MA012 Statistika II

9. Zobecněné lineární modely (GLM)

Ondřej Pokora (pokora@math.muni.cz)

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

(podzim 2015)

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.

Klasický lineární regresní model

$$\boldsymbol{Y} \sim \mathcal{L}(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 \boldsymbol{I}_n)$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- nesystematické chyby:

$$\mathrm{E}\varepsilon_i = 0,\ i = 1, \ldots, n, \quad \text{tzn. } \mathrm{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta},$$

- homogenita rozptylu chyb (měření):

$$\mathrm{D}\varepsilon_i = \sigma^2 > 0,\ i = 1, \ldots, n,$$

- nekorelovanost chyb (měření):

$$\mathrm{C}(\varepsilon_i, \varepsilon_j) = 0,\ i, j = 1, \ldots, n,\ i \neq j, \quad \text{tzn. } \mathrm{D}\boldsymbol{\varepsilon} = \mathrm{D}\boldsymbol{Y} = \sigma^2 \boldsymbol{I}_n$$

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 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ tvoří náhodný výběr.
2. Předpokládá **striktní rovnost mezi střední hodnotou náhodné veličiny Y_i a lineární kombinací prediktorů**: $EY_i = \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, kde

   $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$ 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 rozdělení exponenciálního typu**
2. Zobecnění na nelineární funkce, které **spojují** neznámé střední hodnoty výchozího rozdělení náhodné veličiny Y_i s prediktivními proměnnými.

Základní pojmy a definice

Uvažujeme náhodný výběr $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ náhodné veličiny Y, jejíž rozdělení pravděpodobnosti (reprezentované hustotou pravděpodobnosti či pravděpodobnostní funkcí) závisí na neznámých parametrech $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ z množiny $\boldsymbol{\Theta}$ (tzv. parametrický prostor).

Definice 1

Věrohodnostní funkce (likelihood) je funkce vektorového parametru $\boldsymbol{\theta}$, definovaná jako simultánní hustota (resp. pravděpodobností funkce)

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

Logaritmická věrohodnostní funkce (log-likelihood): $\ell(\boldsymbol{\theta}; \mathbf{y}) = \ln L(\boldsymbol{\theta}; \mathbf{y})$.

Řekneme, že odhad $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ je **maximálně věrohodný odhad (MLE, maximum likelihood estimator)** vektorového parametru $\boldsymbol{\theta}$, pokud platí

$$L(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}; \boldsymbol{Y}) \geq L(\boldsymbol{\theta}; \mathbf{Y}), \quad \text{resp.} \quad \ell(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}; \boldsymbol{Y}) \geq \ell(\boldsymbol{\theta}; \mathbf{Y}), \quad \text{pro všechna } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Vlastnosti MLE vektorového parametru

Věta 2

Mějme náhodný výběr $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *náhodné veličiny* Y *s hustotou pravděpodobnosti* $f(y;\boldsymbol{\theta})$ *závislou na parametrech* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)' \in \boldsymbol{\Theta}$ *a maximálně věrohodný odhad* $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ *na základě* $\boldsymbol{Y}$.

Je-li hustota $f(y;\boldsymbol{\theta})$ **regulární**, *platí*

(1) $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} \overset{as.}{\sim} \mathrm{N}_m\left(\boldsymbol{\theta}, \frac{1}{n}\boldsymbol{J}^{-1}\right)$,

(2) $W = n(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta})'\,\boldsymbol{J}\,(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \overset{as.}{\sim} \chi^2(m)$, *(Waldova statistika)*

kde matice $\boldsymbol{J} = \boldsymbol{J}(\boldsymbol{\theta})$ *je tzv. Fisherova informační matice*

$$\boldsymbol{J}_{ij}(\boldsymbol{\theta}) = \int_{\mathbb{R}^n} \frac{\partial \ln f(\boldsymbol{y};\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f(\boldsymbol{y};\boldsymbol{\theta})}{\partial \theta_j} f(\boldsymbol{y};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{y}, \qquad i,j = 1,\ldots,m.$$

Maximálně věrohodný odhad je tedy asymptoticky nestranný a konzistentní. To však neznamená, že se musí nutně jednat o optimální odhad pro náhodný výběr konečného rozsahu.

Vlastnosti MLE skalárního parametru

Věta 3

Mějme náhodný výběr $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ *náhodné veličiny* Y *s hustotou pravděpodobnosti* $f(y;\theta)$ *závisející na parametru* $\theta \in' heta$ *a jeho maximálně věrohodný odhad* $\widehat{\theta}_{\mathrm{ML}}$ *na základě* $\boldsymbol{Y}$.

Je-li hustota $f(y;\theta)$ **regulární**, *platí*

(1) $\widehat{\theta}_{\mathrm{ML}} \overset{as.}{\sim} N\left(\theta, \frac{1}{nJ}\right)$,

(2) $W = nJ\,(\widehat{\theta}_{\mathrm{ML}} - \theta)^2 \overset{as.}{\sim} \chi^2(1)$, *(Waldova statistika)*

kde $J = J(\theta)$ *je tzv. Fisherova míra informace o parametru* θ,

$$J(\theta) = \int_{\mathbb{R}^n} \left[\frac{\partial \ln f(\boldsymbol{y};\theta)}{\partial \theta}\right]^2 f(\boldsymbol{y};\theta)\,\mathrm{d}\boldsymbol{y}.$$

Rozdělení exponenciálního typu

Definice 4

Rozdělení pravděpodobnosti je **exponenciálního typu (exponential family, exponential class)**, pokud jeho pravděpodobnostní funkce (v případě diskrétních rozdělení) či hustota pravděpodobnosti (v případě spojitých rozdělení) je tvaru

$$f(y) = \exp\Big[a(y)b(\theta) + c(\theta) + d(y)\Big],$$

kde θ je tzv. **přirozený parametr (natural parameter)**, a $a(y)$, $b(\theta)$, $c(\theta)$, $d(y)$ jsou známé funkce.

Pokud $a(y) = y$, hovoříme o **kanonické formě** hustoty, resp. pravděpodobnostní funkce.

V konkrétním rozdělení pravděpodobnosti mohou kromě θ dále figurovat další parametry, které nazýváme **rušivými parametry (nuisance parameters)**.

Škálová forma s jedním rušivým parametrem

V dalším budeme uvažovat pouze **regulární** a **kanonické** formy s jedním rušivým parametrem ϕ.

Definice 5

Škálová forma hustoty, resp. pravděpodobnostní funkce, exponenciálního typu s jedním přirozeným parametrem θ a jedním rušivým parametrem ϕ je tvaru

$$f(y) = \exp\left[\frac{\theta y - \gamma(\theta)}{\phi/\omega} + d(\phi, y)\right],$$

kde $\gamma(\theta)$, $d(\phi, y)$ jsou známé funkce, $\omega > 0$ a $\phi > 0$ je tzv. *scale factor*.

Škálová forma s jedním rušivým parametrem

Věta 6

Pro náhodnou veličinu Y z rozdělení s regulární hustotou (resp. pravděpodobnostní funkcí) exponenciálního typu

$$f(y) = \exp\left[\frac{\theta\, y - \gamma(\theta)}{\phi/\omega} + d(\phi, y)\right]$$

platí

$$\mathrm{E}Y = \gamma'(\theta) = \frac{\partial\gamma(\theta)}{\partial\theta}.$$

Pokud navíc platí $\mathrm{E}\left(\frac{f''(Y;\theta)}{f(Y;\theta)}\right) = 0$, *potom*

$$\mathrm{D}Y = \frac{\phi}{\omega}\,\gamma''(\theta) = \frac{\phi}{\omega}\,\frac{\partial^2\gamma(\theta)}{\partial\theta^2}.$$

Funkce $\gamma''(\theta)$ se nazývá **rozptylová funkce (variance function)**.

Normální rozdělení

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

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

přirozený parametr $\theta = \mu \in \mathbb{R}$ je střední hodnota

- $\gamma(\theta) = \frac{1}{2}\theta^2$
- scale factor $\phi = \sigma^2$ je rozptyl; $\omega = 1$
- $d(\phi, y) = -\frac{y^2}{2\,\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$
- $\gamma(\theta)' = \theta = \mu = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = 1$
- $\frac{\phi}{\omega}\,\gamma(\theta)'' = \sigma^2 = \mathrm{D}Y$

Alternativní rozdělení

$$Y \sim A(p), \qquad p \in [0,1],\ y \in \{0,1\}$$

$$f(y) = p^y(1-p)^{1-y} = \left(\frac{p}{1-p}\right)^y (1-p) =$$

$$= \exp\left[y \ln \frac{p}{1-p} + \ln(1-p)\right] = \exp\left[\theta y - \ln(1+\mathrm{e}^\theta)\right]$$

přirozený parametr $\theta = \ln \frac{p}{1-p} \in \mathbb{R}, \qquad p = \frac{1}{1+\mathrm{e}^{-\theta}},\ 1-p = \frac{1}{1+\mathrm{e}^\theta}$

- $\gamma(\theta) = \ln(1+\mathrm{e}^\theta) = -\ln(1-p)$
- scale factor $\phi = 1$; $\omega = 1$
- $d(\phi, y) = 0$
- $\gamma(\theta)' = \frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta} = p = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = \frac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2} = p(1-p)$
- $\frac{\phi}{\omega}\gamma(\theta)'' = p(1-p) = \mathrm{D}Y$

Binomické rozdělení

$$Y \sim Bi(n,p), \qquad n \in \mathbb{N},\ p \in [0,1],\ y \in \{0,1,\ldots,n\}$$

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} \left(\frac{p}{1-p}\right)^y (1-p)^n =$$

$$= \exp\left[y \ln \frac{p}{1-p} + n \ln(1-p) + \ln \binom{n}{y}\right] = \exp\left[\theta y - n \ln(1+\mathrm{e}^\theta) + \ln \binom{n}{y}\right]$$

přirozený parametr $\theta = \ln \frac{p}{1-p} \in \mathbb{R}, \qquad p = \frac{1}{1+\mathrm{e}^{-\theta}},\ 1-p = \frac{1}{1+\mathrm{e}^\theta}$

- $\gamma(\theta) = n \ln(1+\mathrm{e}^\theta) = -n \ln(1-p)$
- scale factor $\phi = 1$; $\omega = 1$
- $d(\phi, y) = \ln \binom{n}{y}$
- $\gamma(\theta)' = n \frac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta} = np = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = n \frac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2} = np(1-p)$
- $\frac{\phi}{\omega} \gamma(\theta)'' = np(1-p) = \mathrm{D}Y$

Poissonovo rozdělení

$$Y \sim Po(\lambda), \qquad \lambda > 0,\ y \in \mathbb{N}_0$$

$$f(y) = \frac{\lambda^y}{y!}\,\mathrm{e}^{-\lambda} = \exp\Big[y\,\ln\lambda - \lambda - \ln(y!)\Big] = \exp\Big[\theta\,y - \mathrm{e}^{\theta} - \ln(y!)\Big]$$

přirozený parametr $\theta = \ln\lambda \in \mathbb{R}, \qquad \lambda = \mathrm{e}^{\theta}$

- $\gamma(\theta) = \mathrm{e}^{\theta} = \lambda$
- scale factor $\phi = 1$; $\omega = 1$
- $d(\phi, y) = -\ln(y!)$
- $\gamma(\theta)' = \mathrm{e}^{\theta} = \lambda = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = \mathrm{e}^{\theta} = \lambda$
- $\frac{\phi}{\omega}\,\gamma(\theta)'' = \lambda = \mathrm{D}Y$

Exponenciální rozdělení (parametr intenzita)

$$Y \sim Ex(\lambda), \qquad \lambda > 0,\ y \geq 0$$

$$f(y) = \lambda \, \mathrm{e}^{-\lambda y} = \exp\Big[-\lambda y + \ln \lambda\Big] = \exp\Big[\theta y + \ln(-\theta)\Big]$$

přirozený parametr $\theta = -\lambda < 0, \qquad \lambda = -\theta$

- $\gamma(\theta) = -\ln(-\theta) = -\ln(\lambda)$
- scale factor $\phi = 1$; $\omega = 1$
- $d(\phi, y) = 0$
- $\gamma(\theta)' = -\frac{1}{\theta} = \frac{1}{\lambda} = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = \frac{1}{\theta^2} = \frac{1}{\lambda^2}$
- $\frac{\phi}{\omega}\,\gamma(\theta)'' = \frac{1}{\lambda^2} = \mathrm{D}Y$

Exponenciální rozdělení (parametr střední hodnota)

$$Y \sim Ex(\mu), \qquad \mu > 0,\ y \geq 0$$

$$f(y) = \frac{1}{\mu}\,\mathrm{e}^{-y/\mu} = \exp\left[-\frac{y}{\mu} - \ln\mu\right] = \exp\left[\theta\, y + \ln(-\theta)\right]$$

přirozený parametr $\theta = -\frac{1}{\mu} < 0, \qquad \mu = -\frac{1}{\theta}$

- $\gamma(\theta) = -\ln(-\theta) = \ln(\mu)$
- scale factor $\phi = 1$; $\omega = 1$
- $d(\phi, y) = 0$
- $\gamma(\theta)' = -\frac{1}{\theta} = \mu = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = \frac{1}{\theta^2} = \mu^2$
- $\frac{\phi}{\omega}\,\gamma(\theta)'' = \mu^2 = \mathrm{D}Y$

Gama rozdělení

$$Y \sim G(k, \mu), \qquad k > 0,\ \mu > 0,\ y \geq 0$$

$$f(y) = \frac{1}{\Gamma(k)} \left(\frac{\mu}{k}\right)^{-k} y^{k-1} \exp\left[-\frac{ky}{\mu}\right] =$$

$$= \exp\left[\frac{-\frac{y}{\mu} - \ln \mu}{\frac{1}{k}} + k \ln k - \ln \Gamma(k) + (k-1) \ln y\right]$$

přirozený parametr $\theta = -\frac{1}{\mu} < 0, \qquad \mu = -\frac{1}{\theta}$

- $\gamma(\theta) = -\ln(-\theta) = \ln(\mu)$
- scale factor $\phi = \frac{1}{k}$; $\omega = 1$
- $d(\phi, y) = k \ln k - \ln \Gamma(k) + (k-1) \ln y$
- $\gamma(\theta)' = -\frac{1}{\theta} = \mu = \mathrm{E}Y$
- rozptylová funkce $\gamma(\theta)'' = \frac{1}{\theta^2} = \mu^2$
- $\frac{\phi}{\omega}\, \gamma(\theta)'' = \frac{\mu^2}{k} = \mathrm{D}Y$

Zobecněný lineární model I

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

Definice 7 (Zobecněný lineární model)

Říkáme, že $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ se řídí **zobecněným lineárním modelem (GLM, generalized Linear model)**, jestliže platí

(1) rozdělení $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ je exponenciálního typu s **regulární** sdruženou hustotou pravděpodobnosti (resp. sdruženou pravděpodobnostní funkcí) tvaru

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

,

Zobecněný lineární model II

Definice 7 (Zobecněný lineární model)

(2) parametr θ_i závisí na $\boldsymbol{x}_i$ a $\boldsymbol{\beta}$ prostřednictvím tzv. **lineárního prediktoru**

$$\eta_i = \boldsymbol{x}_i' \boldsymbol{\beta},$$

(3) je dána ryze monotónní diferencovatelná funkce g, tzv. **linkovací funkce (link function)**, a pro střední hodnotu $\mathrm{E}Y_i = \mu_i$ platí

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

Definice 8 (Zobecněný lineární model s kanonickou linkovací funkcí)

Linkovací funkce g je **kanonická**, pokud

$$g(\mu_i) = \theta_i = \eta_i, \qquad i = 1, \ldots, n,$$

tzn. když lineárním prediktor je přirozeným parametrem.

Zobecněný lineární model

Lineární prediktor η_i je veličina, která zahrnuje informaci o regresorech do GLM. Je to lineární kombinace neznámých parametrů $\boldsymbol{\beta}$.

Linkovací funkce g popisuje vztah mezi lineárním prediktorem a střední hodnotou pozorovaného rozdělení pravděpodobnosti. Linkovací funkcí může být libovolná ryze monotónní diferencovatelná funkce, v praxi se snažíme uvažovat takové funkce, které mají definiční obor rovný množině množných středních hodnot. Kanonická linkovací funkce vyjadřuje přirozený parametr pomocí střední hodnoty, $\theta_i = g(\mu_i)$.

Pro mnoho v praxi užívaných rozdělení pravděpodobnosti je střední hodnota μ_i přímo parametrem použitého rozdělení pravděpodobnosti. V takovém případě je kanonickou linkovací funkcí funkce g, která převádí hustotu pravděpodobnosti (resp. pravděpodobnostní funkci) na kanonický tvar, $\theta_i = g(\mu_i)$.

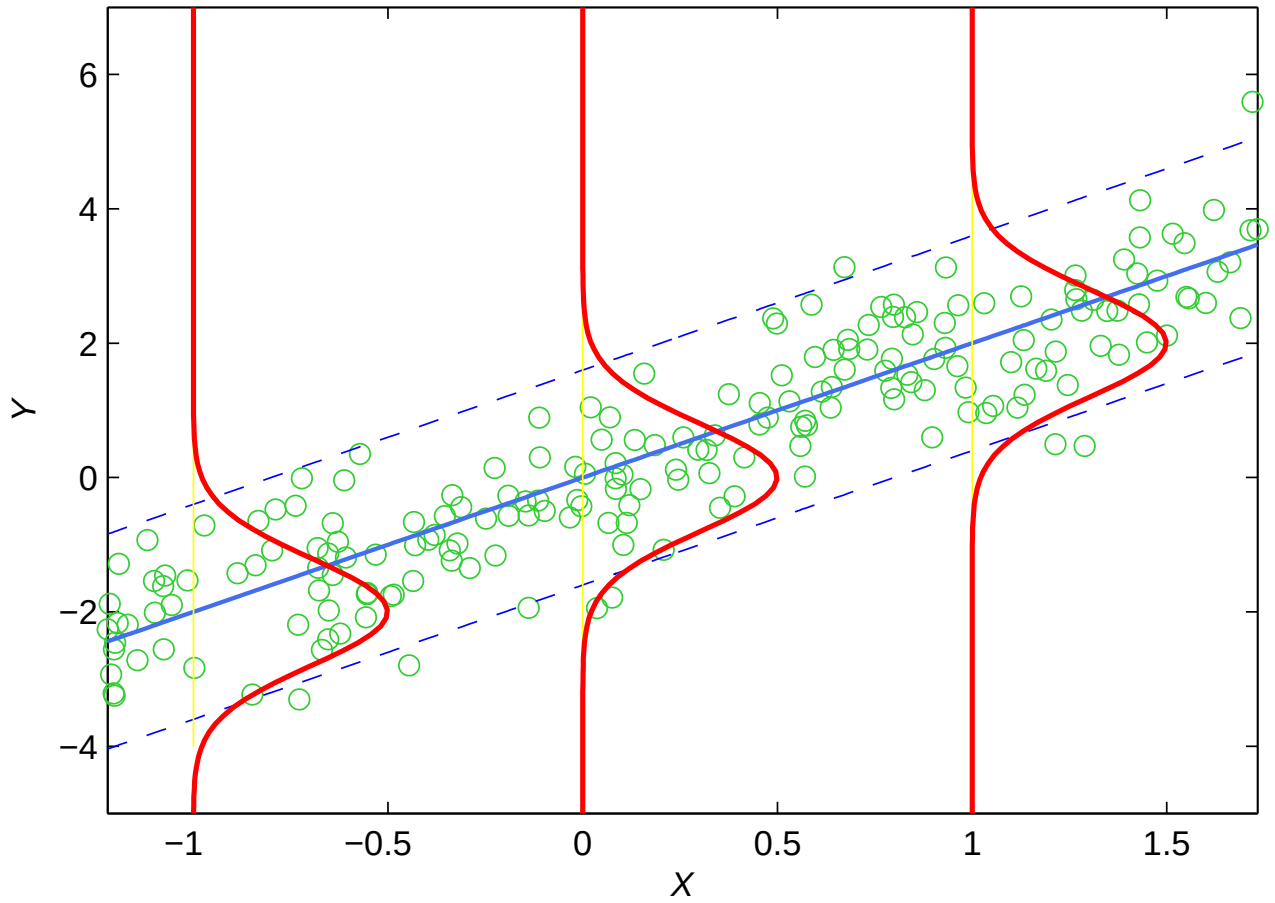
Regresní přímka v klasickém LRM

$$Y_i \sim N(\mu_i, \sigma^2), \quad \mathrm{E}Y_i = \mu_i, \quad i = 1, \ldots, n.$$

Linkovací funkce v GLM je **identita**,

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

β_1, β_2 a σ^2 (rušivý parametr) jsou neznámé parametry, x_i jsou dané kovariáty.



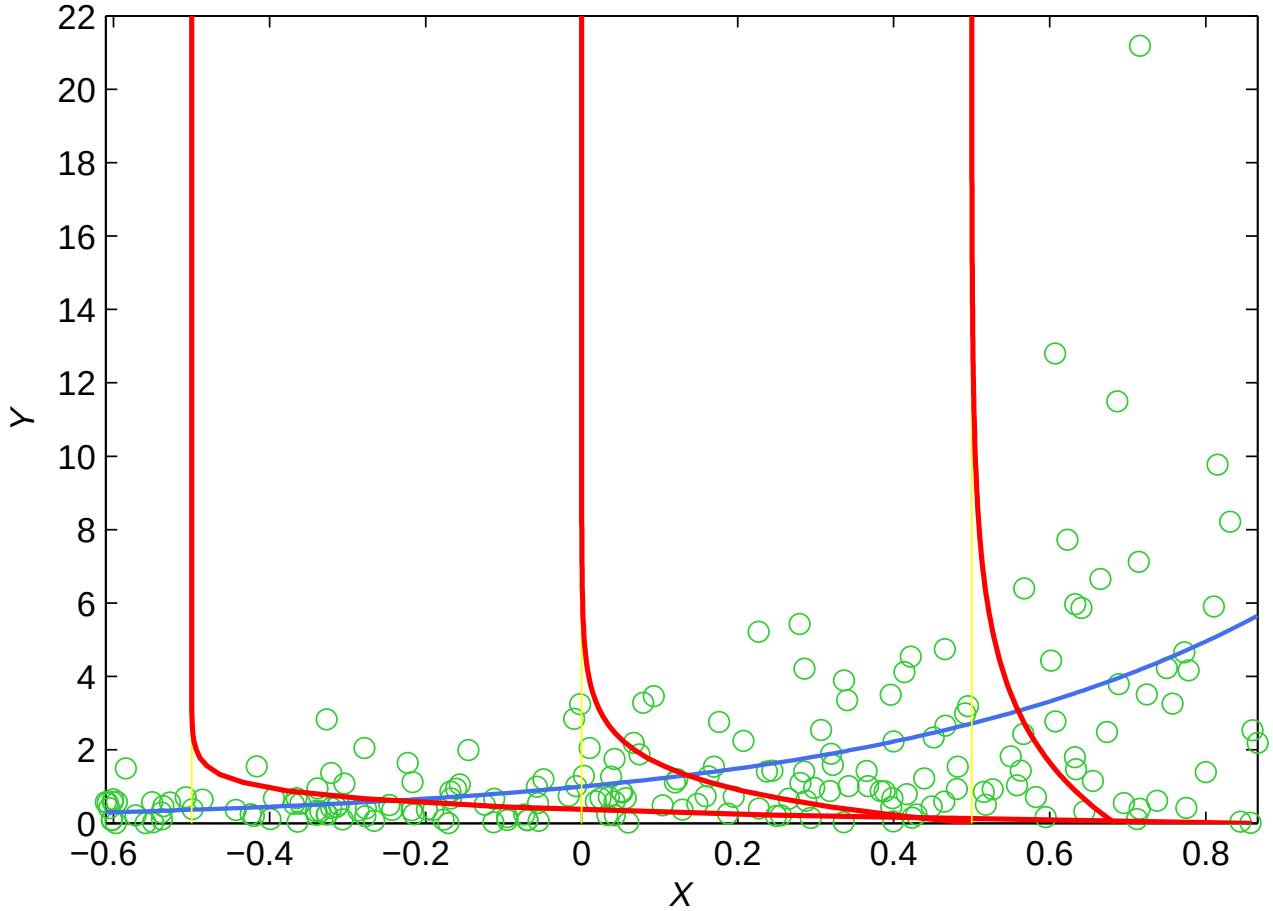
Exponenciální rozdělení, logaritmický link

$$Y_i \sim Ex(\mu_i) \equiv G(1, \mu_i), \quad \mathrm{E}Y_i = \mu_i, \quad i = 1, \ldots, n.$$

Linkovací funkce v GLM je **logaritmus**,

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

β_1, β_2 jsou neznámé parametry, x_i jsou dané kovariáty.



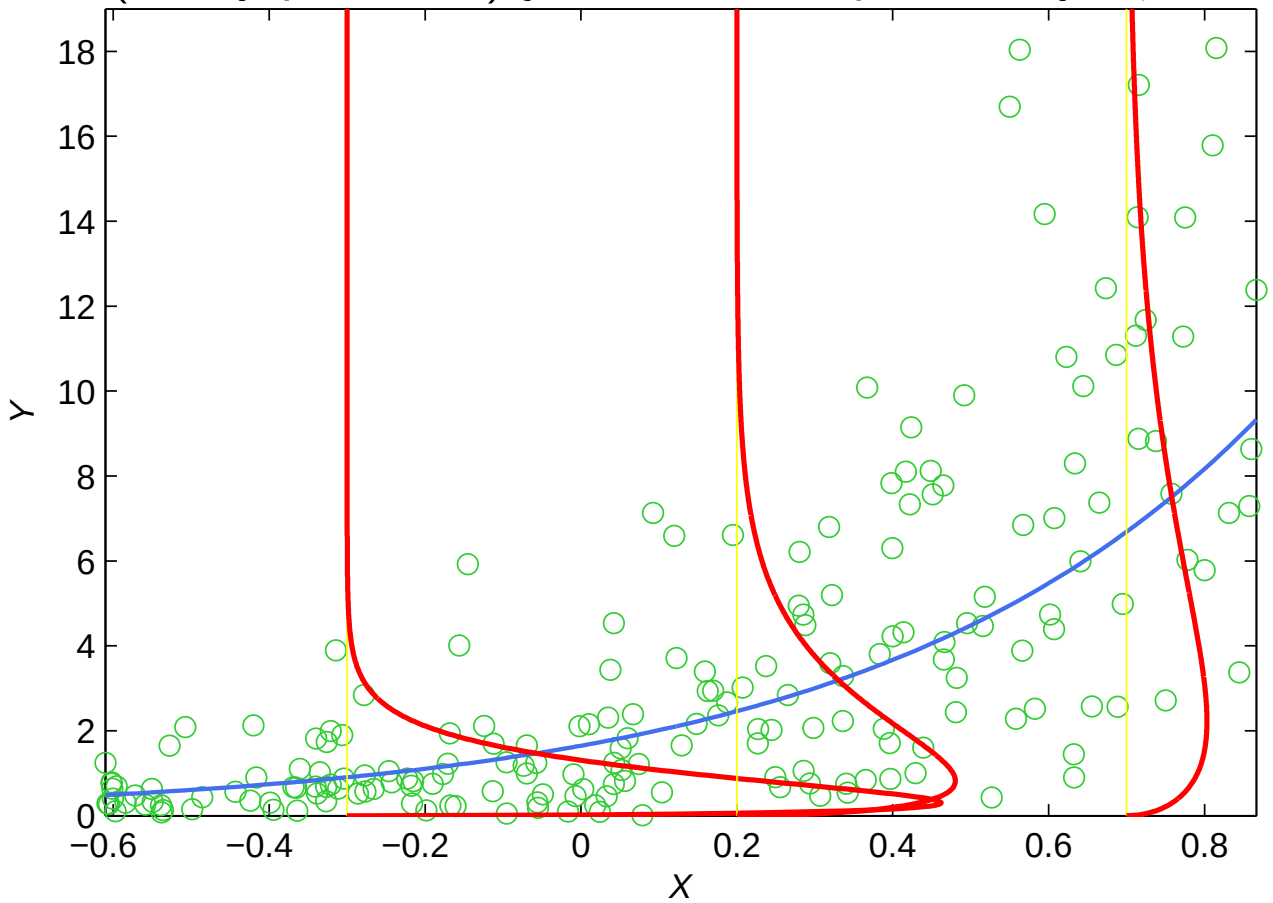
Gama rozdělení, logaritmický link

$$Y_i \sim G(k, \mu_i), \quad \mathrm{E}Y_i = \mu_i, \quad i = 1, \ldots, n.$$

Linkovací funkce v GLM je **logaritmus**,

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

β_1, β_2 a $\phi = 1/k$ (rušivý parametr) jsou neznámé parametry, x_i dané kovariáty.



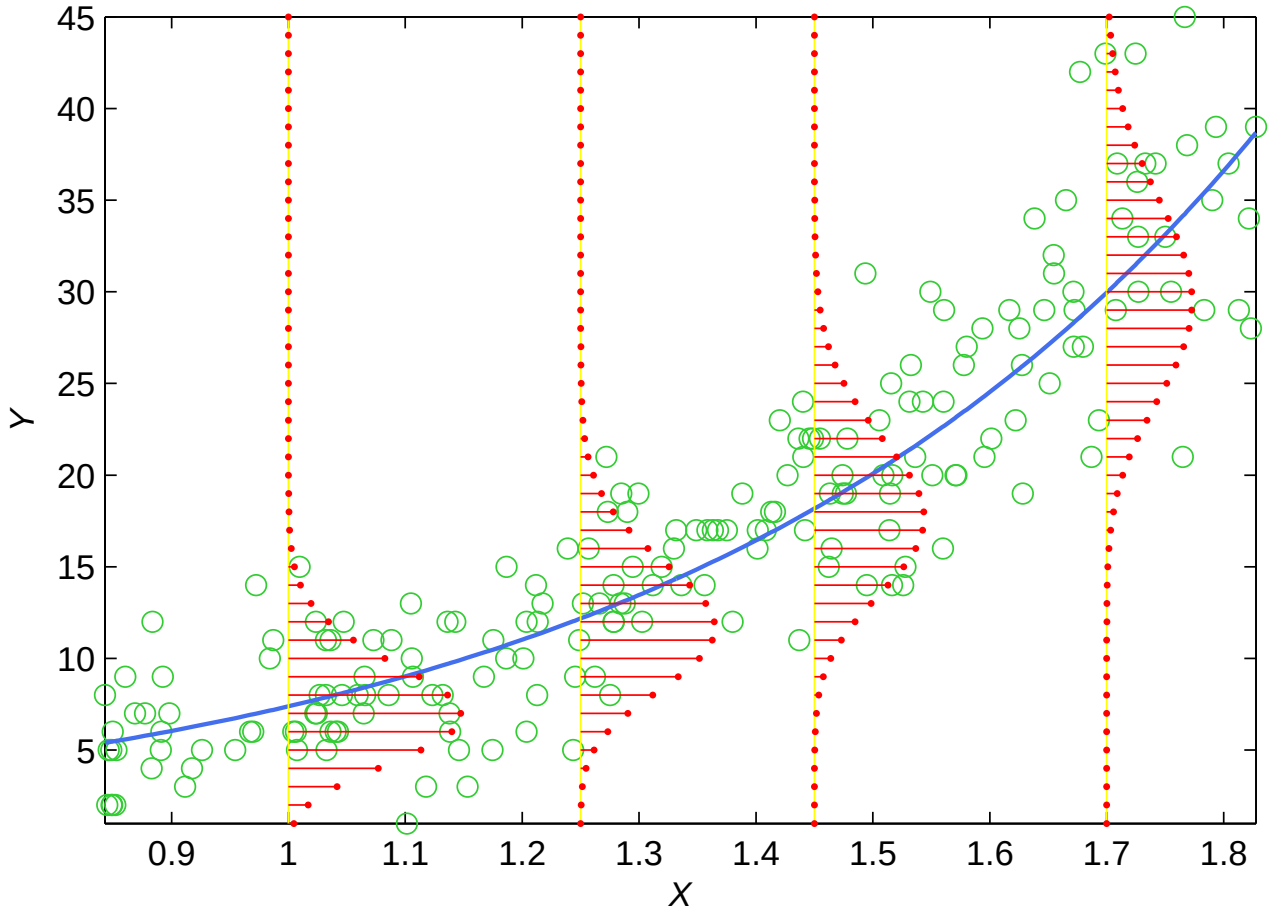
Poissonovo rozdělení, logaritmický link

$$Y_i \sim Po(\lambda_i), \quad \mathrm{E}Y_i = \lambda_i, \quad i = 1, \ldots, n.$$

Linkovací funkce v GLM je **logaritmus**,

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

β_1, β_2 jsou neznámé parametry, x_i jsou dané kovariáty.



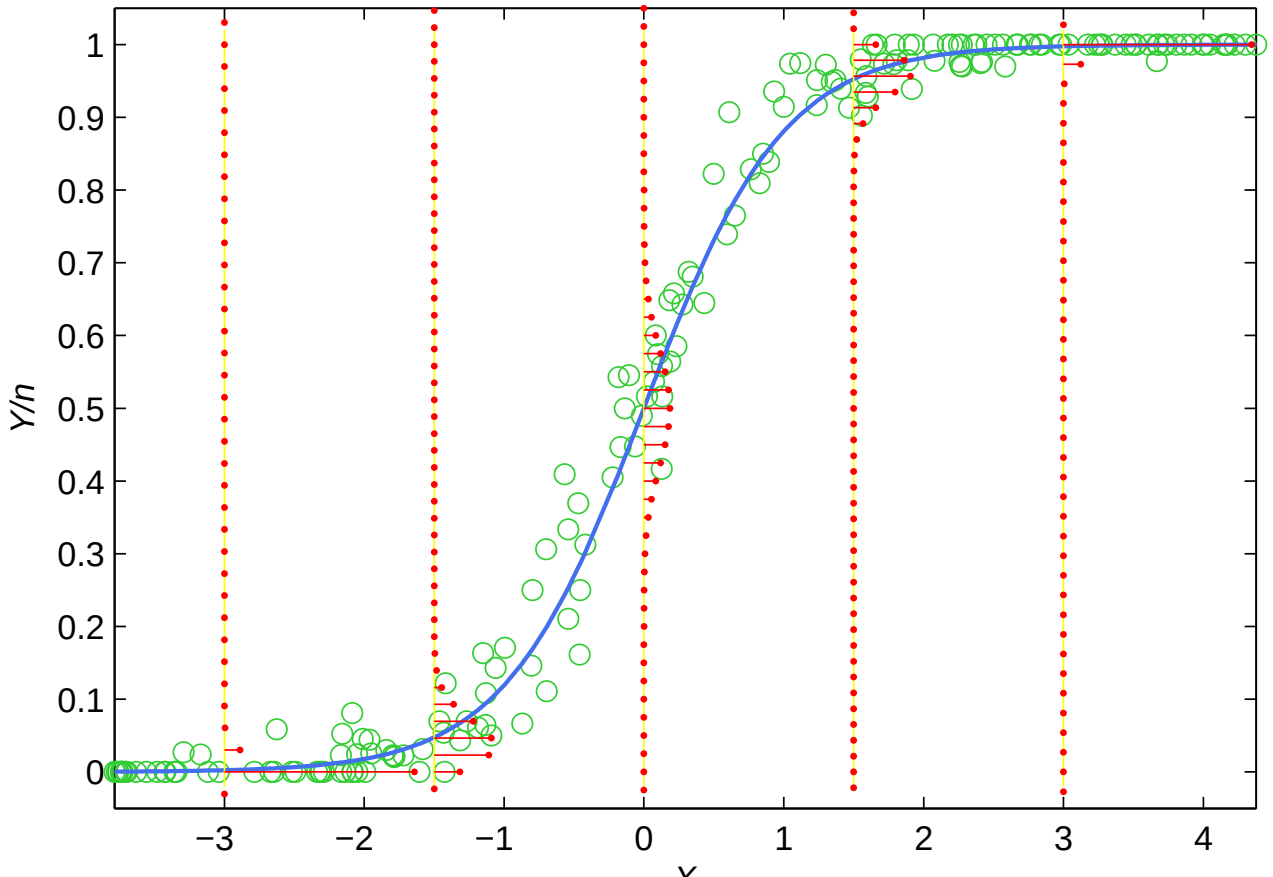
Binomické rozdělení, logitový link

$$Y_i \sim Bi(n_i, p_i), \quad \mathrm{E}\left(\frac{Y_i}{n_i}\right) = \mu_i = p_i, \quad i = 1, \ldots, n.$$

Linkovací funkce v GLM je **logitová funkce**,

$$g(p_i) = \ln \frac{p_i}{1 - p_i} = \eta_i = \beta_1 + \beta_2 x_i,$$

β_1, β_2 jsou neznámé parametry, x_i jsou dané kovariáty.



Odhady neznámých parametrů v GLM

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

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i(\theta_i; y_i) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - \gamma(\theta_i)}{\psi_i(\phi)} + d(y_i, \phi) \right].$$

Odhad neznámých parametrů **metodou maximální věrohodnosti** dostaneme maximalizací logaritmické věrohodnostní funkce $\ell(\boldsymbol{\theta}; \boldsymbol{y})$ vzhledem k parametrům $\boldsymbol{\beta}$, tzn. řešením soustavy věrohodnostních rovnic

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \beta_j} = 0, \qquad j = 1, \ldots, k.$$

Matice druhých parciálních derivací přitom skoro jistě konverguje k matici $-n\,\boldsymbol{J}$, která je při regularitě systému hustot negativně definitní.

Odhady neznámých parametrů v GLM

Uvedenou soustavu věrohodnostních rovnic lze přepsat do tvaru

$$\frac{\partial \ell(\boldsymbol{\theta}; \boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\theta_i; y_i)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}(Y_i - \mu_i)}{DY_i} \frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 1, \ldots, k.$$

Tyto rovnice **nejsou lineární** vzhledem k neznámým parametrům a řeší se proto numericky:

- linearizací pomocí Taylorova rozvoje a výpočtem podle Newtonovy-Raphsonovy metody,
- metodou skórování, kdy se druhých parciálních derivací aproximuje Fisherovou informační maticí.

Testování hypotéz v GLM

Věta 9

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

$$\mathrm{E}Y_i = \mu_i = \gamma'(\theta_i), \qquad \mathrm{D}Y_i = \gamma''(\theta_i)\psi_i(\phi).$$

Dále mějme matici $\boldsymbol{C}_{k\times q}$ *s hodností* $h(\boldsymbol{C}) = q < k$.
Za platnosti $H_0 : \boldsymbol{C}'\boldsymbol{\beta} = 0$ *platí pro* **Waldovu statistiku**

$$W = n\,\widehat{\boldsymbol{\beta}}'_{\mathrm{ML}}\boldsymbol{C}\left(\boldsymbol{C}'\boldsymbol{J}(\boldsymbol{\beta})^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{C}'\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} \overset{as.}{\sim} \chi^2(q).$$

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

Hypotézu $H_0 : \boldsymbol{C}'\boldsymbol{\beta} = 0$ tedy zamítáme na hladině významnosti α, pokud

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

Testování hypotéz v GLM

Testovat hypotézu $H_0 : \beta_j = 0$ pro $j = 1, \ldots, k$ lze následovně:

- Pomocí Waldovy statistiky W, při volbě $\boldsymbol{C}$ tvaru jednotkového vektoru

$$\boldsymbol{C} = (0_1, \ldots, 1_j, \ldots, 0_k)'.$$

- Pomocí vztahu

$$\hat{\beta}_{ML\,j} \overset{as.}{\sim} N\left(\beta_j,\ s^{\star}_{jj}\right), \qquad \text{kde } s^{\star}_{jj} = \frac{1}{n}\left(\boldsymbol{J}(\boldsymbol{\beta})^{-1}\right)_{jj},$$

přičemž hypotézu zamítáme na hladině významnosti α, pokud

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

přičemž opět Fisherovou informační matici $\boldsymbol{J}(\boldsymbol{\beta})$ aproximujeme maticí $\boldsymbol{J}(\widehat{\boldsymbol{\beta}}_{\mathrm{ML}})$.

Maximální a minimální model

Definice 10

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 11

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 12

Mějme zobecněný lineární model s maticí plánu $\boldsymbol{X}_{n\times k}$ a vektorem neznámých parametrů $\boldsymbol{\beta}$. **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 $\boldsymbol{Q}_{n\times q}$, pro kterou platí

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

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

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

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

Deviance

Deviance v zobecněných lineárních modelech je obdobou rozptylu u klasických lineárních regresních modelů. Deviance je tedy kritériem vhodnosti zobecněného lineárního modelu. Metoda maximální věrohodnosti totiž odpovídá hledání minima deviance modelu.

Definice 13 (Škálová deviance)

Mějme modely GLM a GLM_{max}. Nechť náhodný výběr $\boldsymbol{Y}$ se řídí modelem GLM_{max}. **Škálová deviance (scaled deviance)** modelu GLM je statistika

$$D = \ln\left[\frac{L(\widehat{\boldsymbol{\beta}}_{max};\boldsymbol{Y})}{L(\widehat{\boldsymbol{\beta}};\boldsymbol{Y})}\right]^2 = 2\left[\ell(\widehat{\boldsymbol{\beta}}_{max};\boldsymbol{Y}) - \ell(\widehat{\boldsymbol{\beta}};\boldsymbol{Y})\right],$$

kde $\widehat{\boldsymbol{\beta}}_{max}, \widehat{\boldsymbol{\beta}}$ jsou maximálně věrohodné odhady v modelech GLM_{max} a GLM.

Ověřování vhodnosti submodelu

Věta 14

Nechť náhodný výběr $\boldsymbol{Y}$ *se řídí modelem GLM s* $\boldsymbol{\beta} \in \mathbb{R}^k$, $k < n$, *a platí*

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

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

(iii) *existuje* $\mathrm{E}\,\ell(\boldsymbol{\beta};\boldsymbol{Y})$.

Nechť GLM_{sub} *s* $\boldsymbol{\beta}_{sub} \in \mathbb{R}^q$, $q < k < n$, *je submodel modelu GLM.*

Za platnosti hypotézy, že náhodný výběr $\boldsymbol{Y}$ *se řídí modelem* GLM_{sub}, *platí pro rozdíl deviancí těchto modelů*

$$\Delta D = D_{sub} - D \overset{as.}{\sim} \chi^2(k-q).$$

Platnost modelu GLM_{sub} pro náhodný výběr $\boldsymbol{Y}$ tedy zamítáme na hladině významnosti α, pokud

$$\Delta D = D_{sub} - D > \chi^2_{1-\alpha}(k-q).$$

Akaikeovo informační kritérium

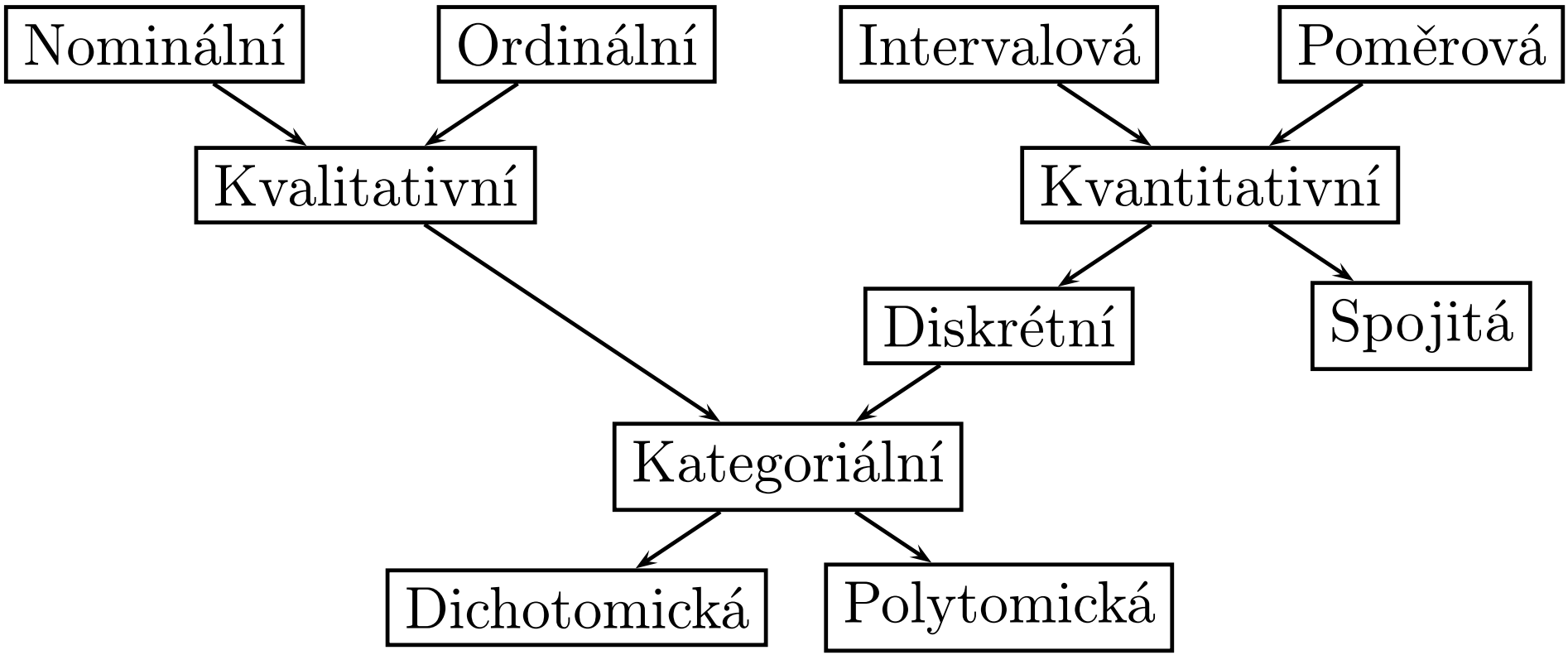
Alternativní mírou relativní kvality modelu je Akaikeovo informační kritérium z teorie informace, založené na relativní Kullbackově-Leiblerově vzdálenosti rozdělení pravděpodobnosti indukované daným GLM vzhledem k GLM_{max}.

Definice 15 (Akaikeovo informační kritérium)

$$AIC = 2k - 2\ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{Y}),$$

kde $\widehat{\boldsymbol{\beta}}$ je maximálně věrohodné odhad v modelu GLM a k je počet parametrů $\boldsymbol{\beta}$.

Typy náhodných veličin



Zobecněné lineární modely v R

Obecná funkce pro řešení GLM v *R* je `glm`.

```
model <-glm (formula, family, data)
```

family	family (link = ...)
gaussian	identity, log, inverse
binomial	logit, probit, cloglog, log, cauchit
poisson	log, sqrt, identity
Gamma	inverse, log, identity
inverse.gaussian	1/mu^2, inverse, log, identity

```
v <- summary (model)
```

S výsledky se pracuje analogicky jako s výsledky funkce `lm` pro LRM.