



FINANCIALSERVICES/ACTUARIAL SERVICES

Zobecněný lineární model (GLM)

Miroslav Šimurda

ADVISORY

AUDIT ■ TAX ■ ADVISORY

Obsah

- **Motivace**
- **Zobecněný lineární model (GLM)**
 - Struktura modelu
 - Vysvětlující proměnné
 - Lineární model
 - Exponenciální rodina rozdělení
 - Metoda maximální věrohodnosti
 - Příklad
- **Sestavení a vyhodnocení modelu**
- **Ukázky**
 - Poškození lodí vlivem počasí
 - Chování pojištěnců – výhody GLM
- **Typické modely, použití**
- **Literatura**

Motivace

- **Formální shoda řady regresních modelů.**
 - Lineární regrese
 - ANOVA
 - Logistická regrese
 - Loglineární modely
 - Multinomické modely
- **Snaha co nejúplněji postihnout vzájemnou souvislost různých jevů:**
 - škodní frekvence v závislosti na segmentaci,
 - průměrná výše škody v závislosti na segmentaci,
 - stornovost v závislosti na čemkoliv,
 - marketing...
- **Metoda schopná správných předpovědí, zohledňující korelace i interakce.**
- **Prakticky použitelná, tj. v běžné praxi nepřiliš složitá.**

GLM – struktura modelu 1

Pozorujeme náhodnou veličinu Y , jejíž každou realizaci y_i (výsledek měření) považujeme za kombinaci systematické složky $E[Y]_i$ a náhodné složky ε_i .

$$y_i = E[Y]_i + \varepsilon_i = \mu_i + \varepsilon_i$$

Systematickou složku se snažíme vyjádřit pomocí vysvětlujících veličin X , náhodná složka je generována podkladovým náhodným dějem, který je zodpovědný za rozdělení $\rho(y_i)$ veličiny Y .

GLM umožňuje na základě historie (n měření) předpovídat systematickou složku pomocí zvolených vysvětlujících veličin a zároveň respektovat náhodnost podkladového děje.

Bohužel ani závislost $\mu_i(x_1, \dots, x_p)$ ani rozdělení $\rho(y_i)$ nemohou být libovolné.

GLM – struktura modelu 2

Předpokládáme, že systematická složka μ_i je prostřednictvím prosté a diferencovatelné funkce g , tzv. *spojovací (link) funkce*, spojena s tzv. *lineárním prediktorem* η_i , tj. lineární funkcí parametrů modelu.

$$g(\mu_i) = \eta_i \Rightarrow \mu_i = g^{-1}(\eta_i)$$

V rámci GLM je tedy systematická složka μ_i funkcí lineárního prediktoru η_i .

Dále předpokládáme, že rozdělení ρ veličiny Y je z tzv. *exponenciální rodiny* rozdělení. Pro tato rozdělení platí, že jsou plně určena střední hodnotou a rozptylem (mají až 2 volné parametry) a rozptyl je funkcí střední hodnoty.

V modelu zvolíme spojovací funkci g , vysvětlující veličiny X , a na základě předpokladu o rozdělení ρ náhodné veličiny Y hledáme takové koeficienty lineárního prediktoru, aby model co nejlépe vystihoval výsledky měření.

GLM – vysvětlující proměnné 1

Lineární prediktor je následující funkce

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j + \xi_i; i = 1, \dots, n$$

$$\vec{\eta} = \mathbf{X} \vec{\beta} + \vec{\xi}$$

$\mathbf{X}$ je tzv. konstrukční matice (*design matrix*) neboli matice $n \times p$, jejíž řádky odpovídají jednotlivým měřením a sloupce tvoří jednotlivé vysvětlující proměnné. Aby byl model jednoznačně definován, musí mít matice $\mathbf{X}$ plnou sloupcovou hodnotu.

β jsou koeficienty, které vyjadřují vliv jednotlivých vysvětlujících proměnných na modelovanou veličinu a jejichž hodnoty hledáme.

ξ je tzv. *offset* neboli člen shrnující vlivy, jejichž efekt na modelovanou veličinu známe a nepotřebujeme tedy, aby jej model odhadoval.

Vysvětlující veličiny, resp. proměnné, mohou být jak kvantitativní (spojité), například hmotnost, tak kvalitativní (kategorické), například barva. Toto rozlišení je však často dáno spíše kontextem a volbou.

GLM – vysvětlující proměnné 2

Kategoriálními proměnnými jsou hladiny (*levels*) jednotlivých kategoriálních veličin, faktorů (*factors*). Například veličina barva může mít několik hladin, které pak tvoří jednotlivé proměnné.

Kategoriální proměnné jsou takové, pomocí nichž sledujeme, zda měření patří nebo nepatří do nějaké kategorie. Nabývají tedy typicky hodnot 1 – patří, 0 – nepatří (Dummy variables).

Hladiny lze zakódovat různě (1,0;-1,1;...) $\Rightarrow$ matice kontrastů (*contrast matrix*).

U kategoriálních proměnných může snadno dojít k lineární závislosti. Například pro proměnné muž a žena, by platilo muž=1-žena. Tyto závislosti ohrožují hodnotu design matrix, a tedy určitost modelu $\Rightarrow$ je třeba správně zvolit kontrasty.

Absolutní člen (*intercept*) β_0 , který v sobě obsáhne všechny základní hladiny faktorů reprezentovaných kategoriálními proměnnými takové obtíže řeší. Všechna měření pak obsahují tento absolutní člen (základní hladinu) a proměnné popisují pouze odlišnost od této reference. Máme pak jen nezávislé proměnné a absolutní člen.

$$\eta_i = \sum_{j=1}^{p'} x_{ij} \beta_j + \beta_0 + \xi_i; i = 1, \dots, n$$

Lineární model a metoda nejmenších čtverců

Obyčejná metoda nejmenších čtverců (n – počet měření, p – počet parametrů modelu)

$$\begin{aligned} \vec{Y} &= \mathbf{X}\vec{\beta} + \vec{\varepsilon} & E\vec{Y} &= \mathbf{X}\vec{\beta} & \text{var } \vec{Y} &= \sigma^2 \mathbf{I} \\ E\vec{\varepsilon} &= 0 & \text{var } \vec{\varepsilon} &= \sigma^2 \mathbf{I} \end{aligned}$$

$$0 = \frac{\partial}{\partial \vec{\beta}} \left[\left(\vec{Y} - \mathbf{X}\hat{\vec{\beta}} \right)^T \left(\vec{Y} - \mathbf{X}\hat{\vec{\beta}} \right) \right] \Rightarrow \hat{\vec{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{Y}$$

$$\hat{\vec{\varepsilon}} = \vec{Y} - \hat{\vec{Y}} = \left(\mathbf{I} - \underbrace{\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T}_{\text{tzv. hat-matrix}} \right) \vec{Y}$$

$$\vec{Y} \approx N(\mathbf{X}\vec{\beta}, \sigma^2) \Rightarrow \hat{\vec{\beta}} \approx N(\vec{\beta}, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2) \quad \frac{\hat{\vec{\varepsilon}}^T \hat{\vec{\varepsilon}}}{\sigma^2} \approx \chi_{n-p}^2 \quad \hat{\sigma}^2 = \frac{\hat{\vec{\varepsilon}}^T \hat{\vec{\varepsilon}}}{n-p}$$

Zobecněná metoda nejmenších čtverců – metoda vážených nejmenších čtverců (w.l.s.)

$$\begin{aligned} \vec{Y} &= \mathbf{X}\vec{\beta} + \vec{\varepsilon} & E\vec{Y} &= \mathbf{X}\vec{\beta} & \text{var } \vec{Y} &= \sigma^2 \Sigma \\ E\vec{\varepsilon} &= 0 & \text{var } \vec{\varepsilon} &= \sigma^2 \Sigma & \Sigma &= \mathbf{S}\mathbf{S}^T \end{aligned}$$

$$\mathbf{S}^{-1} \vec{Y} = \mathbf{S}^{-1} \mathbf{X} \vec{\beta} + \mathbf{S}^{-1} \vec{\varepsilon} \rightarrow \vec{Y}' = \mathbf{X}' \vec{\beta} + \vec{\varepsilon}' \quad \begin{aligned} \text{var } \vec{Y}' &= \sigma^2 \mathbf{I} \\ \text{var } \vec{\varepsilon}' &= \sigma^2 \mathbf{I} \end{aligned}$$

$$\hat{\vec{\beta}} = (\mathbf{X}'^T \Sigma^{-1} \mathbf{X}')^{-1} \mathbf{X}'^T \Sigma^{-1} \vec{Y}' \quad \text{var } \hat{\vec{\beta}} = (\mathbf{X}'^T \Sigma^{-1} \mathbf{X}')^{-1} \sigma^2$$

GLM – exponenciální rodina rozdělání

Hustota pravděpodobnosti exponenciální rodiny rozdělání má obecně tvar

$$\rho(y_i; \theta_i, \phi) = \text{Exp} \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right)$$

θ je *kanonický parametr* související se střední hodnotou, ϕ je *rozptylový parametr* související s rozptylem, $a_i(\phi)$ je spojitá a kladná funkce, $b(\theta)$ (kumulantová funkce) je dvakrát diferencovatelná konvexní funkce a $c(y, \phi)$ je funkce normující ρ , nezávislá na θ .

$$E[y_i] = \mu_i = \left. \frac{db}{d\theta} \right|_{\theta_i} = b'(\theta_i)$$

$$\text{var}(y_i) = a_i(\phi) \left. \frac{d^2 b}{d\theta^2} \right|_{\theta_i} = a_i(\phi) b''(\theta_i) = a_i(\phi) V(\mu_i)$$

V je varianční funkce, obvykle $a_i(\phi) = \phi / w_i$, kde w_i je apriorní váha i -tého měření

GLM – exponenciální rodina rozdělání

	Označení	Definiční obor	ϕ	$b(\theta)$	$c(y, \phi)$	$\mu(\theta)$	$\theta(\mu)$ (kanonický link)	$V(\mu)$
Normální	$N(\mu, \sigma^2)$	$(-\infty, \infty)$	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left(\frac{y^2}{\phi} + \ln(2\pi\phi) \right)$	θ	μ	1
Poissonovo	$P(\mu)$	$0, 1, 2, \dots, \infty$	1	e^θ	$-\ln(y!)$	e^θ	$\ln(\mu)$	μ
Binomické	$Bi(m, \pi)/m$	$\frac{0, 1, 2, \dots, m}{m}$	$\frac{1}{m}$	$\ln(1 + e^\theta)$	$\ln \binom{m}{my}$	$\frac{e^\theta}{1 + e^\theta}$	$\ln \left(\frac{\mu}{1 - \mu} \right)$	$\mu(1 - \mu)$
Gamma	$G(\mu, \nu)$	$(0, \infty)$	$\frac{1}{\nu}$	$-\ln(-\theta)$	$\nu \ln(\nu y) - \ln(y) - \ln(\Gamma(\nu))$	$-\frac{1}{\theta}$	$\frac{1}{\mu}$	μ^2
Inverzní Gaussovo	$IG(\mu, \sigma^2)$	$(0, \infty)$	σ^2	$-\sqrt{-2\theta}$	$-\frac{1}{2} \left(\ln(2\pi\phi y^3) + \frac{1}{\phi y} \right)$	$\frac{1}{\sqrt{-2\theta}}$	$\frac{1}{\mu^2}$	μ^3

„Blízcí příbuzní:“ negativně binomické, Weibulovo, ... (Lognormální NE)

GLM – metoda maximální věrohodnosti

Řešení GLM = nalezení nejvěrohodnějšího odhadu β = hledání maxima věrohodnostní funkce L (*likelihood*), respektive jejího logaritmu ℓ (*loglikelihood*), vůči parametrům β

$$L = \prod_{i=1}^n \rho(y_i; \theta_i, \phi) \qquad \ell = \ln L = \sum_{i=1}^n \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right]$$

$$0 = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{\partial \eta_i}{\partial \mu_i} x_{ij} \rightarrow \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{W_i}{a(\phi)} (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i} x_{ij}$$

Maximum věrohodnostní funkce se hledá numericky (Newton-Raphson, Fisher scoring) $\leftrightarrow$ metoda iteračně vážených nejmenších čtverců.

Odhad parametru ϕ lze provést například pomocí zobecněné Pearsonovy X^2 statistiky nebo pomocí deviance D , resp. škálované deviance D^* .

$$X^2 = \sum \frac{(y - \hat{\mu})^2}{V(\hat{\mu})} \approx \phi \chi_{n-p}^2$$

$$\hat{\phi} = \frac{X^2}{n-p} \qquad \hat{\phi} = \frac{D}{n-p}$$

$$D = \phi D^* \quad D^* = 2\ell(y, \phi|y) - 2\ell(\hat{\mu}, \phi|y) \approx \chi_{n-p}^2$$

GLM – Iterace – nástin...

$$g(\mu) = \eta$$

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu) = \eta + (y - \mu)\frac{\partial \eta}{\partial \mu} \rightarrow z = \hat{\eta} + (y - \hat{\mu})\frac{\partial \eta}{\partial \mu}\bigg|_{\hat{\mu}}$$

$$\text{var } z = E(z^2) - E^2 z$$

$$= E \left(\underbrace{\hat{\eta}^2 + (y - \hat{\mu})^2 \left(\frac{\partial \eta}{\partial \mu} \bigg|_{\hat{\mu}} \right)^2 + 2\hat{\eta}(y - \hat{\mu}) \frac{\partial \eta}{\partial \mu} \bigg|_{\hat{\mu}}}_{E\hat{\eta}^2 + \text{var } y \left(\frac{\partial \eta}{\partial \mu} \bigg|_{\hat{\mu}} \right)^2} - \underbrace{E^2 \left(\hat{\eta} + (y - \hat{\mu}) \frac{\partial \eta}{\partial \mu} \bigg|_{\hat{\mu}} \right)}_{E^2 \hat{\eta}} = \text{var } y \left(\frac{\partial \eta}{\partial \mu} \bigg|_{\hat{\mu}} \right)^2$$

$$\text{var } z \cong V(\hat{\mu}) \left(\frac{\partial \eta}{\partial \mu} \bigg|_{\hat{\mu}} \right)^2 = w^{-1} \rightarrow \mathbf{W} = \Sigma^{-1}$$

$$y \rightarrow \hat{\mu}^{(i)}, \hat{\eta}^{(i)} \rightarrow \bar{z}^{(i)}, \mathbf{W}^{(i)} \xrightarrow{\text{w.l.s.}} \bar{z}^{(i)} = \vec{\eta}^{(i+1)}$$

GLM – exponenciální rodina rozdělení nad-rozptyl (overdispersion)

Za určitých okolností nemusí být splněno, že ϕ u binomického nebo Poissonova rozdělení je 1. Dochází k tzv. nad-rozptylu (*overdispersion*).

V případě binomického rozdělení může nad-rozptyl vznikat například existencí shluků (*clusters*) lišících se pravděpodobnostmi sledovaného jevu (nebo velikostí).

$$Y = Z_1 + Z_2 + \dots + Z_{m/k} \quad Z_i \approx Bi(\pi_i, k) \quad E\pi_i = \pi \quad \text{var } \pi_i = \tau^2 \pi(1-\pi) \quad EY = m\pi$$

$$\text{var } Y = m\pi(1-\pi) \left[1 + (k-1)\tau^2 \right] = \sigma^2 m\pi(1-\pi)$$

V případě Poissonova rozdělení může totéž nastávat například pokud jedna událost přispívá více sledovanými jevy nebo pokud je pravděpodobnost jevu různá pro různé jednotky na nichž výskyt jevu sledujeme.

$$Y = Z_1 + Z_2 + \dots + Z_N \quad Z_1, \dots, Z_N \approx \text{i.i.d.} \quad N \approx Po(n) \quad EY = EN EZ \quad \text{var } Y = EN EZ^2$$

Ve druhém zmíněném případě je náchylnost k jevu u jednotlivých jednotek v souboru různá. Tedy jev má u každé jednotky Poissonovo rozdělení, ale střední hodnota je u každé jednotky jiná (*inter-subject variability*). Pokud mají střední hodnoty v souboru např. gamma rozdělení pak celkové počty jsou rozděleny negativně binomicky.

GLM – kvazi-věrohodnost

Pro specifikaci GLM jsme potřebovali pouze nezávislost měření, spojovací a varianční funkci, bez odkazu na jiné vlastnosti rozdělení.

Pokud pro nezávislá Y_i známe střední hodnotu a rozptyl, pak pro provedení GLM potřebujeme navíc pouze věrohodnostní funkci.

$$u_i = \frac{y_i - \mu_i}{\phi V(\mu_i)}$$

$$Eu_i = 0 \quad \text{var } u_i = \frac{1}{\phi V(\mu_i)} \quad -E \frac{\partial u_i}{\partial \mu_i} = \frac{1}{\phi V(\mu_i)} \quad \longrightarrow \quad u_i \leftrightarrow \frac{\partial \ell}{\partial \mu_i}$$

u_i odpovídá v uvedených vlastnostech derivaci logaritmické věrohodnostní funkce
 $\Rightarrow$

Integrací (pokud lze) získáme „něco jako“ logaritmickou věrohodnostní funkci – tzv. kvazi-věrohodnostní funkci nebo přesněji logaritmickou kvazi-věrohodnostní funkci q (*quasi-likelihood*, *log quasi-likelihood*)

$$q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$

$$q = \sum_{i=1}^n q_i$$

GLM –příklad

	MI	St
M	1000	600
Ž	500	300

$$\vec{y} = \begin{pmatrix} 1000 \\ 600 \\ 500 \\ 300 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad \vec{\beta} = \begin{pmatrix} \beta_{MM} \\ \beta_{\check{Z}} \\ \beta_{St} \end{pmatrix} \quad \vec{\eta} = \begin{pmatrix} \beta_{MM} \\ \beta_{MM} + \beta_{St} \\ \beta_{MM} + \beta_{\check{Z}} \\ \beta_{MM} + \beta_{\check{Z}} + \beta_{St} \end{pmatrix}$$

Rozdělení – Poissonovo
Link – Logaritmus

$$\mu_i = e^{\eta_i} = e^{\beta_{MM} + \check{Z}\beta_{\check{Z}} + St\beta_{St}} = e^{\beta_{MM}} e^{\check{Z}\beta_{\check{Z}}} e^{St\beta_{St}} \quad \ell = \sum_{i=1}^4 y_i \ln(\mu_i) - \mu_i - \ln(y_i!)$$

$$\begin{aligned} \ell = & 1000 \cdot \beta_{MM} - e^{\beta_{MM}} - \ln(1000!) \\ & + 600 \cdot (\beta_{MM} + \beta_{St}) - e^{\beta_{MM} + \beta_{St}} - \ln(600!) \\ & + 500 \cdot (\beta_{MM} + \beta_{\check{Z}}) - e^{\beta_{MM} + \beta_{\check{Z}}} - \ln(500!) \\ & + 300 \cdot (\beta_{MM} + \beta_{\check{Z}} + \beta_{St}) - e^{\beta_{MM} + \beta_{\check{Z}} + \beta_{St}} - \ln(300!) \end{aligned}$$

$$0 = \frac{\partial \ell}{\partial \beta_{MM}} = 2400 - e^{\beta_{MM}} (1 + e^{\beta_{St}} + e^{\beta_{\check{Z}}} + e^{\beta_{St} + \beta_{\check{Z}}})$$

$$0 = \frac{\partial \ell}{\partial \beta_{St}} = 900 - e^{\beta_{St}} (e^{\beta_{MM}} + e^{\beta_{MM} + \beta_{\check{Z}}})$$

$$0 = \frac{\partial \ell}{\partial \beta_{\check{Z}}} = 800 - e^{\beta_{\check{Z}}} (e^{\beta_{MM}} + e^{\beta_{MM} + \beta_{St}})$$

$$\beta_{MM} = 6,9078$$

$$\beta_{\check{Z}} = -0,6932$$

$$\beta_{St} = -0,5108$$

Sestavení a vyhodnocení modelu

- **Rozdělení**

- Analýza rozdělení sledované veličiny, porovnání výsledků modelu se skutečností

- **Spojovací funkce**

- Praktičnost
- Realističnost

- **Vysvětlující proměnné, design matrix**

- Volba veličin
- Volba hladin kategoriálních veličin
- Zahrnutí interakcí
- Analýza vlivu jednotlivých proměnných na výsledky modelu

Sestavení modelu – rozdělení

Volba rozdělení vychází z předchozí znalosti, zkušeností a podstaty podkladového náhodného děje.

Správnost volby lze (ne nezávisle na zbytku modelu) ověřit pomocí různých měr rozdílu, reziduí, mezi měřenými a modelem předpovídanými hodnotami.

Vhodnou volbou jsou tzv. devianční rezidua, která jsou při správné volbě modelu velmi dobře normálně rozdělena.

$$D = \sum_{i=1}^N r_{Di}^2, \quad r_{Di} = \text{sign}(y_i - \mu_i) \sqrt{d_i} = \text{sign}(y_i - \mu_i) \sqrt{2 \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt}$$

Standardizovaná devianční rezidua mají navíc jednotkový rozptyl.

$$r_{DSi} = \frac{r_{Di}}{\sqrt{\phi(1-h_i)}} = \frac{\text{sign}(y_i - \mu_i)}{\sqrt{\phi(1-h_i)}} \sqrt{2 \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt}$$

h_i jsou diagonální prvky vlivové matice (*hat-matrix*) tzv. páky (*leverage*), které popisují vliv i -tého měření na model, 1 – velký vliv, 0 – malý vliv

Sestavení modelu – spojovací funkce

Kanonický link zjednodušuje tvar věrohodnostní funkce, a jeho použití má i jiné příznivé důsledky, které však dnes, díky počítačům, nejsou rozhodující.

Rozhodují data a praktičnost – v pojišťovnictví je zpravidla příjemný multiplikativní model s logaritmem jako spojovací funkcí. Pro binomické modely je třeba link, který zobrazuje hodnoty z intervalu $<0,1>$ na $<-\infty,\infty>$ - např. kvantilové funkce.

Testovat lze maximum věrohodnostní funkce, kterého je možné dosáhnout s různými spojovacími funkcemi.

$$g(x; \lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x), & \lambda = 0 \end{cases}$$

$g(x; \lambda)$ přechází od inverzní, pro $\lambda = -1$, přes logaritmickou, pro $\lambda = 0$, do identické, pro $\lambda = 1$, spojovací funkce, a nabízí tak možnost určit vhodnou spojovací funkci nalezením maxima věrohodnostní funkce v závislosti na λ , a vybrat tak spojovací funkci maximalizující věrohodnost.

Sestavení modelu – proměnné – testování

Přidávány by měly být pouze proměnné, které model signifikantně vylepší.

Standardní mírou dobré shody modelu je deviance D , repektive škálovaná deviance D^*

$$D = 2 \sum_{i=1}^n \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt$$

$$D^* = 2 \sum_{i=1}^n \frac{1}{\phi} \int_{\mu_i}^{y_i} \frac{y_i - t}{V(t)} dt$$

Dva vnořené modely lze tedy porovnávat srovnáním jejich škálovaných deviancí, pokud je parametr ϕ známý (např. u Poissonova rozdělení) (model ω je podmodelem modelu Ω).

$$D_{\omega}^* - D_{\Omega}^* = 2(\ell_{\Omega} - \ell_{\omega}) \sim \chi_{df_{\omega} - df_{\Omega}}^2, \quad df_{\omega} > df_{\Omega}$$

Případně, pokud je i ϕ odhadované,

$$\frac{1}{\hat{\phi}} \frac{D_{\omega} - D_{\Omega}}{df_{\omega} - df_{\Omega}} \sim F_{df_{\omega} - df_{\Omega}, df_{\Omega}}, \quad df_{\omega} > df_{\Omega} \quad \hat{\phi} = \frac{X^2}{df} \quad \hat{\phi} = \frac{D}{df}$$

Porovnávání různých modelů – Akaikeho informační kritérium

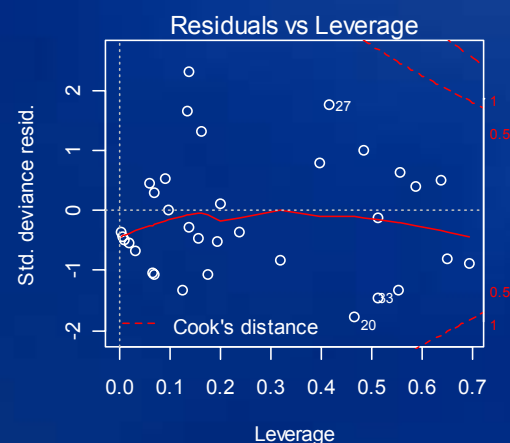
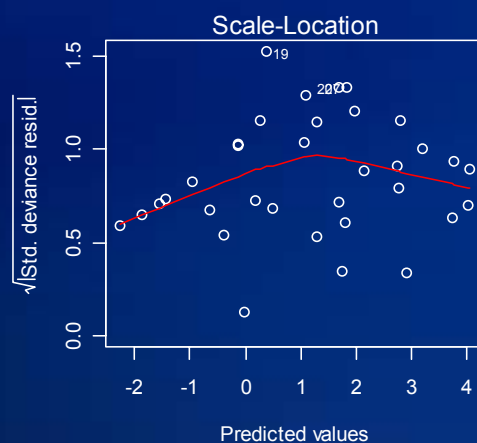
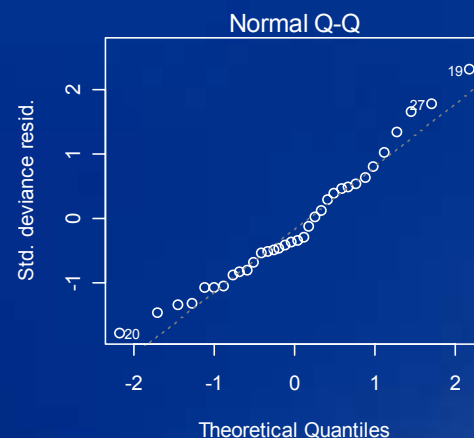
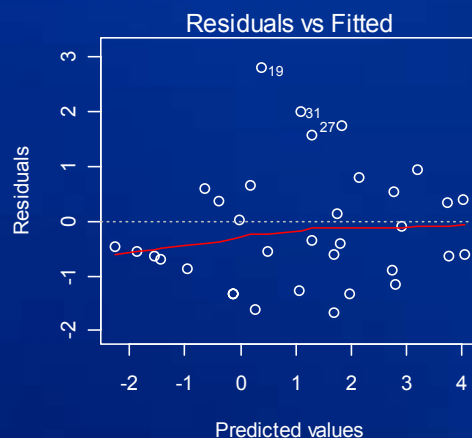
$$AIC = 2[-\ell(\hat{\beta}) + p] \vee 2[-\ell(\hat{\beta}) + p + 1]$$

Sestavení modelu – obecně + příklad

- Dostatečná expozice ve všech kategoriích.
- Rozumné chování reziduí (viz výše).
- Konzistence v čase – koeficienty dané veličiny by neměly jeden rok vypadat úplně jinak než jiný rok.
- Určení parametru příslušného dané proměnné by mělo být přiměřeně přesné.
- Devianční testy modelu.
 - Ilustrační příklad: Poškození způsobené vlnami na předních lodí
 - Typ lodi (TS): A, B, C, D, E
 - Rok stavby (YC): 1960-64, 1965-69, 1970-74, 1975-79
 - Období provozu (OP): 1960-74, 1975-79
 - Vlajka pod kterou loď pluje (FL): 10 kategorií
 - Celková doba provozu v měsících – expozice = offset
 - Počet událostí



Ukázka NL – poškození lodí



Poissonovo rozdělení s nad-rozptylem

Nulový model: YC, OP

$$D_0 = 62,365; df_0 = 29; \phi_0 = 2,85$$

1. model: ST, YC, OP

$$D_1 = 38,695; df_1 = 25; \phi_1 = 1,69$$

$$3.5 = F \sim F_{4,25} \rightarrow p = 0,02$$

ST – signifikantní - zahrnout

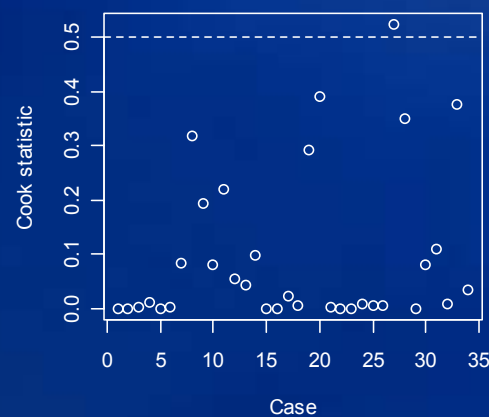
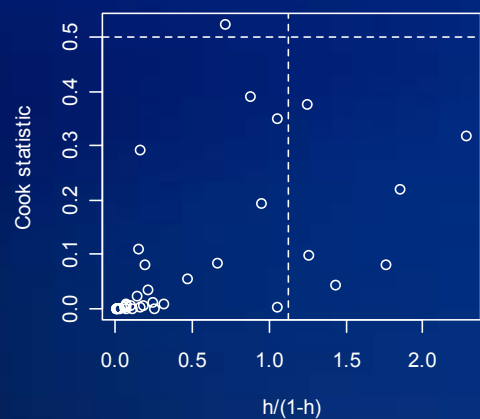
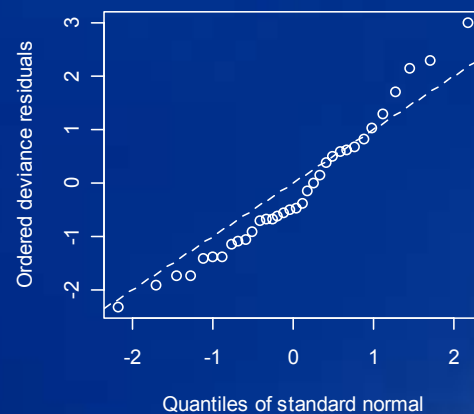
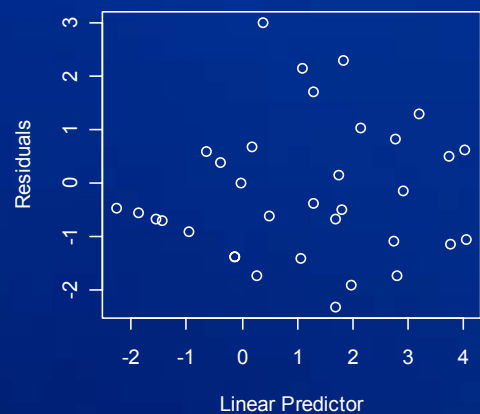
2. model: ST, YC, OP, FL

$$D_2 = 20,965; df_2 = 16; \phi_2 = 1,09$$

$$1.8 = F \sim F_{9,16} \rightarrow p = 0,14$$

FL – nesignifikantní - vyloučit

Ukázka NL – poškození lodí



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-6.40590	0.28276	-22.655	< 2e-16
STB	-0.54334	0.23094	-2.353	0.02681
STC	-0.68740	0.42789	-1.607	0.12072
STD	-0.07596	0.37787	-0.201	0.84230
STE	0.32558	0.30674	1.061	0.29864
YC1965-69	0.69714	0.19459	3.583	0.00143
YC1970-74	0.81843	0.22077	3.707	0.00105
YC1975-79	0.45343	0.30321	1.495	0.14733
OP1975-79	0.38447	0.15380	2.500	0.01935

absolutní člen: STA, YC1960-64, OP1960-74

Ukázka – Analýza storen – vliv interakcí 1

Ukončení životního pojištění během 2 let od sjednání

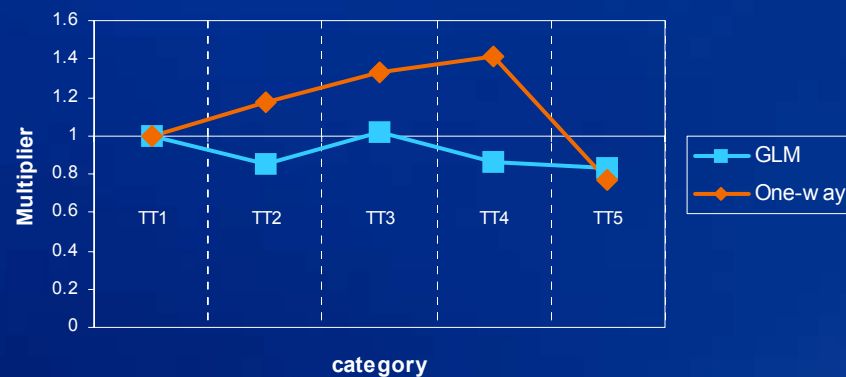
Databáze: 10 let zkušeností – 851 955 smluv.

Zkoumané faktory a jejich kategorizace, tak jak byly uvedeny klientem při uzavření smlouvy:

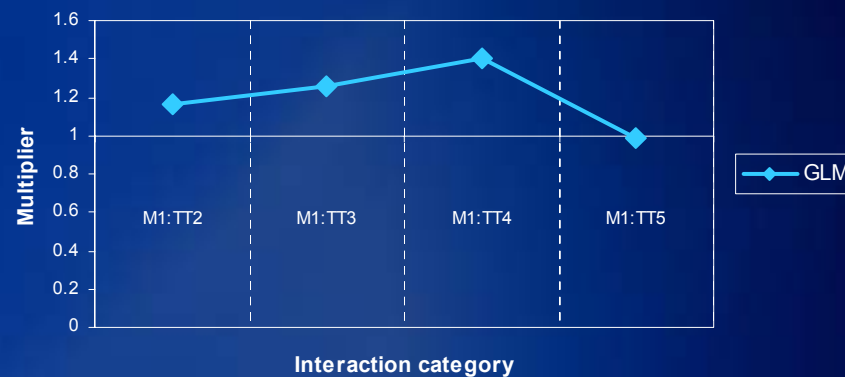
- pohlaví (*Muž, Žena*)
- věk (*A1: 18-29, A2: 30-39, A3: 40-49, A4: 50-59, A5: 60+*),
- manželský stav (*M0: svobodný/rozvedený, M1: ženatý/vdaná*)
- děti (*C0: žádné, C1: 1 a více*)
- výdělek (*tisíce Kč: E1: <10; E2: 10-20; E3: 20-30; E4:30+*)
- sjednané pojištění (*T1: smrt bez podílu na zisku, T2: smrt podíl na zisku, T3: dožití bez podílu na zisku, T4: dožití s podílem na zisku, T5: unit link*)
- pojistná částka (*tisíce Kč I1: 0-500, I2: 500-1000, I3: 1000+*)
- distribuce (*O1, O2, O3, O4, O5*)
- rok sjednání (*kalendářní rok sjednání: Y1: 96-97, Y2: 98-99, Y3: 00-01, Y4: 02-03, Y5: 04-05*)
- sídlo (*obyvatelé: R1: <10000, R2: 10000-50000, R3: 50000-100000, R4:>100000*)

Ukázka – Analýza storen – vliv interakcí 2

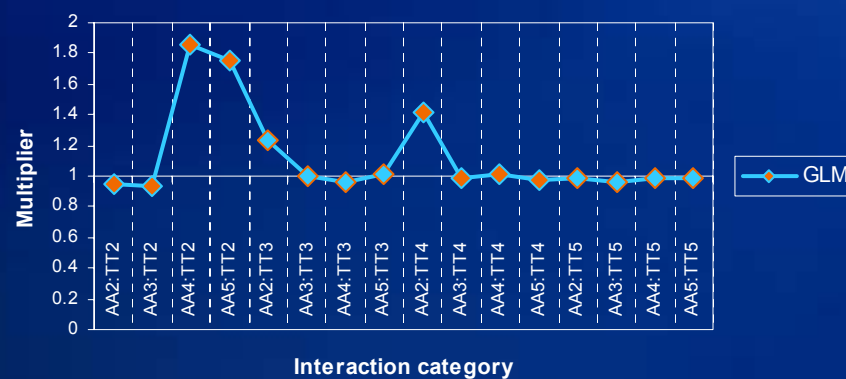
Insurance type



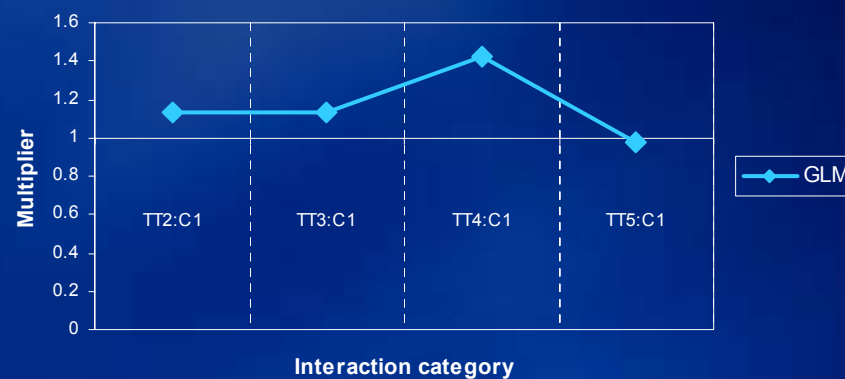
Marriage:Type



Age:Type



Type:Children



Typické modely, použití

Modelovaná veličina Y	Škodní frekvence	Počet škod	Výše škod	Pravděpodobnost - storna
Link	$\ln(x)$	$\ln(x)$	$\ln(x)$	$\ln\left(\frac{x}{1-x}\right)$
Rozdělení	Poissonovo	Poissonovo	gamma	binomické
Škálovací parametr	1	1	odhad	$1/m$
Varianční funkce	x	x	x^2	$x(1-x)$
Apriorní váhy	expozice	1	počet škod / 1	1
Offset	0	$\ln(\text{expozice})$	0	0

Typické modely, použití

- Modelování přežívání, graduace – vyhlazování změřených pravděpodobností.
- Modelování intenzit přechodů mezi stavy ve zdravotním pojištění.
- Fitování rozdělení výše škod v neživotním pojištění.
- Klasifikace rizik – modelování nad-úmrtnosti, storen, ...
- Stanovení pojistného
- Modelování IBNR

Literatura

- P. McCullagh, J. A. Nelder *Generalized Linear Models*, Chapman&Hall/CRC 1997
- J. J. Faraway *Linear Models with R*, Chapman&Hall/CRC 2005
- J. J. Faraway *Extending the linear model with R: Generalized linear, Mixed Effects and Nonparametric Regression Models*, Chapman&Hall/CRC 2006
- S. N. Wood *Generalized Additive Models: An Introduction with R*, Chapman&Hall/CRC 2006
- J. Anděl, *Statistické metody*, MatfyzPress, Praha 2003
- D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi *A Practitioner's Guide to Generalized Linear Models*, CAS 2005
- Zhijin Wu, *BC0251 Generalized Linear Models*, <http://www.stat.brown.edu/~zwu/>
- S. Haberman, A. E. Renshaw, *Generalized Linear Models and Actuarial Science* The Statistician, Vol. 45, No. 4. (1996), pp. 407-436
- Nelder, J.A. & Wedderburn, R.W.M.; J. R. Statist. Soc. A, 135 (1972), 370-384; *Generalized linear models*
- Jorgensen, B.; J. R. Statist. Soc. B, 49 (1987), 2, 127-162; *Exponential Dispersion Models*
- Renshaw, A. E. and Haberman, S.J.; Inst. Act.; 113 (1986), 459-497 *Statistical analysis of life assurance lapses*
- Wright, T.S.J. Inst. Act., 117 (1990), 677-731; *A stochastic method for claims reserving in general insurance*
- Renshaw, A. E. and Haberman, S.J.; Insur. Math. Econ. 17 (1995), 1-17; *On the graduations associated with a multiple state model for permanent health insurance*
- The R Development Core Team, *R: A Language and Environment for Statistical Computing*, 1999 – 2003
- <http://www.statsoft.com/textbook/stathome.html>
- <http://www.wikipedia.org>
- <http://mathworld.wolfram.com>



Miroslav Šimurda

KPMG Česká republika, s.r.o.

+420 222 123 189

msimurda@kpmg.cz

www.kpmg.cz

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation. Informace zde obsažené jsou obecného charakteru a nejsou určeny k řešení situace konkrétní osoby či subjektu. Ačkoliv se snažíme zajistit, aby poskytované informace byly přesné a aktuální, nelze zaručit, že budou odpovídat skutečnosti k datu, ke kterému jsou doručeny, či že budou platné i v budoucnosti. Bez důkladného prošetření konkrétní situace a řádné odborné konzultace by neměla na základě těchto informací být činěna žádná opatření.