M5VM05 Statistické modelování

9. Zobecněné lineární modely

Jan Koláček (kolacek@math.muni.cz)

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

podzim 2013

Motivace

V reálném světě má mnoho procesů jiný, než lineární vztah závislosti. Např. v ekonomii se ukazuje, že mnoho vztahů má logaritmickou závislost, k vysvětlení procesů v přírodních vědách se užívají reciproké, mocninné i další vztahy. Vysvětlovaná veličina popisující pravděpodobnost přežití člověka, v případě určité nemoci a určitého způsobu léčby, může z definice pravděpodobnosti nabývat hodnot pouze z intervalu $[0, 1]$, což by v případě klasického lineárního modelu bylo možné zajistit jen za přijetí určitých omezení na parametry modelu. Také normalita chyb je často nesplněným předpokladem klasického lineárního regresního modelu. Připomeňme, že normalita se vyznačuje nezávislosti střední hodnoty a rozptylu. Typicky např. u ekonomických veličin s rostoucí střední hodnotou obvykle roste rozptyl náhodné veličiny, přičemž náhodné chyby mají v těchto případech často nesymetrická, kladně sešikmená rozdělení.

Základní pojmy a definice I

Definice 1

Mějme parametrický prostor $\Theta \subset \mathbb{R}^m$. Řekneme, že systém m-parametrických hustot

$$\mathcal{F}_{\text{reg}}^m = \{f(\mathbf{y};\boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}} \in \Theta\}$$

je **regulární**, jestliže platí

(1) $\Theta \subset \mathbb{R}^m$ je otevřená borelovská množina.

(2) Množina $M = \{\mathbf{y} \in \mathbb{R}^n : f(\mathbf{y};\boldsymbol{\theta}) > 0\}$ nezávisí na parametru $\boldsymbol{\theta}$.

(3) Pro každé $\mathbf{y} \in M$ existuje konečná parciální derivace

$$f_i'(\mathbf{y};\boldsymbol{\theta}) = \frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} \quad (i = 1, \ldots, m).$$

Základní pojmy a definice II

Definice 1

(4) Pro všechny $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}} \in \boldsymbol{\Theta}$ platí

$$\int_M \frac{f_i'(\mathbf{y};\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta})} dF(\mathbf{y};\boldsymbol{\theta}) = \int_M \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} dF(\mathbf{y};\boldsymbol{\theta}) = 0 \qquad i = 1, \ldots, m,$$

kde $F(\mathbf{y};\boldsymbol{\theta})$ je odpovídající distribuční funkce.

(5) Pro všechny $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}} \in \boldsymbol{\Theta}$ je integrál

$$J_{ij}(\boldsymbol{\theta}) = \int_M \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_j} dF(\mathbf{y};\boldsymbol{\theta}) \qquad i, j = 1, \ldots, m$$

konečný a matice $\mathbf{J} = \mathbf{J}(\boldsymbol{\theta}) = \left(J_{ij}(\boldsymbol{\theta})\right)_{i,j=1}^{m}$ je pozitivně definitní. Matice $\mathbf{J}$ se nazývá **Fisherova informační matice** o parametru $\boldsymbol{\theta}$.

Základní pojmy a definice

Definice 2

Nechť $f \in \mathcal{F}^m_{\text{reg}}$. Pak náhodný vektor

$$\mathbf{U} = \mathbf{U}(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), \ldots, U_m(\boldsymbol{\theta}))^{\mathrm{T}} \qquad \text{se složkami}$$
$$U_i = U_i(\boldsymbol{\theta}) = \frac{\partial \ln f(\mathbf{Y}; \boldsymbol{\theta})}{\partial \theta_i}$$

se nazývá **skórový vektor** příslušný hustotě f.

Základní pojmy a definice I

Věta 3

(1) *Je-li* $f \in \mathcal{F}^m_{reg}$ *a pro* $i, j = 1, \ldots, m$ *existují*

$$f''_{ij}(\mathbf{y}; \boldsymbol{\theta}) = \frac{\partial^2 f(\mathbf{y}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j},$$

pak

$$E\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0} \qquad a \qquad D\mathbf{U}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta}) \ .$$

Základní pojmy a definice II

Věta 3

(2) *Platí-li navíc pro* $i, j = 1, \ldots, m$

$$E\left(\frac{f''_{ij}(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y};\boldsymbol{\theta})}\right) = 0,$$

pak

$$\mathbf{J}(\boldsymbol{\theta}) = -E(\mathbf{U}'(\boldsymbol{\theta})),$$

kde

$$\mathbf{U}'(\boldsymbol{\theta}) = \left(\frac{\partial U_i(\boldsymbol{\theta})}{\partial \theta_j}\right)_{i,j=1}^{m}.$$

Základní pojmy a definice

Uvažujme **náhodný výběr** $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ z rozdělení $f \in \mathcal{F}^m_{\mathrm{reg}}$. Označme $M = \{y \in \mathbb{R} : f(y;\boldsymbol{\theta}) > 0\}$. Pak sdružená hustota

$$f_{\mathbf{Y}_n}(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i,\boldsymbol{\theta}), \qquad \mathbf{y} = (y_1,\ldots,y_n)' \in \mathbb{R}^n.$$

Značení:

funkce:

$$\begin{array}{rcccl} l_k & = & l(\boldsymbol{\theta};y_k) & = & \ln f(y_k;\boldsymbol{\theta}) \\ l^*_n & = & l^*_n(\boldsymbol{\theta};\mathbf{y}) & = & \ln f_{\mathbf{Y}_n}(\mathbf{y};\boldsymbol{\theta}) \end{array}$$

n. vektory:

$$\begin{array}{rcccl} \mathbf{U}_k & = & \mathbf{U}_k(\boldsymbol{\theta}) & = & \left(\frac{\partial \ln f(Y_k;\boldsymbol{\theta})}{\partial\theta_1},\ldots,\frac{\partial \ln f(Y_k;\boldsymbol{\theta})}{\partial\theta_m}\right)^{\mathrm{T}} \\ \mathbf{U}^*_n & = & \mathbf{U}^*_n(\boldsymbol{\theta}) & = & \left(\frac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{Y};\boldsymbol{\theta})}{\partial\theta_1},\ldots,\frac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{Y};\boldsymbol{\theta})}{\partial\theta_m}\right)^{\mathrm{T}} \end{array}$$

matic. fce:

$$\begin{array}{rcccl} \mathbf{J} & = & \mathbf{J}(\boldsymbol{\theta}) & = & \left(J_{ij}(\boldsymbol{\theta})\right)_{i,j=1}^{m} = \left(\int_M \frac{\partial \ln f(y;\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \ln f(y;\boldsymbol{\theta})}{\partial\theta_j} dF(y;\boldsymbol{\theta})\right)_{i,j=1}^{m} \\ \mathbf{J}_n & = & \mathbf{J}_n(\boldsymbol{\theta}) & = & \left(\int_M \cdots \int_M \frac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \ln f_{\mathbf{Y}_n}(\mathbf{y};\boldsymbol{\theta})}{\partial\theta_j} dF_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})\right)_{i,j=1}^{m} \end{array}$$

Základní pojmy a definice I

Věta 4

Uvažujme náhodný výběr $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ *z rozdělení s hustotou* $f \in \mathcal{F}^m_{reg}$.

(1) *Pokud pro* $i, j = 1, \ldots, m$ *existují*

$$f''_{ij}(y; \boldsymbol{\theta}) = \frac{\partial^2 f(y; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j},$$

pak

$$E\mathbf{U}^*_n(\boldsymbol{\theta}) = \mathbf{0} \qquad a \qquad D\mathbf{U}^*_n(\boldsymbol{\theta}) = n\mathbf{J}(\boldsymbol{\theta}) \ .$$

Základní pojmy a definice II

Věta 4

(2) *Platí-li navíc pro* $i,j = 1,\ldots,m$

$$E\left(\frac{f''_{ij}(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta})}\right) = 0,$$

(tj. f *je regulární i v 2. derivacích), pak*

$$E(\mathbf{U}_n^{*'}(\boldsymbol{\theta})) = -n\mathbf{J}(\boldsymbol{\theta}),$$

kde

$$\mathbf{U}_n^{*'}(\boldsymbol{\theta}) = \left(\frac{\partial U_i^*(\boldsymbol{\theta})}{\partial \theta_j}\right)_{i,j=1}^{m}.$$

Základní pojmy a definice I

Věta 5

Mějme náhodný výběr $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ *z rozdělení s regulární hustotou* $f \in \mathcal{F}^m_{reg}$. *Označme* $M = \{y \in \mathbb{R} : f(y;\boldsymbol{\theta}) > 0\}$. *Nechť pro všechna* $y \in M$, $\boldsymbol{\theta} \in \Theta$ *a* $i, j = 1, \ldots, m$ *existují druhé parciální derivace hustoty* $f(y;\boldsymbol{\theta})$.

(1) *Pak platí*

$$\frac{1}{\sqrt{n}}\mathbf{U}^*_n(\boldsymbol{\theta}) \quad \overset{A}{\sim} \quad N_m(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta})).$$

Dále platí

$$\frac{1}{n}\mathbf{U}^*_n(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{J}(\boldsymbol{\theta})^{-1}\mathbf{U}^*_n(\boldsymbol{\theta}) \quad \overset{A}{\sim} \quad \chi^2(m).$$

Základní pojmy a definice II

Věta 5

(2) *Platí-li navíc, že f je regulární i v 2.derivacích, tj.*

$$E\left(\frac{f''_{ij}(\mathcal{Y};\boldsymbol{\theta})}{f(\mathcal{Y};\boldsymbol{\theta})}\right) = 0,$$

pak matice náhodných veličin

$$\frac{1}{n}\mathbf{U}_n^{*\prime}(\boldsymbol{\theta}) = \frac{1}{n}\left(\frac{\partial U_i^*(\boldsymbol{\theta})}{\partial\theta_j}\right)_{i,j=1}^{m} = \frac{1}{n}\left(\frac{\partial^2 l_n(\boldsymbol{\theta};\mathbf{Y})}{\partial\theta_i\partial\theta_j}\right)_{i,j=1}^{m} \xrightarrow{s.j.} -\mathbf{J}(\boldsymbol{\theta}).$$

Základní pojmy a definice

Definice 6

(a) **Věrohodnostní funkcí** rozumíme funkci vektorového parametru $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})$$

(b) **logaritmickou věrohodnostní funkcí** nazýváme funkci

$$l(\boldsymbol{\theta}; \mathbf{y}) = \ln L(\boldsymbol{\theta}; \mathbf{y})$$

(c) Řekneme, že odhad $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}(\mathbf{Y})$ je **maximálně věrohodný odhad (MLE)** vektorového parametru $\boldsymbol{\theta}$, pokud platí

$$L(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}; \mathbf{Y}) \geq L(\boldsymbol{\theta}; \mathbf{Y})$$

pro všechna $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Základní pojmy a definice

Věta 7

Mějme náhodný výběr $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ *z rozdělení s regulární hustotou* $f \in \mathcal{F}^m_{reg}$. *Označme* $M = \{y \in \mathbb{R} : f(y;\boldsymbol{\theta}) > 0\}$. *Nechť pro všechna* $y \in M$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *a* $i, j = 1, \ldots, m$ *existují druhé parciální derivace hustoty* $f(y;\boldsymbol{\theta})$ *a platí* $E\left(\frac{f''_{ij}(Y;\boldsymbol{\theta})}{f(Y;\boldsymbol{\theta})}\right) = 0$. *Pak*

(1) $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta}) \overset{A}{\sim} N_m(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta})^{-1})$

(2) $W = (\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta})^{\mathrm{T}} n\mathbf{J}(\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta}) \overset{A}{\sim} \chi^2(m)$, *tzv.* **Waldova statistika**.

Základní pojmy a definice

Definice 8

Řekneme, že pozorování pochází z rozdělení **exponenciálního typu**, pokud jeho *pravděpodobnostní funkce* (v případě diskrétních rozdělení) či *hustota* (v případě spojitých rozdělení) je tvaru

$$f(y) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

kde

θ je (neznámý) tzv. **přirozený parametr**

a

$a(y), b(\theta), c(\theta), d(y)$ jsou známé funkce.

Pokud

- $a(y) = y$, říkáme že pravděpodobnostní funkce, popř. hustota je v **kanonické formě**.
- v konkrétním rozdělení figurují další neznámé parametry, nazveme je tzv. **rušivými parametry**.

Základní pojmy a definice

V dalším budeme uvažovat pouze **regulární** a **kanonické** formy spolu s podmínkou $b(\theta) = \theta$ a přitom zavedeme do označení jeden rušivý parametr ϕ :

$$\boxed{f(y) = \exp\left\{\frac{y\theta - \gamma(\theta)}{\psi(\phi)} + d(y, \phi)\right\}},$$

kde θ a ϕ jsou parametry
$\gamma(\theta)$, $\psi(\phi) > 0$, $d(y)$ jsou známé funkce,
a pokud

$\psi(\phi) = \frac{\phi}{\omega} > 0$, $\phi > 0$ je tzv. **faktor měřítka** (scale factor)
$\omega > 0$ je známá **apriorní váha**.

Tato forma se také nazývá **škálovou formou hustoty exponenciálního typu**.

Základní pojmy a definice

Věta 9

Mějme náhodnou veličinu Y z rozdělení s regulární hustotou f exponenciálního typu:

$$f(y) = \exp\left\{\frac{y\theta - \gamma(\theta)}{\psi(\phi)} + d(y,\phi)\right\}. \tag{1}$$

Pak

$$EY = \gamma'(\theta)$$

Nechť navíc platí

$$E\left(\frac{f''(Y;\theta)}{f(Y;\theta)}\right) = 0, \tag{2}$$

kde $f''(Y;\theta) = \frac{d^2 f(Y;\theta)}{d\theta^2}$, *pak*

$$DY = \gamma''(\theta)\psi(\phi)$$

Funkce $\gamma''(\theta) = \frac{DY}{\psi(\phi)}$ *se nazývá* **rozptylovou funkcí** *(variance function).*

Základní pojmy a definice

Příklad 10 (Normální rozdělení)

Mějme

$$Y \sim N(\mu, \sigma^2), \qquad \mu \in \mathbb{R},\ \sigma^2 > 0.$$

Pak

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \right\}$$

$$= \exp\left\{ \underbrace{\frac{y\mu - \overbrace{\frac{1}{2}\mu^2}^{\gamma(\theta)}}{\underbrace{\sigma^2}_{\psi(\phi)}}}_{} \underbrace{-\frac{1}{2}\frac{y^2}{\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)}_{d(y,\phi)} \right\}$$

a

$$\begin{aligned} \gamma(\theta) = \tfrac{1}{2}\mu^2 = \tfrac{1}{2}\theta^2 &\Rightarrow \gamma'(\theta) = \theta = \mu \\ &\Rightarrow \gamma''(\theta) = 1 \\ \psi(\phi) = \sigma^2 &\Rightarrow \phi = \sigma^2. \end{aligned}$$

Základní pojmy a definice

Skutečně platí

$$EY = \gamma'(\theta) = \mu$$

a

$$DY = \gamma''(\theta)\psi(\phi) = \sigma^2.$$

Tedy

přirozený parametr	$\theta = \mu$	**scale factor**	$\phi = \sigma^2$
rozptylová funkce	$V(\mu) = 1$	**váhy**	$\omega = 1.$

Základní pojmy a definice

Příklad 11 (Binomické rozdělení)

Mějme

$$Z \sim Bi(n, \pi), \qquad n \in \mathbb{N}, \pi \in (0,1).$$

pak $f_Z(z) = \binom{n}{z}\pi^z(1-\pi)^{n-z} = \exp\left\{z \ln\left(\frac{\pi}{1-\pi}\right) + n\ln(1-\pi) + \ln\binom{n}{z}\right\}$
pro $z = 0, \ldots, n,$
přičemž $EZ = \mu = n\pi$ *a* $DZ = n\pi(1-\pi).$

Pravděpodobnostní funkce **není** *ve škálové formě, proveďme reparametrizaci*

$$\theta = \ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{n\pi}{n-n\pi}\right) = \ln\left(\frac{\mu}{n-\mu}\right)$$

$$\Rightarrow \quad \pi = \frac{e^\theta}{1+e^\theta} \text{ a } 1-\pi = \frac{1}{1+e^\theta}.$$

Základní pojmy a definice

Tedy

$$f_Z(z) = \exp\left\{ z\theta - \underbrace{n \ln\left(1 + e^{\theta}\right)}_{\gamma(\theta)} + \underbrace{\ln \binom{n}{z}}_{d(y,\phi)} \right\}$$

a

$$\begin{aligned}
\gamma(\theta) = n \ln\left(1 + e^{\theta}\right) &\Rightarrow \gamma'(\theta) = n \tfrac{e^{\theta}}{1+e^{\theta}} = n\pi = \mu \\
&\Rightarrow \gamma''(\theta) = n \tfrac{e^{\theta}}{(1+e^{\theta})^2} = n\pi(1-\pi) = \mu\left(1 - \tfrac{\mu}{n}\right) \\
\psi(\phi) = \tfrac{\phi}{\omega} = 1 &\Rightarrow \omega = 1 \quad \phi = 1
\end{aligned}$$

Základní pojmy a definice

Skutečně platí

$$EZ = \gamma'(\theta) = \mu$$

a

$$DZ = \gamma''(\theta)\psi(\phi) = n\pi(1-\pi).$$

Tedy

přirozený parametr	$\theta = \ln\left(\frac{\mu}{n-\mu}\right)$
rozptylová funkce	$V(\mu) = \mu\left(1 - \frac{\mu}{n}\right)$
scale factor	$\phi = 1$
váhy	$\omega = 1.$

Základní pojmy a definice

Další rozdělení

Poissonovo rozdělení	**přirozený parametr**	$\theta = \ln \lambda$
	rozptylová funkce	$V(\mu) = \mu$
	scale factor	$\phi = 1$
	váhy	$\omega = 1$
Gamma rozdělení	**přirozený parametr**	$\theta = -\frac{1}{\mu}$
	rozptylová funkce	$V(\mu) = \mu^2$
	scale factor	$\phi = \frac{1}{\alpha}$
	váhy	$\omega = 1$
Exponenciální rozdělení	**přirozený parametr**	$\theta = -\frac{1}{\mu}$
	rozptylová funkce	$V(\mu) = \mu^2$
	scale factor	$\phi = 1$
	váhy	$\omega = 1.$

Omezení LM

Omezení lineárního modelu:

1. Je **omezen pouze na třídu normálních rozdělení**:
   $Y_i \sim N(\mu_i, \sigma^2) \quad i = 1, \ldots, n$, kde $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ tvoří náhodný výběr.
2. Předpokládá **striktní rovnost mezi střední hodnotou náhodné veličiny** Y_i **a lineární kombinací prediktorů**: $EY_i = \mu_i = \mathbf{x}_i'\boldsymbol{\beta}$, kde

   $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})'$ je vektor prediktorů a
   $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ je vektor neznámých parametrů.

Zobecnění lineárního modelu:

1. Zobecnění na nenormální rozdělení, a to na tzv. **třídu exponenciálních rozdělení**
2. Zobecnění na nelineární funkce, které **spojují** neznámé střední hodnoty výchozího rozdělení náhodné veličiny Y_i s prediktivními proměnnými.

Definice GLM I

Definice 12 (Zobecněný lineární model)

Mějme náhodný výběr $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ a nechť rozdělení Y_i závisí na pevných vektorech $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}} \in \mathbb{R}^k$ prostřednictvím neznámého vektoru parametrů $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$. Matice $\mathbf{X} = \left(\mathbf{x}_1^{\mathrm{T}}, \ldots, \mathbf{x}_n^{\mathrm{T}}\right)^{\mathrm{T}}$ má rozměr $n \times k$ a hodnost $k < n$.

Říkáme, že $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ se řídí **zobecněným lineárním modelem** (Generalized Linear Model), jestliže dále platí:

(1) rozdělení $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ je exponenciálního typu s **regulární** hustotou

$$f(\mathbf{y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i, \theta_i) = \exp\left\{ \sum_{i=1}^{n} \left[\frac{y_i \theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right] \right\} \qquad (3)$$

Definice GLM II

Definice 12 (Zobecněný lineární model)

(2) parametr θ_i závisí na $\mathbf{x}_i$ a $\boldsymbol{\beta}$ prostřednictvím parametru

$$\eta_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} \,, \tag{4}$$

který nazveme **lineární prediktor**.

(3) Existuje známá ryze monotónní diferencovatelná funkce g, tzv. **linkovací funkce** (link function), a platí

$$\eta_i = g(\mu_i) \qquad \mu_i = g^{-1}(\eta_i) \,, \quad \text{kde} \quad \mu_i = \mu(\theta_i) = EY_i. \tag{5}$$

Řekneme, že linkovací funkce je **kanonická**, pokud $\theta_i = \eta_i = g(\mu_i)$.

Příklad

Příklad 13

Regresní přímka v klasickém lineárním regresním modelu:

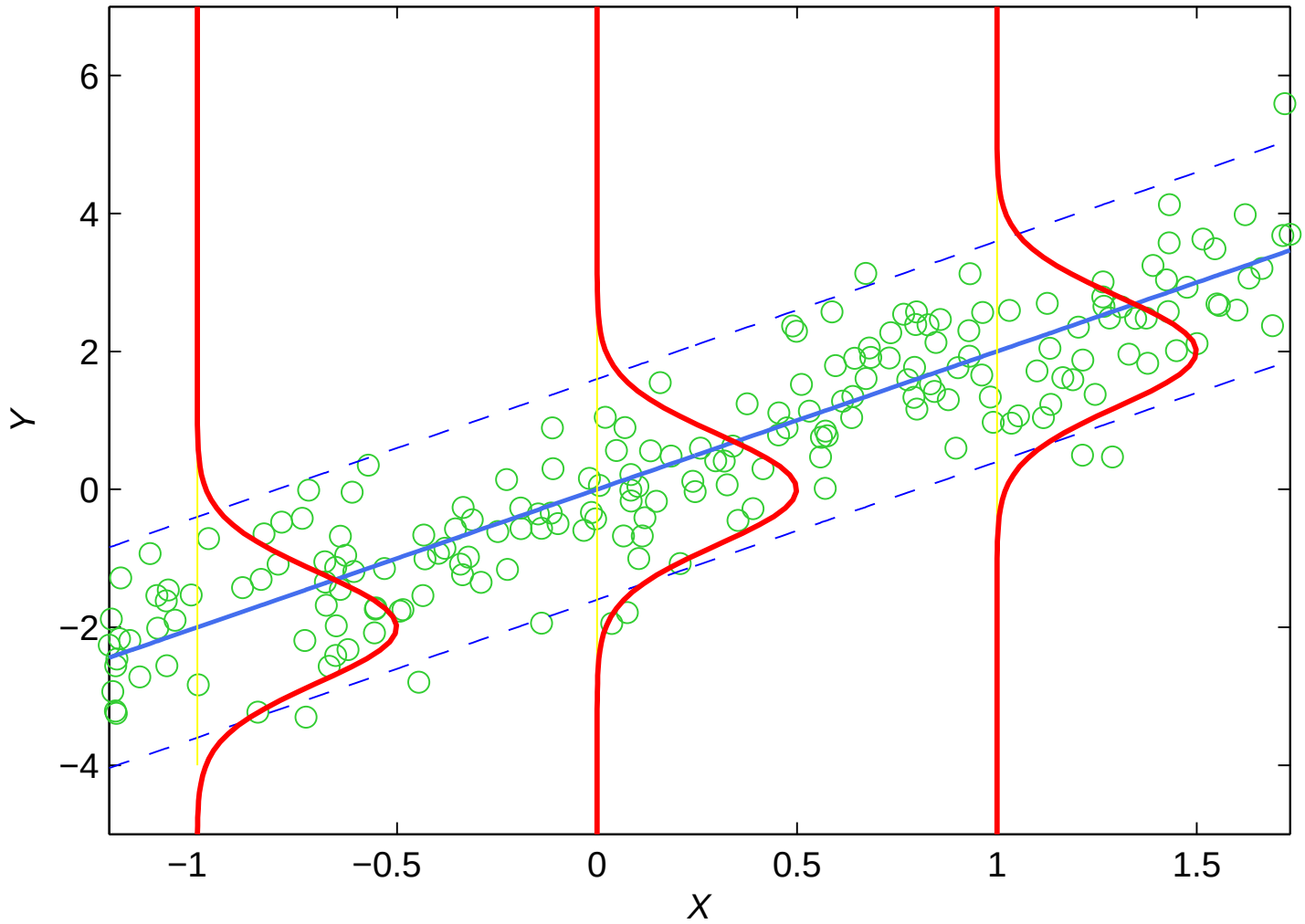
$$Y_i \sim N(\mu_i, \sigma^2)$$

jsou pro $i = 1, \ldots, n$ *nezávislé náhodné veličiny,*

$$g(\mu_i) = \mu_i = \beta_1 + \beta_2 x_i$$

je identická linkovací funkce, β_1, β_2 *a* σ^2 *jsou neznámé parametry (přičemž* σ^2 *je rušivým parametrem) a* x_i *jsou známé kovariáty.*

Příklad



Obrázek : Ukázka klasického regresního modelu s homogenním rozptylem.

Příklad

Příklad 14

Regresní modely s logaritmickou linkovací funkcí pro exponenciálně a gamma rozdělené závisle proměnné:

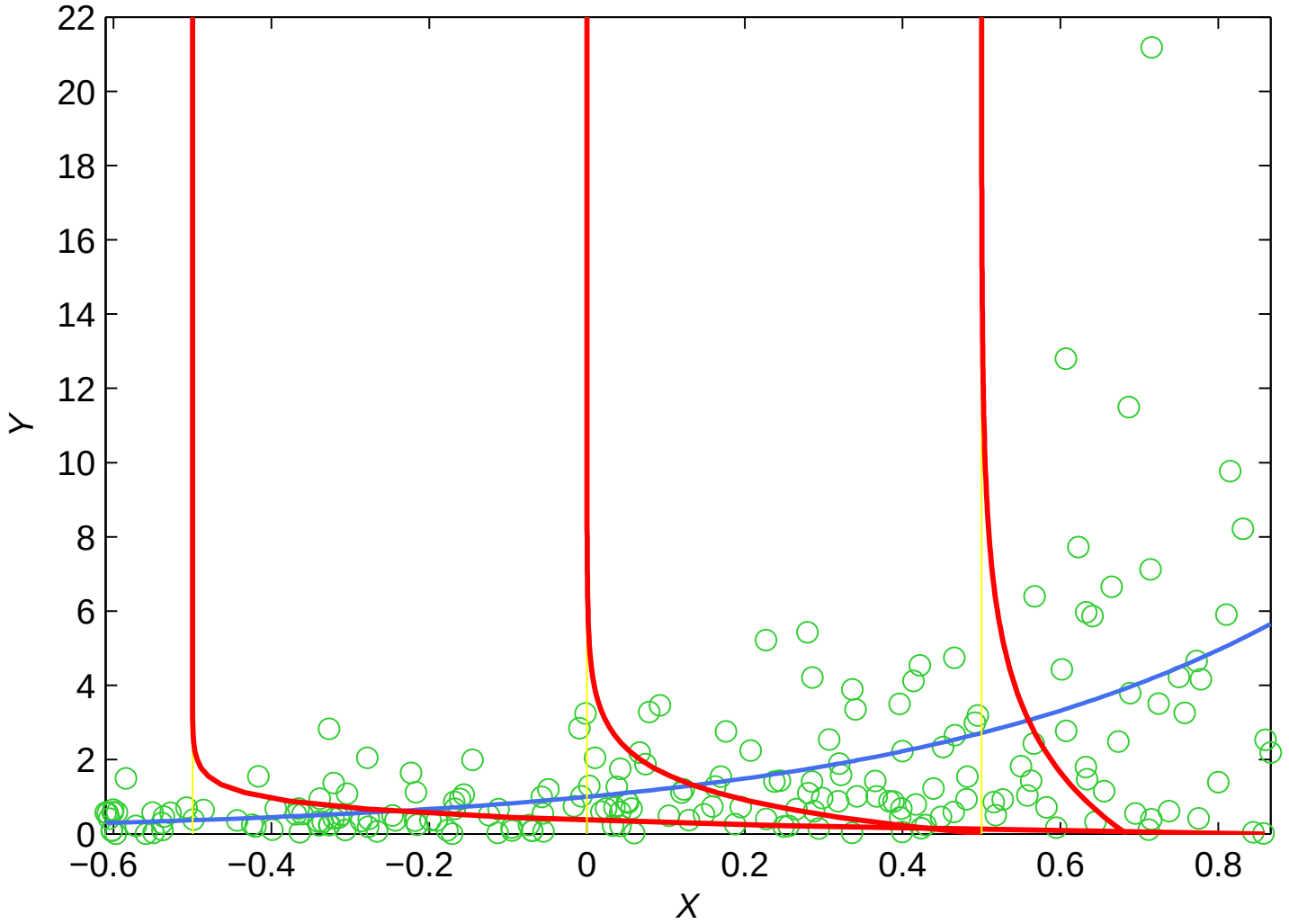
$$Y_i \sim Ex(\lambda_i) \equiv G(1, \lambda_i)$$

jsou pro $i = 1, \ldots, n$ *nezávislé náhodné veličiny* ($EY_i = \mu_i = \lambda_i$),

$$g(\mu_i) = \ln \mu_i = \beta_1 + \beta_2 x_i$$

je logaritmická linkovací funkce, β_1, β_2 *jsou neznámé parametry a* x_i *jsou známé kovariáty.*

Příklad

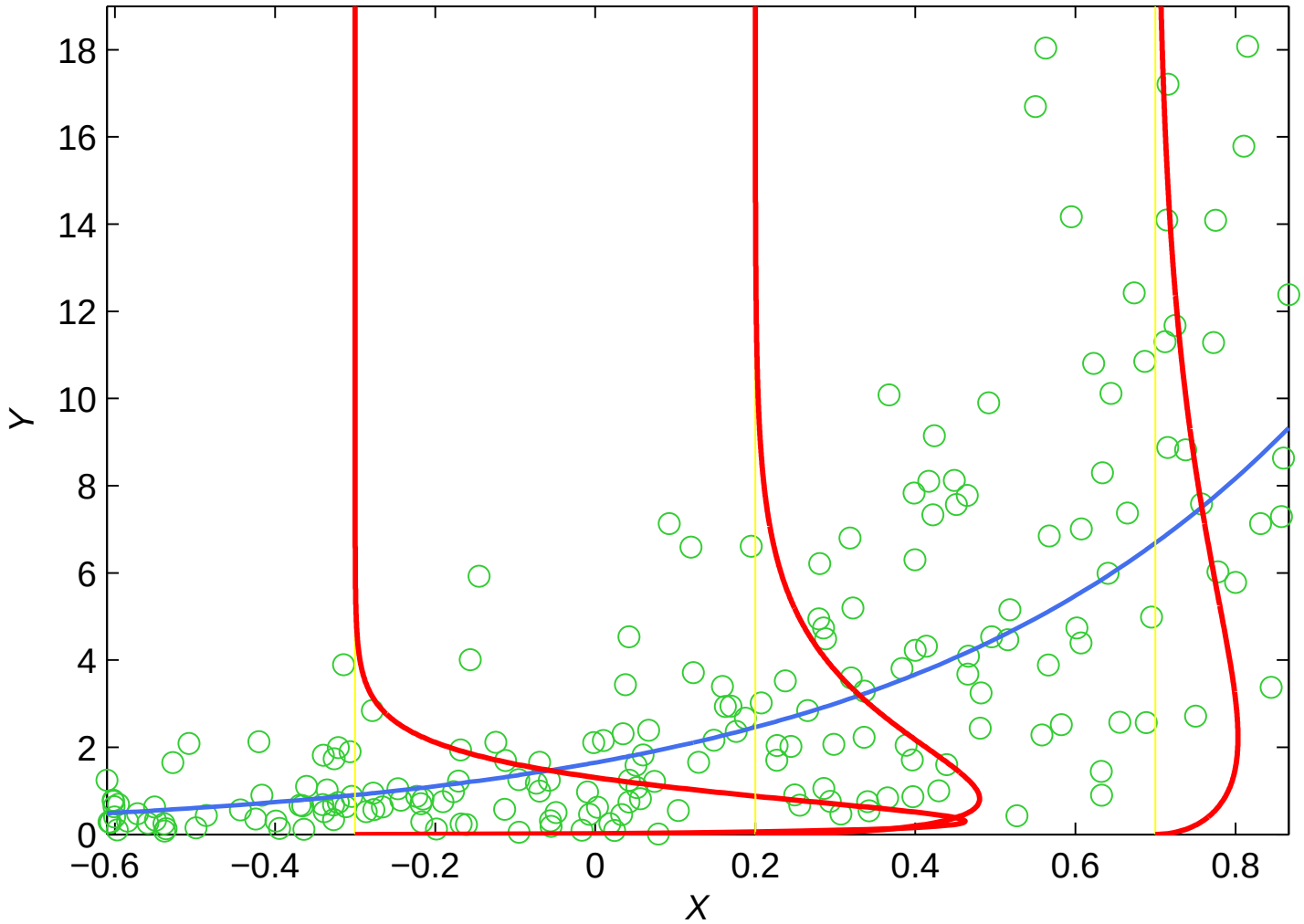


Obrázek : Ukázka GLM modelu s linkovací funkcí $g(\mu) = \ln \mu$ pro exponenciálně rozdělenou náhodnou veličinu Y.

Příklad

Jestliže $Y_i \sim G(\alpha, \beta_i = \frac{\mu_i}{\alpha})$ jsou pro $i = 1, \ldots, n$ nezávislé náhodné veličiny $(EY_i = \mu_i = \alpha\beta_i)$, $g(\mu_i) = \ln \mu_i = \beta_1 + \beta_2 x_i$ je logaritmická linkovací funkce, β_1, β_2 a $\alpha = \frac{1}{\phi}$ jsou neznámé parametry (α je rušivý parametr) a x_i jsou známé kovariáty.

Příklad



Obrázek : Ukázka GLM modelu s linkovací funkcí $g(\mu) = \ln \mu$ pro náhodnou veličinu Y s gamma rozdělením.

Příklad

Příklad 15

Poissonovská regrese:

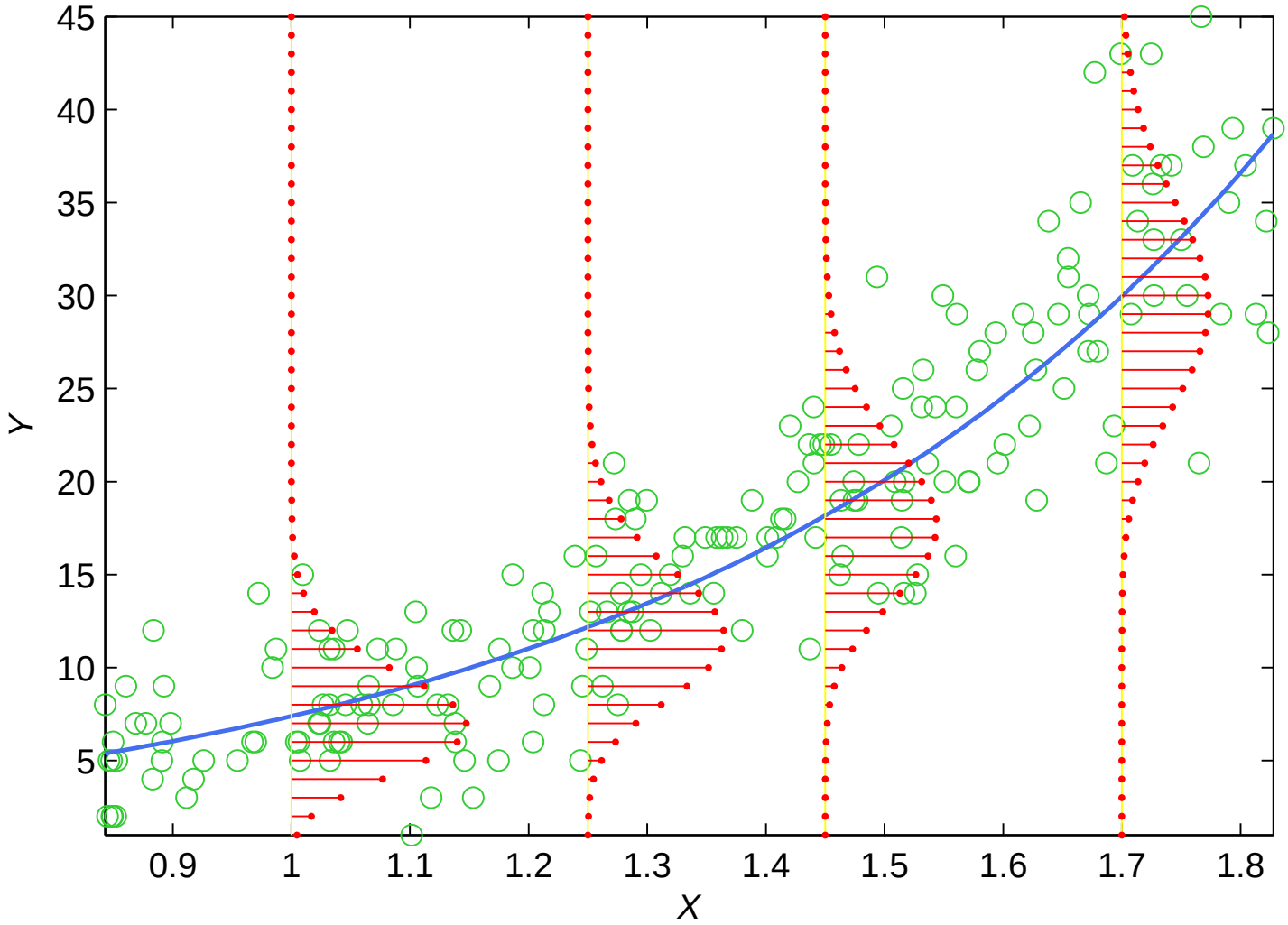
$$Y_i \sim Po(\mu_i)$$

jsou pro $i = 1, \ldots, n$ *nezávislé náhodné veličiny* ($EY_i = \mu_i$),

$$g(\mu_i) = \ln \mu_i = \beta_1 + \beta_2 x_i$$

je logaritmická linkovací funkce, β_1, β_2 *jsou neznámé parametry a* x_i *jsou známé kovariáty.*

Příklad



Obrázek : Ukázka poissonovské regrese s linkovací funkcí $g(\mu) = \ln \mu$.

Příklad

Příklad 16

Binomická regrese:

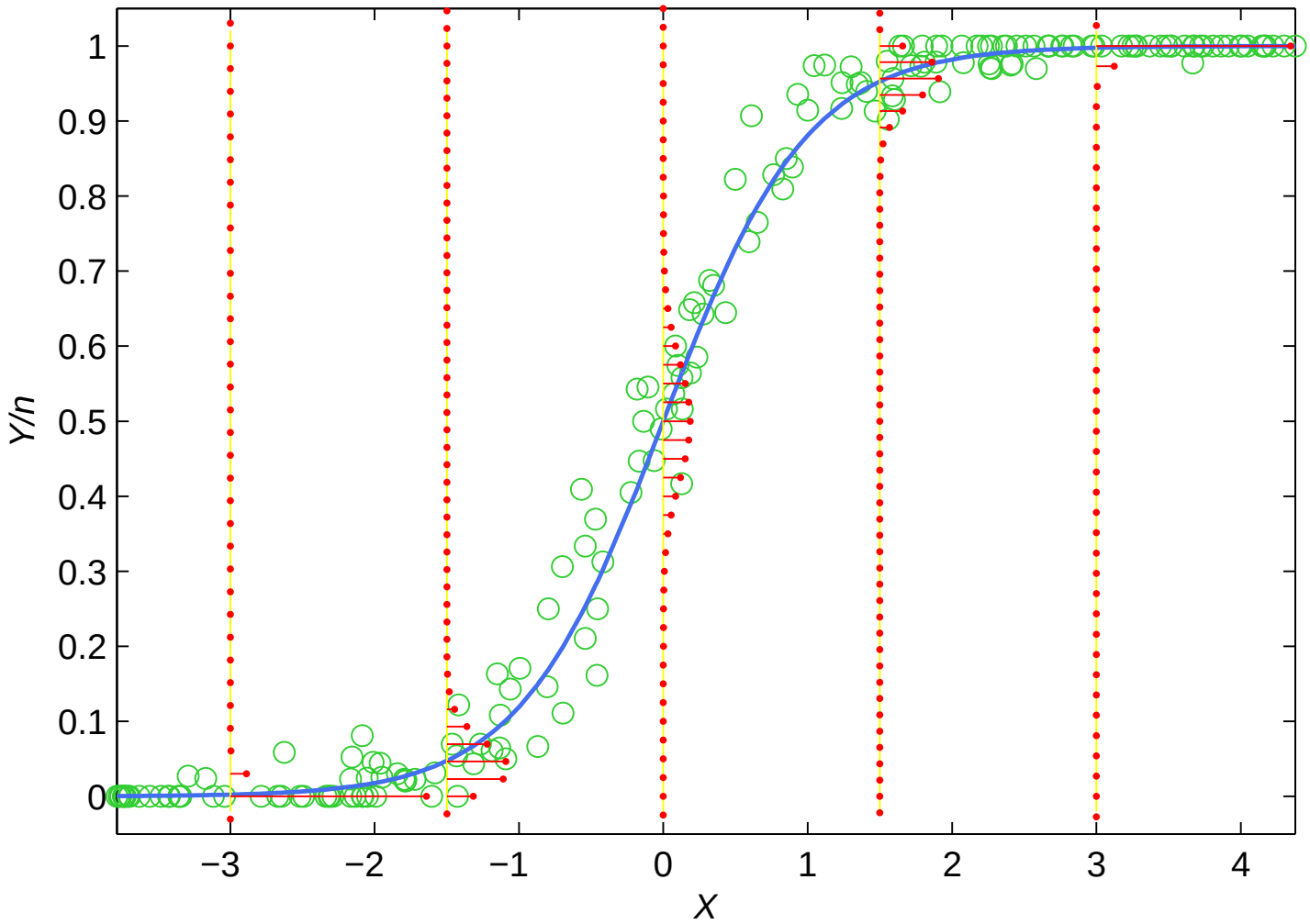
$$Y_i \sim Bi(n_i, \pi_i)$$

jsou pro $i = 1, \ldots, n$ *nezávislé náhodné veličiny, kde*

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

je logistická linkovací funkce, β_1, β_2 *jsou neznámé parametry a* x_i *jsou známé kovariáty.*
Například ve farmaceutickém experimentu může být n_i *počet pacientů, kterým byla podána dávka* x_i *nového léku a* Y_i *počet pacientů dávající pozitivní odpověď na danou dávku* x_i *nového léku.*
Jestliže pozorujeme, že $\frac{Y_i}{n_i}$ *roste spolu s* x_i*, hledáme model, ve kterém* π_i *je funkcí* x_i*, hodnot* $0 < \pi_i < 1$*. Proto model* $\pi_i = \beta_1 + \beta_2 x_i$ *není vhodný, avšak* $\beta_1 + \beta_2 x_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ *obvykle pracuje dobře.*

Příklad



Obrázek : Ukázka binomické regrese s linkovací funkcí $g(\pi) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$.

Praktický příklad

Příklad 17

V souboru „`motak.Rdata`“ jsou uložena data o lovu tetřeva dravcem jménem Moták pilich (Circus cyaneus) v závislosti na výskytu tetřeva. Označme Y_i procento zkonzumovaných tetřevů a x_i počet tetřevů v dané oblasti. Teorie zabývající se chováním těchto dravců navrhují k modelování použít vztahu

$$E(Y_i) = \mu_i = \frac{\alpha x_i^3}{\delta + x_i^3},$$

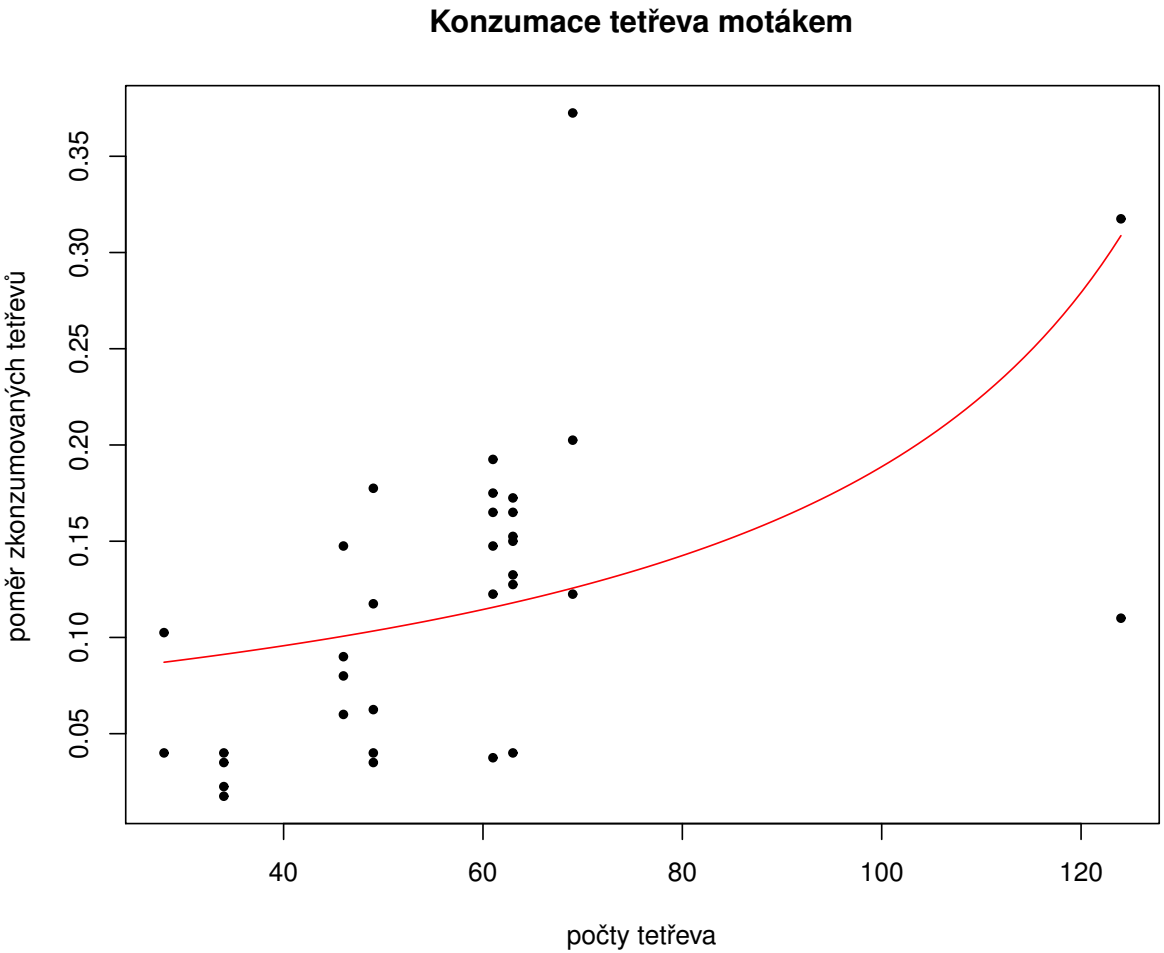
kde Y_i má Gamma rozdělení. Je tedy třeba odhadnout neznámé parametry α a δ. Užitím linkovací funkce `inverse` dostáváme

$$\frac{1}{\mu_i} = \frac{1}{\alpha} + \frac{\delta}{\alpha x_i^3}.$$

Definování nových parametrů $\beta_0 = 1/\alpha$ a $\beta_1 = \delta/\alpha$ dostáváme lineární vztah

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 \frac{1}{x_i^3}.$$

Praktický příklad



Obrázek : Aplikace Gamma regrese s linkovací funkcí $g(\mu) = 1/\mu$ na data `motak`.

Odhady neznámých parametrů v GLM

Všimněme si, že rozdělení náhodných veličin Y_i jsou stejného typu a **logaritmus sdružené věrohodnostní funkce** má tvar

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

Odhad neznámých parametrů metodou **maximální věrohodnosti** dostaneme řešením rovnic typu

$$\frac{\partial l^*}{\partial \boldsymbol{\beta}} = \mathbf{U}_n^*(\boldsymbol{\beta}) = \mathbf{0}$$

Podle věty 5 konverguje matice druhých parciálních derivací skoro jistě k matici $-\mathbf{J}_n$, která je při regularitě systému hustot negativně definitní.

Řešení věrohodnostních rovnic I

Věta 18

Mějme náhodný výběr $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, *který se řídí* **zobecněným lineárním modelem** *s linkovací funkcí*

$$g(\mu_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} = \eta_i \qquad i = 1, \ldots, n.$$

Předpokládejme, že pro $i = 1, \ldots, n$ *existují příslušné derivace* $\gamma'(\theta_i), \gamma''(\theta_i)$ *a platí*

$$EY_i = \mu_i = \gamma'(\theta_i) \qquad DY_i = \gamma''(\theta_i)\psi_i(\phi).$$

Pak

$$U_j^* = U_j^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{x_{ij}(Y_i - \mu_i)}{DY_i} \frac{\partial \mu_i}{\partial \eta_i} \tag{6}$$

a

$$J_{jk}^* = J_{jk}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{DY_i} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2, \tag{7}$$

Řešení věrohodnostních rovnic II

Věta 18

což lze zapsat maticově

$$\mathbf{U}_n^* = \mathbf{U}_n^*(\boldsymbol{\beta}) = (U_1^*, \ldots, U_k^*)^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{Q}\mathbf{r} \tag{8}$$

a

$$\mathbf{J}_n = \mathbf{J}_n(\boldsymbol{\beta}) = \left(J_{ij}^*\right)_{i,j=1}^{k} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\,, \tag{9}$$

kde

$$\mathbf{r} = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))^{\mathrm{T}} \qquad r_i = Y_i - \mu_i = Y_i - g^{-1}(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$$

$$\mathbf{W} = diag\{w_1(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta})\} \qquad w_i = \frac{1}{DY_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

$$\mathbf{Q} = diag\{q_1(\boldsymbol{\beta}), \ldots, q_n(\boldsymbol{\beta})\} \qquad q_i = \frac{\partial \eta_i}{\partial \mu_i}.$$

Řešení věrohodnostních rovnic

Řešíme tedy **věrohodnostní rovnice**

$$U_j^* = \frac{\partial l^*}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}(Y_i - \mu_i)}{DY_i} \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad j = 1, \ldots, k.$$

Ty **nejsou lineární** vzhledem k neznámým parametrům, musí se řešit numerickou iterací.

Newton–Raphsonova metoda

Chceme-li najít řešení systému nelineárních rovnic $\mathbf{U}_n^*(\boldsymbol{\beta}) = \mathbf{0}$, lze použít následující iterativní postup:

1. Nejprve provedeme **linearizaci pomocí Taylorova rozvoje** v okolí bodu $\boldsymbol{\beta}_0$, kde $\boldsymbol{\beta}_0$ je nějaký počáteční odhad: $\mathbf{U}_n^*(\boldsymbol{\beta}) \approx \mathbf{U}_n^*(\boldsymbol{\beta}_0) + \mathbf{U}_n^{*\prime}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Protože $\mathbf{U}_n^*(\boldsymbol{\beta}) = 0$, pak po jednoduchých úpravách dostaneme $\boldsymbol{\beta} \approx \boldsymbol{\beta}_0 - \left[\mathbf{U}_n^{*\prime}(\boldsymbol{\beta}_0)\right]^{-1} \mathbf{U}_n^*(\boldsymbol{\beta}_0)$.
2. Odhady parametrů v s-tém kroku jsou získány ze vztahu

$$\widehat{\boldsymbol{\beta}}^{(s)} = \widehat{\boldsymbol{\beta}}^{(s-1)} - \left[\mathbf{U}_n^{*\prime}\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right)\right]^{-1} \mathbf{U}_n^*\left(\widehat{\boldsymbol{\beta}}^{(s-1)}\right).$$

3. Iterační proces popsaný v předchozím bodě pokračuje tak dlouho, dokud

$$\widehat{\boldsymbol{\beta}}^{(s+1)} - \widehat{\boldsymbol{\beta}}^{(s)} \approx 0.$$

Metoda skórování

Alternativní procedurou k Newton-Raphsonově metodě je tzv. **metoda skórování**, kdy se matice druhých parciálních derivací $\mathbf{U}_n^{*'}(\boldsymbol{\beta})$ nahradí její střední hodnotou, tj. maticí $-\mathbf{J}_n(\boldsymbol{\beta})$, kde $\mathbf{J}_n(\boldsymbol{\beta})$, je **informační matice**.
Druhý iterační krok pak upravíme takto:

$$\widehat{\boldsymbol{\beta}}^{(s)} = \widehat{\boldsymbol{\beta}}^{(s-1)} + \left[\mathbf{J}_n(\widehat{\boldsymbol{\beta}}^{(s-1)})\right]^{-1} \mathbf{U}_n^*(\widehat{\boldsymbol{\beta}}^{(s-1)}).$$

Využijme vztahů:

$$\mathbf{U}_n^*(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\beta})\mathbf{Q}(\boldsymbol{\beta})\mathbf{r}(\boldsymbol{\beta}) \quad \text{a} \quad \mathbf{J}_n(\boldsymbol{\beta}) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(\boldsymbol{\beta})\mathbf{X}$$

a dostáváme iterační rovnici

$$\mathbf{X}^{\mathrm{T}}\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s)} = \mathbf{X}^{\mathrm{T}}\mathbf{W}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{Z}(\widehat{\boldsymbol{\beta}}^{(s-1)}),$$

kde

$$\mathbf{Z}(\widehat{\boldsymbol{\beta}}^{(s-1)}) = \left[\mathbf{X}\widehat{\boldsymbol{\beta}}^{(s-1)} + \mathbf{Q}(\widehat{\boldsymbol{\beta}}^{(s-1)})\mathbf{r}(\widehat{\boldsymbol{\beta}}^{(s-1)})\right].$$

Testování hypotéz v GLM modelech

Věta 19

Mějme náhodný výběr $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, *který se řídí* **zobecněným lineárním modelem** *s maticí vysvětlujících proměnných* $\mathbf{X}_{n\times k}$. *Předpokládejme, že pro* $i = 1, \ldots, n$ *existují příslušné derivace* $\gamma'(\theta_i), \gamma''(\theta_i)$ *a platí*

$$EY_i = \mu_i = \gamma'(\theta_i) \qquad DY_i = \gamma''(\theta_i)\psi_i(\phi).$$

Dále mějme matici $\mathbf{C}_{k\times q}$ *s hodností* $h(\mathbf{C}) = q < k$. *Platí-li hypotéza:* $H_0 : \mathbf{C}^{\mathrm{T}}\boldsymbol{\beta} = 0$, *pak* **Waldova statistika**

$$W = \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{MLE}}\mathbf{C}\left(\mathbf{C}^{\mathrm{T}}\mathbf{J}_n(\boldsymbol{\beta})^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}} \quad \overset{A}{\sim} \quad \chi^2(q),$$

kde $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ *je maximálně věrohodným odhadem vektorového parametru* $\boldsymbol{\beta}$.

Důsledek

Hypotézu

$$H_0 : \mathbf{C}^{\mathrm{T}}\boldsymbol{\beta} = 0$$

zamítáme na hladině významnosti α, pokud platí

$$W > \chi^2_{1-\alpha}(q).$$

Protože odhad $\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ konverguje za předpokladu existence $E(l^*(\boldsymbol{\beta}))$ skoro jistě k $\boldsymbol{\beta}$, aproximujeme při výpočtu Waldovy statistiky W Fisherovou informační matici $\mathbf{J}_n(\boldsymbol{\beta})$ maticí $\mathbf{J}_n(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})$.

Prakticky

Testovat hypotézu

$$H_0 : \beta_j = 0$$

pro $j = 1, \ldots, k$ lze více způsoby:

- Pomocí Waldovy statistiky W, a to při speciální volbě

$$\mathbf{C} = \mathbf{c}_{k\times 1} = (0, \ldots, \overset{j}{1}, \ldots, 0)^{\mathrm{T}}.$$

- Pomocí vztahu

$$\hat{\beta}_{MLE,j} \quad \overset{A}{\sim} \quad N\left(\beta_j, s^{\star}_{jj} = \left(\mathbf{J}_n(\boldsymbol{\beta})^{-1}\right)_{jj}\right),$$

přičemž hypotézu **zamítáme**, pokud

$$\frac{|\hat{\beta}_{MLE,i}|}{\sqrt{s^{\star}_{jj}}} > u_{1-\frac{\alpha}{2}},$$

kde opět Fisherovou informační matici $\mathbf{J}_n(\boldsymbol{\beta})$ aproximujeme maticí $\mathbf{J}_n(\widehat{\boldsymbol{\beta}}_{\mathrm{MLE}})$.

Ověřování vhodnosti modelu

Definice 20

Maximální GLM, který označíme GLM_{max}, splňuje následující podmínky

(1) Maximální model je zobecněný lineární model se stejným typem rozdělení jako zkoumaný GLM model.

(2) Maximální model a zkoumaný mají stejnou linkovací funkci.

(3) Počet parametrů maximálního modelu je roven počtu vysvětlovaných veličin n, maximálně věrohodný odhad parametru $\boldsymbol{\beta}_{max}$ je n-rozměrný vektor $\widehat{\boldsymbol{\beta}}_{max}$.

Definice 21

Minimální GLM, který označíme GLM_{min}, splňuje následující podmínky

(1) Minimální model je zobecněný lineární model se stejným typem rozdělení jako zkoumaný GLM model.

(2) Minimální model a zkoumaný mají stejnou linkovací funkci.

(3) Počet parametrů minimálního modelu je roven 1, maximálně věrohodný odhad parametru $\boldsymbol{\beta}_{min}$ je skalár $\hat{\beta}_{min}$.

Submodel

Definice 22

Mějme zobecněný lineární model s maticí plánu $\mathbf{X}_{n\times k}$ a vektorem neznámých parametrů β. **Submodel**, který označíme GLM_{sub}, splňuje následující podmínky

(1) Submodel je zobecněný lineární model se stejným typem rozdělení jako zkoumaný GLM model.

(2) Submodel a zkoumaný model mají stejnou linkovací funkci.

(3) Vektor neznámých parametrů $\boldsymbol{\beta}_{sub} \in \mathbb{R}^q$a matice plánu $\mathbf{Q}_{n\times q}$, pro kterou platí

$$\mathbf{Q}_{n\times q} = \mathbf{X}_{n\times k}\mathbf{T}_{k\times q}.$$

Aby GLM_{sub} byl submodelem modelu GLM, musí každý sloupec matice $\mathbf{Q}$ patřit do obalu sloupců matice $\mathbf{X}$. To bude splněno právě tehdy, bude-li $\mathbf{Q}$ typu

$$\mathbf{Q}_{n\times q} = \mathbf{X}_{n\times k}\mathbf{T}_{k\times q}.$$

Je třeba si uvědomit, že GLM_{sub} je speciálním případem modelu GLM. Platí-li tudíž pro náhodný výběr $\mathbf{Y}$ model GLM_{sub}, platí pro $\mathbf{Y}$ také model GLM.

Deviace

Deviace v zobecněných lineárních modelech je obdobou rozptylu u klasických lineárních regresních modelů. Deviace je tedy kritériem vhodnosti zobecněného lineárního modelu. Jak bude patrné z definice, metoda maximální věrohodnosti totiž odpovídá hledání minima deviace modelu.

Definice 23

Mějme modely GLM a GLM_{max}. Nechť náhodný výběr $\mathbf{Y}$ se řídí modelem GLM_{max}. **Škálovou deviací modelu GLM** (scaled deviance) rozumíme statistiku

$$D = 2\left[l^*(\widehat{\boldsymbol{\beta}}_{max}; \mathbf{Y}) - l^*(\widehat{\boldsymbol{\beta}}; \mathbf{Y})\right],$$

kde $\widehat{\boldsymbol{\beta}}_{max}, \widehat{\boldsymbol{\beta}}$ jsou odpovídající maximálně věrohodné odhady.

Ověřování vhodnosti submodelu

Věta 24

Mějme základní model GLM *s* $\boldsymbol{\beta} \in \mathbb{R}^k$ *a jeho submodel* GLM_{sub} *s* $\boldsymbol{\beta}_{sub} \in \mathbb{R}^q$, *přičemž* $q < k < n$. *Dále nechť náhodný výběr* $\mathbf{Y}$ *se řídí modelem* GLM *a platí*

(i) *existují druhé parciální derivace hustoty* $f(\mathbf{y};\boldsymbol{\beta})$ *podle složek* $\boldsymbol{\beta}$,

(ii) *platí* $E\left(\frac{f''_{\beta_i\beta_j}(\mathbf{y};\boldsymbol{\beta})}{f(\mathbf{y};\boldsymbol{\beta})}\right) = 0 \qquad (i,j = 1,\ldots,k)$

(iii) *a existuje* $E\, l^*(\boldsymbol{\beta};\mathbf{Y})$.

Platí-li hypotéza, že **submodel $\mathbf{GLM_{sub}}$ je vhodný**, *pak asymptoticky lze rozdělení statistiky* $\Delta D = D_{sub} - D$ *aproximovat rozdělením* $\chi^2(k-q)$, *tj.* $\Delta D = D_{sub} - D \overset{A}{\sim} \chi^2(k-q)$.

Analýza reziduí

Nejznámější typy reziduí používaných v GLM :

(a) **Standardizovaná rezidua** (*linear*): též *Pearsonova*

(b) **Standardizovaná transformovaná rezidua** (*transformed linear*)

(c) **Deviační rezidua** (*deviance residual*)

$$r_i^D = sign(Y_i - \hat{\mu}_i)\sqrt{d_i},$$

přičemž

$$D = 2[l^*(\mathbf{b}_{max}; \mathbf{Y}) - l^*(\mathbf{b}; \mathbf{Y})] = \sum_{i=1}^{n} d_i.$$

Ještě lepší vlastnosti mají tzv. **korigovaná deviační rezidua** (bias-adjusted deviance residual)

$$r_i^{AD} = sign(Y_i - \hat{\mu}_i)\sqrt{d_i} + \rho_3(\hat{\mu}_i)/6,$$

kde

$$\rho_3(\mu) = E\left[\left(\frac{Y - \mu}{\sqrt{DY}}\right)^3\right].$$

Úlohy k procvičení

Příklad 1.1

V souboru „`toxic.RData`“ jsou uvedeny hodnoty množství jedovaté látky, která vzniká jako vedlejší produkt při určitém chemickém procesu. Datový soubor obsahuje tyto proměnné:

`VOL`	*objem vzniklé jedovaté látky (litry)*
`TEMP`	*teplota při chemickém procesu (°C)*
`CAT`	*hmotnost katalyzátoru (kg)*
`METHOD`	*metoda použitá při výrobě (kategoriální proměnná – A,B)*

Hledejte vhodný model pro popis závislosti objemu jedovaté látky na podmínkách procesu. Testujte nejprve, zda použitá metoda má vliv na výsledný objem jedovaté látky. Pomocí stepwise procedury najděte nejvhodnější lineární model a nejvhodnější zobecněný lineární model. U obou modelů ověřte normalitu reziduí.

[Metoda má vliv, vhodný model: $\texttt{VOL} = \beta_0 + \beta_1 \texttt{METHODB} + \beta_2 \texttt{TEMP}$, residua jsou normální]

Úlohy k procvičení I

Příklad 1.2

V balíku „`car`“, proměnné „`SLID`“ jsou uvedeny výsledky průzkumu z roku 1994 v kanadské provincii Ontario. Průzkum se zabýval vlivem některých faktorů na mzdu respondentů. Datový soubor obsahuje tyto proměnné:

`wages`	*hodinová mzda (kanadské dolary)*
`education`	*počet let vzdělávání (roky)*
`age`	*věk (roky)*
`sex`	*pohlaví (1 – žena, 2 – muž)*
`language`	*jazyk (1 – angličtina, 2 – francouzština, 3 – ostatní)*

Úlohy k procvičení II

Příklad 1.2

Hledejte vhodný model pro popis závislosti platu respondenta na ostatních faktorech.

1. *Zkuste nejprve použít klasický lineární model, najděte nejvhodnější model a proveďte analýzu residuí. Jsou splněny předpoklady modelu?*
2. *Stále uvažujte lineární model. Místo proměnné* `wages` *uvažujte* `log(wages)`*. Opět nalezněte nejvhodnější model. Zkuste také přidat dvojné či trojné interakce proměnných. Zlepší se kvalita modelu?*
3. *Pomocí stepwise procedury najděte nejvhodnější zobecněný lineární model.*

[(1) Vhodný model: $\texttt{wages} = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{education} + \beta_3\texttt{sex}$, residua nejsou normální, (2) kvalita se zlepší přidáním dvojných interakcí, (3) vhodný model: $\texttt{wages} = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{education} + \beta_3\texttt{sexMale} + \beta_4\texttt{age:sexMale} + \beta_5\texttt{education:sexMale} + \beta_6\texttt{age:education}$.]

Úlohy k procvičení I

Příklad 1.3

V souboru „`novorozenci.RData`“ jsou uvedeny porodní hmotnosti novorozenců a informace o jejich rodičích. Datový soubor obsahuje tyto proměnné:

`hmnov`	*porodní hmotnost novorozence (g)*
`vyska`	*výška matky (cm)*
`hmmat`	*hmotnost matky (kg)*
`prir`	*váhový přírůstek matky během těhotenství (kg)*
`pohlavi`	*pohlaví dítěte (0 – dívka, 1 – chlapec)*
`stav`	*stav matky při porodu (1 – svobodná, 2 – vdaná, 3 – rozvedená, 4 – vdova)*
`vzdmat`	*vzdělání matky (1 – zákl., 2 – vyuč., 3 – středošk., 4 – vysokošk.)*
`vzdot`	*vzdělání otce (0 – neuved., 1 – zákl., 2 – vyuč., 3 – středošk., 4 – vysokošk.)*

Úlohy k procvičení II

Příklad 1.3

Hledejte vhodný model pro popis závislosti hmotnosti novorozence na jeho rodičích. Testujte nejprve, zda pohlaví má vliv na porodní hmotnost. Pomocí stepwise procedury najděte nejvhodnější model. U modelu ověřte normalitu reziduí.

[Pohlaví má vliv, vhodný model:
$\texttt{hmmat} = \beta_0 + \beta_1\texttt{prir} + \beta_2\texttt{pohlavi1} + \beta_3\texttt{vzdot1} + \beta_4\texttt{vzdot2} + \beta_5\texttt{vzdot3} + \beta_6\texttt{vzdot4} + \beta_7\texttt{vyska} + \beta_8\texttt{hmmat:pohlavi1}$, residua jsou normální]