\frac{s-\mu}{\sigma}\right) \exp\left[-\exp\left(\frac{s-\mu}{\sigma}\right)\right] ds = 1 - \exp\left[-\exp\left(\frac{x-\mu}{\sigma}\right)\right]$$

s tzv. **komplementární log-log linkovací funkcí**

$$g_3(\pi) = \log[-\log(1 - \pi)] = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \quad \text{tj.} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

CLogLog model



Obrázek : Log-Weibullovo rozdělení.

LogLog model

Pokud jako toleranční funkci zvolíme **zobecněné Gumbelovo rozdělení** (*extreme-miaximal-value distribution*) ve tvaru

$$f_4(s) = \frac{1}{\sigma} \exp\left(-\frac{s-\mu}{\sigma}\right) \exp\left[-\exp\left(-\frac{s-\mu}{\sigma}\right)\right],$$

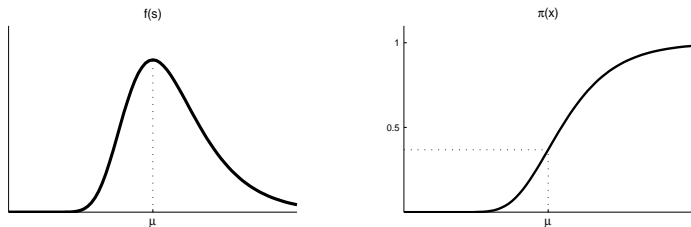
dostaneme

$$\pi_4(x) = \int_{-\infty}^x \frac{1}{\sigma} \exp\left[-\frac{s-\mu}{\sigma} - \exp\left(-\frac{s-\mu}{\sigma}\right)\right] ds = \exp\left[-\exp\left(-\frac{x-\mu}{\sigma}\right)\right]$$

s tzv. **log-log linkovací funkcí**

$$g_3(\pi) = -\log[-\log(\pi)] = \frac{x-\mu}{\sigma} = \beta_0 + \beta_1 x \quad \text{tj.} \quad \beta_0 = -\frac{\mu}{\sigma} \quad \beta_1 = \frac{1}{\sigma} > 0.$$

LogLog model



Obrázek : Zobecněné Gumbelovo rozdělení.

Nejčastěji se používá **logit** linkovací funkce

$$g_2(\pi) = \log\left(\frac{\pi}{1-\pi}\right).$$

Zajímá nás vztah pravděpodobností úspěchu či neúspěchu k hodnotám regresorů (kovariát) $\mathbf{x} = (x_1, \dots, x_k)^T$, tj.

$$P(Y = 1|x_1, \dots, x_k) = \pi(\mathbf{x}) = \frac{\exp\{\eta(\mathbf{x})\}}{1 + \exp\{\eta(\mathbf{x})\}} = \frac{1}{1 + \exp\{-\eta(\mathbf{x})\}}$$

a

$$P(Y = 0|x_1, \dots, x_k) = 1 - \pi(\mathbf{x}) = \frac{1}{1 + \exp\{\eta(\mathbf{x})\}} = \frac{\exp\{-\eta(\mathbf{x})\}}{1 + \exp\{-\eta(\mathbf{x})\}}$$

Předpokládejme, že lineární prediktor je roven

$$\eta(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}.$$

Logistická regrese

Všimněme se nejprve, že podíl

$$\frac{\text{odds}(1)}{\text{odds}(0)} = \frac{P(Y = 1|x_1, \dots, x_k)}{P(Y = 0|x_1, \dots, x_k)} = \frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})$$

má bezprostřední interpretaci. Porovnává pravděpodobnost jedničky (tj. výskyt sledovaného jevu při daných hodnotách kovariát) a nuly (nevýskyt sledovaného jevu při daných hodnotách kovariát). Anglickému označení **odds** odpovídá české označení **šance**.

Pro $k = 1$ jsou šance

$$\text{odds}(0) = \exp(\beta_0),$$

$$\text{odds}(1) = \exp(\beta_0 + \beta_1).$$

Poměr šancí (anglicky *odds ratio*) pro binární x je pak

$$OR = \frac{\text{odds}(1)}{\text{odds}(0)} = \exp(\beta_1),$$

takže parametr β_1 je roven logaritmu poměru šancí.

Příklad 1

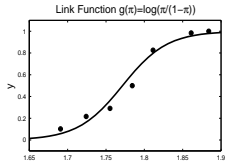
V souboru „*beetle.RData*“ jsou uvedeny údaje o úmrtnosti Potemníka skladištního (*Tribolium confusum*) v reakci na sirouhlík CS_2 . Datový soubor obsahuje tyto proměnné

<i>dose</i>	množství sirouhlíku (mg/l)
<i>population</i>	počet kusů ve zkoumaném vzorku
<i>killed</i>	počet mrtvých kusů ve zkoumaném vzorku

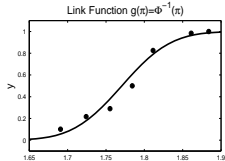
Modelujte závislost úmrtnosti na množství CS_2 .

Řešení. Pro modelování závislosti použijeme logistický model, probitový model a model s komplementární log-log linkovací funkcí.

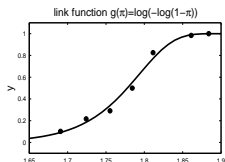
Příklad



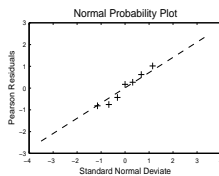
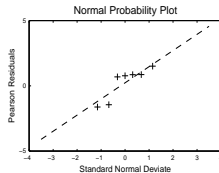
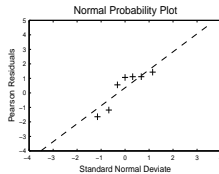
$$y = \exp(-61.05 + 34.461x)$$



$$y = \Phi(-35.127 + 19.838x)$$



$$y = 1 - \exp(-\exp(-40.647 + 22.656x))$$



Obrázek : Modely pro úmrtnost *Potemníka skladištního*.

Modely pro poissonovská data

Předpokládejme, že náhodný výběr rozsahu n je z Poissonova rozdělení, tj

$$Y_i \sim Po(\lambda_i), \quad P(Y_i = y) = \begin{cases} \frac{\lambda_i^y e^{-\lambda_i}}{y!} & \lambda_i > 0; \quad y = 0, 1, 2, \dots \\ 0 & \text{jinak} \end{cases}$$

přičemž

$$EY_i = DY_i = \lambda_i.$$

Y_i ... **počet výskytu sledovaného jevu v určitém časovém intervalu** (na ploše velikosti t apod.).

Jestliže jsou splněny následující podmínky

- jevu může nastat v kterémkoliv časovém okamžiku,
- počet výskytů jevu během časového intervalu závisí jen na jeho délce a ne na jeho počátku ani na tom, kolikrát jevu nastoupil před jeho počátkem,
- pravděpodobnost, že jevu nastoupí více než jednou v intervalu délky t , konverguje k nule rychleji než t ,
- λ je **střední hodnota počtu výskytů jevu za časovou jednotku**

pak uvedená náhodná veličina má rozdělení $Po(\lambda)$.

Náhodnou veličinou, která má Poissonovo rozdělení, je tedy např.

- počet vadných výrobků ve velké sérii, jestliže pravděpodobnost vyrobení vadného výrobku je velmi malá
- počet těžkých dopravních úrazů za den v určitém městě
- počet zákazníků v prodejně během nějakého časového intervalu
- počet částic v jednotce plochy nebo objemu, např. počet částic v zorném poli mikroskopu
- počet telefonních volání v časovém intervalu t
- počet létavic pozorovaných během intervalu délky t

Modely pro poissonovská data

Předpokládejme, že Y_i závisí na k veličinách x_{i1}, \dots, x_{ik} , a úkolem je najít vztah mezi nimi, tj. hledáme funkci

$$\lambda_i = \lambda(\mathbf{x}_i) = \lambda(x_{i1}, \dots, x_{ik}).$$

GLM modely \Rightarrow modelujeme pravděpodobnosti λ_i pomocí linkovacích funkcí

$$g(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Nejjednodušším je **lineární model**

$$\lambda_i = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Tento model má řadu nevýhod, především je třeba zajistit, aby $\mathbf{x}_i^T \boldsymbol{\beta}$ nabývala pouze kladných hodnot. Nejčastěji se volí tyto dvě možnosti:

log-lineárních model:

$$EY_i = \mu_i = \lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \quad \Rightarrow \quad g_1(\mu_i) = \eta_i = g_1(\lambda_i) = \log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

odmocninový model:

$$EY_i = \mu_i = \lambda_i = (\mathbf{x}_i^T \boldsymbol{\beta})^2 \quad \Rightarrow \quad g_2(\mu_i) = \eta_i = g_2(\lambda_i) = \sqrt{\lambda_i} = \mathbf{x}_i^T \boldsymbol{\beta}$$

Modelování binomických dat pomocí poissonovského modelu

Pomocí Poissonova rozdělení $Po(\lambda)$ lze dobře aproximovat binomické rozdělení $Bi(n, \pi)$ za podmínek

$$n \rightarrow \infty \quad \& \quad \pi \rightarrow 0 \quad \& \quad n\pi \rightarrow \lambda < \infty,$$

obvykle se doporučuje $n > 30$ a $\pi < 0,1$.

Chceme-li tedy aproximovat binomické rozdělení $Bi(n_i, \pi_i)$ pomocí Poissonova rozdělení $Po(\lambda_i = n_i\pi_i)$ a přitom použijeme logaritmickou linkovací funkci, platí

$$\lambda_i = n_i\pi_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \Rightarrow \log(\lambda_i) = \underbrace{\log(n_i)}_{\text{tzv. „offset“}} + \log(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Příklad 2

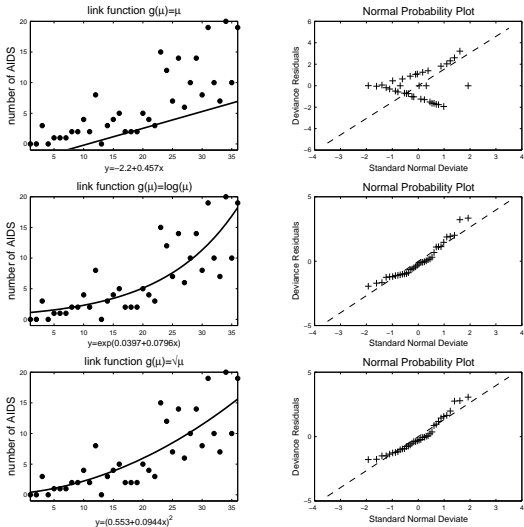
V souboru „*aids.RData*“ jsou uvedeny údaje o počtech nových případů AIDS ve Velké Británii za období prosinec 1982 až listopad 1985. Datový soubor obsahuje tyto proměnné

<i>month</i>	<i>měsíc</i>
<i>year</i>	<i>rok</i>
<i>number</i>	<i>počet nových případů AIDS</i>

Modelujte závislost počtu nových případů AIDS na čase.

Řešení. Pro modelování závislosti použijeme lineární model, log-lineární model a odmocninový model.

Příklad



Obrázek : Modely pro výskyt nových onemocnění AIDS ve Velké Británii.

Overdispersion, underdispersion

Předpokládáme, že náhodný výběr $\mathbf{Y}_n = (Y_1, \dots, Y_n)^T$ 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 deviance

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

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

Overdispersion, underdispersion

Protože platí

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

pak

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

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

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

a proto dalším momentovým odhadem založeným na této statistice je

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

Overdispersion, underdispersion

Přehled rušivých parametrů

<i>Rozdělení</i>	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/nížší variabilitu v datech a zabraňuje tak nadbytku/nedostatku falešně pozitivních výsledků testů hypotéz o parametrech modelu.

Příklad 3

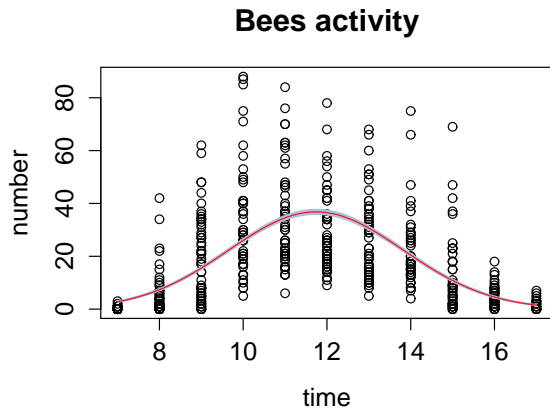
V souboru „*bees.RData*“ jsou uvedeny údaje o aktivitě včel v závislosti na čase. Jednou z důležitých charakteristik při zkoumání včelí aktivity je počet včel, které opustí úl kvůli práci ve vnějším prostředí. Studie se zabývala měřením této veličiny během několika slunečných dní v závislosti na čase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

<i>number</i>	počet včel, které opustily úl
<i>time</i>	čas, kdy byl tento údaj zaznamenán

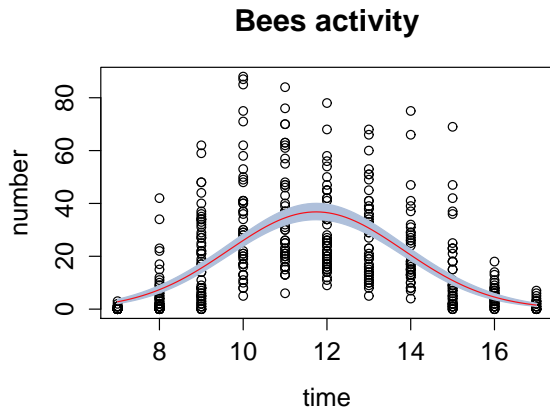
Modelujte závislost počtu včel, které opustí úl, na čase během dne.

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 deviance (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.



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



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

Modely pro multinomická data

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

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

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

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

s tzv. **celkovou** dodatečnou podmínkou

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

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

Modely pro multinomická data

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

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

tj.

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

přičemž

$$EY_{jk} = N\pi_{jk}$$

$$DY_{jk} = N\pi_{jk}(1 - \pi_{jk})$$

$$C(Y_{jk}, Y_{j'k'}) = -N\pi_{jk}\pi_{j'k'}$$

Kontingenční tabulky

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

Kontingenční tabulka četností

faktor A	faktor B				Σ
	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}	$N_{J.}$
Σ	$N_{.1}$	$N_{.2}$	\cdots	$N_{.K}$	$N = N_{..}$

Kontingenční tabulky

Kontingenční tabulka pravděpodobností

faktor A	faktor B				Σ
	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}	$\pi_{J.}$
Σ	$\pi_{.1}$	$\pi_{.2}$	\cdots	$\pi_{.K}$	$\pi_{..} = 1$

Kontingenční tabulky

Nejčastěji se v kontingenčních tabulkách testuje hypotéza, že

faktory A a B jsou nezávislé

tj.

faktor A	faktor B			Σ
	...	B_k	...	
⋮	⋮	⋮	⋮	⋮
A_j	...	$\pi_j \cdot \pi_k$...	π_j
⋮	⋮	⋮	⋮	⋮
Σ	...	π_k	...	1

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

Log-lineární modely

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

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

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

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

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.

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

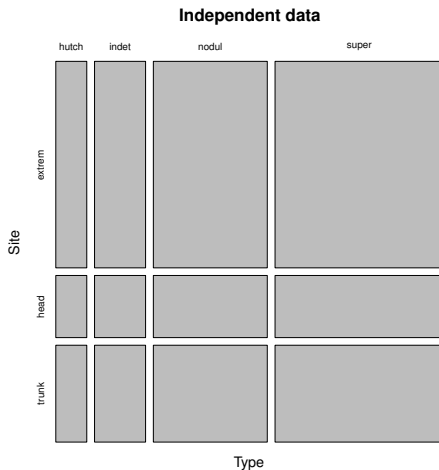
Příklad 4

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.

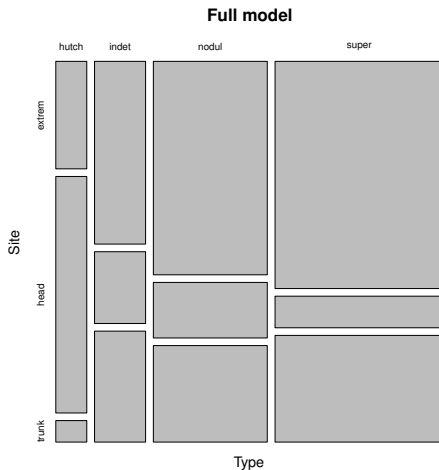
Typ rakoviny	Část těla		
	končetiny	hlava a krk	trup
<i>Hutchinson's melanotic freckle</i>	10	22	2
<i>neurčitý</i>	28	11	17
<i>Nodular</i>	73	19	33
<i>Superficial spreading melanoma</i>	115	16	54

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

Řešení Nejprve definujeme oba log-lineární modely, tj. model m_1 , který předpokládá nezávislost obou faktorů a model m_2 , který počítá i s interakcemi. Model m_1 je tedy submodelem modelu m_2 . K testování využijeme analýzu deviance, 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.



Obrázek : Mozaikový graf pro model, který předpokládá nezávislost.



Obrázek : Mozaikový graf pro model s interakcemi.

Příklad 1.1

V souboru „heart.RData“ jsou uvedena data o přítomnosti infarktu myokardu v závislosti na věku pacienta. Datový soubor obsahuje tyto proměnné:

age věk pacienta (roky)

chd indikátor infarktu (1 – nastal, 0 – nenastal)

Pro modelování závislosti použijte logistický model, probitový model a model s komplementární log-log linkovací funkcí. Výsledky vykreslete do obrázku.

Příklad 1.2

V souboru „*nemocnice.RData*“ jsou uvedeny údaje o zotavení pacientů v závislosti na závažnosti onemocnění a nemocnici, ve které se léčili. Datový soubor obsahuje tyto proměnné:

<i>Infection_Severity</i>	vážnost onemocnění
<i>Treatment_Outcome</i>	indikátor uzdravení (1 – zdravý, 0 – smrt)
<i>Hospital</i>	typ nemocnice (1, 2, 3)

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku.

Příklad 1.3

V souboru „*cancer.RData*“ jsou uvedeny údaje o počtu onemocnění rakovinou kůže u žen v závislosti na věku a oblasti v USA, ve které pacientky žily. Datový soubor obsahuje tyto proměnné:

<i>Cases</i>	počet onemocnění
<i>Town</i>	město (0 – Minneapolis (Minnesota), 1 – Dallas (Texas))
<i>Age</i>	věková skupina pacientky
<i>Population</i>	celkový počet žen dané věkové skupiny v příslušném městě

Pro modelování závislosti nalezněte vhodný logistický model. Výsledky vykreslete do obrázku. Porovnejte pravděpodobnost vzniku onemocnění u 60-ti leté pacientky žijící v Minneapolisu s pravděpodobností pro stejně starou pacientku žijící v Dallasu.

[Minneapolis: 0.00117, Dallas: 0.00276.]

Příklad 1.4

V souboru „*car_income.RData*“ jsou uvedeny údaje o koupi nového auta během posledních 12-ti měsíců v závislosti na příjmu domácnosti a stáří původního auta. Datový soubor obsahuje tyto proměnné:

<i>purchase</i>	indikátor nákupu nového auta (1 – ano, 0 – ne)
<i>income</i>	roční příjem domácnosti (v tis. dolarů)
<i>age</i>	stáří původního auta (roky)

Nejprve vykreslete závislosti proměnné *purchase* na ostatních. Pro modelování závislosti nalezněte vhodný logistický model. Jsou všechny proměnné statisticky významné? Znovu modelujte s použitím proměnné *age* jako *factor*. Opět sledujte statistickou významnost *age*. Vyzkoušejte tuto proměnnou zakomponovat do modelu jako *factor* s méně úrovněmi. Výsledky vykreslete do obrázku.

Úlohy k procvičení I

Příklad 1.5

V souboru „*druhy.RData*“ jsou k dispozici data, která se týkají dlouhodobého zemědělského experimentu. Bylo sledováno 90 pozemků (pastvin) o rozloze 25 m × 25 m, lišících se v biomase, pH půdy a druhové bohatosti (počet rostlinných druhů na celém pozemku). Je dobře známo, že s rostoucí biomasou dochází k poklesu druhové bohatosti. Ale zůstává otázka, zda rychlost poklesu nesouvisí s úrovní pH v půdě. Proto byly jednotlivé pozemky klasifikovány podle hodnoty pH v půdě do tří úrovní (nízká, střední a vysoká úroveň) a do experimentu bylo vybráno vždy po 30 pozemcích pro každou úroveň. Spojitá veličina *Biomass* je dlouhodobým průměrem naměřených červnových hodnot biomasy. Datový soubor obsahuje tyto proměnné:

<i>pH</i>	úroveň pH v půdě (<i>low</i> – nízká, <i>mid</i> – střední, <i>high</i> – vysoká)
<i>Biomass</i>	množství biomasy
<i>species</i>	počet rostlinných druhů

Příklad 1.5

Nejprve vykreslete závislosti proměnné $species$ na ostatních. Pro modelování závislosti nalezněte vhodný poissonovský model. Vyzkoušejte postupně logaritmickou, identickou a odmocninovou linkovací funkci. Jsou všechny proměnné statisticky významné? Pokud ne, zkuste modely zjednodušit a pomocí analýzy deviace rozhodněte, zda takové zjednodušení je možné. Získané výsledné modely vykreslete do obrázku. Pomocí všech modelů odhadněte počet rostlinných druhů na pozemku s hodnotou biomasy 9 a střední úrovní pH v půdě.

[Odhady počtu druhů pro log link: 8,895, identity link: 4,513, sqrt link: 7,414.]

Úlohy k procvičení I

Příklad 1.6

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:

<i>Year</i>	rok
<i>Population</i>	velikost populace
<i>Attacks</i>	počet napadení žraloky
<i>Fatalities</i>	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*.

Příklad 1.6

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, predefinujte model a výsledky znovu vykreslete do obrázku. Pomocí výsledného modelu odhadněte, kolik útoků (na 1 milion 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 milion obyvatel, interval spolehlivosti: [3,207;359,55].]

Úlohy k procvičení

Příklad 1.7

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

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

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

[závisí]