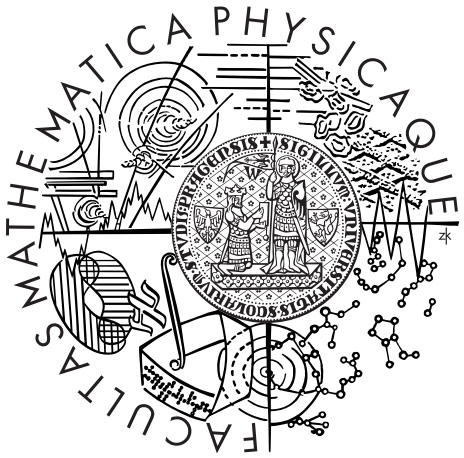
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravděpodobnosti a matematické statistiky



Zobecněné lineární modely v pojišťovnictví

RNDr. Martin Branda, Ph.D.

Zpracováno v rámci projektu

Fondu pro podporu vzdělávání v pojišťovnictví

Praha 2013

Obsah

1 Úvod

Zobecněné lineární modely nacházejí široké uplatnění v pojišťovnictví, například při sazbování a rezervování v neživotním pojištění nebo při podpoře obchodu[1]. Logistická regrese se využívá k modelování pravděpodobnosti sledovaného jevu, např. pojistné události, storna smlouvy, nákupu (při)pojištění. Pomocí Poissonovské regrese můžeme modelovat očekávaný počet pojistných událostí během určitého období, resp. škodní frekvenci. Gamma regrese je pak vhodná pro odhad očekávané výše vyplacených škod z pojistné události, doby do storna, doby do (následující) pojistné události apod.

Zobecněné lineární modely by měly patřit mezi základní znalosti absolventa magisterského oboru Finanční a pojistná matematika na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze. Text předkládá základní poznatky, které jsou nutné pro pochopení zobecněných lineárních modelů, především s ohledem na volbu vhodného modelu, metody a výpočetní náročnost odhadu parametrů a interpretaci výsledků, vše s přihlédnutím na četné aplikace v pojišťovnictví zmíněné výše. Cílem textu tedy není poskytnout hluboký pohled do teorie, ta je často pouze naznačena s příslušným odkazem do odborné literatury.

Před vznikem předmětu pro výše zmíněný obor bude přednáška o zobecněných lineárních modelech součástí vzdělávací části cyklu v rámci Semináře z aktuárských věd[2], který se koná již tradičně od 8:10 každý pátek v semestru na matfyzu[3].

Na závěr poznamenejme, že text bude dále rozšiřován a aktuální verze bude dostupná na webu autora, který bude vděčný za jakékoliv náměty a připomínky.

[1] Modely mohou sloužit k navýšení prodeje produktů (up-selling) nebo prodeji dalších produktů stávajícím zákazníkům pojišťovny (cross-selling).

[2] Program je dostupný na www.actuaria.cz

[3] Na adrese Sokolovská 83, Praha 8, v učebně K1.

2 Data

Zobecněné lineární modely bývají budovány na datech, které jsou získány z rozsáhlých databází pojišťovny. Využívají se například v **data-miningu**, který se zabývá získáváním netriviálních skrytých a potenciálně užitečných informací z dat. Proto se nejprve budeme věnovat právě datům a zaměříme se na úpravu hrubých dat do podoby vhodné pro práci s regresními modely.

V tomto textu využíváme následující značení:

- **Závisle proměnná** (odezva): $\mathbf{Y}^T = (Y_1, \ldots, Y_n)$, např. počet pojistných událostí v daném období, výše vyplacené škody, příznak storna.
- **Nezávisle proměnné** (prediktory, regresory): označíme-li i-té pozorování nezávisle proměnných $\mathbf{x}_i^T = (X_{i1}, \ldots, X_{im})$, můžeme n pozorování seřadit do matice

$$\mathbf{X} = \begin{pmatrix} X_{11} & \ldots, & X_{1m} \\ \vdots & & \vdots \\ X_{n1} & \ldots, & X_{nm} \end{pmatrix}$$

  Předpokládáme, že matice má plnou sloupcovou hodnost. Proměnné dále klasifikujeme na

  - **kvantitativní** - např. věk, počet aktivních smluv, počet najetých kilometrů a další. Často jsou kategorizovány kvůli nevhodnému rozdělení, odlehlým pozorováním nebo nelinearitě vztahu mezi nimi a závisle proměnnou.
  - **kvalitativní** (kategoriální) - kódovány pomocí 0-1 "dummy" proměnných, např. pohlaví, region (kraj, okres) a další.

V databázi máme, případně nad databází vytvoříme, data ve struktuře uvedené v následující tabulce, kde každý řádek představuje jednu pojistnou smlouvu v určitém období, například jednom roce. Závisle proměnnou je pro nás počet škod na smlouvě za jeden rok. Jako vysvětlující proměnné slouží například pohlaví pojistníka, počet obyvatel žijících v místě bydliště pojistníka a jeho věk k datu počátku období:

Y	Data			
Počet škod	**Pohlaví**	**Počet obyvatel**	**Věk**	
			(v letech)	
2	muž	15 423	21	
0	muž	1 205 321	44	
1	žena	20 893	35	
0	žena	580	51	
⋮	⋮	⋮	⋮	⋱

Z kategoriálních proměnných obvykle vytváříme binární (0-1, dummy) proměnné, kde každá proměnná odpovídá jedné kategorii původní kategoriální proměnné. Softwarové balíky jsou obvykle schopny vytvořit dummy proměnné automaticky při označení původních proměnných jako kategoriálních. Velikost místa bydliště je kategorizována na základě jednoduchého pravidla. V reálných aplikacích se často využívá

optimální kategorizace[4] vytvořená pomocí vhodných metod.

Y	Data						
Počet škod	**Pohlaví**		**Region**			**Věk**	
	žena	muž	velká města	malá města	venkov	(v letech)	
2	0	1	0	1	0	21	
0	0	1	1	0	0	44	
1	1	0	0	1	0	35	
0	1	0	0	0	1	51	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Je-li přidán absolutní člen, je z každé kategoriální proměnné odebrána jedna dummy proměnná, jíž odpovídající kategorie slouží potom jako referenční. Výsledek je zobrazen v tabulce, která představuje data vhodná pro odhad regresního modelu.

Y	**X**					
Počet škod	**Abs.člen**	**Pohlaví**	**Region**		**Věk**	
		žena	velká města	malá města	(v letech)	
2	1	0	0	1	21	
0	1	0	1	0	44	
1	1	1	0	1	35	
0	1	1	0	0	51	
⋮	⋮	⋮	⋮	⋮	⋮	⋱

2.1 Chybějící a chybné hodnoty v datech

Při práci s reálnými daty je potřeba věnovat pozornost chybějícím a chybným hodnotám. Příklad chybných hodnot je uveden v následující tabulce:

Y	Data			
Počet škod	**Pohlaví**	**Počet obyvatel**	**Věk**	
			(v letech)	
2	muž	15 423	21	
-1	muž	1 205 321	44	
1	žež	20 893	138	
0	žena	-112	51	
⋮	⋮	⋮	⋮	⋱

Chybějící hodnoty jsou obvykle v náhledu dat reprezentovány speciálním znakem[5], případně prázdným místem[6]:

[4] Optimal binning.
[5] Zde u numerických proměnných tečkou.
[6] Obvykle u textových řetězců.

Y	Data			
Počet škod	**Pohlaví**	**Počet obyvatel**	**Věk**	
			(v letech)	
2	muž	15 423	21	
·	muž	1 205 321	44	
1		20 893	·	
0	žena	·	51	
⋮	⋮	⋮	⋮	⋱

Bez ošetření by po kategorizaci vznikla následující data:

Y	**X**					
Počet škod	**Abs.člen**	**Pohlaví**	**Region**		**Věk**	
		žena	velká města	malá města	(v letech)	
2	1	0	0	1	21	
·	1	0	1	0	44	
1	1	·	0	1	·	
0	1	1	·	·	51	
⋮	⋮	⋮	⋮	⋮	⋮	⋱

S chybějícími a chybnými hodnotami pracujeme dle jejich výskytu:

- **V závisle proměnné** - pozorování obvykle vypadnou z odhadu modelu, je však možné dopočítat očekávanou odezvu.
- **V nezávisle proměnných**
  - Kvantitativní - nahrazení chybějících hodnot, např. průměrem nebo pomocí sofistikovanějších metod[7].
  - Kvalitativní (kategoriální) - vytvoření speciální kategorie.

Použitím uvedených metod jsme zavedly speciální kategorii pro chybějící informaci o počtu obyvatel a nahradili chybějící věk průměrným věkem klientů:

Y	**X**						
Počet škod	**Abs.člen**	**Pohlaví**	**Region**			**Věk**	
		žena	velká města	malá města	missing	(v letech)	
2	1	0	0	1	0	21	
·	1	0	1	0	0	44	
1	1	1	0	1	0	38.43	
0	1	1	0	0	1	51	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

[7] Klasifikační nebo regresní stromy apod.

3 Lineární regrese

V této části velice stručně shrneme základní poznatky o modelu lineární regrese, který zobecněné lineární modely zahrnují jako speciální případ. Lineární regrese však obvykle nebývá vhodná pro aplikace v pojišťovnictví. Více o modelu je možné se dočíst ve Zvára (2008).

Model lineární regrese můžeme zapsat ve tvaru

$$Y_i = \sum_{j=1}^{m} X_{ij}\beta_j + \varepsilon_i,\ i = 1, \ldots, n,$$

kde předpokládáme

1. chyby (disturbance) ε_i jsou nezávislé,
2. $\mathbb{E}[\varepsilon_i] = 0$,
3. reziduální rozptyl $var\varepsilon_i = \sigma^2 > 0$.

Často se využívá maticový zápis pomocí symbolů zavedených v předešlé části

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

kde $\beta^T = (\beta_1, \ldots, \beta_m)$ a $\epsilon^T = (\varepsilon_1, \ldots, \varepsilon_n)$.

Odhad parametrů β probíhá nejčastěji metodou nejmenších čtverců (LS), když za předpokladu plné sloupcové hodnosti $\mathbf{X}$ dostáváme

$$\begin{aligned}
\hat{\beta} &= \arg\min_{\beta\in\mathbb{R}^m} \sum_{i=1}^{n} (Y_i - \sum_{j=1}^{m} X_{ij}\beta_j)^2 \\
&= \arg\min_{\beta\in\mathbb{R}^m} (\mathbf{Y} - \mathbf{X}^T\beta)^T(\mathbf{Y} - \mathbf{X}^T\beta) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y}).
\end{aligned}$$

Odhad též splňuje soustavu normálních rovnic

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$$

a je nestranný, tj. $\mathbb{E}\hat{\beta} = \beta$, s rozptylem $var\hat{\beta} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Vyrovnané hodnoty spočteme pomocí vztahu

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

a rezidua jako

$$\mathbf{u} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y},$$

kde $\mathbf{I}$ je jednotková matice rozměrů $n \times n$. Nestranný odhad reziduálního rozptylu σ^2 získáme poté pomocí vztahu:

$$\hat{\sigma}^2 = \frac{\mathbb{E}[\mathbf{u}^T\mathbf{u}]}{n - m}.$$

Za předpokladu normality $\varepsilon_i \sim N(0, \sigma^2)$ navíc platí $\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 I)$
a $\hat{\beta} \sim N_m(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

3.1 Aitkenův model a vážené nejmenší čtverce

V této části stručně popíšeme model lineární regrese s porušeným předpokladem na rozptyl, tzv. Aitkenův model. Nechť pro rozptyl chyb platí

$$var\ \epsilon = \mathbf{W}\sigma^2,$$

kde $\mathbf{W}$ je obecná pozitivně definitní matice, tj. chyby nemusí být nezávislé se stejným rozptylem. Pomocí rozkladu $\mathbf{W}^{-1} = \mathbf{C}^T\mathbf{C}$, kde $\mathbf{C}$ je regulární odmocninová matice, přepíšeme model do tvaru

$$\mathbf{CY} = \mathbf{CX}\beta + \mathbf{C}\epsilon,$$

který již odpovídá předešlému modelu lineární regrese s nezávislými chybami. Odhad β je v tomto případě řešením soustavy normálních rovnic

$$\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}\beta = \mathbf{X}^T\mathbf{W}^{-1}\mathbf{Y}.$$

Tedy dostaneme odhad parametrů

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{Y},$$

kde $\hat{\beta} \sim (\beta, \sigma^2(\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1})$. S analogickým vztahem se setkáme při odhadu parametrů v zobecněných lineárních modelech. Další odhady a statistiky získáme analogicky jako v modelu bez porušeného předpokladu na rozptyl chyb.

4 Zobecněné lineární modely

Zobecněné lineární modely jsou definovány pomocí tří stavebních elementů:

1) Závisle proměnná Y_i má **rozdělení z exponenciální rodiny s hustotou**[8]

$$f(y;\theta_i,\phi) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y,\phi)\right\},\ y \in \mathbb{R} \tag{1}$$

pro známé funkce a, b, c, neznámý kanonický parametr θ_i závisející na pozorování a neznámý disperzní parametr ϕ společný pro celý model.

2) **Lineární prediktor** vzniká jako lineární kombinace

$$\eta_i = \sum_{j=1}^{m} X_{ij}\beta_j = \mathbf{x}_i^T\beta, \tag{2}$$

kde β_j jsou neznámé parametry a X_{ij} jsou známé hodnoty regresorů.

3) Striktně monotónní a dvakrát diferencovatelná **linková funkce** propojující střední hodnotu závisle proměnné a lineární prediktor:

$$\mathbb{E}[Y_i] = \mu_i = g^{-1}(\eta_i). \tag{3}$$

Při budování modelu a odvozování teoretických výsledků se využívají následující předpoklady:

- Rozdělení Y_i závisí na $\mathbf{x}_i$.
- Pozorování $(Y_i, \mathbf{x}_i)$ jsou nezávislé náhodné vektory nebo Y_i jsou nezávislé náhodné veličiny a $\mathbf{x}_i$ jsou měřené konstanty. My budeme nadále uvažovat druhý případ.

4.1 Rodina exponenciálních rozdělení

Obecný tvar hustoty rozdělení z exponenciální rodiny můžeme zapsat jako

$$f(z;\xi,\phi) = \exp\left\{\frac{T(z)A(\xi) + B(\xi)}{a(\phi)} + C(z,\phi)\right\}$$

s disperzním parametrem ϕ a parametrem polohy ξ. **Kanonický tvar** hustoty dostaneme, položíme-li $y = T(z)$, $\theta = A(\xi)$

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

kde $\theta \in \mathbb{R}$, $a(\phi) \in (0,\infty)$ a $a : \mathbb{R}_+ \to \mathbb{R}_+$. Často se využívá následující přepis pomocí $a(\phi) = \varphi \in (0,\infty)$

$$f(y;\theta,\varphi) = \exp\left\{\frac{y\theta - b(\theta)}{\varphi} + c(y,\varphi)\right\}.$$

[8] Uvádíme jednu z parametrizací, další popíšeme dále v textu.

Pozn. Při studiu různých zdrojů je nutné věnovat pozornost použité parametrizaci. V literatuře se běžně objevují různé parametrizace, například pro známé funkce $a, b, \tilde{c}$, a neznámé parametry θ, ϕ:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}\right\} \cdot \tilde{c}(y,\phi).$$

Tento tvar využívají v knize de Jong a Heller (2008), my jej nebudeme dále uvažovat.

Pro náhodnou veličinu Y patřící do rodiny exponenciálních rozdělení platí: Je-li b dvakrát spojitě diferencovatelná, potom

$$\begin{aligned} \mathbb{E}[Y] &= b'(\theta), \\ var(Y) &= a(\phi)b''(\theta) = \varphi b''(\theta). \end{aligned}$$

Pro parciální derivaci hustoty podle parametru θ totiž platí

$$\frac{\partial f(y;\theta,\phi)}{\partial \theta} = f(y;\theta,\phi)\frac{y - b'(\theta)}{a(\phi)}$$

integrací obou stran podle y dostaneme (za předpokladu, že je možné zaměnit pořadí derivace a integrálu)

$$\begin{aligned} 0 &= \int \frac{\partial f(y;\theta,\phi)}{\partial \theta} dy \\ &= \frac{\partial}{\partial \theta} \int f(y;\theta,\phi) dy \\ &= \frac{\mathbb{E}Y - b'(\theta)}{a(\phi)}. \end{aligned}$$

Pro druhou parciální derivaci hustoty platí

$$\frac{\partial^2 f(y;\theta,\phi)}{\partial \theta^2} = f(y;\theta,\phi)\left(\frac{y - b'(\theta)}{a(\phi)}\right)^2 - \frac{b''(\theta)}{a(\phi)}$$

integrací obou stran podle y dostaneme (za předpokladu, že je možné zaměnit pořadí derivace a integrálu)

$$\begin{aligned} 0 &= \int \frac{\partial^2 f(y;\theta,\phi)}{\partial \theta^2} dy \\ &= \frac{\partial^2}{\partial \theta^2} \int f(y;\theta,\phi) dy \\ &= \frac{\mathbb{E}[(Y - b'(\theta))^2]}{(a(\phi))^2} - \frac{b''(\theta)}{a(\phi)}. \end{aligned}$$

Obecný důkaz je možné provést pomocí momentové vytvořující funkce.

Pomocí **rozptylové funkce** definované jako

$$V(\mu) = b''[(b')^{-1}(\mu)]$$

můžeme vztah pro rozptyl přepsat jako

$$var(Y) = a(\phi)V(\mu) = \varphi V(\mu).$$

Rozptylová funkce tedy vyjadřuje vztah mezi střední hodnotou a rozptylem. Zároveň jednoznačně identifikuje rozdělení z exponenciální rodiny.

Rodina exponenciálních rozdělení zahrnuje:

- Normální, gamma, inverzní Gaussovo, Poissonovo, alternativní,
- Chí-kvadrát, exponenciální, binomické, geometrické, multinomické, beta,
- se známým parametrem: Weibullovo, negativně binomické, Paretovo.

V následujících částech podrobně probereme jednotlivé členy z první uvedené skupiny rozdělení.

4.1.1 Normální rozdělení

Značíme $Y \sim N(\mu, \sigma^2)$: Pro $y \in \mathbb{R}$ můžeme hustotu vyjádřit jako

$$\begin{aligned} f(y;\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{ \frac{y\mu - \overbrace{\mu^2/2}^{b(\theta)}}{\underbrace{\sigma^2}_{\varphi}} \underbrace{-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)}_{c(y,\varphi)} \right\}, \end{aligned}$$

kde $\theta = \mu$, $b(\theta) = \mu^2/2$ a $\varphi = \sigma^2$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = \mu$,
- $var(Y) = \varphi b''(\theta) = \sigma^2$, tj. rozptyl nezávisí na střední hodnotě $V(\mu) = 1$ (jako jediné rozdělení z exponenciální rozdělení).

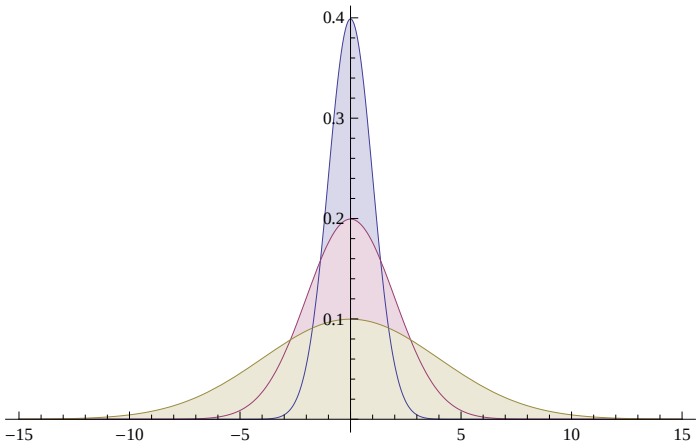
4.1.2 Gamma rozdělení

Značíme $Y \sim \Gamma(a, p)$: Pro $0 < y < \infty$ můžeme hustotu vyjádřit jako

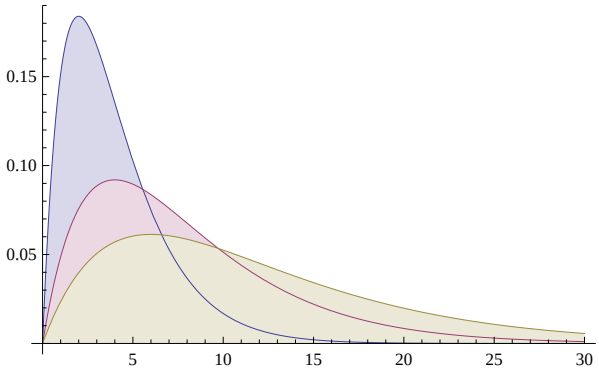
$$\begin{aligned} f(y;a,p) &= \frac{a^p}{\Gamma(p)} y^{p-1} \exp\{-ay\} \\ &= \exp\{(p-1)\log y - ay + p\log a - \log\Gamma(p)\} \\ &= \exp\left\{ \frac{y(-a/p) + \log a/p}{1/p} \right. \\ &\quad \left. + p\log p - \log\Gamma(p) + (p-1)\log y \right\} \end{aligned}$$

kde $\theta = -a/p$, $\varphi = 1/p$, $b(\theta) = -\log(-\theta)$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = -1/\theta = p/a = \mu$,



Obrázek 1: Hustoty $N(0,1),\ N(0,2),\ N(0,4)$



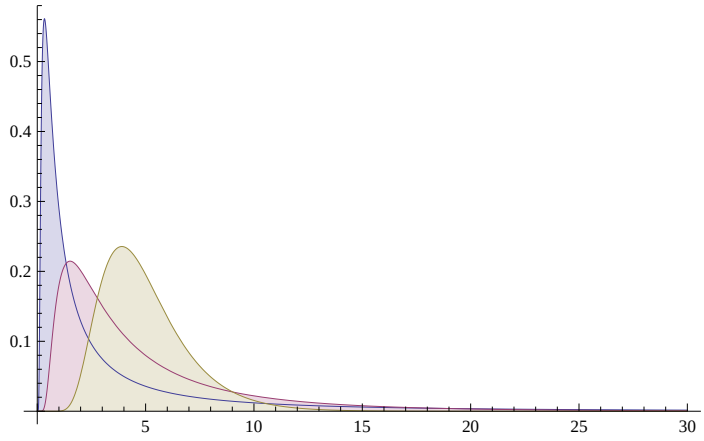
Obrázek 2: Hustoty $\Gamma(2,2),\ \Gamma(4,2),\ \Gamma(6,2)$

- $var(Y) = \varphi b''(\theta) = p/a^2 = \mu^2/p$, tj. rozptyl závisí na střední hodnotě $V(\mu) = \mu^2$.

Parametrizace v SASu je odlišná $Y \sim \Gamma(\mu, \nu)$: Pro $0 < y < \infty$ můžeme hustotu vyjádřit jako

$$f(y;\mu,\nu) = \frac{1}{\Gamma(\nu)y}\left(\frac{y\nu}{\mu}\right)^{\nu} \exp\left\{-\frac{y\nu}{\mu}\right\},$$

kde $a = \nu/\mu$ a $p = \nu$, $\varphi = \nu^{-1}$, $var(Y) = \mu^2/\nu$



Obrázek 3: Hustoty $IG(5,1),\ IG(5,5),\ IG(5,30)$

4.1.3 Inverzní Gaussovo rozdělení

Značíme $Y \sim IG(\mu, \lambda)$: Pro $0 < y < \infty$ můžeme hustotu vyjádřit jako

$$\begin{aligned} f(y;\mu,\lambda) &= \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\} \\ &= \exp\left\{\frac{-\lambda y^2}{2\mu^2 y} + \frac{\lambda\mu y}{\mu^2 y} - \frac{\lambda\mu^2}{2\mu^2 y} + \frac{1}{2}\log\lambda - \frac{1}{2}\log 2\pi y^3\right\} \\ &= \exp\left\{\frac{y/(-2\mu^2) + 1/\mu}{1/\lambda} - \frac{\lambda}{2y} + \frac{1}{2}\log\lambda - \frac{1}{2}\log 2\pi y^3\right\}, \end{aligned}$$

kde $\theta = -1/(2\mu^2)$, $b(\theta) = -\sqrt{-2\theta}$ a $\varphi = 1/\lambda$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = 1/\sqrt{-2\theta} = (-2\theta)^{-1/2} = \mu$,
- $var(Y) = \varphi b''(\theta) = (-2\theta)^{-3/2}/\lambda = \mu^3/\lambda$, tj. rozptyl závisí na střední hodnotě $V(\mu) = \mu^3$.

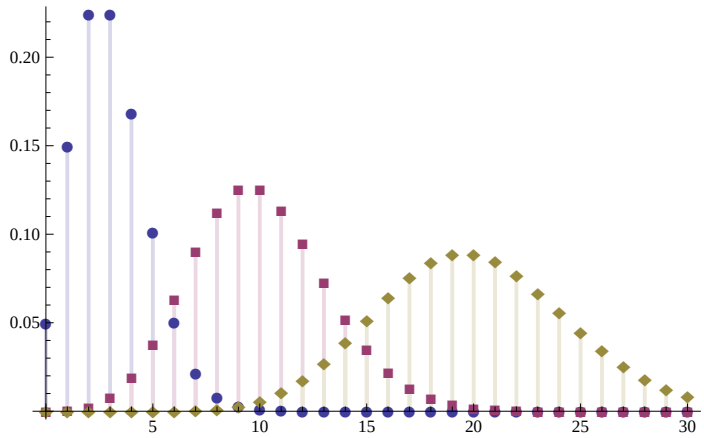
4.1.4 Poissonovo rozdělení

Značíme $Y \sim Po(\lambda)$: Pro $y = 0, 1, 2, \ldots$ můžeme hustotu vyjádřit jako

$$\begin{aligned} f(y;\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \exp\left\{\frac{y\log\lambda - \lambda}{1} - \log y!\right\}, \end{aligned}$$

kde $\theta = \log\lambda$, $b(\theta) = e^\theta$ a $\varphi = 1$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = e^\theta = \lambda$,
- $var(Y) = \varphi b''(\theta) = e^\theta = \lambda$, tj. rozptyl závisí na střední hodnotě $V(\mu) = \mu$.



Obrázek 4: Hustoty Po(3), Po(10), Po(20)

4.1.5 Alternativní rozdělení

Značíme $Y \sim Alt(p)$: Pro $y \in \{0, 1\}$ můžeme hustotu vyjádřit jako

$$\begin{aligned} f(y;p) &= p^y(1-p)^{1-y} \\ &= \exp\left\{\frac{y \log \frac{p}{1-p} + \log(1-p)}{1} + 0\right\}, \end{aligned}$$

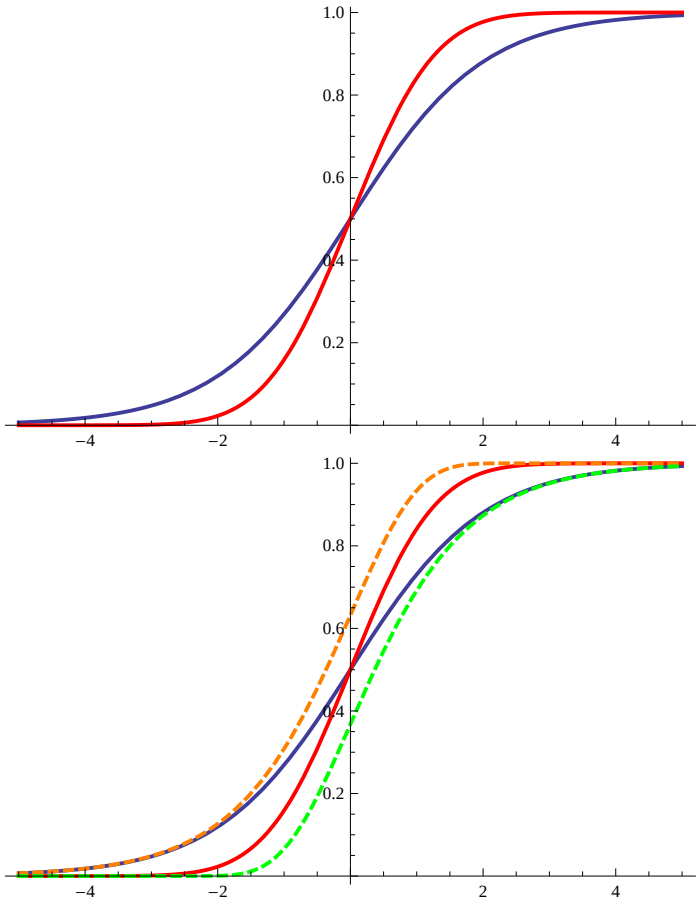
kde $\theta = \log \frac{p}{1-p}$, $b(\theta) = \log(1 + e^\theta)$ a $\varphi = 1$. Potom dostaneme

- $\mathbb{E}Y = b'(\theta) = \frac{e^\theta}{1+e^\theta} = p$,
- $var(Y) = \varphi b''(\theta) = p(1-p)$, tj. rozptyl závisí na střední hodnotě $V(\mu) = \mu(1-\mu)$.

4.2 Linkové funkce

V této části uvádíme nejčastěji používané linkové funkce:

- identita: $g(\mu) = \mu$
- logaritmus: $g(\mu) = \log(\mu)$
- logit: $g(\mu) = \log(\mu/(1-\mu))$
- probit: $g(\mu) = \Phi^{-1}(\mu)$, kde Φ je distribuční funkce standardního normálního rozdělení
- log-log: $g(\mu) = -\log(-\log(\mu))$
- komplementární log-log: $g(\mu) = \log(-\log(1-\mu))$
- mocninný: $g(\mu) = \mu^p$ pro $p \neq 0$ (pro $p = 0$ logaritmický)



Obrázek 5: Porovnání inverzí linků: Logit (modrá), Probit (červená), kompl. (oranžová), log-log (zelená)

Důležitým pojmem především pro teorii je **kanonický link**, který splňuje $g(\mu) = \theta$, tedy musí platit $g(\mu) = (b')^{-1}(\mu)$ a také

$$g'(\mu) = \frac{1}{V(\mu)}.$$

V části o odhadu parametrů uvedeme zjednodušení vztahů při užití kanonického linku. Současné softwarové balíky však umí pracovat s libovolným linkem bez omezení na kanonický.

4.3 Přehled rozdělení

V následující tabulce uvádíme přehled rozdělení z exponenciální rodiny spolu s jejich hlavními charakteristikami:

Rozdělení	Hustota	Disperze φ	Kanonický link $\theta(\mu)$	Střední hodnota $\mu(\theta)$	Rozptylová funkce $V(\mu)$
$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	σ^2	μ	θ	1
$Po(\mu)$	$\frac{\mu^y e^{-\mu}}{y!}$	1	$\log(\mu)$	e^θ	μ
$\Gamma(\mu, \nu)$	$\frac{1}{\Gamma(\nu) y} \left(\frac{y\nu}{\mu}\right)^\nu e^{-\frac{y\nu}{\mu}}$	$\frac{1}{\nu}$	$-\frac{1}{\mu}$	$-\frac{1}{\theta}$	μ^2
$IG(\mu, \lambda)$	$\sqrt{\frac{\lambda}{2\pi y^3}} e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$	$\frac{1}{\lambda}$	$-\frac{1}{2\mu^2}$	$\frac{1}{\sqrt{-2\theta}}$	μ^3
$Alt(\mu)$	$\mu^y (1-\mu)^{1-y}$	1	$\log \frac{\mu}{1-\mu}$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$

4.4 Srovnání regresních modelů

V této části krátce srovnáme model lineární regrese se zobecněným lineárním modelem.

	Lineární regrese	**Zobecněný lineární model**
Rozdělení:	$Y_i \sim N(\mu_i, \sigma^2)$	$Y_i \sim \mathcal{EF}(\theta_i, \varphi)$
Závislost:	$\mathbb{E}[Y_i] = \mathbf{x}_i^T \beta$	$\mathbb{E}[Y_i] = g^{-1}(\mathbf{x}_i^T \beta)$
Rozptyl:	$varY_i = \sigma^2$	$varY_i = \varphi V(\mu_i)$

Za předpokladu normality a identického linku, tj. $g(\mu) = \mu$, dostaneme lineární regresi jako speciální případ zobecněného lineárního modelu.

4.5 Vážení

Zobecněné lineární modely umožňují dva způsoby vážení, které jsou vhodné pro různé situace.

4.5.1 Offset

Offset je člen v lineárním prediktoru s pevně daným koeficientem. V pojišťovnictví obvykle slouží jako korekce modelu s ohledem na expozici v riziku (počet rizik, délka platnosti smlouvy apod.). Například pro expozici n_i $i-$tého řádku a logaritmický link položíme lineární prediktor roven

$$\eta_i = \ln n_i + \mathbf{x}_i' \beta,$$

kde $\ln n_i$ slouží jako offset. Dostaneme tedy

$$\mu_i = e^{\eta_i} = n_i \cdot e^{\mathbf{x}_i' \beta}.$$

4.5.2 Váhy pozorování

Při zahrnutí apriorních vah pozorování w, kdy v parametrizaci pokládáme $a(\phi) = \varphi/w$, platí

$$\begin{aligned} \mathbb{E}[Y] &= b'(\theta), \\ var(Y) &= \frac{a(\phi)b''(\theta)}{w} = \frac{\varphi b''(\theta)}{w}. \end{aligned}$$

Pomocí rozptylové funkce můžeme vztah pro rozptyl přepsat

$$var(Y) = \frac{a(\phi)V(\mu)}{w} = \frac{\varphi V(\mu)}{w}.$$

Tyto váhy jsou využívány při modelování průměrné výše škody (w = počet škod) nebo škodní frekvence (w = délka expozice).

4.6 Odhad parametrů

4.6.1 Metoda maximální věrohodnosti

Nadále předpokládáme, že má závisle proměnná Y_i rozdělení s hustotou $f(y; \theta_i, \varphi)$, která závisí na prediktorech a neznámých koeficientech β skrze vztah $\theta_i = (b')^{-1}(g^{-1}(\mathbf{x}_i'\beta))$. Věrohodnostní funkce je pak pro nezávislá pozorování definována jako

$$L(\mathbf{Y}; \beta, \varphi) = \prod_{i=1}^{n} f(Y_i; \theta_i, \varphi)$$

Obvykle pracujeme s logaritmickou věrohodnostní funkcí

$$l(\mathbf{Y}; \beta, \varphi) = \sum_{i=1}^{n} \log(f(Y_i; \theta_i, \varphi)),$$

kterou je možné díky obecnému tvaru hustoty dále přepsat

$$l(\mathbf{Y}; \beta, \varphi) = \sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{\varphi} + c(Y_i, \varphi).$$

Praktický odhad parametrů je založen na derivacích logaritmické věrohodnostní funkce. V neobecnější formě můžeme parciální derivaci prvního řádu dle parametru β_j vyjádřit jako

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial f}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)X_{ij}}{g'(\mu_i)\varphi V(\mu_i)},$$

kde jsme využili

$$\begin{aligned} \frac{\partial f}{\partial \theta_i} &= \frac{Y_i - b'(\theta_i)}{\varphi} = \frac{Y_i - \mu_i}{\varphi}, \\ \frac{\partial \theta_i}{\partial \mu_i} &= \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}, \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{g'(\mu_i)}, \\ \frac{\partial \eta_i}{\partial \beta_j} &= X_{ij}. \end{aligned}$$

Poznamenejme, že obecně platí $V(\mu_i) > 0$ (nenulovost rozptylu) a $g'(\mu_i) > 0$, což vyplývá z ryzí monotonie linkové funkce. Pro přehlednost uveďme vztahy mezi parametry

$$\eta = \mathbf{x}'\beta \qquad \eta = g(\mu) \qquad \mu = b'(\theta)$$

$$\beta \quad \longrightarrow \quad \eta \quad \longleftrightarrow \quad \mu \quad \longleftrightarrow \quad \theta$$

$$\mu = g^{-1}(\eta) \qquad \theta = (b')^{-1}(\mu)$$

Pro maximalizaci věrohodnostní funkce jsou využívány následující dvě iterační metody[9]:

- **Metoda iterativních vážených nejmenších čtverců**

  $$\hat{\beta}^{(k)} = (\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{Z}^{(k-1)},$$

  kde $\mathbf{W}$ je váhová matice a $\mathbf{Z}$ je linearizovaná odezva, které budou definovány níže.

- Iterační **Newtonův-Raphsonův algoritmus**

  $$\hat{\beta}^{(k)} = \hat{\beta}^{(k-1)} - (\mathbf{H}^{(k-1)})^{-1}\nabla^{(k-1)},$$

  kde ∇ značí gradient logaritmické věrohodnostní funkce a $\mathbf{H}$ její Hessovu matici.

Detailnější popis algoritmů je obsahem následujících sekcí.

4.6.2 Metoda iterativních vážených čtverců

Zvolte počáteční odhady jako $\hat{\mu}_i^{(0)} = Y_i$ a pomocí níže uvedených vztahů dopočteme $\mathbf{W}^{(0)}$ a $\mathbf{Z}^{(0)}$. Pro $k \geq 1$ opakuj následující kroky, dokud není splněno kritérium konvergence $\left\|\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\right\| < \varepsilon$:

1. Spočti nový odhad parametrů

   $$\hat{\beta}^{(k)} = (\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{Z}^{(k-1)}.$$

2. Spočti nový odhad vektoru středních hodnot

   $$\hat{\mu}_i^{(k)} = g^{-1}(\mathbf{x}_i^T\hat{\beta}^{(k)}).$$

3. Aktualizuj váhy $\mathbf{W}^{(k)}$ a linearizovanou odezvu $\mathbf{Z}^{(k)}$

   $$\mathbf{W}^{(k)} = diag\left\{\frac{1}{[g'(\hat{\mu}_i^{(k)})]^2 V(\hat{\mu}_i^{(k)})}\right\},$$

   $$\mathbf{Z}^{(k)} = g(\hat{\mu}^{(k)}) + g'(\hat{\mu}^{(k)})(\mathbf{Y} - \hat{\mu}^{(k)}).$$

[9] Horní index k značí iteraci.

Poznamenejme, že není třeba znát odhad disperzního parametru φ.

Pozn. V popsané metodě se využívá následující tvar derivace věrohodnostní funkce

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\varphi V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \beta_j}\right).$$

Definujeme-li váhy

$$w(\mu_i) = \frac{1}{[g'(\mu_i)]^2 V(\mu_i)},$$

pak můžeme parciální derivace zapsat jako

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\varphi}\sum_{i=1}^{n} w(\mu_i) g'(\mu_i)(Y_i - \mu_i) X_{ij} = 0.$$

Pokusíme se vyjádřit vztah pro odhad koeficientů β. Pro přehlednost je možné využít maticový zápis. Nechť $\mathbf{W} = diag\{([g'(\mu_i)]^2 V(\mu_i))^{-1}\}$, $\mathbf{G} = diag\{g'(\mu_i)\}$, potom dostaneme

$$\mathbf{X}^T\mathbf{W}\mathbf{G}(\mathbf{Y} - \mu) = 0.$$

K oběma stranám přičteme $(\mathbf{X}^T\mathbf{W}\mathbf{X})\beta = \mathbf{X}^T\mathbf{W}g(\mu)$, tedy máme

$$\mathbf{X}^T\mathbf{W}[g(\mu) + \mathbf{G}(\mathbf{Y} - \mu)] = (\mathbf{X}^T\mathbf{W}\mathbf{X})\beta,$$

a za předpokladu regularity matice $(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})$ získáme vztah pro odhad parametrů splňující

$$\hat{\beta} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z},$$

kde $\mathbf{Z} = g(\mu) + g'(\mu)(\mathbf{Y} - \mu)$ bývá nazýváno linearizovaná odezva. Všimněte si podobnosti se vztahem pro odhad parametrů v Aitkenově modelu lineární regrese. Přestože vypadá i tento exaktně, není možné jej přímo využít pro výpočet odhadu parametrů, neboť vektor $\mathbf{Z}$ a matice $\mathbf{W}$ závisí na aktuálním odhadu vektoru μ a ten závisí na odhadu parametrů β. Je tedy nutné aplikovat iterační metodu uvedenou výše.

Pozn. Pro kanonický link dochází ke zjednodušení předešlých vztahů

$$w(\mu_i) = \frac{1}{[g'(\mu_i)]^2 V(\mu_i)} = V(\mu_i) = \frac{1}{g'(\mu_i)},$$

tedy

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\varphi}\sum_{i=1}^{n}(Y_i - \mu_i) X_{ij} = 0,$$

což můžeme přepsat maticově

$$\mathbf{X}^T(\mathbf{Y} - \mu) = 0.$$

4.6.3 Newtonův-Raphsonův algoritmus

Abychom mohli aplikovat Newtonův-Raphsonův algoritmus, je nutné spočíst **druhé parciální derivace** logaritmické věrohodnostní funkce:

$$\frac{\partial}{\partial \beta_{j'}}\left(\frac{\partial l}{\partial \beta_j}\right) = \sum_{i=1}^{n} \frac{\partial}{\partial \mu_i}\left(\frac{(Y_i-\mu_i)X_{ij}}{g'(\mu_i)\varphi V(\mu_i)}\right)\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_{j'}},$$

kde

$$\begin{aligned}
\frac{\partial}{\partial \mu_i}\left(\frac{(Y_i-\mu_i)}{g'(\mu_i)\varphi V(\mu_i)}\right) &= \frac{-1}{g'(\mu_i)\varphi V(\mu_i)} - (Y_i-\mu_i)\frac{g''(\mu_i)V(\mu_i)+g'(\mu_i)V'(\mu_i)}{(g'(\mu_i))^2\varphi(V(\mu_i))^2},\\
\frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{g'(\mu_i)},\\
\frac{\partial \eta_i}{\partial \beta_{j'}} &= X_{ij'}.
\end{aligned}$$

Definujeme-li diagonální matici

$$\mathbf{V} = diag\left\{\frac{-1}{(g'(\mu_i))^2\varphi V(\mu_i)} - (Y_i-\mu_i)\frac{g''(\mu_i)V(\mu_i)+g'(\mu_i)V'(\mu_i)}{(g'(\mu_i))^3\varphi(V(\mu_i))^2}\right\},$$

můžeme **Hessovu matici** zapsat ve tvaru

$$\mathbf{H} = \mathbf{X}^T\mathbf{V}\mathbf{X}.$$

Označíme vektor prvních parciálních derivací logaritmické věrohodnostní funkce

$$\nabla^T = \left(\frac{\partial l}{\partial \beta_1}, \ldots, \frac{\partial l}{\partial \beta_m}\right).$$

Algoritmus: Počáteční odhady $\hat{\mu}_i^{(0)} = Y_i$, $\hat{\nabla}^{(0)}$ a $\hat{\mathbf{H}}^{(0)}$. Pro $k \geq 1$ opakuj následující kroky, dokud není splněno kritérium konvergence $\left\|\hat{\beta}^{(k)} - \hat{\beta}^{(k-1)}\right\| < \varepsilon$:

1. Spočti nový odhad parametrů
$$\hat{\beta}^{(k)} = \hat{\beta}^{(k-1)} - (\mathbf{H}^{(k-1)})^{-1}\nabla^{(k-1)}$$
2. Spočti
$$\hat{\mu}_i^{(k)} = g^{-1}(\mathbf{x}_i^T\hat{\beta}^{(k)}).$$
3. Aktualizuj $\nabla^{(k)}$ a $\mathbf{H}^{(k)}$.

4.7 Testování hypotéz

4.7.1 Testy významnosti parametrů

Waldův test se využívá pro test hypotézy $H_0 : \beta_j = c$ (nejčastěji $c = 0$), kde testová statistika je tvaru

$$\frac{(\beta_j - c)^2}{\varphi\ var(\beta_j)} \sim \chi_1^2.$$

Pro test obecnější hypotézy $H_0 : \mathbf{C}\beta = \mathbf{c}$ a $\mathbf{c} \in \mathbb{R}^q$, kde matice $\mathbf{C}$ má q řádků, slouží

$$(\mathbf{C}\beta - \mathbf{c})^T[\varphi\mathbf{C}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\beta - \mathbf{c}) \sim \chi_q^2.$$

4.7.2 Konfidenční intervaly

Konfidenční intervaly pro závisle proměnnou jsou založeny na asymptotické normalitě odhadu parametrů, kdy za určitých předpokladů platí $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, \varphi(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1})$. Potom interval spolehlivosti (y_l, y_u) pro střední hodnotu závisle proměnné dostaneme pomocí

$$\begin{aligned} g(y_l) &= \mathbf{x}^T\hat{\beta} - z\sqrt{\hat{\varphi}\mathbf{x}^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}}, \\ g(y_u) &= \mathbf{x}^T\hat{\beta} + z\sqrt{\hat{\varphi}\mathbf{x}^T(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}}, \end{aligned}$$

kde z je příslušný kvantil standardního normálního rozdělení, $\mathbf{x}$ je vektor regresorů a váhovou matici $\mathbf{W}$ jsme nahradili jejím odhadem $\hat{\mathbf{W}}$.

4.8 Kvalita modelu a testy podmodelů

Významnou roli hraje **saturovaný model**, v němž je počet parametrů roven počtu pozorování[10] a platí

$$\hat{\mu}_i = Y_i, \ \hat{\theta}_i^* = (b')^{-1}(Y_i).$$

Věrohodnost saturovaného modelu je totiž rovna největší dosažitelné věrohodnosti pro daná data

$$l^*(\mathbf{Y}) = \sum_{i=1}^{n} \frac{Y_i\hat{\theta}_i^* - b(\hat{\theta}_i^*)}{\varphi} + c(Y_i, \varphi).$$

Slouží tedy jako (nedosažitelná) hranice „nejvyšší kvality“ při daném rozdělení. **Škálovaná deviance** poté udává ztrátu na logaritmické věrohodnosti vůči saturovanému modelu

$$\begin{aligned} D^*(\mathbf{Y}, \hat{\beta}) &= 2(l^*(\mathbf{Y}) - l(\mathbf{Y}, \hat{\beta})) \\ &= \frac{1}{\varphi}\sum_{i=1}^{n} Y_i(\hat{\theta}_i^* - \hat{\theta}_i) - (b(\hat{\theta}_i^*) - b(\hat{\theta}_i)), \end{aligned}$$

kde $\hat{\theta}_i = (b')^{-1}(g^{-1}(\mathbf{x}_i'\hat{\beta}))$. Někdy je využívána (neškálovaná) **deviance** definovaná jako

$$D(\mathbf{Y}, \hat{\beta}) = \varphi D^*(\mathbf{Y}, \hat{\beta}).$$

Poznamenejme, že existují explicitní vztahy pro devianci pro konkrétní rozdělení.

4.8.1 Testování podmodelů

Je-li $\hat{\beta}$ odhad parametrů **v modelu** a $\hat{\beta}'$ odhad parametrů **v podmodelu**, potom asymptoticky platí

$$D^*(\mathbf{Y}, \hat{\beta}') - D^*(\mathbf{Y}, \hat{\beta}) \sim \chi^2_d,$$

[10] Obecně pro model neplatí vlastnosti ML odhadů.

kde d je rozdíl počtu parametrů v porovnávaných modelech. Tento test vlastně odpovídá testu poměrem věrohodností.
Další test je založen na F-statistice, kde asymptoticky

$$\frac{D(\mathbf{Y}, \hat{\beta}') - D(\mathbf{Y}, \hat{\beta})}{d\hat{\varphi}} \sim F_{d,n-m}$$

a m je počet parametrů v modelu, ze kterého byl odhadnut disperzní parametr $\hat{\varphi}$.

4.8.2 Akaikeho informační kritérium

Akaikeho informační kritérium slouží pro porovnání více modelů, když zohledňuje nejen hodnotu věrohodnostní funkce ale i počet parametrů:

$$AIC = -2(l(\mathbf{Y}; \hat{\beta}, \hat{\varphi}) - m).$$

Preferujeme model s minimální hodnotou AIC.

4.9 Odhad disperzního parametru

Odhad disperzního parametru není obvykle součástí metody maximální věrohodnosti. Vhodné vlastnosti má **Pearsonův odhad** ve tvaru

$$\hat{\varphi} = \frac{1}{n-m} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Využívá se taky **odhad založený na devianci**

$$\hat{\varphi} = \frac{D(\mathbf{Y}, \hat{\beta})}{n-m}.$$

4.10 Korelovaná data, náhodné efekty a GEE modely

Jsou-li data korelovaná v rámci shluku i o velikosti n_i, může využít model s **náhodným absolutním členem** ve tvaru

$$g(\mu_{ik}) = \alpha_i + \mathbf{x}_{ik}^T \beta, \ k = 1, \ldots, n_i,$$

kde $\alpha_i \sim N(0, \sigma^2)$. Při daném α_i jsou pozorování ve shluku nezávislá.
Další možností jsou **Generalized Estimating Equations**, kde platí

$$\sum_{i=1}^{n} \mathbf{d}_i^T \mathbf{U}_i^{-1} (\mathbf{Y}_i - \mu_i) = 0,$$

kde $\mathbf{d}_i = \{\partial \mu_i / \partial \beta_j\}_j$ je sloupcový vektor derivací a $\mathbf{U}_i$ je zobecněná varianční matice shluku i zahrnující strukturu závislosti pozorování.

5 Příklady zobecněných lineárních modelů

5.1 Data

Začneme popisem dat, na něž aplikujeme několik zobecněných lineárních modelů. Uvažujeme 50 000 smluv **povinného ručení** (pojištění odpovědnosti z provozu motorového vozidla) simulované dle mírně upravených reálných charakteristik:

- Závisle proměnné: **počet a výše škod** za poslední rok, příznak **storna**
- Nezávisle proměnné:
  - **tarifní skupina** dle objemu motoru vozidla (TS): 5 kategorií (do 1000, do 1350, do 1850, do 2500, nad 2500 ccm),
  - **stáří pojistníka** spojité (veks): 18-75 let,
  - **stáří pojistníka** (vek): 3 kategorie (18-30, 30-65, 65 a více),
  - **velikosti místa bydliště** (region): 4 kategorie (nad 500 000 obyvatel, nad 50 000, nad 5 000, do 5 000),
  - **pohlaví** (pohlavi): 2 kategorie (1 - žena, 2 - muž).

5.2 Dostupný software

Procedury a funkce pro práci se zobecněnými lineárními modely je možné nalézt například v následujících softwarech:

- **SAS**: procedura GENMOD
- **Statistica**: Generalized Linear Models (GLZ)
- **IBM SPSS**: GENLIN (ne GLM!!!)
- **Mathematica**: GeneralizedLinearModelFit
- **R**: glm
- a další.

My budeme dále využívat SAS a proceduru GENMOD.

5.2.1 Lineární regrese

Pro přehlednost shrneme každý zobecněný lineární regresní model seznamem, který udává základní stavební kameny každého modelu. Na úvod uvádíme model lineární regrese, který však nebudeme na pojistná data dále aplikovat.

- Závisle proměnná: spojitá
- Rozdělení: normální $Y_i \sim N(\mu_i, \sigma^2)$
- Střední hodnota: $\mathbb{E}Y_i = \mu_i$
- Linková funkce: identita $g(\mu) = \mu$
- Rozptylová funkce: $V(\mu) = 1$
- Disperzní parametr: $\varphi = \sigma^2$

```
proc genmod data=ccc;
class ts vek pohlavi;
model pocet_skod = ts vek pohlavi/

noint
dist=poisson
link=log
/*offset=w*/
type1 type3;
quit;
```

Obrázek 6: Syntax v SASu

5.3 Regresní model očekávaného počtu pojistných událostí

5.3.1 Poissonovská regrese (log-lineární model)

V příkladu využití Poissonovské regrese budeme modelovat očekávaný počet pojistných událostí na smlouvě během jednoho roku v závislosti na tarifní skupině, stáří pojistníka a pohlaví. Využijeme následující stavební prvky, resp. vlastnosti Poissonovské regrese:

- Závisle proměnná: počet pojistných událostí na smlouvě za 1 rok
- Rozdělení: Poissonovo $Y_i \sim Po(\lambda_i)$
- Střední hodnota: $\mathbb{E}Y_i = \lambda_i$
- Linková funkce: $g(\mu) = \log(\mu)$
- Rozptylová funkce: $V(\mu) = \mu$
- **Disperzní parametr**: $\varphi = 1$

Kritéria pro hodnocení dobré shody, resp. kvality modelu, najdeme v následující tabulce. Uvedeny jsou deviance, Pearsonovy statistiky a hodnota logaritmické věrohodnostní funkce. Názvy sloupců uvádíme vždy tak, jak jsou obsaženy ve výstupu ze SASu.

Kritérium	DF	Hodnota	Hodnota/DF
Deviance	5E4	18582.5892	0.3717
Scaled Deviance	5E4	18582.5892	0.3717
Pearsonuv Chí-kvad	5E4	50208.1517	1.0043
Scaled Pearson X2	5E4	50208.1517	1.0043
Log verohodnost		-12571.1203	

Další výstup ze softwaru uvádí odhady parametrů (Odhad) spolu s chybou odhadu (Stand. chyba), intervaly spolehlivosti (Waldovy meze interv. spol.), Waldovou testovou statistikou významnosti parametrů (Chí-kv.) a odpovídající p-hodnotou (Pr > ChíKv).

Par.		DF	Odhad	Stand. chyba	Waldovy meze interv. spol.		Chí-kv.	Pr > ChíKv
Int		0	0.0000	0.0000	0.0000	0.0000	.	.
TS	1	1	-2.9646	0.0521	-3.0666	-2.8625	3243.56	<.0001
TS	2	1	-2.9421	0.0517	-3.0435	-2.8407	3235.33	<.0001
TS	3	1	-2.9016	0.0512	-3.0019	-2.8013	3216.13	<.0001
TS	4	1	-2.7451	0.0490	-2.8411	-2.6491	3141.87	<.0001
TS	5	1	-2.7284	0.0488	-2.8240	-2.6329	3131.40	<.0001
vek	1	1	0.5700	0.0426	0.4865	0.6535	178.95	<.0001
vek	2	1	0.2183	0.0456	0.1289	0.3076	22.94	<.0001
vek	3	0	0.0000	0.0000	0.0000	0.0000	.	.
pohlavi	1	1	-0.2278	0.0342	-0.2948	-0.1607	44.32	<.0001
pohlavi	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Škála		0	1.0000	0.0000	1.0000	1.0000		

Interpretace parametrů a výpočet odhadu očekávaného počtu škod probíhá takto:

- Pro TS = 1 (do 1000 ccm), vek = 1 (18–30 let), pohlavi = 1 (žena) dostáváme hodnotu lineárního prediktoru a odhad střední hodnoty

$$\begin{aligned} \eta &= -2,9646 + 0,5700 - 0,2278 = -2,6224 \\ \mu &= \exp\{-2,9646 + 0,5700 - 0,2278\} = \exp\{-2,6224\} \\ &= 0,0516 \cdot 1,7683 \cdot 0,7963 = 0,0726, \end{aligned}$$

  kde poslední zápis může být interpretován jako multiplikativní příspěvek kategorie k výslednému odhadu očekávaného počtu pojistných událostí.

- Pravděpodobnosti počtu událostí na smlouvě můžeme snadno spočíst po dosazení odpovídajícího odhadu parametru λ do hustoty Poissonova rozdělení, například pro výše spočtený $\lambda = 0,0726$ máme
  - $P(Y = 0) = 0.9300$,
  - $P(Y = 1) = 0.0675$,
  - $P(Y = 2) = 0.0025$,
  - $P(Y = 3) = 5.93\ 10^{-5}$,
  - $P(Y = 4) = 1.07\ 10^{-6}$,
  - ...

Pro další hodnoty prediktorů dostáváme

- TS = 1 (do 1000 ccm), vek = 1 (18–30 let), pohlavi = 2 (muž)

$$\begin{aligned} \eta &= -2,9646 + 0,5700 + 0 = -2,3946 \\ \mu &= \exp\{-2,9646 + 0,5700 + 0\} = \exp\{-2,3946\} \\ &= 0,0516 \cdot 1,7683 \cdot 1 = 0,0912. \end{aligned}$$

- TS = 5 (nad 2500 ccm), vek = 1 (18–30 let), pohlavi = 2 (muž)

$$\begin{aligned} \eta &= -2,7284 + 0,5700 + 0 = -2,1584 \\ \mu &= \exp\{-2,7284 + 0,5700 + 0\} = \exp\{-2,1584\} \\ &= 0,0653 \cdot 1,7683 \cdot 1 = 0,1155. \end{aligned}$$

Výsledky testování významnosti regresorů jsou uvedeny v následujících tabulkách. Statistiky LR pro analýzu typu 1 odpovídají postupnému přidávání regresorů, tedy záleží na pořadí regresorů v zadání.

Zdroj	Deviance	DF	Chí-kvadrát	Pr > ChíKv
TS	18822.16			
vek	18627.20	2	194.96	<.0001
pohlavi	18582.59	1	44.61	<.0001

Statistiky LR pro analýzu typu 3 testují významnost regresoru při ponechání všech ostatních regresorů v modelu, tedy nezáleží na pořadí, v jakém jsou zadány.

Zdroj	DF	Chí-kvadrát	Pr > ChíKv
TS	4	34.98	<.0001
vek	2	194.41	<.0001
pohlavi	1	44.61	<.0001

Vidíme, že regresory jsou významné na všech obvykle využívaných hladinách.

5.3.2 Overdispersed Poissonův model

Základní vlastností Poissonova rozdělení je rovnost střední hodnoty a rozptylu. To však bývá v praxi často porušeno a my pozorujeme rozptyl větší než je střední hodnota, což vede k jevu nazývanému overdispersion. Existují dva přístupy, jak tento jev zohlednit v zobecněných lineárních modelech. První je využití negativně binomického modelu s dalším neznámým parametrem, druhý poté využití overdisperzed Poissonova modelu, kde je hodnota disperzního parametru uvolněna a odhadnuta.

Overdisperzed Poissonův zobecněný lineární model je charakterizován takto:

- Závisle proměnná: počet pojistných událostí na smlouvě za 1 rok
- Rozdělení: Overdispersed Poissonovo[11] $Y_i \sim O\text{-}Po(\lambda_i, \varphi)$
- Střední hodnota: $\mathbb{E}Y_i = \lambda_i$
- Linková funkce: $g(\mu) = \log(\mu)$
- Rozptylová funkce: $V(\mu) = \mu$
- **Disperzní parametr**: $\varphi \in (0, \infty)$

[11] Nejedná se o skutečné pravděpodobnostní rozdělení.

Parciální derivace dle parametrů má následující tvar

$$\frac{\partial ql}{\partial \beta_j} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\varphi V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \beta_j} \right)$$

potom odpovídají **kvazi**-(logaritmické-)**věrohodnostní funkci** pro obecnou rozptylovou funkci V a disperzní parametr φ

$$ql = \sum_{i=1}^{n} \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\varphi V(t)} dt.$$

Poznamenejme, že „umělé“ navýšení rozptylu se využívá i pro binomické, resp. alternativní rozdělení.

Dále uvedeme tabulky bez podrobnějšího komentáře, upozorníme pouze na změny. Kritéria pro hodnocení dobré shody:

Kritérium	DF	Hodnota	Hodnota/DF
Deviance	5E4	18582.5892	0.3717
Scaled Deviance	5E4	49992.0000	1.0000
Pearsonuv Chí-kvad	5E4	50208.1517	1.0043
Scaled Pearson X2	5E4	135072.9917	2.7019
Log verohodnos		-33819.5845	

Analýzu odhadů parametrů uvádí následující tabulka, kde je v posledním řádku odhad disperzního parametru:

Par.		DF	Odhad	Stand. chyba	Waldovy meze intrv. spol.		Chí-kv.	Pr > ChíKv
Int		0	0.0000	0.0000	0.0000	0.0000	.	.
TS	1	1	-2.9646	0.0521	-3.0666	-2.8625	3243.56	<.0001
TS	2	1	-2.9421	0.0517	-3.0435	-2.8407	3235.33	<.0001
TS	3	1	-2.9016	0.0512	-3.0019	-2.8013	3216.13	<.0001
TS	4	1	-2.7451	0.0490	-2.8411	-2.6491	3141.87	<.0001
TS	5	1	-2.7284	0.0488	-2.8240	-2.6329	3131.40	<.0001
vek	1	1	0.5700	0.0426	0.4865	0.6535	178.95	<.0001
vek	2	1	0.2183	0.0456	0.1289	0.3076	22.94	<.0001
vek	3	0	0.0000	0.0000	0.0000	0.0000	.	.
pohlavi	1	1	-0.2278	0.0342	-0.2948	-0.1607	44.32	<.0001
pohlavi	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Škála		0	1.6097	0.0000	0.6097	0.6097		

Statistiky LR pro analýzu typu 1 při postupné přidávání regresorů, kdy záleží na pořadí v zadání, v tomto případě využíváme F-testy:

Zdroj	Odchylka	DF cit	DF jmen	F hodnota	Pr > F
TS	18822.16				
vek	18627.20	2	49992	262.25	<.0001
pohlavi	18582.59	1	49992	120.02	<.0001

Statistiky LR pro analýzu typu 3 významnosti regresoru při ponechání všech ostatních regresorů v modelu:

Zdroj	DF cit	DF jmen	F hodnota	Pr > F
TS	4	49992	23.53	<.0001
vek	2	49992	261.51	<.0001
pohlavi	1	49992	120.02	<.0001

5.4 Regresní model výše škod – Gamma regrese

Pomocí Gamma regrese budeme modelovat očekávanou výši škody z pojistné události na smlouvě v závislosti pouze na tarifní skupině. K redukci počtu regresorů dochází kvůli úbytku dat, kdy výše modelujeme pouze na základě nastalých škod, kterých je díky nízké škodní frekvenci obvykle znatelně menší počet. Gamma regrese má následující vlastnosti:

- Závisle proměnná: spojitá kladná výše škody
- Rozdělení: $Y_i \sim \Gamma(\mu, \nu)$
- Střední hodnota: $\mathbb{E}Y_i = \mu$
- Linková funkce: $g(\mu) = \log(\mu)$ (není kanonický link)
- Rozptylová funkce: $V(\mu) = \mu^2$
- Disperzní parametr: $\varphi = 1/\nu$

Kritéria pro hodnocení dobré shody (ML odhad parametru měřítka):

Kritérium	DF	Hodnota	Hodnota/DF
Deviance	3458	0.0007	0.0000
Scaled Deviance	3458	3464.2364	1.0018
Pearson Chi-Square	3458	0.0007	0.0000
Scaled Pearson X2	3458	3466.7934	1.0025

Analýzu odhadů parametrů a různé odhady parametru měřítka najdeme v následující tabulce:

Par.		DF	Odhad	Stand. chyba	Waldovy meze intrv. spol.		Chí-kv.	Pr > ChíKv
Int		0	0.0000	0.0000	0.0000	0.0000	.	.
TS	1	1	10.3127	0.0033	10.3062	10.3192	9613383	<.0001
TS	2	1	10.3592	0.0033	10.3528	10.3656	9966668	<.0001
TS	3	1	10.4662	0.0032	10.4599	10.4725	1.061E7	<.0001
TS	4	1	10.5388	0.0030	10.5329	10.5447	1.244E7	<.0001
TS	5	1	10.7211	0.0030	10.7153	10.7269	1.306E7	<.0001
Scale		0	146.0294	0.0000	146.0294	146.0294		

Tabulky pro testování významnosti jednotlivých regresorů neuvádíme, neboť v našem modelu uvažujeme jen jeden regresor.

5.5 Regresní model stornovosti – logistická regrese

V této části uvedeme model pravděpodobnosti storna smlouvy během jednoho roku v závislosti na tarifní skupině, velikosti místa bydliště, pohlaví, stáří pojistníka. Model logistické regrese má obecně následující vlastnosti:

- Závisle proměnná: binární – jev nastal/nenastal, tj. storno během jednoho roku – ano/ne
- Rozdělení: binomické (alternativní): $Y_i \sim Alt(p_i)$
- Střední hodnota: $\mathbb{E}Y_i = p_i$
- Linková funkce: logit $g(\mu) = \log(\mu/(1-\mu))$
- Rozptylová funkce: $V(\mu) = \mu(1-\mu)$
- Disperzní parametr: $\varphi = 1$

Střední hodnota alternativní proměnné je rovna pravděpodobnosti, tedy

$$\mathbb{E}Y_i = p_i = \frac{e^{\mathbf{x}_i'\beta}}{1+e^{\mathbf{x}_i'\beta}}.$$

Kritéria pro hodnocení dobré shody

Kritérium	DF	Hodnota	Hodnota/DF
Deviance	5E4	56802.0249	1.1363
Scaled Deviance	5E4	56802.0249	1.1363
Pearsonuv Chí-kvad	5E4	49969.3190	0.9996
Scaled Pearson X2	5E4	49969.3190	0.9996
Log verohodnost		-28401.0124	

Analýza odhadů parametrů

Par.		DF	Odhad	Stand. chyba	Waldovy meze intrv. spol.		Chí-kv.	Pr > ChíKv
Intercept		1	-1.6157	0.0429	-1.6998	-1.5316	1417.00	<.0001
TS	1	1	-0.3326	0.0323	-0.3959	-0.2692	105.90	<.0001
TS	2	1	-0.2814	0.0322	-0.3445	-0.2183	76.36	<.0001
TS	3	1	-0.2248	0.0320	-0.2874	-0.1622	49.51	<.0001
TS	4	1	-0.0711	0.0314	-0.1326	-0.0095	5.12	0.0237
TS	5	0	0.0000	0.0000	0.0000	0.0000	.	.
region	1	1	0.4820	0.0290	0.4252	0.5389	275.76	<.0001
region	2	1	0.2633	0.0296	0.2053	0.3214	79.06	<.0001
region	3	1	0.1272	0.0300	0.0683	0.1860	17.96	<.0001
region	4	0	0.0000	0.0000	0.0000	0.0000	.	.
pohlavi	1	1	0.5584	0.0206	0.5180	0.5989	731.75	<.0001
pohlavi	2	0	0.0000	0.0000	0.0000	0.0000	.	.
veks		1	0.0058	0.0006	0.0046	0.0071	82.36	<.0001
Škála		0	1.0000	0.0000	1.0000	1.0000		

Interpretace parametrů je možné provést pomocí **šance**

$$\frac{p_i}{1-p_i} = \exp\{\mathbf{x}_i'\beta\} = \exp\left\{\sum_{j=1}^{m} X_{ij}\beta_j\right\}.$$

Pokud zvýšíme regresor $\tilde{j}$ o jednotku $X_{i\tilde{j}} + 1$ a ostatní neměníme, potom pro šanci platí

$$\frac{\tilde{p}_i}{1-\tilde{p}_i} = \exp\left\{\sum_{j=1, j\neq\tilde{j}}^{m} X_{ij}\beta_j + (X_{i\tilde{j}}+1)\beta_{\tilde{j}}\right\} = \exp\left\{\sum_{j=1}^{m} X_{ij}\beta_j\right\}\exp\{\beta_{\tilde{j}}\},$$

tj. $e^{\beta_{\tilde{j}}}$ vyjadřuje změnu šance při zvýšení příslušného regresoru o jednotku.

Predikovanou hodnotu, tedy pravděpodobnost storna během jednoho roku, pro TS = 5 (nad 2500 ccm), region = 4 (do 5000), pohlavi = 2 (muž), veks = 22 let spočteme jako

$$\begin{aligned} \eta &= -1,6157 + 0 + 0 + 0 + 22 \cdot 0.0058 = -1,4881 \\ \mu &= \frac{\exp\{-1,4881\}}{1+\exp\{-1,4881\}} = 0,1842. \end{aligned}$$

Statistiky LR pro analýzu typu 1 při postupné přidávání regresorů, kdy záleží na pořadí v zadání:

Zdroj	Deviance	DF	Chí-kvadrát	Pr > ChíKv
Intercept	58087.7242			
TS	57937.9201	4	149.80	<.0001
region	57626.8576	3	311.06	<.0001
pohlavi	56884.5504	1	742.31	<.0001
veks	56802.0249	1	82.53	<.0001

Statistiky LR pro analýzu významnosti regresoru při ponechání všech ostatních regresorů v modelu:

Zdroj	DF	Chí-kvadrát	Pr > ChíKv
TS	4	154.02	<.0001
region	3	309.14	<.0001
pohlavi	1	743.64	<.0001
veks	1	82.53	<.0001

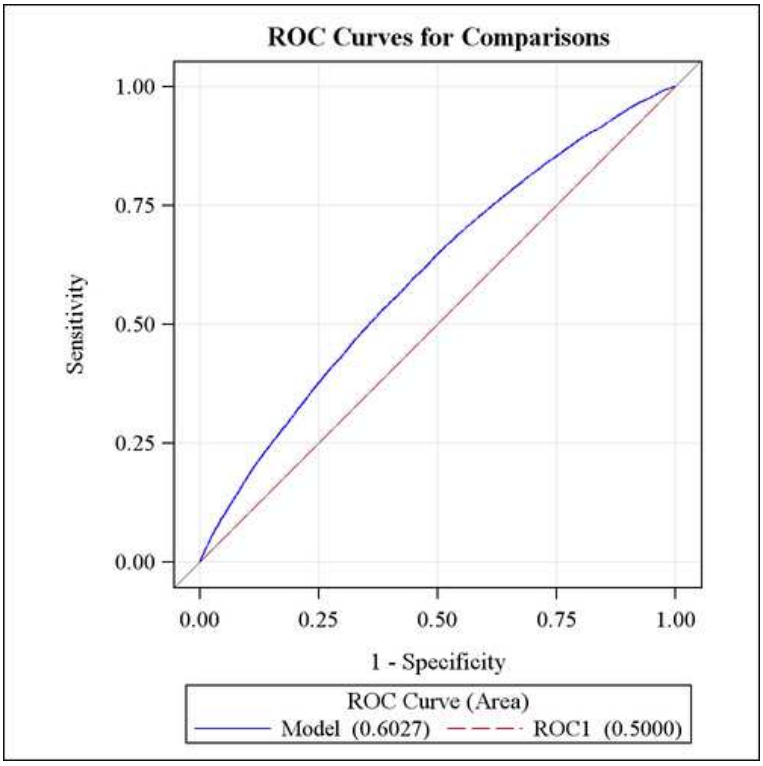
Zvláštní pozornost věnujeme ROC křivce sloužící k posouzení kvality modelu a nastavení prahové hodnoty. Pro predikované pravděpodobnosti, které jsou vyšší než prahová hodnota, očekáváme, že sledovaný jev spíše nastane, u hodnot nižších naopak. ROC křivka poté zakresluje:

- **Na svislé ose** grafu relativní četnost skutečně pozitivních případů TP, tedy pravděpodobnost, že jako správný bude vyhodnocen pozitivní případ: Sensitivity = TP/(TP+FN).
- **Na vodorovné ose** relativní četnost falešně pozitivních případů FP, tedy pravděpodobnost, že jako správný bude vyhodnocen negativní případ: 1-Specificity = FP/(TN+FP).

Vycházíme přitom z následující tabulky, kde značíme True (T), False (F), Positive (P), Negative (N):

skutečnost/predikce	1	0
1	TP	FP
0	FN	TN

Čím je větší plocha pod ROC křivkou, resp. čím více je křivka vypouklá nahoru, tím lepší má model predikční schopnost. Křivka pro náš model je zakreslena na následujícím obrázku.



5.6 Postup konstrukce zobecněného lineárního modelu

Obecně může být zobecněný lineární model konstruován v následujících krocích:

1. Vyberte rozdělení
2. Vyberte link
3. Vyberte nezávisle proměnné
4. Odhadněte parametry
5. Posuďte kvalitu modelu
6. Iterujte od vhodného kroku

Často si nemusí být jisti, které regresory do modelu zahrnout a které naopak vyloučit. Pro výběr nejvhodnějších regresorů jsou používány následující sekvenční postupy:

- **Vzestupný výběr** (forward selection) - začneme od prázdného modelu, postupně přidáváme statisticky významné regresory.
- **Sestupný výběr** (backward selection) - začneme od modelu se všemi regresory, postupně odebíráme statisticky nevýznamné.
- **Krokový výběr** (stepwise selection) - začneme od prázdného modelu, v každém kroku přidáme jeden statisticky významný regresor a poté se pokusíme odebírat statisticky nevýznamné (i více). Hladina pro přidávání musí být menší než hladina pro odebírání, jinak může dojít k zacyklení.

Při praktickém použití zobecněných lineárních modelů máme často k dispozici rozsáhlý soubor dat. Ten je možné náhodně rozdělit na „trénovací“ a „testovací“ podsoubor. Na prvním je model odhadnut, na druhém potom ověřena jeho kvalita, resp. predikční schopnost. Jako kritérium může sloužit například střední čtvercová chyba $1/n \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$, kde $\hat{Y}_i$ značí predikci pomocí odhadnutého modelu.

6 Reference

- M. Denuit, X. Maréchal, S. Pitrebois, J.-F. Walhin: *Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems*. John Wiley & Sons, Chichester, 2007.
- C.-C. Günthera, I.F. Tvete, K. Aas, G.I. Sandnes and O. Borgan: Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*. DOI:10.1080/03461238.2011.636502
- P. de Jong, G. Z. Heller: *Generalized Linear Models for Insurance Data*. Cambridge University Press, 2008.
- P. McCullagh, J.A. Nelder: *Generalized Linear Models*. 2nd Ed. Chapman and Hall, London, 1989.
- E. Ohlsson, B. Johansson: *Non-Life Insurance Pricing with Generalized Linear Models*. EAA Series, Springer-Verlag Berlin Heidelberg, 2010.
- K. Zvára: *Regrese*. Matfyzpress, Praha, 2008.
- *Zápisky z přednášky Zobecněné lineární modely* (NSTP196), 2010, MFF UK, přednášející Doc. Mgr. Michal Kulich, Ph.D.
- SAS/STAT 9.3: User's Guide.